OPEN ACCESS

BIO-complexity

Research Article

A Single-Couple Human Origin is Possible

Ola Hössjer[1]*and Ann Gauger[2]
[1]Department of Mathematics, Stockholm University, Sweden
[2]Biologic Institute, Redmond, Washington, USA

Abstract

The problem of inferring history from genetic data is complex and underdetermined; there are many possible scenarios that would explain the same data. It can be made more tractable by making reasonable simplifications to the model, but it is continually important to remember what has been demonstrated and what is merely a parsimonious working assumption. In this paper we have chosen to model the demographic ancestry of humanity using the simplest of assumptions, with a homogeneous population whose size can vary over time. All other assumptions such as the mutation rates were standard, and no natural selection was in operation. Using a previously published backwards simulation method and some newly developed and faster algorithms, we run our single-couple origin model of humanity and compare the results to allele frequency spectra and linkage disequilibrium statistics from current genetic data. We show that a single-couple origin of humanity as recent as 500kya is consistent with data. With only minor modifications of our parsimonious model assumptions, we suggest that a single-couple origin 100kya, or more recently, is possible.

Cite as: Hössjer O, Gauger A (2019) A single-couple human origin is possible. BIO-Complexity 2019 (1):1-20. **doi:**10.5048/BIO-C.2019.1.

Editor: Jonathan Bartlett

Received: January 1, 2019; **Accepted:** September 17, 2019; **Published:** October 21, 2019

Notes: A *Critique* of this paper, when available, will be assigned **doi:**10.5048/BIO-C.2019.1.c.

*Email: ola.hossjer@gmail.com

1. INTRODUCTION

It has often been claimed that science shows that humans evolved from a large population over a period of millions of years, and therefore that the notion of one human race descended from just two people, must be false.

For example, Francisco Ayala argued that the genetics of HLA-DRB1 of the major histocompatibility complex necessitates a large population continuing back tens of millions of years [1–3]. Other scientists have made other estimates using different methods [4–7]. In addition, several authors of popular-level books have claimed that the human race must have evolved from a large population [8–11]. Dennis Venema, for example, has claimed that the human population has always been in the thousands, and has further stated that this is a fact which can be known with the same certainty as heliocentrism [12]. However, recent debates have shown that this representation can be disputed. It seems that certain common assumptions used for convenience have been misinterpreted as if they were data-driven conclusions, without testing the single-couple origin hypothesis scientifically. These debates can be found on several web pages [13–15]. For instance, some recent qualitative discussions suggest that genetic data might be compatible with an extreme bottleneck from a previously evolving large population to a single-couple around 500kya ago [16], whereas other authors suggest a primordial couple that lived much more recently [17–19].

In this paper we continue work along this line and explore the question whether the single-couple hypothesis is compatible with genetic data. We previously developed a backwards simulation method [20, 21] in order to test the possibility that the human population arose from a single couple, either by the way of bottleneck or a single first pair. We will call the first model the Bottleneck to Single Couple (BTSC) model and the second one the Single-Couple Origin (SCO) model. It is preferable to test these two models by means of backward simulation, in order to have a faster algorithm, capable of handling larger populations and more variables. In this paper we therefore implemented this algorithm, and also developed other faster ones that use alternate mathematical methodology but produce the same results, lending confidence in the output of each method. We have chosen to use the most parsimonious conditions possible for this study, because we want to determine under the strictest possible conditions, with few parameters, whether a single pair is possible. In more detail, we regard the human population as homogeneous (no

geographic subdivision), dioecious and diploid (males and females are distinguished, and each individual has two copies of a non-sex chromosome), and reproduction is selectively neutral. The only parameters of our SCO model, which will be explained in detail in the paper, are the time-varying effective population size, the constant germline mutation rate, the constant recombination rate and one parameter associated with the genomic diversity of the first single couple.

2. BACKGROUND

2.1 Genomic Data

Modern human genomes derive from our earliest ancestors, and contain traces of that history. Scientists hope to use these traces to reconstruct our history. Great progress has been made in recent decades through major projects such as HapMap [22] and the 1000 Genomes project [23] which collated genetic sequence data sampled from the genomes of many people around the world; two haploid genomes per person sampled. In this paper we focus on the 1000 Genomes Project. This data set has conveniently reduced the mass of raw data into the variant-call-format, which is a list of variants, along with a list of which genomes each variant is found in. It is possible to cross-reference the location and ethnicity of each individual in the sample. Even in its reduced form, this is still an enormous matrix of data: there are tens of millions of genetic variants cataloged, multiplied by 5008 haploid genomes in which each variant is 'called' as either present or absent.

2.1.1 Allele Frequency Spectrum

One way to represent 1000 Genomes data in a comprehensible form is to graph the allele frequency spectrum (AFS). The frequency of a variant or allele is the proportion of genomes in a sample which carry that variant, and the AFS shows the number of variants that exist at each proportion. See also references [24–26].

Figure 1 shows the AFS for each of the African, East Asian, and European superpopulations. The scale along the y-axis is a number-density, so that for instance for the African population, the area under the African curve over a certain interval, corresponds to the number of variants with minor allele frequency within that interval, and similarly for the East Asian and European populations. Therefore, the number density conveys information, not only about the distributions of allele frequencies, but also on the total genetic diversity of the population.

Note three important features: (i) Most variants are rare; they have a small frequency in the sample population. (ii) The population of Africa has more genetic variation than the others. (iii) The three populations have differing numbers of rarer alleles (left end) but very similar numbers of more common alleles (right end).

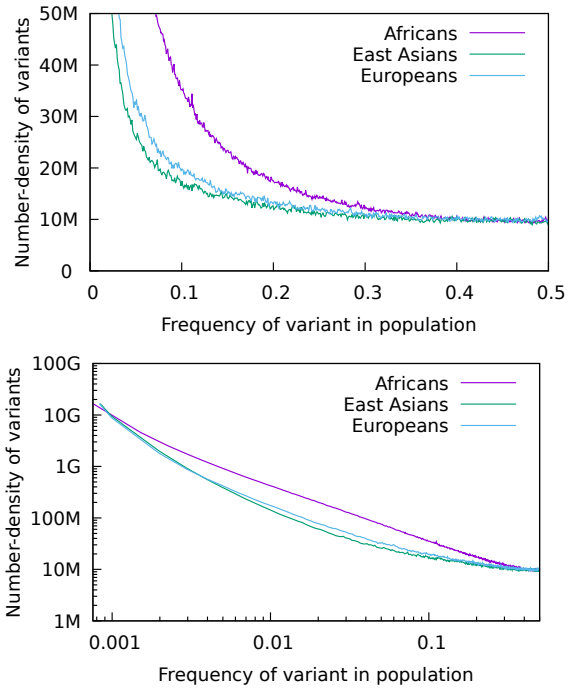

Figure 1: Folded allele frequency spectrum for Africans, East Asians, and Europeans using 1000 Genomes data. The x-axis is the minor allele frequency (MAF) of biallelic single nucleotide polymorphisms (SNPs) in a given population. The y-axis shows the number of variants that have that frequency. The number of variants has been transformed into a number-density (the number in each bin divided by the width of the bin) so that results from different bin widths and samples can be compared. Biallelic means there are only two variants observed in our sample of genomes at that position (locus). The minor allele is the least prevalent of the two, and so the MAF cannot be greater than 0.5. The upper (lower) graph plots the same folded spectra on a linear-linear (logarithmic-logarithmic) scale. The corresponding unfolded allele frequency spectra (not shown) have the frequency of one reference allele of each SNP along the horisontal axis. Its x-axis therefore ranges between 0 and 1. Often the reference allele is the derived allele, caused by a germline mutation at a previously unpolymorphic site. **doi:**10.5048/BIO-C.2019.1.f1

2.1.2 Linkage Disequilibrium Statistics

A complementary way to summarise the data is by plotting measures of correlation against distance along the chromosome. This is a way to measure the amount of recombination that has occurred since the common ancestor.

Linkage occurs because genetic information is stored together on a limited number of linear chromosomes, and so some variants tend to be copied together. Recombination happens in diploid organisms when the two copies of a chromosome are brought together during meiosis and cross over, producing gametes with slightly shuffled chromosomes. Over many generations this shuffling breaks up the linkage between genetic variants, eventually becoming completely uncorrelated, which is called linkage equilibrium. The linkage or correlation, which is named *linkage disequilibrium* (or LD), falls off with

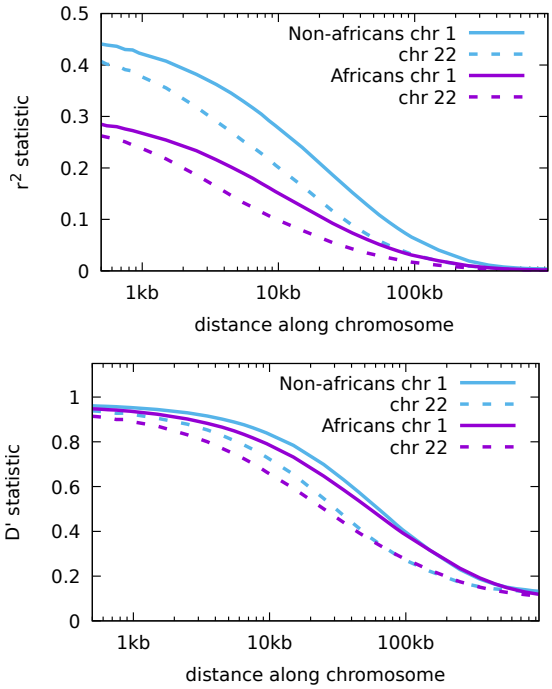

Figure 2: Linkage disequilibrium (LD) statistics r^2 (top) and D' (bottom) against distance on chromosome 1 and chromosome 22, for Africans and Non-Africans, using 1000 Genomes data. See [21, 27, 28] for definitions of r^2 and D'. Variants with minor frequency less than 0.05 were excluded from the analysis. **doi:**10.5048/BIO-C.2019.1.f2

distance between genetic variants, and typically falls off faster with distance for an older population.

Figure 2 illustrates averaged values of the two most commonly used measures of LD, r^2 and D', for pairs of markers at different distances. Both measures have values between 0 and 1, and these values are higher the stronger the LD is. D' equals exactly 1 when there is complete LD between two biallelic markers (at least one of four possible combinations of the two pairs of variants is missing), whereas r^2 is usually less than 1 even for complete LD, unless the two markers happen to have the same minor allele frequency.

2.2 Genetic Diversity Due to Mutation and Drift

All mutations start as single copy-errors but some of them increase in the population by random processes. Figure 3 below shows the ancestry graph with mutations for a small population of constant size, and for the purpose of illustration, the population is haploid. We will discuss the principles of ancestry, and how it can be analyzed, for such a population, where each individual inherits only from a single parent and there is no recombination. A haploid population could represent a female population of mitochondrial DNA or a male population of Y-chromosome DNA. The same principles of ancestry and mutation also apply for a single locus

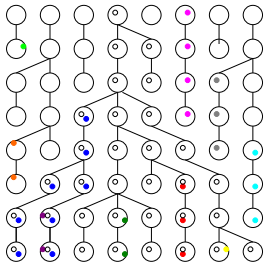

Figure 3: Genealogy of a halpoid population of constant size with non-overlapping generations. Each generation has 8 haploid individuals (corresponding to 4 diploid individuals). The ancestry of the individuals from the present (bottom) generation can be traced backward in time as marked with lines, over the course of 7 generations. Mutations are shown as colored dots. **doi:**10.5048/BIO-C.2019.1.f3

of DNA (such as haplotype blocks[1]) in a two-sex population, where each individual is diploid and carries two haploid copies of DNA from that locus; one from each parent. However, our actual model, as described later in the paper, is diploid with recombinations, when ancestry is followed for many loci over larger sections of a chromosome. But that is more complicated to illustrate[2]. In the genealogy of a haploid population, as in Figure 3, each individual carries a single copy of a specific gene or chromosome segment. Ancestors are at the top of the graph, and descendants at the bottom. Random differences in reproductive success cause some lineages to branch, and others to go extinct. Mutations (dots) happen randomly at a more-or-less constant rate and accumulate as they are inherited by descendants. Historical individuals who don't have descendants in the final generation are called *non-ancestral*. They have no direct effect on the genetic data we have in the present, and we have no direct genetic evidence that they even existed. They can be ignored and removed from the graph, as shown in Figure 4.

It is important to consider what Figure 4 implies. All living individuals (of a haploid population) ultimately derive from a single ancestor at some point in the past. This is not only true for a population that grew from a single individual, but also for any population of constant large size, as in Figure 4. Indeed, in the latter case we can trace the ancestry of the individuals of the present generation as far back in time as we wish, and sooner or later, with a very high probability their ancestral lineages will merge at the so called most recent common

[1]Haplotype blocks are small chunks of recombinant DNA of non-sex chromosomes or the X-chromosome. Such a portion of DNA has not yet experienced recombinations, and therefore it can be regarded as one single locus, with an ancestry following the patterns of inheritance of a haploid population.

[2]Due to recombinations, different haplotype blocks have different ancestral trees. The whole collection of trees along one non-sex chromosome is referred to as an ancestral recombination graph [21, 29–31].

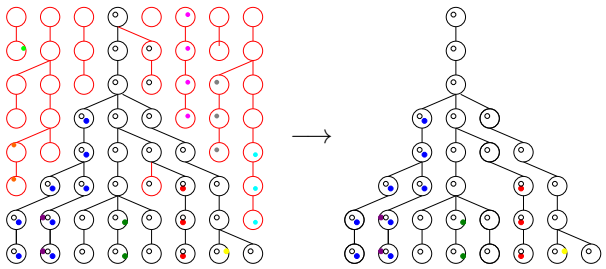

Figure 4: Genealogy of the haploid population of Figure 3 (left), and a pruning where all non-ancestral individuals are removed (right). **doi:**10.5048/BIO-C.2019.1.f4

ancestor. Therefore, we cannot always distinguish a constant population size scenario, as in Figure 4, from a single individual origin scenario, if their ancestral trees are similar. Thus it may not always be possible to tell the difference between the two scenarios from genetic data.

Scientists stand in the present and look back in time to try to figure out what happened in the past. For this reason, it is conventional to count generations backwards in time, where $t = 0$ is the present. The process of lineages branching, when looked at backwards in time, becomes a process of lineages *coalescing*. Coalescence is an important concept in modern population genetics.

There are several other things worth noting from Figures 3-4:

1. Many mutations go extinct almost immediately (going forward in time); in a growing population, fewer do; in a shrinking population, more do. We care only about *ancestral mutations*: those that did not go extinct but were passed on from ancestors to the present day.

2. Ancestral mutations that happened more recently are greater in number (the number of distinct mutations is greater), but each one is found in proportionally fewer members of the final generation, on average. These tend to appear on the left end of the allele frequency spectrum (see Figure 1). Conversely, ancestral mutations that are more ancient are fewer in number (the number of distinct mutations is fewer), but each one is found in proportionally more members of the final generation, on average. These can appear anywhere on the AFS, and are the dominant kind at the right end.

3. Ancestral mutations that occurred before the common ancestor are present in *every* member of the final sample and are thus *fixed* in the population, and are no longer variants. Evolutionary *neutralists* believe that the vast majority of permanent sequence evolution comes from the random fixation of mutations by drift rather than fixation due to selection, so that evolution is mostly random [32, 33]. Here we will assume evolution is predominantly neutral (no natural selection), and we will not concern ourselves with fixation. Instead we are interested in what the variation can tell us.

2.3 Bottlenecks

When a population gets very small, or starts very small in the case of a single-couple origin, this is described as a bottleneck. A bottleneck can cause a significant reduction in the number of lineages. The loss of lineages can mean loss of genetic diversity, and an increase in double recessives, leading to inbreeding depression. However, the greatest damage to genetic diversity does not come from the size of the bottleneck alone but from how long it lasts. Extreme bottlenecks do not necessarily cause extreme harm, if they are of short duration. Even an extreme bottleneck of two individuals (the BTSC model) reduces heterozygosity by only 25% [34], so long as the population rebounds quickly enough. Although we will consider neutral variation, it is important to emphasize that the frequency of some harmful variants may increase during a bottleneck. On the other hand, it also causes a very large number of rare but potentially harmful recessive mutations to be lost. One example is a single pair of mouflon that rapidly colonised a previously uninhabited island [35]. In the same way, even an extreme bottleneck of only two people, from a previously evolving population of hominids or chimps, would not necessarily lead to extinction. But if the bottleneck lasted for long, the genetic threat is substantial.

There is also the further possibility that the original pair were instantiated with only helpful or neutral variation, and no harmful recessive mutations, and thus might not see any inbreeding depression even if the population started small. This brings us to the topic of the next section.

3. MODEL DEVELOPMENT

3.1 Primordial Diversity

In the evolutionary model of a diploid population every portion of a chromosome without recombination events (a haplotype block) could trace back to one ancestor, but there is also the possibility of a discontinuous origin (a 'Big Bang') with more than one ancestor. In the case of a single-couple origin (the SCO model), there could be up to four original versions of each autosomal (non-sex) chromosome, and up to four original versions of every haplotype block [17, 18, 20, 21, 36]. If a model with a first couple is true, when seen through the lens of evolutionary theory, any such primordial variation would be misinterpreted as being due to ancient mutations. The

[2]This assumes traditional models of reproduction. There are more complex models which may have different consequences, see for instance [19].

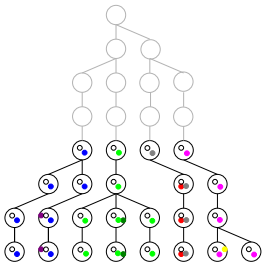

Figure 5: Genealogy of 8 haploid chromosome segments in a diploid population founded three generations back in time by two individuals (corresponding to four chromosome segments). If the model with a founding generation is correct, the chromosome segments of the founding generation have no further ancestry, and then the shaded lines above them represent non-existing ghost lineages. **doi:**10.5048/BIO-C.2019.1.f5

latter assumption then leads to four ghost lineages and a period of microevolution that never actually happened (see Figure 5). In Section 3.2 we will analyze the diversity of a single couple in more detail, both when the couple has a unique origin, the primordial variation scenario, or SCO model, and when the couple is the result of bottleneck from a large population, the Bottleneck to Single Couple (BTSC) model.

3.2 Primordially Diverse Genetic Diversity Compared to Bottleneck Diversity

Assume a selectively neutral model, where no individuals have higher expected reproductive fitness than others. Recall from Figure 1 that a folded allele frequency spectrum has the frequency of the derived allele of a biallelic SNP along the x-axis. A diploid and neutral Wright-Fisher population that has had constant effective size N_e for long enough to reach an equilibrium between mutation and genetic drift, should have an expected unfolded allele frequency spectrum $\mathrm{AFS}_{\mathrm{before},i} = 4N_e\mu/i$ [26, 37], before any bottleneck occurs. This is the expected number of sequences with frequency $i = 1, \ldots, S-1$ for the derived allele, for a sample of size S. We will see what happens if this population goes through a bottleneck of two diploid individuals in the next generation, corresponding to the BTSC model. It is convenient to regard the whole population as our sample ($S = 2N_e$ haploid copies for a diploid population of size N_e) and express the frequencies $0 < f = i/(2N_e) < 1$ of the derived allele on a scale between 0 and 1, as explained in the caption of Figure 1. This gives a normalized number-density $\mathrm{AFS}_{\mathrm{before}}(f) = S \cdot \mathrm{AFS}_i = 4N_e\mu/f$ before the bottleneck (see Appendix A.5). Since the bottleneck corresponds to four haploid individuals (on a non-sex chromosome), there can be 0-4 copies of each allele at the bottleneck. If the frequency of a particular allele before the bottleneck is f, the number of copies k of this allele among the the two diploid individuals has a binomial distribution with parameters 4 and f. As we are interested in variation, we ignore the cases where variants go extinct (k=0) or fix (k=4). By summing over all possible alleles before the bottleneck, according to the expected unfolded allele frequency spectrum, the expected unfolded allele frequency spectrum at the bottleneck becomes

$$\begin{aligned}
\mathrm{AFS}_k &= \sum_{i=1}^{2N_e-1} \mathrm{AFS}_{\mathrm{before},i} \cdot \binom{4}{k} f^k (1-f)^{4-k} \\
&= \int_{f=0}^{1} \mathrm{AFS}_{\mathrm{before}}(f) \cdot \binom{4}{k} f^k (1-f)^{4-k} \mathrm{d}f \\
&= 4N_e\mu \times \begin{cases} 1, & k=1, \\ 1/2, & k=2, \\ 1/3, & k=3. \end{cases} \qquad (1)
\end{aligned}$$

In the same way that we may not be able to tell which variant is ancestral and which is derived, we may not be able to distinguish between k=1 and k=3. The folded distribution at the bottleneck, using the minor allele frequency k, is[3]

$$\mathrm{AFS}_k = 4N_e\mu \times \begin{cases} 4/3, & k=1, \\ 1/2, & k=2. \end{cases} \qquad (2)$$

We notice from equation (2) that the bottleneck scenario only has one free parameter: the prior heterozygosity $4N_e\mu$, determined by the prior effective population size N_e and mutation rate μ. This is essentially true, even if the population size before the bottleneck varied over time, although the AFS will then look slightly different [38, 39]. In any case, if there are prior assumptions on the effective population size history and the mutation rate, the single parameter of the bottlenecked AFS in (1), might not be freely variable.

In contrast, the SCO model with primordially diverse genomes has two free parameters in the allele frequency spectrum of the founding couple (for non-sex chromosomes). These two parameters, AFS_1 and AFS_2, correspond to the number of variants within the first couple with a minor allele frequency of 1/4 and 1/2 respectively. In this model, both AFS_1 and AFS_2 of the folded primordial allele distribution will therefore be a side-effect of a different kind of mechanism; for instance a design process.[4] Therefore, the only difference between

[3] It is also possible to express (2) as a number density $\mathrm{AFS}(f) = 4/3 \cdot \delta(f-1/4) + 1/2 \cdot \delta(f-1/2)$, where $\delta(f-f_0)$ is a delta function centered at frequency f_0. Therefore, the folded AFS of the BTSC model, right after the bottleneck, has two spikes at minor allele frequencies $f = 1/4$ and $f = 1/2$. Likewise, the unfolded spectrum of the BTSC model in (1) has a number density with three spikes for the frequency of the derived allele, at $f = 1/4$, $f = 1/2$, and $f = 3/4$.

[4] Consider Ewert [40] for an example of what the results of a design process might look like. The primordial gametes model of Sanford et al [19] has many more free parameters, which allows for greater flexibility, but at a cost of reduced parsimony.

the bottleneck and primordial AFS for a neutral model, is the extra parameter associated with the primordial AFS. In this paper (see Section 3.3) we will simplify the primordial AFS and put $\mathrm{AFS}_2 = 0$ in equation (2). The primordial AFS is therefore instantiated with a MAF of 1/4, and this makes the BTSC and SCO models very similar.

Recall though from Section 2.3 that a large number of germline mutations are believed to be deleterious or slightly deleterious [32]. Consequently, the bottleneck diversity of a non-neutral BTSC model should incorporate deleterious variants, whereas this is not necessary for a SCO model in which the single couple has primordial diversity. For this reason, for non-neutral models, the bottleneck model faces a greater challenge to handle inbreeding depression due to recessive disorders that spread in the population, at least if the bottleneck lasts for long [20, 41].

3.3 Primordial Haplotype Block Structure

A primordially diverse SCO model has further choices regarding how the variation of the first single couple should be distributed among their four chromosome copies, and how to place boundaries between primordial haplotype blocks. Let us assume that all nucleotide variants are biallelic, that is, they have at most two variants. These can be distributed on the four initial chromosomes with a 1:3 or a 2:2 frequency distribution, like cases $k = 1$ and $k = 2$ in Equation (2). Let us call these one-copy and two-copy variants respectively. One-copy variants are either exclusively located on the same chromosome copy in the founding couple, or they appear on different chromosomes (in this paper they are evenly distributed on all four chromosomes). In terms of the D' linkage statistic, it is always true that $D' = 1$ between a one-copy variant and any other variant, whether it has one-copy or two-copies (see [21, 27, 28] for a definition of D'). This means that one-copy variants give no information about initial linkage or initial haplotype blocks. On the other hand, two-copy variants do give information: Between two-copy variants, there is $D' = 1$ if the variants are always together on the same two primordial chromosomes out of four, or never together because the two variants appear on opposite pairs of chromosomes; or $D' = 0$ if they are together on only one chromosome but not on the others. Groups of adjacent variants all having $D' = 1$ between them indicate a haplotype block.

Figure 6 below shows variants distributed on the founding couple's four copies of a specific non-sex chromosome. Each variant is color-coded according to which copy or pair of copies it is found on. The vertical positions along the four chromosome copies represent different loci and the horizontal, dashed lines represent boundaries between different haplotype blocks. The top-left case shows one-copy variants. The top-right, bottom-left, and bottom-right cases show two-copy variants randomly

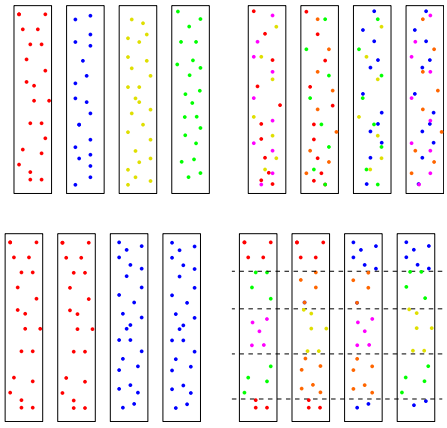

Figure 6: Different Single Couple Origin (SCO) models of primordial diversity, with biallelic variants of the founding couple marked as dots along their four chromosomes. Each allele (variant) is color-coded according to which chromosome or pair of chromosomes it is found on, and dashed horisontal lines mark haplotype boundaries. The four graphs correspond to one-copy variants (top-left), and two-copy variants distributed randomly (top-right), on the same two chromosome copies (bottom-left) and on random haplotype blocks (bottom-right). **doi:**10.5048/BIO-C.2019.1.f6

distributed, distributed on the same two pairs of chromosome copies, and distributed on random haplotype blocks respectively. The top-left case suggests there are four haplotypes at every section of the chromosome, but gives no indication where block boundaries might be. The top-right case suggests blocks so small that they cannot be discerned. The bottom-left case suggests one single haplotype block spanning the length of a chromosome. Only the bottom-right case shows evidence of multiple clear haplotype blocks.

However, for the purposes of this paper, we made the simplifying assumption that each variant was present on only one initial copy of the chromosome, as in the upper left case. This implies we do not need to define primordial haplotype blocks. Since $\mathrm{AFS}_2 = 0$, there is only one free parameter AFS_1 of the primordial allele frequency spectrum. Therefore, our primordial diversity SCO model of human origins is similar but slightly different from an evolutionary bottleneck (BTSC) model (2) of human origins.

4. METHODS

4.1 Model Parameters

We restricted our investigations to homogeneous populations. We chose a generation time of 20 years, and a germline mutation rate of 1.6×10^{-8} per nucleotide per generation, or 48 per haploid genome per generation, with a genome of size 3×10^9 bases. The literature gives a wide range of estimates for the germline mutation rate ranging from 1×10^{-8} to 2.5×10^{-8} per nucleotide per generation [42–53]. For the recombination rate we used a value of 1.0×10^{-8} per nucleotide per generation [54, 55].

4.2 Simulation Methods

We implemented four different simulation methods. The primary simulation code is called Haplo - an implementation of the simulation model described by Hössjer, Gauger and Reeves [20, 21]. Haplo is the only code we used that can calculate linkage disequilibrium statistics.

The other three are simpler codes for calculating the allele frequency spectrum. They are described in the appendices: a Stochastic Forward Simulation (Appendix A), a Matrix Forward Calculation Method (Appendix B) and Coaly - an Approximate Coalescence Backwards Calculation Method (Appendix C). They are in large parts the work of Andrew Jones. Of the three smaller codes, the stochastic simulation method generates allelic diversity, forward in time, for a haploid Wright-Fisher population of time-varying size. Among the three algorithms in the appendix, it is the one that most closely approximates what really happens in an evolving population. However, it takes very long to compute results for large populations, large timescales or large mutation rates. It turns out that tiny biases in the random number generator could accumulate over very long simulations. To supplement this, Coaly was invented as a very efficient and approximate method for computing the expected AFS, recursively back in time. It can be used to test ideas and get qualitative results quickly, but it is inaccurate close to a bottleneck, or close to the time of the common ancestor. To bridge that gap between Stochastic Forward Simulation and Coaly, Matrix Forward Simulation was implemented in order to eliminate the stochastic element of the simulation. It calculates expected values of the AFS forward in time, using transition probabilities of the Wright-Fisher model. The Matrix method is most accurate but not as fast as Coaly and particularly struggles with very large population sizes.

Haplo and Coaly can be found at the two web addresses https://github.com/DiscoveryInstitute/coaly and https://github.com/DiscoveryInstitute/haplo. Coaly also contains an implementation of the Stochastic method and the Matrix method.

5. RESULTS

5.1 Evolution of the Allele Frequency Spectrum

One possible model of human origins is a couple instantiated with perfectly homozygous genomes - that is, without any primordial diversity - and have slowly been accumulating mutations ever since. The left column of Figure 7 shows what this scenario would look like at various time points. The middle column of Figure 7 shows the distribution of mutations from different times in the past. Notice that the newest mutations appear on the left, meaning they are the rarest mutations in the population, and the oldest mutations are most spread out, ranging from rare to very common in the population.

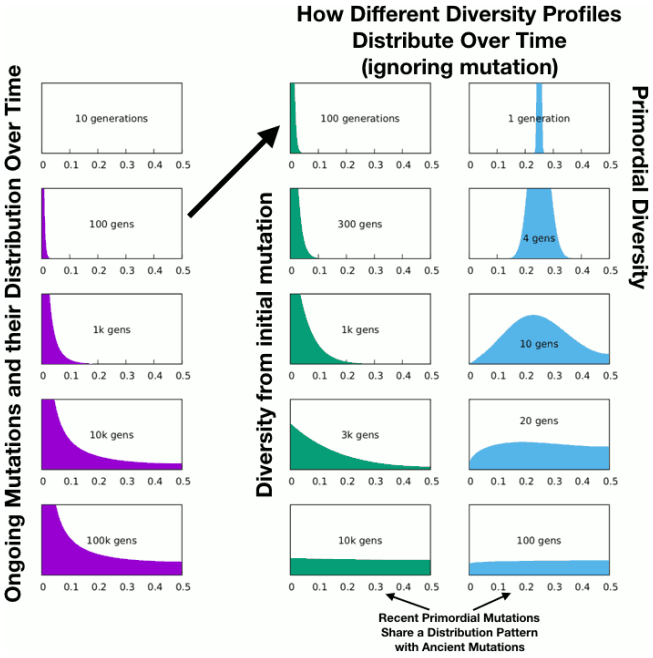

Figure 7: Time-evolution of folded allele frequency spectrum: mutations vs primordial variants. Folded allele frequency spectra (AFS) showing number density of alleles vs minor allele frequency. In the three columns of the figure, the time-evolution of the form of the AFS is compared for three different scenarios. The vertical axes have been scaled in each subplot, to make it easier to compare the shapes of different curves. The folding or symmetrization at f=0.5 takes into account the potential ambiguity as to whether the current major variant is the ancestral variant (or the primordial major variant) or not. The left column (purple) shows how mutations accumulate over time to form the folded allele frequency spectrum. Each mutation variant starts as a single copy; the minimum possible frequency at the left of the graph. Most variants quickly go extinct but a random few drift rightwards and become established in the population, while new mutations accumulate to the left. Eventually the AFS reaches equilibrium where the number of new mutations is matched by the number of variants that go extinct or fix in the population. The population size is 10,000 (the effective population size of the human origins is believed to be in this ballpark [56]) and the mutation rate is 48 per haploid genome per generation. The middle and right columns show how the distribution of existing variation changes over time (that is, the impact of new mutations is excluded). The middle column (green) shows how mutations change in frequency over time by genetic drift in a population of 10,000. The right column (blue) shows how primordial diversity would change over time in a population growing sigmoidally from 4 to 10,000 over the course of the given number of generations in each graph, so that there is a different growth rate in each graph. Genetic drift occurs much faster in small population and so the degree of genetic drift has an inverse dependence on how quickly the population grows. See also [36] where genetic diversity is initialised at 0.5 instead of 0.25. Well mixed primordial diversity looks similar to ancient mutational diversity; since the AFS of existing variants evolves towards a flat distribution in either case. See also [57].
doi:10.5048/BIO-C.2019.1.f7

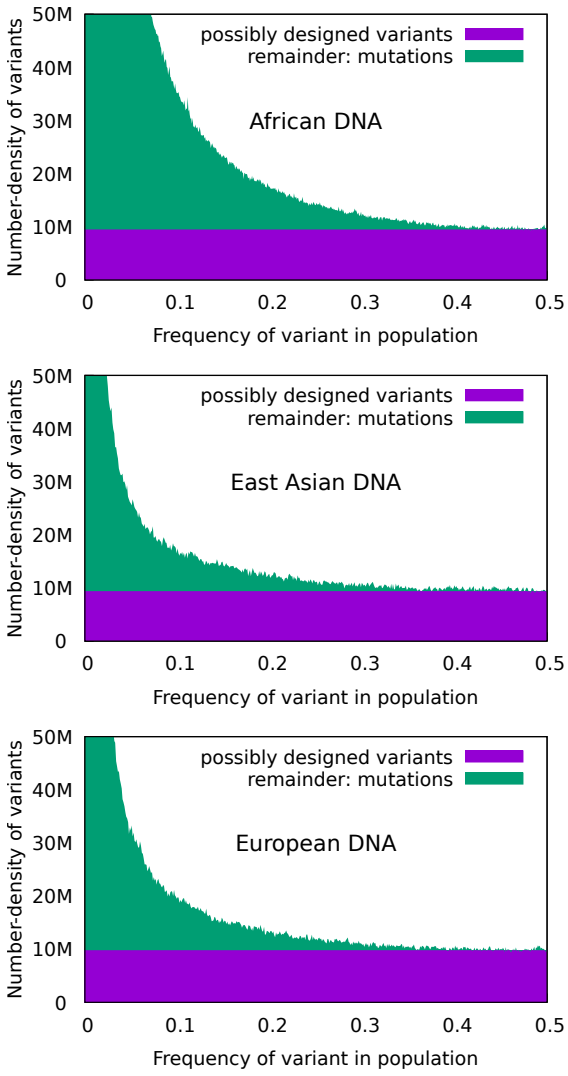

Figure 8: Visual estimates of which genetic variants *could* be primordial in three global superpopulations, on the assumption that the final distribution of primordial variants will be approximately flat. African DNA exhibits a lot more variation in total than for the other populations, but curiously does not show more of the oldest (possibly primordially diverse) variants. All three superpopulations appear to have roughly the same number of the oldest variants.
doi:10.5048/BIO-C.2019.1.f8

Another possible model (SCO) is a couple with four heterogeneous versions of each non-sex chromosome, each having primordially diverse variants. The right column of Figure 7 shows how this variation would become distributed over time. This process can happen in just a few generations, because genetic drift is fastest when the population is still small. Notice that if enough genetic drift has happened, the distribution of primordially diverse variants ends up looking similar to the distribution of very ancient mutations that have not yet been fixed or removed. However, due to the large population size in the middle column of Figure 7, it takes much longer for the ancient mutations to reach the flat (quasi) equilibrium distribution.

There is a caveat to be noted here: the resulting flat distribution of primordial diversity depends strongly on the assumption that the population is homogeneous and growing uniformly. It is easy to imagine alternative scenarios which would give different distributions. Firstly the population splits into several subpopulations which experience different growth patterns, or waves of migration across continents, or a single wave of colonisation. The latter is like a series of bottlenecks at the front, leaving more and more variation behind. Secondly, for a non-neutral model with directional selection with selective sweeps (cf. Section 6.1 of [26]), one of the two alleles of some SNPs will drift towards fixation. This will cause that part of the AFS that originates from primordial variation to be skewed towards the left.

5.2 Possibly-Primordial Variation Visual Estimate

If we assume that the distribution of primordial variation would be approximately flat, and that the allele frequency spectrum is non-increasing, we can infer the portion of the AFS that could be due to primordial variation based on the number-density AFS(0.5) at $f = 0.5$ (except for the caveat mentioned above that this only applies to homogeneous and selectively neutral populations). Figure 8 shows this breakdown for each of the three global superpopulations. Curiously, since AFS(0.5) $\approx 10M$ for all three populations, they have approximately the same number 5M of possibly primordial variants. This number is obtained by summing the purple part of the AFS over all frequencies ($\int_{f=0.0}^{0.5} 10\text{M}\,\mathrm{d}f = 5M$) in each subplot of Figure 8.

5.3 Two Human History Scenarios Simulated

We present two different diploid parsimonious scenarios that fit the African DNA data reasonably well (see Figures 9-10), both in terms of allele frequency spectra and linkage disequilibrium statistics. We fit to African DNA because Africans have greater heterozygosity, and this might correspond to a more ancient past.

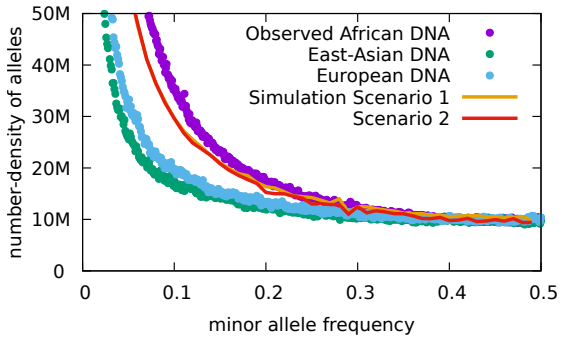

Figure 9: Folded allele frequency spectrum: Simulation results from Scenarios 1 and 2, compared to actual human genetic data from three global superpopulations. The two very different scenarios give similar results, both in good agreement with data from the African superpopulation. Simulations were performed using the full Haplo algorithm. A genome of length 3×10^9bp was simulated in 60 equal chunks. **doi:**10.5048/BIO-C.2019.1.f9

Scenario 1 A single couple 100,000 generations ago (about 2mya) having zero heterozygosity (identical homozygous chromosomes), grows rapidly to a population of 10,000 people, then grows slowly and linearly to 16,000 people near the present.

Scenario 2 A single couple 25,000 generations ago (about 500kya) having primordial heterozygosity of 0.012,[5] grows rapidly to a population of 16,000 people, then holds steady.

Scenarios 1 and 2 are both instances of a Single Couple Origin (SCO) model, without or with primordial diversity. The parameters of Scenarios 1 and 2 were based on their fits to the allele frequency spectrum, using the Haplo algorithm. Then these parameters were used to fit Scenarios 1 and 2 for the linkage disequilibrium plots as well. For both scenarios, the population doubles to four in the first generation, and then doubles every ten generations until it reaches the specified plateau. The real life human population has increased in size recently and is now close to 8 billions, but this makes only a very small difference to the extreme left of the allele frequency spectrum and for simplicity we did not include this in the model. Both models were kept as simple as possible, with a focus on explaining the right end of the AFS, which is more due to ancient history,[6] and not overfitting the left end of the distribution. It is likely that there are many other scenarios that would explain the data equally well.

[5]In Scenario 2, most of the primordial heterozygosity is lost in the first few generations due to the slow growth rate. It is also possible to imagine scenarios with lower primordial diversity compensated for by a faster growth rate.

[6]Under the parsimonity assumption of a single homogeneous population.

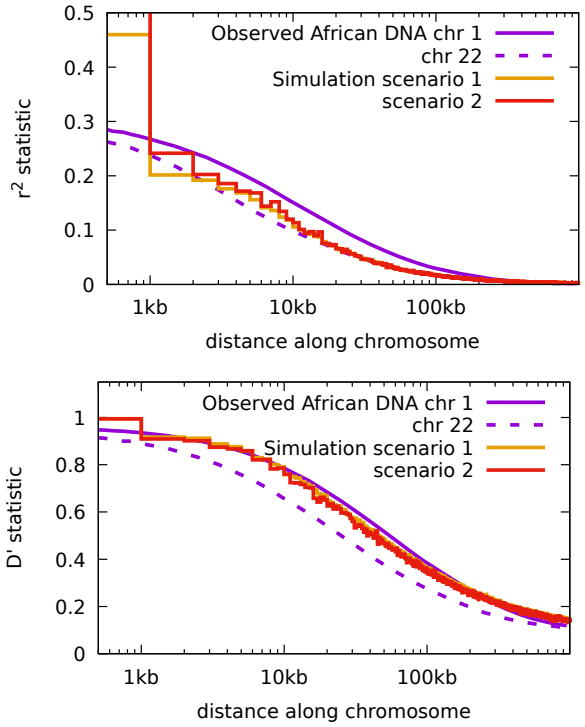

Figure 10: Linkage disequilibrium (LD) statistics r^2 (top) and D' (bottom): Simulation results from Scenarios 1 and 2, compared to actual human genetic data from the African superpopulation. The two very different scenarios give similar results, both in good agreement with data. Simulations were performed using full Haplo algorithm. A section of chromosome of length 50Mbp was simulated. Variants with minor frequency less than 0.05 were excluded from the analysis. **doi:**10.5048/BIO-C.2019.1.f10

5.4 Effective Population Size in the Allele Frequency Spectrum

Given the parsimonity assumption of a single homogeneous population, it is possible to estimate population size from fitting to the allele frequency spectrum [37]. Figure 11 shows why the AFS of African DNA suggests the very slowly growing population of Scenario 1: The left hand side of the AFS corresponds to more recent times. The observed and simulated AFS are then closer to the equilibrium for a population of effective size 16,000 - the more recent population size of Scenario 1. On the right hand side, the AFS corresponds to time points further back. The observed and simulated AFS are then closer to the equilibrium for a population of effective size 10,000, the ancient population size of Scenario 1. Notice however that the fit is not perfect between either of the two dashed curves and the left or right parts of the African AFS. This is the reason why we included this graph, to illustrate how the frequency of alleles changes over time is a consequence of population size dynamics.

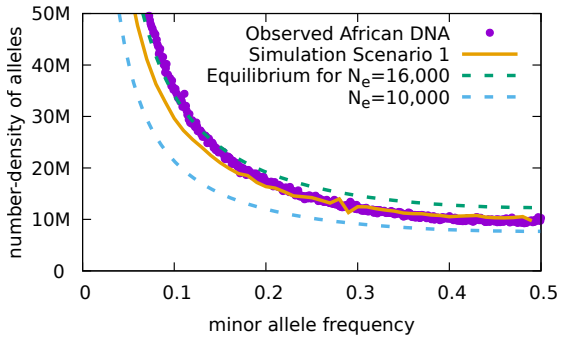

Figure 11: Folded allele frequency spectrum of Scenario 1 shows evidence of the slow change in diploid population size from 10,000 to 16,000. The Scenario 1 data set is the same used in Figure 9. The equilibrium distributions (blue and green dashed) are theoretical results, based on an effective population size N_e of 10,000 and 16,000 respectively. **doi:**10.5048/BIO-C.2019.1.f11

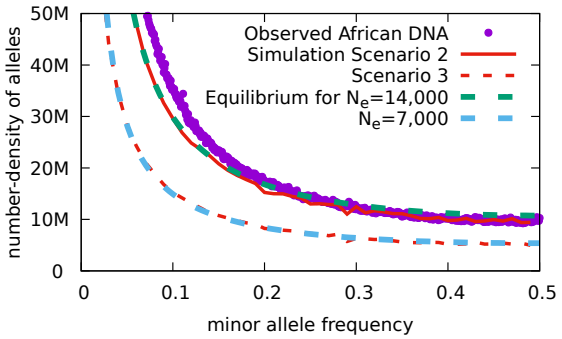

Figure 12: Simulation results showing the effect of mating patterns on effective population size. Scenario 3 is identical to Scenario 2 except for non-random mating; it has identical population sizes, but exhibits a lower *effective* population size N_e. The Scenario 2 data set is the same used in Figure 9. Scenario 3 (red dashed) was simulated using the full Haplo algorithm in a similar way: A genome of length 3×10^9 bp in 60 equal chunks. The equilibrium distributions (green and blue dashed) are theoretical results. **doi:**10.5048/BIO-C.2019.1.f12

5.5 Why Models can be Underdetermined

A definite population history cannot always be easily determined from the genetic data for the simple reason that different causes can sometimes produce the same results. The following subsections give several examples of this. These are meant to be illustrative, not exhaustive.

5.5.1 Mating Patterns Compensate for Population Size

The full simulation algorithm allows for simulation of non-random mating patterns. There are two parameters α and β which were explained in detail in [21, 58, 59]. In Scenario 2 they are set to $\alpha{=}\infty$, $\beta{=}\infty$, which means completely random relationships between parents and children, with the consequence that the effective population size and actual population size are equal. Scenario 3 is identical to Scenario 2 except that $\alpha{=}2$, which means some women are more likely to have children than others, and $\beta{=}0$, which means each woman mates with only one man.

Figure 12 shows that Scenario 2 yields an AFS similar to the equilibrium expected AFS for a population of 14,000, whereas Scenario 3 yields an AFS similar to the equilibrium expected AFS for a population of only 7,000. This demonstrates that non-random mating has reduced the effective population size by half (the precise factor will depend on the values of α and β). This suggests that, since human reproduction is not at all random in reality, the effective population sizes used elsewhere in the paper might underestimate the actual population sizes.

5.5.2 Primordial Genetic Diversity Compensates for Shorter Timescale

To accumulate genetic diversity by mutation alone takes a long time, see Figure 13. Recall that Scenario 1 and 2 give similar AFS, even though Scenario 1 is 2mya and Scenario 2 is 500kya. The reason is that Scenario 2 has primordial diversity. If one takes away primordial diversity there is a major shift in the AFS that must be compensated by a longer time frame.

5.5.3 Mutation Rate Compensates for Population Size and Timescale

The mutation rate is generally assumed to be a known constant, and the population size is assumed to be variable over time. If we were to allow for the possibility of non-constant mutation rates instead, that would make population sizes and timescales much more uncertain.

Figure 14 illustrates that the shape of the allele frequency spectrum depends on the amount of genetic drift that occurs, and the rate of genetic drift depends on the population size. Therefore reducing the population size reduces the timescale required to get the right shape of AFS. The height or density of the allele frequency spectrum also depends on the amount of mutation. Therefore a larger (smaller) mutation rate reduces (increases) the timescale to get the right shape of the AFS. Figure 14 shows the effect of a smaller population size and a shorter time, and second the effect of a larger mutation rate with smaller population size and shorter time.

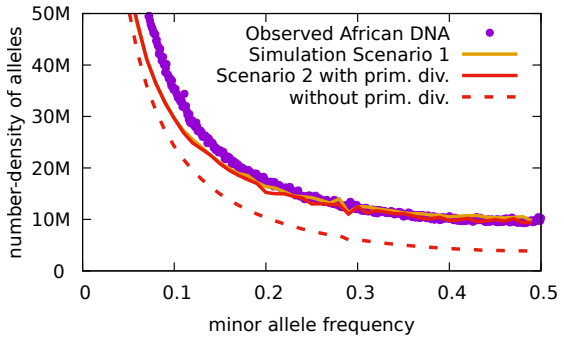

Figure 13: Simulation results showing effect of primordial diversity in Scenario 2. In Scenario 2 primordial diversity makes a significant difference. In Scenario 1 the same distribution of genetic variation is made up by more time accumulating mutations with genetic drift. The Scenario 1 and 2 data sets are the same used in Figure 9. Scenario 2 without primordial diversity (red dashed) was calculated using the Matrix method. **doi:**10.5048/BIO-C.2019.1.f13

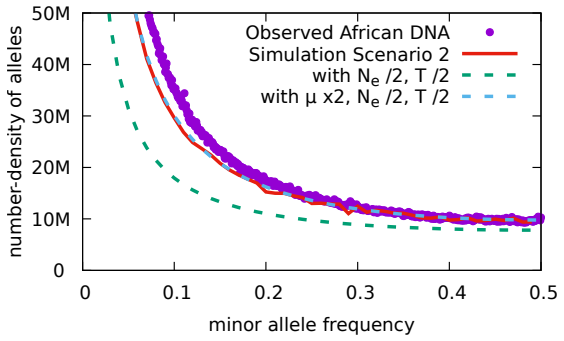

Figure 14: Simulation results showing the effect of population size and mutation rate on Scenario 2. The green dashed line shows the effect of the effective population size N_e and time T since the origin both reduced by a factor of 2. The light blue dashed line is same as green, but with a compensating effect of increasing the mutation rate μ by a factor of 2. The Scenario 2 data set is the same used in Figure 9, using the Haplo method. The green and blue lines were generated using the Matrix method. **doi:**10.5048/BIO-C.2019.1.f14

6. DISCUSSION

The simulation method for human origins proposed by Hössjer, Gauger, and Reeves [20, 21] has now been implemented. A number of simpler models were developed (Appendices A-C) for exploratory research and to validate the full model (Appendices D-E).

We then performed two large simulations, effectively simulating an entire genome in a population of up to 16,000 people for up to 100,000 generations. This demonstrates that human genetic data (at least as summarised in the allele frequency spectrum and simple linkage disequilibrium statistics) from non-sex chromosomes is consistent with at least two different but parsimonious models of human origins from a single couple. The model without diversity of the first couple dates to about 2mya ago, whereas the model with primordial diversity has a first couple that lived about 500kya ago. Thus, we show that using assumptions commonly used by evolutionary geneticists, a single-couple origin is possible, despite claims to the contrary.

A general principle of model selection is to choose a parsimonious model that explains data as well as possible, see for instance [60] and references therein. The intent of this paper is to make a limited point in the most forceful way: that a single-couple origin is certainly possible. This conclusion is made more forceful by the fact that our models are very simple and parsimonious, yet match the observed data very closely. The SCO model we consider differs from the prevailing interpretation of human ancestry in only two assumptions: (i) That we evolved continuously from non-humans in a large population. (ii) That genetic diversity is always due to germline mutations. This raises at least two questions, firstly whether it is possible to discriminate between models where humanity descended from a large population and a first couple, and secondly whether including additional considerations in the model would bring the first couple forward in time. These are the topics of the next two subsections.

6.1 Discrimination Between the Single Couple Origin and the Large Population Ancestry Models

Although typing of ancient DNA has exploded in recent years [61], population genetics still faces the problem of reconstructing human ancestry, mostly from current data. This makes it possible for several quite different demographic models of human history to fit genomic data equally well. In particular, we showed in this paper that a unique origin model of humanity fits allele frequency spectra and linkage disequilibrium plots at least as well as a model for which humanity descends from a large population. Is it possible then to discriminate between these two models, using only genetic data? This is discussed in detail in Section 3 of [20], but here we highlight four possibilities. The first option is to include genetic data from Neanderthals and Denisovans, which have intermixed with the ancestors of humans alive today [62–64]. The task is then to compare models where these archaic populations descend from a founding human couple or not. The second option is to study more carefully the consequences of inbreeding depression (see Sections 2.3 and 3.2) between a unique origin model where primordial diversity of the first couple is selectively neutral, and a model where humanity descends from a large population in which individuals have many deleterious or slightly deleterious variants passed on from their ancestors [41, 65]. The third option is to consider ancestry of DNA without recombination, from mitochondria and Y-chromosomes, where one single family tree is reconstructed, for females and males respectively [18, 66–69]. These female and male trees will only go back to the most recent common ancestor (MRCA), whether

a unique origin model of humanity holds or not. It is however possible to test a single origin hypothesis by comparing ages of the female and male MRCAs.

The fourth option is to study a first couple hypothesis using haplotype blocks of recombining DNA. The objective would be to test whether these blocks can be clustered into four and three groups, for non-sex chromosomes and X-chromosomes respectively [70–72]. The issue of haplotype blocks is of particular interest for the HLA-DRB1 gene of the major compatibility complex of chromosome 6, where previous research indicates perhaps as few as four ancestral lineages [73–76]. In fact, the entire HLA complex could be examined to see if no more than four haplotype blocks are present in any given genome segment. The HLA complex is one of the most polymorphic regions of the human genome and as such an excellent test for a unique origin hypothesis of humanity. However, it is also one of the most complex, with gene conversion on one hand, and stretches where recombinations are suppressed or its products are lethal on the other, complicating the evaluation of haplotype blocks. One must also consider the evidence of shared sequence with other primates, though there is evidence that HLA genes are subject to convergent evolution [77, 78]. In order to test whether or not the first pair hypothesis is correct, at any position along the HLA complex, all lineages should coalesce to at most four ancestral lineages [75].

6.2 Extensions of the Single Couple Origin Model

It is very likely possible to extend our first couple origin model of humanity in order to find more complex scenarios in which the data is compatible with a more recent origin, but we deliberately circumvented fine tuning of the model to data. We avoided introducing any additional hypotheses, beyond the idea of a single-couple origin with or without primordial diversity. Let us however mention three possible extensions of our model that possibly could warrant a more recent time point of the founding couple. First, the most obvious extension of our model is to generalize the parameter that (apart from primordial diversity) determines the timescale - the germline mutation rate. Our models assume a constant mutation rate and we use values that have been estimated in specific populations in the present [42–53]. But the mutation rate is also known to be variable, with distinct families and populations exhibiting different rates of mutation [44, 79–81]. Mutator lineages exist in many species, including mice. It is even logically possible there exists some kind of targeted adaptive mutagenesis, and this could also skew the distribution of alleles. It should be noted however that variable mutation rates impose challenges, since they are easily confounded with other demographic parameters.

As a second extension, it is worth noting that the human population has probably been very non-homogeneous, with several more or less interconnected subpopulations, which could skew the distribution of alleles. Population subdivision has in fact been accounted for in common descent models of human history [82], and the same could be done for a Single Couple Origin model. This would add several parameters to our SCO model that capture geographic movements, such as colonization of new regions and occasional migration between partially isolated populations. In this context it is of interest to fit not only the allele frequency spectrum of the metapopulation, but also the multipopulation allele frequency spectrum of several subpopulations simultaneously [83]. It is also worth noting that population substructure and 'stirring' effects are well suited for phylogenetic reconstruction, including algorithms such as the ARGweaver [84, 85].

The third and possibly most promising extension of our model is to include natural selection, most notably directional selection with selective sweeps. Notice however that Haplo is a backward simulation algorithm that does not allow for selection. It is possible that a genomewide forward simulation approach, such as Mendel's Accountant [86], must be employed in order to incorporate directional selection. On one hand the neutral theory of evolution [32] postulates that genetic drift of neutral variants is more important than fixation of advantageous alleles, in particular when genomewide statistics such as allele frequency spectra and linkage disequilibrium statistics, are used. On the other hand, this neutral view of evolution has recently been challenged [87–89]. Models with directional selection are particularly relevant if the first single pair lived quite recently, since then there are fewer recombinations in human history. As this increases the impact of selective sweeps to drive more alleles towards fixation, we conjecture that directional selection will increase the number of SNPs with a fairly small minor allele frequency, so that the left part of the allele frequency spectrum is elevated. Interestingly, it was the left part of the AFS that gave the largest mismatch between our parsimonious model of human ancestry and data, when the first couple was chosen more recently than for Scenario 2 (500 kya ago).

It is beyond the remit of this paper to explore these three and other possible expansions of our parsimonious single-couple origin model for humanity. But in light of the many possible extensions, we suggest that it is possible to fit a model to genetic data, for which the founding couple lived 100kya ago or even more recently. In any case, the critical point that we wish to make is that, as far as we know scientifically from the genetic data, the human species could have come from as a single couple, so that all humans alive today could have descended uniquely from that first pair.

APPENDIX A. STOCHASTIC SIMULATION

In this appendix we describe a simple but inefficient algorithm for generating the allele frequency spectrum using forward simulation. In a forward simulation all individuals are included, not just the ancestral ones.

A.1 Reproduction

Assume that generations are numbered $t = 0, 1, \ldots, T$, where $t = 0$ is the present and $t = T$ the founder generation. Consider a haploid population of size N_t at generation t, reproducing like the Wright-Fisher model, and let $n_{a,t}$ be the number of copies of allele a at generation t. The probability that an individual in generation t is a child of a parent having allele a in generation $t+1$ is

$$p \equiv n_{a,t+1}/N_{t+1}. \tag{3}$$

The expected number of copies of that allele in generation t is

$$\bar{n}_{a,t} = N_t\, p = n_{a,t+1} N_t / N_{t+1}, \tag{4}$$

but the actual number of copies $n_{a,t}$ is sampled from a binomial distribution with parameters $n = N_t$ and $p = n_{a,t+1}/N_{t+1}$, using standard code libraries. Sampling from successive random distributions simulates the process of genetic drift. If the number of copies falls to zero, the allele is deleted from the simulation. This is what happens to most alleles under most circumstances; they go extinct and are thus non-ancestral to the final sample. If the number of copies reaches N_t, which happens less often, then the allele is fixed, and it can also be deleted from the simulation.

A.2 Mutation

Suppose that the mutation rate per individual per generation is μ, and that the infinite sites model holds [90]. By this we mean a model where every mutation is unique; the chance of a mutation happening at the site of a previous mutation is zero (one in infinity). This model is good enough for many purposes, and is also used in the full simulation code Haplo. Since each new mutation produces a new allele, the expected number of new alleles in generation t is μN_t. The actual number of new alleles is sampled from a Poisson distribution with parameter μN_t. This number of new alleles is added at generation t, each as one copy: $n_{a,t} = 1$.

A.3 Founding Diversity

The founding generation T is at the start of the simulation. Suppose the allele frequency spectrum of the founding population is a vector $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_{N_T-1})$, if all alleles are recorded, or $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_{N_T/2})$, if only alleles with the minor frequency are recorded. In any case, simply add ν_1 new alleles a, each with frequency $n_{a,T} = 1$, then ν_2 new alleles each with frequency $n_{a,T} = 2$, and so on.

A.4 Sampling

At the end of the simulation, the variants must be sampled without replacement to see how many in the final total population make it into the final sample population. This is implemented at the final generation step, by replacing the true population size N_0 of the present generation with the sample size S. Likewise, the true frequency $n_{a,0}$ of allele a at time $t = 0$ is replaced by the number of copies of a in the sample.

A.5 AFS - Approximate Continuous Function

The allele frequency spectrum is taken from the distribution of alleles at t=0, after the sampling, by summing the number of alleles a at each frequency from 1 to $S-1$. The number of alleles with frequency i is

$$\mathrm{AFS}_i \;=\; \sum_a \delta\left(n_{a,0}, i\right), \tag{5}$$

where

$$\delta(n,i) = \begin{cases} 1, & n = i, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Finally, transform this distribution into a continuous function $\mathrm{AFS}(f) = S \cdot \mathrm{AFS}_i$, defined for frequencies $f = i/S$ between 0 and 1. Since $\mathrm{d}f = \mathrm{d}i/S$, it follows that

$$\textstyle\sum_{i=1}^{S} \mathrm{AFS}_i = \int_0^1 \mathrm{d}f \mathrm{AFS}(f). \tag{7}$$

The program outputs the results in the following form:

f	$\mathrm{AFS}(f)$
$1/S$	$S \times \mathrm{AFS}_1$
$2/S$	$S \times \mathrm{AFS}_2$
$\cdots$	
$(S-1)/S$	$S \times \mathrm{AFS}_{S-1}$

Suppose summation in (5) is taken over *all* alleles a at all SNPs. Then the allele frequency spectrum (5) and (7) is *not* the unfolded spectrum. If also all polymorphic sites are biallelic, it follows that $\mathrm{AFS}(f)$ is symmetric around $f = 0.5$. The folded spectrum of Figure 1 is then simply the restriction of $\mathrm{AFS}(f)$ to the interval $(0, 0.5]$. On the other hand, if summation in (5) is over *all derived* alleles a, we obtain the unfolded spectrum. The folded spectrum of Figure 1 then equals $\mathrm{AFS}(f) + \mathrm{AFS}(1-f)$ for all frequencies $f \in (0, 0.5]$. As mentioned below Figure 1, the folded spectrum is a number-density such that its integral over any subset of $(0, 0.5]$ gives the number of SNPs with minor allele frequency within that set. It turns out that $\mathrm{AFS}(f)$, as well as the folded spectrum, is more or less independent of the sample size $S \leq N_0$ for frequencies f not very close to 0. For instance, in Section 3.2 we indicate that a diploid Wright-Fisher model, with effective size N_e and mutation rate μ, has an unfolded spectrum $\mathrm{AFS}(f) = 4N_e\mu/f$ when equilibrium between genetic drift and mutations has been attained, independently of sample size S.

APPENDIX B. MATRIX METHOD FOR CALCULATING EXPECTED AFS

Consider the same haploid Wright-Fisher model as in Appendix A. The algorithm in this appendix gives the expected (that is, average) allele frequency spectrum exactly with better efficiency and better numerical accuracy than the stochastic approach of Appendix A. Instead of storing an actual number of copies for each allele separately, we store the expected number of alleles that should have that number of copies; that is, we store the expected AFS at each generation. This AFS is propagated forward through the generations using a binomial distribution transition matrix instead of binomial sampling [91].

B.1 Reproduction

Let $m_{i,t}$ be the expected number of alleles that have i copies at time t, where $0 \leq i \leq N_t$. These can be calculated from the preceding generation $t+1$ as follows:

$$m_{i,t} \quad = \quad \sum_{j=0}^{N_{t+1}} m_{j,t+1} \, \frac{N_t!}{i!(N_t-i)!} \left(\frac{j}{N_{t+1}}\right)^i \left(1-\frac{j}{N_{t+1}}\right)^{N_t-i} \tag{8}$$

for $i = 0, 1, \ldots, N_t$. Note that the transition matrix is a set of $N_{t+1}+1$ binomial distributions having $n = N_t$ trials with probability of success $p = j/N_{t+1}$, for $j = 0, 1, \ldots, N_{t+1}$. Also note that if $N_t \neq N_{t+1}$, then vector m_t and m_{t+1} have different sizes and the transition matrix is not square. Finally, note that $m_{i=0,t}$ and $m_{i=N_t,t}$ are the total expected number of mutations that have been lost or fixed respectively. They are absorbing states which can be treated a special cases.

B.2 Mutation

Again assuming the infinite sites model [90], new mutations appear as single copies ($i = 1$). The expected number of new alleles is μN_t. Adding this number to (8), we update the expected AFS according to

$$m_{i=1,t} \quad \leftarrow \quad m_{i=1,t} + \mu N_t. \tag{9}$$

B.3 Founding Diversity

At the start of the simulation, at the founding generation T, the allele frequency spectrum m is simply the founding diversity $\boldsymbol{\nu}$, i.e.

$$m_{i,T} = \nu_i. \tag{10}$$

B.4 Sampling

As in the stochastic simulation, at the end of the forward simulation, the allele frequency spectrum must be sampled from the final total population. This is again implemented at the final Generation step, by replacing the true population size N_0 with the sample size S. The allele frequency spectrum is transformed in the same way as described in Section A.5.

APPENDIX C. COALY: APPROXIMATE COALESCENT BACKWARD METHOD FOR ESTIMATING EXPECTED AFS

Assume that the haploid Wright-Fisher model of Appendices A-B holds. In this appendix we present an approximate but very fast method of calculating the expected AFS for such a model. Recall that t is the generation back in time from the present, and that N_t is the size of the population at time t. A sample is taken from the population at the present time, so let $S \leq N_0$ be the size of that sample. Let A_t be the size of the *ancestral* population at time t; that is, the number of individuals who have descendants in the sample and are thus ancestors of the sample. The number of 'ancestors' in the extant generation is just $A_0 = S$ but beyond this they may diverge, with the ancestral population shrinking as lineages coalesce.

C.1 Number of Ancestors

The expected number of ancestral children per individual living at time $t+1$ is

$$\lambda \equiv A_t/N_{t+1}. \tag{11}$$

The number of ancestors at time $t+1$ is the number of parents who have ancestral children. If we assume the relationships between parents and children are independent and random, then the probability P_n that any particular individual at time $t+1$ will have n ancestral children can be approximated by a Poisson distribution (this is not quite accurate for small populations) with expected value λ. That is,

$$P_n = \tfrac{\lambda^n}{n!} e^{-\lambda}, \tag{12}$$

for $n = 0, 1, 2, \ldots$. The probability of any particular individual having children is $1-P_0 = 1-e^{-\lambda}$, and thus the expected number of ancestors in the parental generation $t+1$ is

$$A_{t+1} = N_{t+1}\,(1 - e^{-\lambda}). \tag{13}$$

This expression can be used iteratively to calculate the ancestral population at each generation back from the present. If we expand the expression in a Taylor series, we get a close approximation to a familiar expression for pairwise coalescence [92]:

$$A_{t+1} = A_t - \frac{A_t^2}{2N_{t+1}} + \mathcal{O}(\lambda^3). \tag{14}$$

However, we can go beyond pairwise to multiple coalescence in a single generation.

C.2 Number of Descendants per Ancestor

Define $\omega_{t,n}$ to be the probability that an *ancestral* individual in generation t is ancestor to n individuals in the

sample, for $n = 0, 1, \ldots, S$. In particular, in the present generation each individual is 'ancestral' to only itself:

$$\omega_{0,n} = \begin{cases} 1 & n = 1 \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

Now define the coalescence probabilities C_n which are equal to the probability of an ancestral individual at time $t+1$ having n children, which is the probability of any individual having n ancestral children given that he/she has any at all:

$$C_n = \frac{P_n}{1-P_0} = \frac{\lambda^n}{n!}\frac{e^{-\lambda}}{1-e^{-\lambda}}. \tag{16}$$

Now we make a simplifying assumption that the different ancestral individuals of the same generation have an independent number of offspring in the sample. In reality they would be negatively correlated, but the more ancestral individuals there are, the smaller is this negative correlation and the better this approximation will be. Then $\omega_{t+1,n}$ may be calculated iteratively from $\omega_{t,1}, \ldots, \omega_{t,n}$ by conditioning on the number of ancestral children an individual at time $t+1$ has:

$$\begin{aligned} \omega_{t+1,n} \quad = \quad & C_1\, \omega_{t,n} \\ & + C_2 \textstyle\sum_{i=1}^{n-1} \omega_{t,i}\, \omega_{t,n-i} \\ & + C_3 \textstyle\sum_{i=1}^{n-2} \sum_{j=1}^{n-i-1} \omega_{t,i}\, \omega_{t,j}\, \omega_{t,n-i-j} \\ & + \cdots . \end{aligned} \tag{17}$$

There is a self term, a pairwise coalescent, but then also a threewise coalescent and more. This is a computationally expensive polynomial sum, but it can be made more efficient. The first strategy is to define a weight vector that can be calculated iteratively:

$$\omega_{t,n}^{(k)} \quad \equiv \quad \begin{cases} \omega_{t,n}, & k = 1, \\ \sum_{i=1}^{n-1} \omega_{t,i}\, \omega_{t,n-i}^{(k-1)}, & 1 < k \leqslant n, \\ 0, & \text{otherwise.} \end{cases} \tag{18}$$

Then the sum in equation (17) simplifies to

$$\omega_{t+1,n} = C_1\, \omega_{t,n}^{(1)} + C_2\, \omega_{t,n}^{(2)} + \cdots + C_n\, \omega_{t,n}^{(n)}. \tag{19}$$

The second strategy is to truncate the summation when the size of each term approaches the machine precision. Finally, noting that the weights do not change greatly from generation to generation, a third efficiency strategy is to not update all the weights every generation, but instead multiply only the small number of coalescent probabilities, and then apply those to the full set of weights only when a certain threshold is reached. The scheme to do this is analogous to equation (17) and can itself be accelerated by analogy to equations (18)-(19). Together, these strategies speed up the calculation by several orders of magnitude.

C.3 AFS - Mutational Diversity

If a mutation happens at generation t, it happens in exactly one ancestor, and propagates forward to its descendants. We assume each mutation is unique and happens at a different site (the *infinite sites* approximation). Thus the probability that a particular mutation affects n descendants in the sample is $\omega_{t,n}$. Let μ denote the mutation rate per individual per generation. Then the expected number of mutations at time t that affect exactly n individuals in the sample is $\mu A_t \omega_{t,n}$. From this it follows that the total expected allele frequency spectrum due to mutations is

$$\mathrm{AFS}_n = \textstyle\sum_{t=0}^{T} \mu\, A_t\, \omega_{t,n}\,, \tag{20}$$

for $n = 1, 2, \ldots, S-1$.

C.4 AFS - Founding Diversity

To simulate a population founding event at time T, it will often be the case that N_T is small. In this case, we might want to specify the number of minor alleles that are present in $1, 2, \cdots, N_T/2$ members of the founding population. Let ν_n be the number of alleles that are present in exactly n members of the founding population. However, not all of the founding population is ancestral, so this distribution must be *sampled* to reflect the ancestral population, using the binomial distribution this time because it is more correct for small populations. We find that

$$\nu'_m \quad = \quad \textstyle\sum_n \frac{n!}{m!(n-m!)}\, \lambda^m\, (1-\lambda)^{(n-m)}\, \nu_n, \tag{21}$$

with $\lambda = A_T/N_T$. The ancestral founding diversity at time T propagates through time to the sample generation at time 0. The contribution to the allele frequency spectrum can then be calculated as

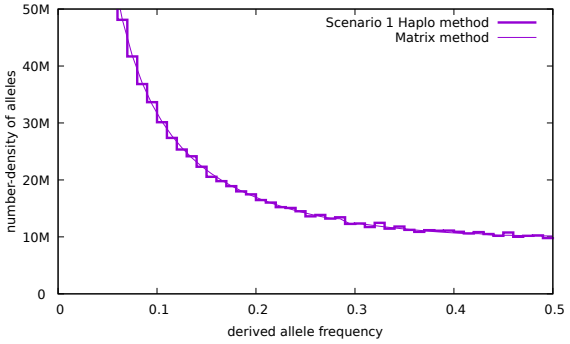

Figure 15: Haplo agrees with the accurate forward matrix calculation method for Scenario 1: A single-couple origin scenario at 100,000 generations ago (about 2mya), where there is no primordial diversity, the population rapidly rises to 10,000 people and then grows linearly to 16,000 people near the present day. The genome-wide mutation rate per individual per generation is $\mu = 48$.
doi:10.5048/BIO-C.2019.1.f15

$$\begin{aligned}\mathrm{AFS}_n \;=\;& \nu_1'\,\omega_{T,n} \\ &+ \nu_2'\,\textstyle\sum_{i=1}^{n-1}\,\omega_{T,i}\,\omega_{T,n-i} \\ &+ \nu_3'\,\textstyle\sum_{i=1}^{n-2}\sum_{j=1}^{n-i-1}\,\omega_{T,i}\,\omega_{T,j}\,\omega_{T,n-i-j} \\ &+ \cdots\,, \end{aligned} \tag{22}$$

for $n = 1, 2, \ldots, S-1$. The ν_1' term is just like mutational diversity, but the ν_2' term is like a pairwise coalescence followed by mutation, and the ν_3' term is like threewise coalescence followed by mutation, and so on. Notice the similarity to equation (17). It can also be accelerated using the scheme in equations (18)-(19). Finally, the allele frequency spectrum (the sum of (20) and (22) is transformed in the same way as described in Appendix A.5, with allele frequencies given on scale $0 < f < 1$.

APPENDIX D. METHOD TESTS FOR AFS COMPUTATION

In this appendix, we compare the results of different methods of calculating of the allele frequency spectrum. The purpose is to validate that Haplo and the Matrix method are working as expected, to demonstrate the speed of Coaly, but also to show the accuracy limitations of the Stochastic Method and Coaly. See Figures 15-20 for details.

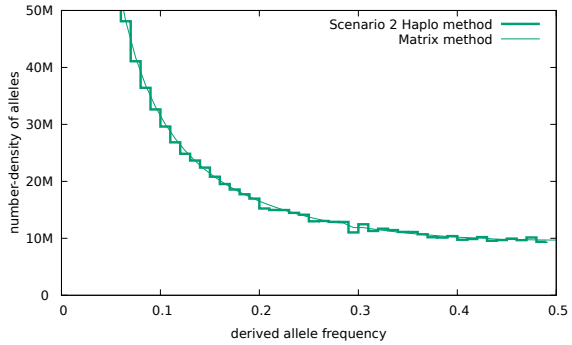

Figure 16: Haplo agrees with the accurate forward matrix calculation method for Scenario 2: A single-couple origin scenario at 25,000 generations ago (about 500kya), where there is primordial diversity with initial heterozygosity of 0.012, the population doubles to 4 in one generation and then doubles every 10 generations until it reaches 16,000 people, then stays constant up to near the present day. The genome-wide mutation rate per individual per generation is $\mu = 48$. **doi:**10.5048/BIO-C.2019.1.f16

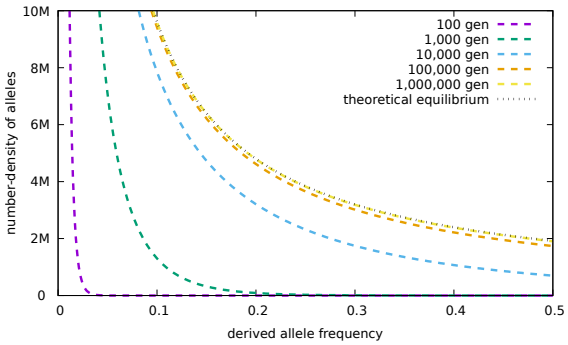

Figure 17: Coaly converges on the theoretical equilibrium distribution [37] for constant population size scenarios. Parameters: effective population size $N_e = 10,000$ and genomewide mutation rate per individual per generation $\mu = 48$. The theoretical equilibrium distribution for this haploid population is $2\mu N_e/f$, where f is the derived allele frequency. There is zero founding diversity in these scenarios. These AFS are not 'folded': they show the final frequency $0 < f < 1$ of alleles that began as minor alleles, although the plot is truncated at $f = 0.5$. **doi:**10.5048/BIO-C.2019.1.f17

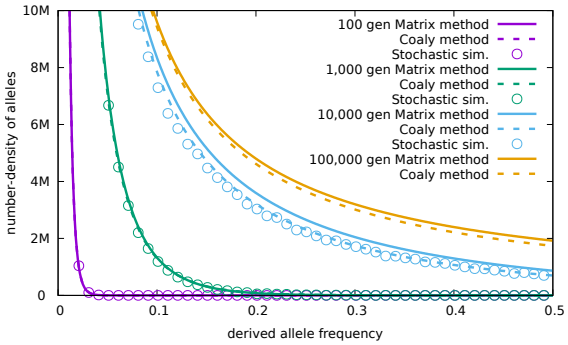

Figure 18: Constant population size scenarios without founding diversity: Coaly roughly agrees with the Stochastic simulation and the Matrix method. Parameters: effective population size $N_e = 10,000$ and genome-wide mutation rate per individual per generation $\mu = 48$. The three methods diverge for long simulations (10,000 gens or more). These AFS are not 'folded': they show the frequency $0 < f < 1$ of the derived allele (not all minor alleles), although the plot is truncated at $f = 0.5$. **doi:**10.5048/BIO-C.2019.1.f18

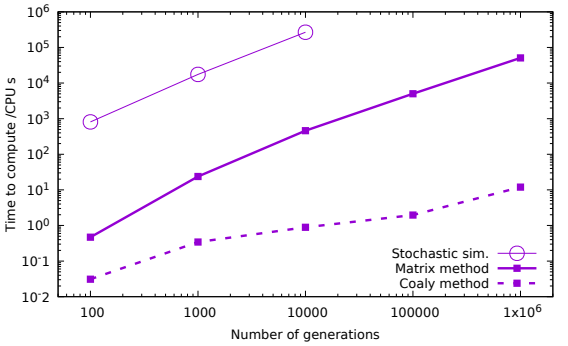

Figure 19: Timing comparisons for the constant population size calculations in Figures 17&18. **doi:**10.5048/BIO-C.2019.1.f19

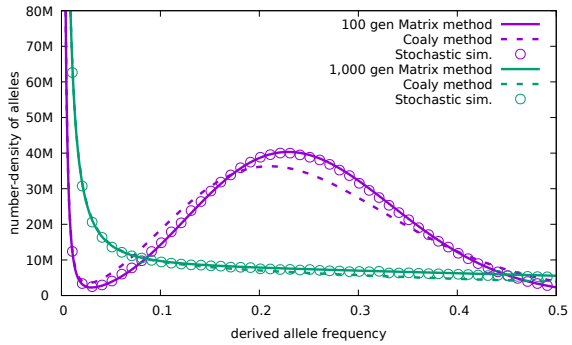

Figure 20: Extreme bottleneck scenarios with founding diversity: Coaly agrees with the Stochastic simulation and the Matrix method. Population size grows linearly from 4 to 10,000, the genome-wide mutation rate per individual per generation is $\mu = 48$, and the founding diversity (variants each present in exactly one of the four founders) is $\nu = 10^7$, corresponding to a heterozygosity of about 0.003. As in Figure 18, the approximations used by Coaly begin to fail when the number of remaining lineages approaches 1. These AFS are not 'folded': they show the final frequency $0 < f < 1$ of alleles that began as minor alleles, although the plot is truncated at $f = 0.5$.
doi:10.5048/BIO-C.2019.1.f20

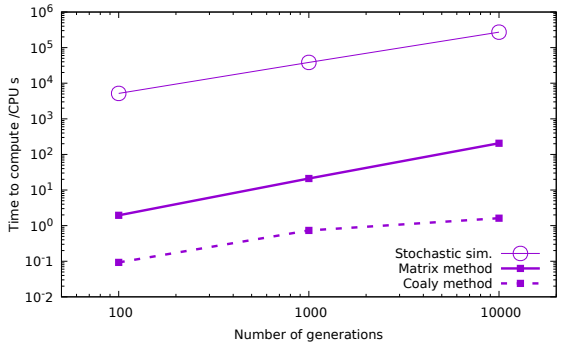

Figure 21: Timing comparisons for the bottleneck calculations in Figure 20. **doi:**10.5048/BIO-C.2019.1.f21

APPENDIX E. METHOD TESTS FOR LD STATISTICS

In this appendix we describe a simple test of the linkage disequilibrium statistics using Haplo. The approximate theoretical expression for the σ^2 statistic at physical distance d along the chromosome in a diploid population of effective size N_e is

$$\sigma^2(d) \;=\; \frac{10+\rho(d)}{22+13\rho(d)+\rho(d)^2}, \qquad (23)$$

where $\rho(d) \equiv 4N_e r d$, see [93, 94] for details. We simulated a constant diploid population of effective size $N_e = 24,000$ for 100,000 generations (about 2mya) with a recombination rate $r = 1\times10^{-8}$ and a mutation rate $\mu = 1.6\times10^{-8}$ per nucleotide per generation, on a section of chromosome of length 50Mbp.

From Figure 22 we see that this analytical expression for σ^2 agrees well with simulated values. No corresponding formulas exist for the expected values of r^2 and D', but for comparison, simulated averages of these LD statistics are shown in Figures 23-24.

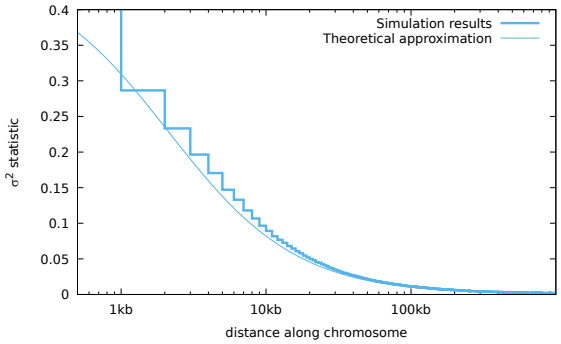

Figure 22: Haplo roughly agrees with theoretical approximation for the σ^2 linkage disequilibrium (LD) statistic. The population scenario is described in the text. Variants with minor frequency less than 0.05 were excluded from the analysis.
doi:10.5048/BIO-C.2019.1.f22

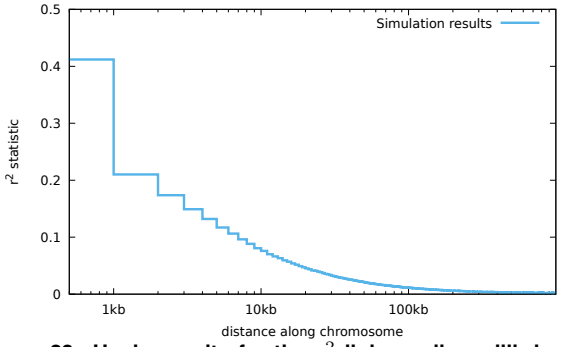

Figure 23: Haplo results for the r^2 linkage disequilibrium (LD) statistic. The population scenario is described in the text. Variants with minor frequency less than 0.05 were excluded from the analysis.
doi:10.5048/BIO-C.2019.1.f23

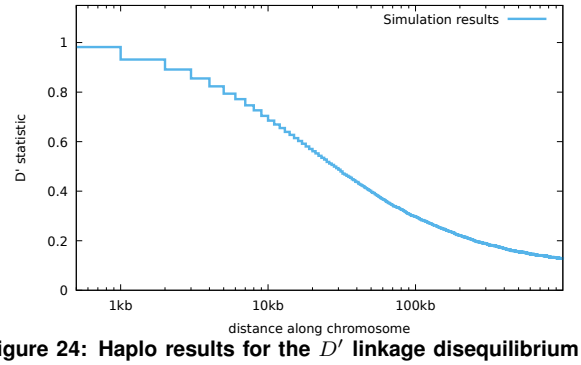

Figure 24: Haplo results for the D' linkage disequilibrium (LD) statistic. The population scenario is described in the text. Variants with minor frequency less than 0.05 were excluded from the analysis.
doi:10.5048/BIO-C.2019.1.f24

ACKNOWLEDGEMENTS

This article would not have been written without the work of Andrew Jones, who invented the Coaly method,

developed the software and significantly contributed in writing a first draft of the paper. We are also grateful to Danny Crookes, Peter Loose, Colin Reeves, Chris Shaw, and Blesson Varghese for very valuable discussions and support. We wish to thank a section editor and three anonymous reviewers for helpful comments that considerably improved the quality of the paper. Finally, both authors want to thank J.C. for inspiration and AG wants to thank I.C. for inspiration as well.

1. Ayala FJ, Escalante A, O'Huigin C, Klein J (1994) Molecular genetics of speciation and human origins. Proc Natl Acad Sci 91(15):6787-6794. doi:10.1073/PNAS.91.15.6787
2. Ayala FJ (1995) The myth of eve: molecular biology and human origins. Science 270(5244):1930-1936. doi:10.1126/science.270.5244.1930
3. Ayala FJ, Escalante AA (1996) The evolution of human populations: a molecular perspective. Mol Phylogenet Evol 5(1):188-201. doi:10.1006/mpev.1996.0013
4. Klein J, Sato A, Nikolaidis, N (2007) MHC, TSP, and the origin of species: From immunogenetics to evolutionary genetics. Annu. Rev. Genet. 41: 281-304. doi:10.1146/annurev.genet.41.110306.130137
5. Blum MGB, Jakobsson M (2011) Deep divergences of human gene trees and models of human origins. Mol Biol Evol 28: 889-898. doi:10.1093/molbev/msq265
6. Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. Nature 475:493-496. doi:10.1038/nature10231
7. Yang MA, Harris K, Slatkin M (2014) The projection of a test genome onto a reference population and applications to humans and archaic hominins. Genetics 198:1655-1670. doi:10.1534/genetics.112.145359
8. Miller KR (2002) Finding Darwin's God. 272. Harper Perennial (New York).
9. Alexander D (2008) Creation or Evolution: Do We Have to Choose? Lion Hudson (Oxford).
10. Enns P (2012) The Evolution of Adam: What the Bible Does and Doesn't Say About Human Origins. 147. Brazos Press (Grand Rapids).
11. Giberson K (2015) Saving the Original Sinner: How Christians Have Used the Bible's First Man to Oppress, Inspire, and Make Sense of the World. 170-171. Beacon Press.
12. McKnight S, Venema DR (2017) Adam and the Genome: Reading Scripture after Genetic Science. Brazos Press (Grand Rapids).
13. https://natureecoevocommunity.nature.com/channels/522-journal-club/posts/22075-adam-and-eve-a-tested-hypothesis Last accessed September 28, 2019.
14. https://biologos.org/blogs/dennis-venema-letters-to-the-duchess/adam-eve-and-population-genetics-a-reply-to-dr-richard-buggs-part-1 Last accessed September 28, 2019.
15. https://natureecoevocommunity.nature.com/users/24561-richard-buggs/posts/32171-adam-and-eve-lessons-learned Last accessed September 28, 2019.
16. https://discourse.biologos.org/t/adam-eve-and-population-genetics-a-reply-to-dr-richard-buggs-part-1/37039/48 Last accessed September 28, 2019.
17. Sanford JC, Carter RW (2014) In light of genetics ... Adam, Eve and the Creation/Fall. Christ Apol J 12(2):51-98.
18. Jeanson NT (2016) On the origin of eukaryotic species' genotypic and phenotypic diversity: Genetic clocks, population growth curves, and comparative nuclear genome analyses suggest created heterozygosity in combination with natural processes as a major source mechanism. Answ Res J 9:81-122.
19. Sanford JC, Carter RW, Baumgardner J, Potter B (2018) Adam and Eve, designed diversity, and allele frequencies. In 8th Int Conf Creationism, 200-216. doi:10.15385/jpicc.2018.8.1.20
20. Hössjer O, Gauger AK, Reeves C (2016) Genetic modeling of human history Part 1: Comparison of common descent and unique origin approaches. BIO-Complexity 2016(3):1-15. doi:10.5048/BIO-C.2016.3
21. Hössjer O, Gauger AK, Reeves C (2016) Genetic modeling of human history Part 2: A unique origin algorithm. BIO-Complexity 2016(4):1-36. doi:10.5048/BIO-C.2016.4
22. International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature, 449:851-862. doi:10.1038/nature06258
23. The 1000 Genomes Project Consortium (2015) A global reference for human genetic variation, Nature 526, 68-74. doi:10.1038/nature15393
24. Crow J, Kimura M (1970) An Introduction to Population Genetics Theory. The Blackburn Press, Caldwell (New Jersey).
25. Ewens WJ (2004) Mathematical Population Genetics. I. Theoretical introduction, 2nd ed, Springer (New York). doi:10.1007/978-0-387-21822-9
26. Durrett R (2008) Probability models for DNA sequence evolution, 2nd ed, Springer (New York).
27. Devlin B, Risch N (1995) A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 29(2):311-322. doi:10.1006/geno.1995.9003
28. Thomas DC (2004) Statistical Methods in Genetic Epidemiology. Oxford University Press (Oxford).
29. Hudson RR (1983) Properties of a neutral allele model with intragenic recombination. Theor Popul Biol 23:183-201. doi:10.1016/0040-5809(83)90013-8
30. Hudson RR (1990) Gene genealogies and the coalescent process. Oxford Surv Evol Biol 7:1-44.
31. Griffiths RC, Marjoram P (1997) An ancestral recombination graph. In: Donnelly P, Tavaré S, eds. Progress in Population Genetics and Human Evolution. IMA Volumes in Mathematics and its Applications, vol 87, Springer (New York). doi:10.1007/978-1-4757-2609-1_16
32. Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press (Cambridge). doi:10.1017/CBO9780511623486
33. Lynch M (2007) The frailty of adaptive hypotheses for the origins of organismal complexity. Proc Natl Acad Sci 104:8597-8604. doi:10.1073/pnas.0702207104
34. Nei M, Maruyama T, Chakraborty R (1975) The bottleneck effect and genetic variability in populations. Evolution 29(1):1. doi:10.2307/2407137
35. Keuffer R, et al (2007) Unexpected heterozygosity in an island mouflon population founded by single pair of individuals. Proc R Soc B 274:527-533. doi:10.1098/rspb.2006.3743
36. Carter RW, Powell M (2016) The genetic effects of the population bottleneck associated with the Genesis Flood. J Creat 30(2).

37. Fu YXX (1995) Statistical properties of segregating sites. Theor Popul Biol 48(2):172-197. doi:10.1006/tpbi.1995.1025
38. Wooding S, Rogers A (2002) The matrix coalescent and an application to human single-nucleotide polymorphisms. Genetics 161:1641-1650.
39. Rafajlovic M, Klassman M, Eriksson E, Wiehe T, Mehlig B (2014) Demographic-adjusted tests of neutrality based on genome-wide SNP data. Theor Popul Biol 95:1-12. doi:10.1016/j.tpb.2014.05.002
40. Ewert W (2018) The dependency graph of life. BIO-Complexity 2018(3):1-27. doi:10.5048/BIO-C.2018.3
41. Lynch M, Conery J, Burger J (1995) Mutation accumulation and extinction of natural populations. Am Nat 146:489-518. doi:10.1086/285812
42. Nachman MW, Crowell SL (2000) Estimate of the mutation rate per nucleotide in humans. Genetics 156(1):297-304. PMID:10978293.
43. Roach JC, et al (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328(5978):636-9. doi:10.1126/science.1186802
44. Conrad DF, et al (2012) Variation in genome-wide mutation rates within and between human families. Nat Genet 43(7):712-714. doi:10.1038/ng.862.Variation
45. Kong A, et al (2012) Rate of de novo mutations and the importance of father's age to disease risk. Nature 488(7412):471- 475. doi:10.1038/nature11396
46. Campbell CD, et al (2012) Estimating the human mutation rate using autozygosity in a founder population. Nat Genet 44(11):1277-1281. doi:10.1038/ng.2418
47. Michaelson JJ, et al (2012) Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. Cell 151(7):1431-1442. doi:10.1016/j.cell.2012.11.019
48. Sun JX, et al (2012) A direct characterization of human mutation based on microsatellites. Nat Genet 44(10):1161-1165. doi:10.1038/ng.2398
49. Awadalla P, et al (2010) Direct measure of the de novo mutation rate in autism and schizophrenia cohorts. Am J Hum Genet 87(3):316-324. doi:10.1016/j.ajhg.2010.07.019
50. Neale BM, et al (2012) Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature 485(7397):242-245. doi:10.1038/nature11011
51. O'Roak BJ, et al (2012) Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature 485(7397):246-250. doi:10.1038/nature10989
52. Sanders SJ, et al (2012) De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature 485(7397):237-241. doi:10.1038/nature10945
53. Scally A, Durbin R (2012) Revising the human mutation rate: implications for understanding human evolution. Nat. Rev Genet 13(10):745-753. doi:10.1038/nrg3295
54. Collins A, Frézal J, Teague J, Morton NE (1996) A metric map of humans: 23 500 loci in 850 bands. Proc Natl Acad Sci USA 93:14771-14775. doi:10.1073/pnas.93.25.14771
55. Kong a et al (2002) A high-resolution recombination map of the human genome. Nat Genet 31:241-247. doi:10.1038/ng917
56. Takahata N, Satta Y, Klein J (1995) Divergence time and population size in the lineage leading to modern humans. Theor Popul Biol 48(2):198-221. doi:10.1006/TPBI.1995.1026
57. Kimura M (1964) Diffusion models in population genetics. J Appl Probab 1(02):177-232. doi:10.2307/3211856
58. Gasbarra D, Sillanpää MJ, Arjas E (2005) Backward simulation of ancestors of sampled individuals. Theor Popul Biol 67:75-83. doi:10.1016/j.tpb.2004.08.003
59. Gasbarra D, Pirinen M, Sillanpää MJ, Salmela E, Arjas E (2007) Estimating genealogies from unlinked marker data: a Bayesian approach. Theor Popul Biol 72(3):305-322. doi:10.1016/j.tpb.2007.06.004
60. Bartlett J, Halloway E (2019) Generalized information. A straightforward method for judging machine learning methods. Comm Blyth Inst 1(2):13-21. doi:10.33014/issn.2640-5652.1.2.bartlett.1
61. Haber M, Mezzavilla M, Xue Y, Tyler-Smith C (2016) Ancient DNA and the rewriting of human history: Be sparing with Occam's razor. Gen Biol 17:1, 8 pages.
62. Sankararaman S et al (2014) The genomic ancestry of Neanderthal ancestry in present-day humans. Nature 505:43-49. doi:10.1038/nature12961
63. Dechamps M, Leval G, Fagny M, Itan Y, Abel L et al (2016) Genetic signatures of selective pressures and introgression from archaic hominins at human innate immunity genes. Am J Hum Genet 98:5-21. doi:10.1016/j.ajhg.2015.11.014
64. Danneman M, Andrés AM, Kelso J (2016) Introgression of Neandertal- and Denisovan-type haplotypes contributes to adaptive variation in human toll-like receptors. Am J Hum Genet 98:22-33. doi:10.1016/j.ajhg..2015.11.015
65. Allendorf FW, Ryman N (2002) The role of genetics in population viability analysis In Population Viability Analysis, ed Bessinger SR, McCullogh DR. The University of Chicago Press (Chicago).
66. Howell N (2003) The pedigree rate of sequence divergence in the human mitochondrial genome: There is a difference between the phylogenetic and pedigree rates. Am J Hum Genet 72, 659-670. doi:10.1086/368264
67. Wilson Sayres MA, Lohmuller KE, Nielsen R (2014) Natural selection reduced diversity in human Y chromosomes. PLoS Genet 10(2014):e1004064. doi:10.1371/journal.pgen.1004064
68. Stoeckle MY, Thaler DS (2018) Why should mitochondria define species? Hum Evol 33(1-2):1-30. doi:10.14673/HE2018121037
69. Carter RW, Lee SS, Sanford JC (2018) An overview of the independent histories of the human Y chromosome and the human mitochondrial chromosome. Proc 8th Int Conf Creat, ed. Whitmore JH, pp 133-151, Creation Science Fellowship (Pittsburgh). doi:10.15385/jpicc.2018.8.1.15
70. The mosaic that is our genome. Nature 421:409-412. doi:10.1038/nature01400
71. Myers S et al (2005) A fine-scale map of recombination rates and hotspots across the human genome. Science 310:321-324. doi:10.1126/science.1117196
72. Rosenfeld JA, Mason CE, Smith TM (2012) Limitations of the human reference genome for personalized genomics. PLoS ONE 7:e40294. doi:10.1371/journal.pone.0040294
73. Titus-Trachtenberg, EA (1994) Analysis of HLA Class II haplotypes in the Cayapa Indians of Ecuador: A novel DRB1 allele reveals evidence for convergent evolution and balancing selection at position 86. Am J Hum Genet. 55:160-167.
74. Bergström TF, Josefsson A, Erlich HA, Gyllensten U (1998) Recent origin of HLA-DRB1 alleles and implications for human evolution. Nat Genet 18(3):237-242. doi:10.1038/ng0398-237
75. von Salomé J, Gyllensten U, Bergström T (2007) Full-length sequence analysis of the HLA-DRB1 locus suggests a recent origin of alleles. Immunogenetics 59:261-271. doi:10.1007/s00251-007-0196-8

76. Gauger AK, Luskin C, Axe D (2012) Science and Human Origins. Discovery Institute Press (Seattle).
77. de Groot NG, Otting N, Robinson J et al. (2012) Nomenclature report on the major histocompatibility complex, genes, and alleles of Great Ape, Old and New World monkey species. Immunogenetics 64:615-631. doi:10.1007/s00251-012-0617-1
78. Vahdati R, Wagner A (2016) Parallel or convergent evolution in human population genomic data revealed by genotype networks. BMC Evolutionary Biology16:154. doi:10.1186/s12862-016-0722-0
79. Amos W (2013) Variation in heterozygosity predicts variation in human substitution rates between populations, individuals and genomic regions. PLoS One 8(4):e63048. doi:10.1371/journal.pone.0063048
80. Harris K (2015) Evidence for recent, population-specific evolution of the human mutation rate. Proc Natl Acad Sci 112(11):3438-3444. doi:10.1073/pnas.1418652112
81. Harris K, Pritchard JK (2017) Rapid evolution of the human mutation spectrum. eLife 6. doi:10.7554/eLife.24284
82. Scerri EML, Tomas MG et al (2018) Did our species evolve in subdivided populations across Africa, and why does it matter? Trends Ecol Evol 33(8):582-594. doi:10.1016/j.tree.2018.05.005
83. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet 5(10), e:1000695. doi:10.1371/journal.pgen.1000695
84. Rasmussen MD, Hubisz MJ, Gronau I, Siepel A (2014) Genome-wide inference of ancestral recombination graphs. PLoS Genet 10(5):e1004342. doi:10.1371/journal.pgen.1004342
85. Spence JP, Steinrücken M, Terhorst J, Song YS (2018) Inference of population history using coalescent HMMs: review and outlook. Curr Opin Genet Dev 53:70-76. doi:10.1016/J.GDE.2018.07.002
86. Sanford JC, Baumgardner J, Brewer W, Gibson P, ReMine W (2007) Mendel's accountant: A biologically reasonable forward-time population genetics program. Scalable Computing: Practice and Experience 8(2):147-165.
87. Sella G, Petrov DA, Przeworski M, Andolfatto P (2009) Pervasive natural selection in the Drosophila genome? PLoS Genet 5(6):1-13. doi:10.1371/journal.pgen.1000495
88. Neher RA, Shraiman, BI (2011) Genetic draft and quasineutrality in large facultatively sexual populations. Genetics 188:975-966.
89. Charlesworth B (2012) The role of background selection in shaping patterns of molecular evolution and variation: Evidence from variability on the Drosophila. Genetics 191:233-246. doi:10.1534/genetics.111.138073
90. Kimura M (1969) The number of heterozygous nucleotide sites maintained in a finite population due to a steady flux of mutations. Genetics 61:893-903. PMID:5364968.
91. Evans SN, Shvets Y, Slatkin M (2007) Non-equilibrium theory of the allele frequency spectrum. Theor Popul Biol 71(1):109-119. doi:10.1016/J.TPB.2006.06.005
92. Kingman JFC (1982) The Coalescent. Stoch Proc Appl 13:235-248.
93. Ohta T, Kimura M (1971) Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. Genetics 68(4):571-80. PMID:5120656
94. McVean GAT (2002) A genealogical interpretation of linkage disequilibrium. Genetics 162(2):987-91. PMID:12399406