Research Article

Genetic Modeling of Human History Part 1: Comparison of Common Descent and Unique Origin Approaches

Ola Hössjer,*[1] Ann Gauger,[2] and Colin Reeves[3]

[1] Department of Mathematics, Stockholm University, Sweden
[2] Biologic Institute, Redmond, WA, USA
[3] Applied Mathematics Research Centre, Coventry University, United Kingdom

Abstract

In a series of two papers (Part 1 and 2) we explore what can be said about human history from the DNA variation we observe among us today. Population genetics has been used to infer that we share a common ancestry with apes, that most of our human ancestors emigrated from Africa 50 000 years ago, that they possibly had some mixing with Neanderthals, Denisovans and other archaic populations, and that the early Homo population was never smaller than a few thousand individuals. Population genetics uses mathematical principles for how the genetic composition of a population develops over time through various forces of change, such as mutation, natural selection, genetic drift, recombinations and migration. In this article (Part 1) we investigate the assumptions about this theory and conclude that it is full of gaps and weaknesses. We argue that a unique origin model where humanity arose from one single couple with created diversity seems to explain data at least as well, if not better. We finally propose an alternative simulation approach that could be used in order to validate such a model. The mathematical principles of this model are described in more detail in our second paper (Part 2).

Cite as: Hössjer O, Gauger A, 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.

Editor: Douglas D. Axe

Received: June 4, 2016; **Accepted:** October 5, 2016; **Published:** November 4, 2016

Notes: A *Critique* of this paper, when available, will be assigned **doi:**10.5048/BIO-C.2016.3.c.

*Email: ola.hossjer@gmail.com

INTRODUCTION

We all have a genetic fingerprint. The more closely related we are the more similar our fingerprints are. Genetic data can be used for a number of purposes: to find out whether we carry a risk gene of an inheritable disease, to ascertain that a young man is the father of a newborn child, to find evidence against a suspect of a crime, or to retrieve our ethnic mix at ancestry.com. In this paper and a subsequent one we will investigate what human genetic data has to say about common ancestry. In academia it has more or less been taken for granted that humans have a common ancestor with chimps. The prevailing view of this common descent scenario is that the first steps to humanity arose from ape-like ancestors, whose number were never less than a few thousand individuals at any time in history. But there are virtually no attempts to check the results of a scenario by which humanity descends from a single first couple. We will call this second scenario the unique origin model.

From a scientific point of view it is important to compare and test both scenarios. In this first article (Part 1) we will describe the results of comparing and testing them against each other. The conclusion of our qualitative argument is that a unique origin scenario is in fact more plausible. We end by suggesting a quantitative model by which such scenarios can be tested more formally. The mathematics of this quantitative model is described in more detail in our second article (Part 2).

In more detail, this paper (Part 1) is organized as follows: In Section 1 we introduce some basic concepts from population genetics. Although the content is well known to many readers, it makes the article more self contained and easier to follow. Then in Section 2 we describe in more detail different versions of the two competing common descent and unique origin models. They are compared and evaluated in Section 3, using several different criteria. Finally, in Section 4 we briefly discuss how the hypotheses of the unique origin model can be tested with data, as a preparation for our accompanying paper (Part 2).

1. POPULATION GENETICS

Population genetics is a discipline that describes how the genetic makeup of a group of people changes over time [1]. It has numerous applications, but here we will use it as a tool for comparing different scenarios of human history.

Before that, we will first review some basic principles of genetics [2].

1A. Human DNA and Its Inheritance

Each one of us has genetic information stored in our cells in terms of DNA. Most of this information is contained in the cell nucleus as 46 chromosomes, 23 of which are inherited from our father and 23 from our mother. Forty-four of these are non-sex (autosomal) chromosomes and come in almost identical (homologous) pairs. The remaining two chromosomes determine our sex. Females have two copies of an X-chromosome, one from each parent, whereas males have one Y-chromosome from the father, and one X-chromosome from the mother. There is also some DNA in the mitochondria.

The chromosomes of the cell nucleus and the mitochondria of the cell cytoplasm are DNA molecules, whose structure is a double-stranded helix. Both strands are written in an alphabet that consists of the four letters adenine (A), guanine (G), cytosine (C) and thymine (T). Each such letter is usually referred to as a nucleotide or base, and they are connected along each strand by strong chemical covalent bonds. Between the two strands, the nucleotides pair up by weaker hydrogen bonds, A connecting to T, and G to C. Each pair, A-T or G-C, is referred to as a base pair, and our genome can be thought of as a book with three billion base pairs of nuclear DNA along with sixteen thousand base pairs of mitochondrial DNA (Fig. 1).

The human genome is estimated to contain about 20,000–25,000 genes. These genes comprise only a small part of DNA, and each one of them carries information about various types of proteins through the genetic code. The mechanism of this code is that triplets (codons) of nucleotides are translated into one of twenty possible amino acids, the building blocks of all proteins. There is some redundancy in the genetic code. Since most of the $4^3 = 64$ possible nucleotide triplets code for amino acids, some codons will correspond to the same amino acid. Non-coding DNA exists both between and within genes. It has several important functions, one of which is to regulate tissue-specific gene activity. This is achieved through epigenetic changes, for instance, or when various transcription factors attach to the DNA molecule and either activate or suppress expression of a gene in order to control how much of its protein is produced.

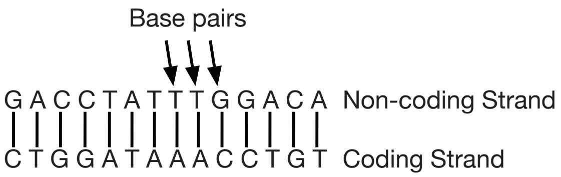

Figure 1. A small part of a double-stranded DNA-molecule with 14 base pairs. Since A always connects to T, and C to G, both strands carry the same information. For population geneticists one strand is treated as the coding strand, because it carries the information for the majority of genes; thus the nucleotide of the coding strand specifies the identity of the base pair. **doi:**10.5048/BIO-C.2016.3.f1

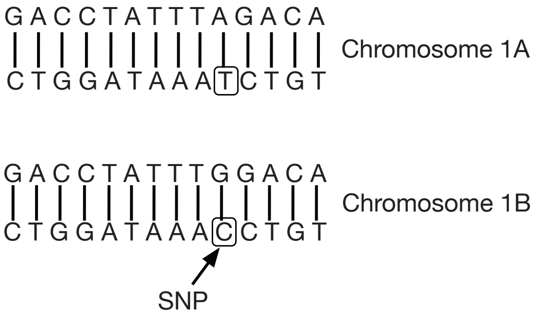
G A C C T A T T T A G A C A
C T G G A T A A A T C T G T
Chromosome 1A
G A C C T A T T T G G A C A
C T G G A T A A A C C T G T
Chromosome 1B
SNP

Figure 2. A small part of a DNA molecule, with 14 base pairs, exists in two copies in the worldwide human population. The upper copy is the same as in Figure 1, whereas the lower differs in that the 10th base pair is C-G rather than T-A. This position is a single nucleotide polymorphism (SNP), with two possible alleles T and C at the coding strand. **doi:**10.5048/BIO-C.2016.3.f2

1B. DNA Variation Among Individuals

We don't all look the same, and a major reason for this is the fact that our genomes are not identical. There are small differences that make each one of us genetically unique. Most parts of DNA are the same for all humans, but those that vary are called polymorphisms. A Single Nucleotide Polymorphism (SNP) is the most common kind of variation.[1] For each base pair of the DNA molecule, the convention is to refer to only one of its two nucleotides, the one located on the coding strand.[2] This nucleotide usually exists in two variants at a SNP, for instance C or T, also referred to as the two alleles of the SNP (Fig. 2). There are several million bi-allelic SNPs along the human genome, and for those located on one of the 22 autosomes, each person will have two copies of it, one on the chromosome inherited from the father and the other from the chromosome that the mother passed on. For a SNP with alleles C and T, there are three possible pairs of alleles, C on both chromosomes (CC), C on one chromosome and T on the other (CT), and T on both chromosomes (TT). They are referred to as the three genotypes. (Since the two alleles of a genotype sit on different but homologous chromosomes they are not the same thing as a base pair!)

1C. Origin, Purpose and Limitations of Population Genetics

Since our genomes are not identical, genetics can tell us something about human variation and history. It is possible to study how big the genetic differences in the worldwide human population are, and indirectly how these differences have changed in the past. One may also compare the genetic makeup of individuals from different regions like Europe, Africa, Middle East or East Asia, or study smaller groups of people, like the inhabitants of Sardinia, Iceland or Polynesia. Population genetics is a discipline that uses mathematical methods to quantify how genetic differences vary among individuals, between geographic

[1] Other types of polymorphisms include, for instance, indels, short tandem repeats, copy number variation of larger genomic regions, and Alu insertions. Reference [3] contains statistics for their relative occurrence in human DNA.

[2] This is the strand involved in protein coding. Less than 2% of the coding strand of human DNA actually codes for proteins. This coding part consists of a number of exons that are first transcribed into mRNA and then translated into proteins.

regions and over time. Starting in the 1920s, much of its theory was built by prominent mathematicians and geneticists like Ronald Fisher, Sewell Wright, John Haldane, Motoo Kimura and Samuel Karlin. A very comprehensive summary of the first 50 years of work can be found in the book by Crow and Kimura [4]. The mathematical theory was quickly picked up by biologists, and it had a major role to play in the neo-Darwinian (or Modern) Synthesis that took place between 1930 and 1950 [5,6]. Darwinian theory was before that very speculative, since a quantitative method for analyzing the spread of genetic differences via natural selection was lacking. Population genetics seemed to fill this gap by giving the tools for describing genetic variation in populations, which was thought would eventually explain how humans and all other species came to share a common ancestry. Researchers assumed that such evolution was only guided by genetic changes in germ cells, effectively making us vessels of our genomes.

But population genetics was founded at a time when very little was known about the complexity of the cell and its hereditary mechanisms. The knowledge of molecular biology has expanded enormously since then. As we will see below, the mathematics is built upon the principle of small stepwise changes, and as such is a good tool for describing microevolution within species, and some limited degree of speciation. This has found many applications, for instance animal breeding and plant breeding, and wildlife management and conservation biology, where the viability of species is studied and their adaptation to environmental changes and inbreeding is quantified and estimated over time [7].

Although population genetics does not require common ancestry per se, it was largely founded in order to support macroevolution and common descent, based on the idea that macroevolution was merely microevolution writ large. But today it is possible to use it for the opposite purpose: to show how unlikely macroevolution is (at least when it is based on the population genetics of neo-Darwinism). Indeed, macroevolution requires formation of new kinds of genes, new proteins, new organs and other irreducibly complex structures.[3] Although the mathematical theory of population genetics has continued to develop until this day [9,10], it has so far not been able to deal with macroevolution in a convincing way. An increasing number of biologists have realized the severe limitations of neo-Darwinism [11,12].

1D. Mechanisms of Genetic Change

Human DNA contains information about our history, because our genomes are scrambled images of our ancestors' genomes. In order to understand how history has scrambled ancestral DNA, we first need to describe the mechanisms that population geneticists use to explain how the genetic composition of a population changes over time. These five mechanisms reflect demography as well as genetic inheritance, and thus can tell us something about a population's history. They can be summarized as follows:

I. *Mutations* are changes of DNA. Germline mutations are the ones of most interest to population genetics, since they are the ones that are inherited. These mutations are typically copying errors. Suppose for instance that a single nucleotide is changed from an A to a T in a sperm cell that later unites with an ovum during fertilization. Because of ordinary cell division, all the cells of the child will carry the mutated allele T (apart from occasional somatic back-mutations). It is obvious that mutations increase genetic diversity, and the molecular clock gives the speed at which this happens. There is a sophisticated DNA copying repair mechanism, and for this reason the probability is very small, of the order 10^{-8} per nucleotide per generation for mutations in nuclear DNA to occur, whereas it is several orders of magnitude larger for mitochondrial DNA.[4]

II. *Genetic drift.* Whereas mutations are changes of DNA, already existing polymorphisms will change in their frequency from one generation to the next for at least two reasons. First, reproductive success of parents varies. For instance, for a SNP with alleles A and T it may happen strictly by chance that parents with genotypes AA on average have more children than those with genotype TT. Second, Mendel's law of inheritance implies that a parent with genotype AT is equally likely to pass the A and T when a sperm or ova cell is formed. Due to chance, though, more A sperm or eggs may be passed on. The overall effect in both cases is that the frequency of allele A increases in the next generation, a phenomenon known as genetic drift.

This random change of allele frequency is more rapid in small populations than in large ones (for the same reason as when we toss a fair coin a few times, the frequencies of heads and tails will deviate more from 50%, compared to when we toss it many times). In a bottleneck, when the size of a population is radically reduced for some time before it recovers, many rare alleles will be lost during the near extinction phase, because of an increased amount of genetic drift.

III. *Natural selection* is similar to genetic drift. The varying reproductive success of parents causes allele frequencies to change in the next generation. The difference is that these changes are not only caused by chance, but to some extent expected to happen. Several types of selection exist, and we will only mention some of them here. To this end, consider again a single nucleotide polymorphism with two alleles. For directional selection, one allele (say A) will have a reproductive advantage over the other (T). The reason could either be that parents with an AA genotype will have a higher reproductive fitness (be expected to have more children) than those that carry an AT, and the parents with an AT will have a higher fitness than those with a TT genotype. Another possibility is that fertilized eggs with an AA genotype have the highest survival probability, and those with a TT genotype have the lowest survival probability. Directional selection tends to decrease the amount of variation, since it is likely that the frequency of A will increase over time, so that A

[3] A biological structure or activity that is irreducibly complex is composed of several parts well-suited for each other and designed for a particular function, where removal of any of the parts causes the biological structure or activity to cease. For more discussion and description of such biological processes or structures, see [8].

[4] Mutations of mtDNA are more complicated than for nuclear DNA. There are 100 to 10,000 mitochondria per cell, and some mutations only affect a subset of them, see for instance [13].

eventually becomes fixed in the whole population (becomes the only version of nucleotide at that position). In the early days of evolutionary theory, directional selection was considered to be a very important mechanism of genetic change. But its power to alter the genetic composition of a population is limited [14,15], since most alleles are selectively neutral or slightly deleterious. And the remaining small number of beneficial mutations often have a fitness that is only marginally higher than for neutral alleles, since selection does not operate on genes, but on phenotypes like body strength or speed. Explaining the origin of new organs by these sorts of processes is therefore a stretch, because a large number of mutations are required to increase in frequency in a coordinated manner.

There is a closely related form of natural selection called purifying selection, by which deleterious alleles are removed in the population, since their carriers either die or have fewer offspring. Like directional selection it tends to reduce diversity. But since most alleles are selectively neutral or slightly deleterious, purifying selection is unable to counterbalance mutations and remove all new slightly deleterious alleles in a coordinated way. The effect of this is that the entropy of the genome increases over time.[5]

A third type of selection has the opposite effect of increasing diversity. It is called balancing selection, and it happens when AT has a selective advantage over AA and TT. This selective force, which maintains many individuals with an AT genotype, automatically creates a balance between the frequencies of A and T. Balancing selection may be important for explaining the high diversity in some parts of the human genome, like the HLA system on Chromosome 6. Several of the genes in this region are crucial for the immune system, and a high genetic variability could increase the ability of a population to recognize harmful intruders and thereby survive an epidemic better, see for instance Chapter 5 of [18] and references therein.

IV. *Recombination* is another way of generating more diversity for non-sex chromosomes and X-chromosomes. No novel DNA is produced, but existing DNA is combined in new ways.

When DNA is inherited from one generation to the next, the number of chromosomes is halved from 46 to 23 in the sperm cell from the father and the ovum from the mother. This process, called meiosis, is necessary so that when the sperm and egg come together a full complement of 46 is restored. During meiosis, recombination between homologous chromosomes takes place (see Fig. 3). After recombination, each chromosome of a sperm or ovum (a germ cell) is a mosaic of the homologous grandmaternal and grandpaternal chromosomes. The only parts of the genome that are inherited without recombination are DNA from mitochondria, the non-recombining parts of Y-chromosomes and X-chromosomes from a father.

Recombinations make variation at different parts of the chromosome more independent. In order to illustrate this, suppose a new mutation T→A occurs in a chromosome that has a preexisting G at another SNP on the same chromosome. (Other chromosomes carry a C at that position.) Since there is initially only one chromosome copy with A at the first SNP, an A will always coexist with a G at the other SNP at the time when the mutation first arrives. There are no AC versions in the population, and therefore the two SNPs are completely associated, meaning that an A has always been inherited together with a G. Population geneticists use a concept called linkage disequilibrium (LD) in order to quantify how much association there is between alleles of two SNPs (this will be explained in more detail in Section 1E). When the mutation arrives, so that A and G are always linked together, the amount of LD between the two SNPs is maximal. Recombinations between them will later on appear in descendants of the first mutated individual, so that some germ cell chromosomes get an A from the grandparent with a mutation at the first SNP, and a C from the other grandparent at the second SNP. These recombinations will gradually break up the association between A at the first SNP with G at the second (reducing LD), and since they happen more often between remote parts of a chromosome, the number of recombinations between a distant SNP pair will tend to be greater, and the LD therefore gets smaller.

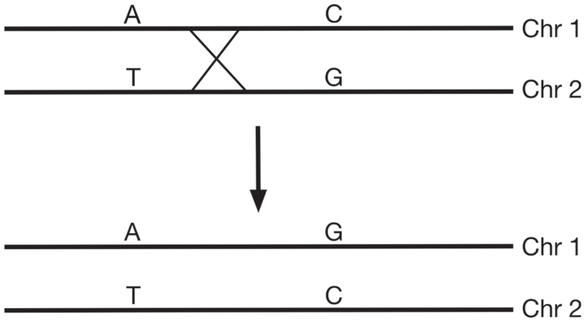

Figure 3. Recombination between two homologous chromosomes results in shuffling of alleles. The two recombined chromosomes below the arrow end up in separate germ cells. **doi:**10.5048/BIO-C.2016.3.f3

There are also closely located double recombinations called gene conversions. They will only affect the LD pattern for SNP pairs if only one of them is between the two points of recombination. The implication of this is that gene conversions affect the LD pattern of very nearby SNPs, but not those that are further apart.

V. *Colonization, isolation and migration.* Especially in the past, humans have been more or less isolated by distance, with mating couples living nearby. Sometimes new subpopulations are formed (colonization), and occasionally men or women migrate over longer distances to find their mates. It is evident that migration will decrease subpopulation differentiation, whereas isolation will increase it. Colonization will also increase local geographic differences, especially if the founding population is small, and then quickly expands after it has settled.

Neo-Darwinism accounts for the above-mentioned mechanisms I–V, and among them germline mutations are essentially the only way by which novel DNA can arise. The theory does not allow for large amounts of new and suddenly appearing diversity. The reason is that neo-Darwinism is framed within methodological naturalism. This prevailing approach to science

[5] This is usually referred to as Haldane's dilemma, see for instance Chapters 7-9 of [16] and [17].

only allows for natural hypotheses. But if an intelligent designer is invoked as a possible explanation, and if humanity originates from one single couple, it is possible that their chromosomes were created with considerable diversity from the beginning [19,20]. This gives us a sixth mechanism of genetic change:

VI. ***Created founder diversity*** is biologically plausible for DNA of non-sex chromosomes. Since there are four copies of each non-sex chromosome in two individuals, we may think of one of the founding male's two copies as a reference or template for the other three. All differences between his reference chromosome and the other three founder chromosomes can be thought of as a very large number of mutations, all of which occurred in one generation. Since the founding pair had three copies of the X-chromosome (one in the man and two in the woman), for the same reason they may also have been different. Some founder diversity is possible for mitochondrial DNA as well. Since the founding female carried hundreds of mitochondria that could have been diverse, it is possible that she passed on some of that diversity on to her daughters.

1E. Summary Measures of DNA

In order to describe how population genetics theory works, we need to introduce some concepts. This material, though abstruse, is important because it is in part by these measurements that population geneticists determine the past history of populations, in our case, the past history of humanity back to the time of the first founding couple.

Since the size of our genome is huge, it is helpful to first summarize the genetic composition of all individuals in a population in some convenient way. To make it simple we assume here that all polymorphisms are SNPs with two alleles. This is by far the most common type of variation, and data from bi-allelic SNPs provide the following four summary measures:

i. ***Nucleotide diversity.*** A typical genome differs from the human reference sequence (which has the most common SNP at all positions) at about four to five million nucleotides, slightly more than 0.1% of the human genome [3]. A more commonly used measure of diversity is the total number of SNPs adjusted for the size of the population (since a large population will have more polymorphic variants). Another measure of variation, which is easier to interpret, is the nucleotide diversity. It is defined as the fraction of nucleotides at which the genomes of two randomly chosen individuals differ. For the worldwide human population it has been estimated to 0.08% for the whole genome, although it is highest for non-sex chromosomes, smaller for X-chromosomes, and even lower for Y-chromosomes [21,22]. The nucleotide diversity of mitochondrial DNA is higher, about 0.25% [23].

ii. ***Allele frequency spectrum.*** Consider a single SNP with two alleles A and T in a population of size 500. If this SNP is located on one of the 22 autosomes, there will be 500x2=1000 copies of it in the population. Suppose 400 of these copies have allele A and the remaining 600 ones have the other allele T. Then A has a minor allele frequency (MAF) 400/1000=40%. In other words, 40% of the population carries an A at that position, and it is called the minor allele frequency, since it is less than the frequency 60% of T. The MAF will always be a number between 0% and 50%. The allele frequency spectrum is a histogram of MAFs for all known SNPs (almost 100 million for the worldwide human population), where the height of each bar corresponds to the fraction of variants within a certain range of the minor allele frequency. Suppose 0–50% is divided into 10 equally large intervals. Then the leftmost bar shows the number of all rare variants, those with a MAF between 0% and 5%, the next bar consists of all SNPs with a MAF between 5% and 10% and so on. For instance, if there are 10 million SNPs in a population, and 2.2 million have a rare variant, the height of the leftmost bar is 2,200,000/10,000,000=22% (see Fig. 4). But for the worldwide human population, the fraction of rare SNPs is much larger than this [24], and consequently the fraction of common variants (MAF between 5% and 50%) is much smaller than 100–22=78%.

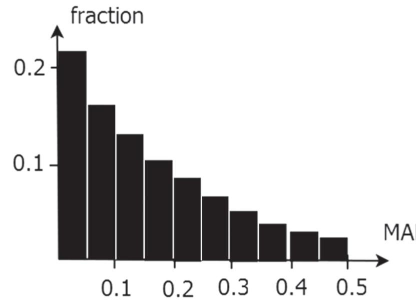

Figure 4: Schematic illustration of the allele frequency spectrum of a population. The height of each bar gives the fraction of SNP with a minor allele frequency within the interval that the base of the bar extends over. The fraction of rare variants (MAF between 0 and 5%) of the human population is much larger than the 22% that the figure indicates though. **doi:**10.5048/BIO-C.2016.3.f4

iii. ***LD plots.*** In Section 1D we introduced linkage disequilibrium (LD), a word geneticists use to describe how tightly associated alleles of two SNPs are, that is, how often different combinations of them are found together. In order to illustrate this concept, consider a pair of SNPs on the same non-sex chromosome, of which the first has alleles A and T with frequencies 40% and 60% (as in the example of II), and the second one has alleles C and G, with frequencies 30% and 70%. In a population of size 500, there are 2x500=1000 copies of this chromosome. The alleles from the two SNPs form a haplotype, which can have four possible variants AC, AG, TC or TG. If the two SNPs vary independently in the population, the frequency of AC is 0.4x0.3=12%. This number is obtained by multiplying the frequency of A at the first SNP with the frequency of C at the second. The frequencies of the other three haplotypes are found similarly as 0.4x0.7=28% for AG, 0.6x0.3=18% for TC and 0.6x0.7=42% for TG. Since alleles at the two SNPs vary independently in the population, geneticists refer to such a scenario as absence of linkage disequilibrium (LD=0). But it may also happen that a chromosome with an A at the first SNP is more likely to have a G at the second SNP, compared to a chromosome with a T at the first SNP. Then there is linkage disequilibrium between the two SNPs. The maximal amount

No LD

SNP2 SNP1	C	G	Sum
A	0.12	0.28	0.4
T	0.18	0.42	0.6
Sum	0.3	0.7	1

Max LD

SNP2 SNP1	C	G	Sum
A	0	0.4	0.4
T	0.3	0.3	0.6
Sum	0.3	0.7	1

Figure 5: Two possible scenarios of allelic association (linkage disequilibrium, LD) between two SNPs. The first one has alleles A and T with frequencies 40% and 60%, and the second one alleles C and G, with frequencies 30% and 70%. In the left table there is no association (LD=0) and in the right table the amount of association is maximal (LD=1), since one haplotype (AC) is absent. **doi:**10.5048/BIO-C.2016.3.f5

of linkage disequilibrium (LD=1) occurs when one of the four haplotypes is completely absent from the population.[6] This happens for instance if there are no AC and the other three haplotypes have frequencies 40% for AG, 30% for TC and 30% for TG (see Fig. 5). Another example of complete LD was found in Section 1D, when a mutation occurs that defines a new SNP.

An LD plot gives the average amount of LD for SNP pairs at various distances (Fig. 6). This average LD will in general decrease with distance, since nearby SNPs tend to be inherited together, and ones farther apart tend to be inherited independently of one another. It turns out that LD will both depend on when the mutations at the two SNPs first arrived and the amount of recombination between them after that.[7] The main question is over how long distances LD exists, since this provides information about past demography (population behavior). The longer the two SNPs have been around, and the more distant they are, the more likely they will have been separated by recombination—the more shuffled the chromosomal region between them will be.

iv. ***Subpopulation differentiation***. There is no rationale for the concept of human races, since genetic differences within subpopulations (like Africa, Europe, Middle East, East Asia and native Americans) are much larger then differences between them. It is still the case that subpopulations whose inhabitants have lived isolated for a long time will on average have different genetic profiles. The most straightforward way to check this is to compare nucleotide diversities, allele frequency spectra and LD plots for the subpopulations. A more quantitative approach is to compute a so-called fixation index F_{ST} for each pair of subpopulations. This is a kind of distance measure, a number that ranges between 0 and 1. The smaller it is, the more similar the two subpopulations are. The fixation indices for all pairs of subpopulations can be translated into a two-dimensional scatterplot, in which each subpopulation is drawn as a dot, and those from the same continent tend to cluster (left part of Fig. 7).[8] There are other even more sophisticated methods that don't require individuals to self-report their ethnicity or subpopulation membership. They automatically infer, for each individual what mixture of proportions from different ancestral groups the individual has (right part of Fig. 7).[9]

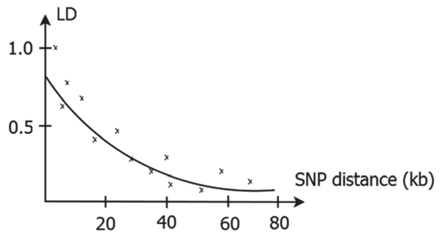

Figure 6: Schematic illustration of an LD plot of a population. Each x corresponds to a SNP pair, whose distance is given in units of 1000 nucleotides or kilobase pairs (kb). The smoothed curve gives the average amount of LD for SNP pairs at various distances. If a least one of the two SNPs has a very recent mutation, its x will be above this curve. **doi:**10.5048/BIO-C.2016.3.f6

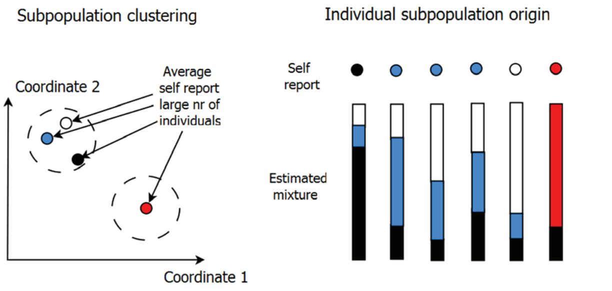

Figure 7: The left plot shows clustering of four subpopulations (white, light blue, black, red) based on their pairwise F_{ST} distances and two-dimensional scaling. All individuals have self-reported their subpopulation membership. The first cluster contains the white, light blue and black population, and the second one consists of the red population. The right plot shows the estimated fractions of ancestral DNA for six individuals, without requiring their self-reported ethnicity. **doi:**10.5048/BIO-C.2016.3.f7

In the next two sections we will use the mechanisms of genetic change (Section 1D) and the summary measures of DNA (Section 1E) in order to address the question that was asked in the beginning of this article: In view of genetic data, is the common descent or the unique origin model of human history most plausible?

2. HUMAN HISTORY MODELS AND HOW TO RECONSTRUCT THE PAST

In this section we will juxtapose two models of human history: the standard model that assumes common descent, and another model that assumes we came from two first parents (unique origin). As shown in Figure 8, the two models have different starting assumptions, but follow the same rules of inheritance and population genetics in their working out. The standard model assumes we came from an initial population of about 10,000 at our separation from chimps, an estimate based

[6] Here we tacitly assume that LD is measured in terms of Lewontin's D', see for instance Chapter 8 of [25] for an exact definition.

[7] See Section 1D.

[8] This can be done in a variety of ways, for instance through principal component analysis, see [26].

[9] This includes for instance the STRUCTURE program [27].

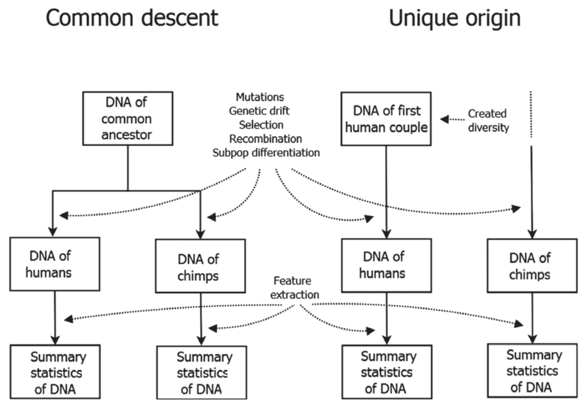

Figure 8: Illustration of the two competing models of human history; common descent and unique origin. The figure shows how the different mechanisms of genetic change (Section 1D) affect DNA of humans and chimps. These are condensed into summary measures of DNA (Section 1E). The challenge is to find the best fitting genealogy of each scenario, in order to see which one fits the summary statistics the best. **doi:**10.5048/BIO-C.2016.3.f8

on matching current genetic diversity with population histories without any created diversity. Our unique origin model assumes we started from two first parents with initial created diversity in our chromosomes, in other words a starting array of SNPs scattered throughout. Our model is a forward projection in time of the outworking of that diversity.

2A. Reconstruction of History

When population geneticists try to reconstruct human history from genetic data, they compare different demographic scenarios in order to see how well mutations, recombinations, genetic drift and selection are able to reconstruct the genetic variation we see today in terms of nucleotide diversity (i), allele frequency spectra (ii), LD plots (iii) and subpopulation differentiation (iv). There are highly sophisticated mathematical methods for doing this, but it is still very difficult to reconstruct history. The main reason for this is lack of data from the past. It is indeed possible to sequence DNA from some of our ancient relatives, and in recent years this line or research has exploded [28]. In spite of these advances, ancient DNA is still so sparse that to a large extent, future genetic analyses of human history will continue to rely on DNA samples from the most recent generations. With little historical data, any reconstructed genealogy is an estimate only with assumptions embedded in it.

What kind of history is it that genetic data provides estimates of? It is not a usual type of pedigree, where each individual has two parents. It is rather a genealogy showing which ancestors passed on DNA to persons alive today. The form of this genealogy will vary along with the DNA being studied. For example, since mitochondrial DNA is inherited almost always only from the mother, its genealogy will be a tree of females, where each individual has the mother as the only parent.[10] For the worldwide human population, the root of this tree is usually referred to as mitochondrial Eve. She is the female ancestor of all people alive today, usually called our most recent common ancestor. Mitochondrial Eve exists whether our common descent from ape-like ancestors is true or not.

This is because genealogies tend to coalesce over time. Not every woman alive at the time of mitochondrial Eve passed on her genes to daughters, and the same was true in every generation. Generation by generation the number of lineages remaining was whittled down. Eventually only one female's descendants remain (Fig. 9). Therefore mitochondrial Eve is not necessarily a proof of a unique female origin of humanity, since other women may have lived at the same time. But on the other hand, she may be, it is only that this data alone cannot tell us that.

The ancestry of Y-chromosomal DNA is similarly a tree of males in which each individual has a single parent, his father.[11] The root of this tree, Y-chromosome Adam, is not a proof of a unique origin either, for the same reasons.

The genealogy of autosomal and X-chromosome DNA is more complicated, since shuffling between homologous chromosomes (recombination) will divide the autosomes and the X-chromosome into blocks with different ancestral trees, each one with a different root. The lineages of all autosomal and X-chromosome ancestral trees may be followed further back in time until a "grand most recent common ancestral chromosome" is found (with huge error bars, we might add).

In order to explain how the standard population genetics model and genetic data is used to generate information about the ancestral size of the human population, let us first consider nucleotide diversity (i). *If humanity is very old* (for instance if

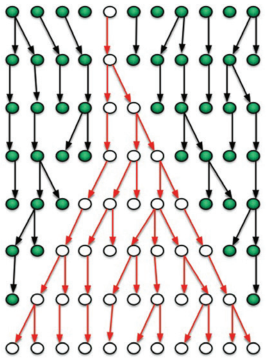

Figure 9: A hypothetical genealogy (for instance a mitochondrial DNA tree). There are 11 individuals in each generation, with the fifth individual in generation 1 (G1) the eventual progenitor of every one in G8, due to genetic drift and/or natural selection. The root of the tree, or the most recent common ancestor (MRCA) of the individuals in G8, is the fifth individual of G2. If there are no mutations among the descendants of the MRCA, there is no genetic diversity in G8. The reason is that all (mt DNA) variants of the MRCA have been fixed, whereas the variants that existed in other G2 individuals have been lost. For a larger population, it typically takes a lot more than seven generations to reach the MRCA when tracing ancestry backwards in time. **doi:**10.5048/BIO-C.2016.3.f9

[10] See http://www.phylotree.org/ for the latest update of the worldwide female gene tree.

[11] See http://www.phylotree.org/Y/ for the latest update of the worldwide male gene tree.

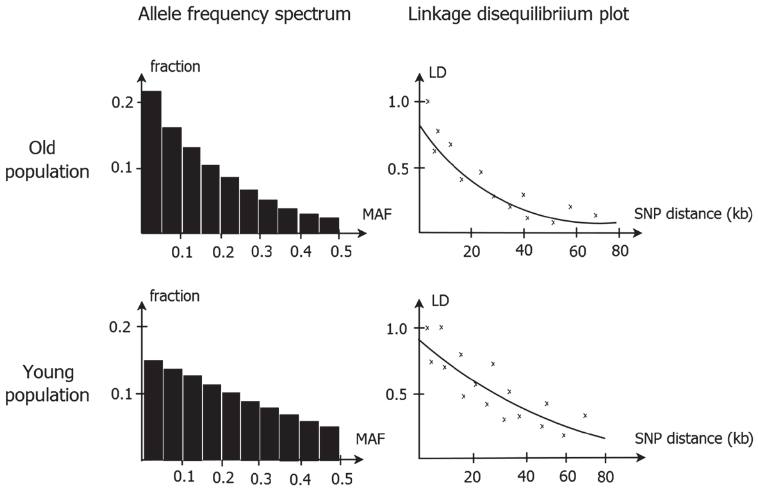

Figure 10. Illustration of allele frequency spectra and LD plots of an old population without created diversity (upper), and a young population founded by a single couple with created diversity (lower). The height of each bar of the allele frequency spectra gives the fraction of SNPs with a minor allele frequency within the interval that the base of the bar extends over. If the population expanded recently, the fraction of rare variants (MAF between 0 and 5%) is much larger for both allele frequency spectra. Each x of the LD plots corresponds to a SNP pair, whose distance is given in units of 1000 nucleotides or kilobase pairs (kb). The comparison of the two rows is only schematic. The younger the population of the lower subplots is, the more its allele frequency spectrum and LD plots differ from the upper population. **doi:**10.5048/BIO-C.2016.3.f10

we share ancestry with chimps) the diversity we see today is mainly produced by a balance among mutations in germ cells, genetic drift, and to some extent selection. A population with little diversity may have had a small mutation rate, so that few new variants arose in the past. But a low diversity can also be due to a small ancestral population with a young most recent common ancestor, since many variants that existed in the past have then been lost or fixed due to genetic drift (see Figure 8). Alternatively, a population with a large amount of diversity either had a large mutation rate or a large ancestral size/old most recent common ancestor. Consequently, if the estimated mutation rate is chosen too high, the ancestral size will be underestimated and the root of the tree will be dated too recently. On the other hand, if the mutation rate chosen is too small, the most recent common ancestor will be pushed too far back in time. It is therefore crucial to have a reliable molecular clock in order to estimate the age of humans correctly by this method. Since the mutation rates are so small, it is only recently that researchers have been able estimate them *de novo,* by comparing the DNA of children with their parents' and counting the fraction of nucleotides that differ [29]. Before that, the molecular clock was often calibrated under an assumption that we shared common ancestry with chimps. A typical estimate of the mutation rate was based on the postulated divergence time between humans and chimps, and the observed amount of divergence between parts of the two species' DNA.

However, *if humanity was founded recently by a single couple*, then the age of the founding generation and its amount of created genetic variation will also impact the diversity we see today. Y-chromosome DNA should then have much less diversity, since all Y-chromosomes descend from one singly copy, and few germline mutations have occurred since the founding generation. But this is not necessarily the case for autosomal and X-chromosome DNA, if the founding couple was created with diversity, or for mtDNA, if the first woman's mitochondria were diverse. That is, a high amount of nucleotide diversity in DNA other than Y-chromosomes, does not necessarily indicate an old population.

Second, the allele frequency spectrum (ii) gives additional information about the past. If the population has many rare variants, there are many possible explanations, for instance a rapidly expanding population (since most individuals lived recently, and therefore most mutations happened in the near past), a highly subdivided population with many more or less isolated tribes (since different alleles get fixed in different subpopulations, and each such locally fixed allele tends to have a small frequency in the whole population), or an excess of deleterious alleles that have not been removed by purifying selection. A flatter allele frequency spectrum with more common variants may indicate a recent bottleneck, a small and recent founding population or balancing selection, where the two SNP variants will tend to stabilize.

Considering the unique origin, created diversity model, if the founders lived recently and with genetic diversity, most of this diversity would remain today, unless the population has experienced an extreme bottleneck. The number of common variants would then be high, in particular if the founding population is young (Fig. 10).

Third, the LD plots (iii) provide further knowledge, but they are affected by several factors. Mutations, genetic drift, subpopulation division and certain types of directional selection will all generate linkage disequilibrium. It is still possible to observe some general patterns, for instance that LD extends over larger SNP distances for a small population, or one that is fragmented into a number of subpopulations (Fig. 10).

Fourth, subpopulation differentiation (iv) indicates partial isolation between different people groups. But it is difficult to reconstruct exactly how this division came about [30]. If, for instance, two subpopulations are related but yet genetically distinct, it is hard to distinguish a scenario where one smaller population was first colonized by members of another larger one, and then kept in isolation, from another scenario where this exchange was more gradual, with migration to the smaller group extending over a longer period of time.

2B. Common Descent

The common descent theory of humanity currently holds that our ancestors diverged from chimps about six million years ago. Then a hominid species *Homo erectus* evolved in Africa. It spread to Europe and Asia about 2 million years ago. Various archaic species are believed to have evolved from *Homo erectus* the last 500,000–800,000 years, including Neanderthals in Europe and Denisovans in Asia. There are two main variants within this framework for how *Homo sapiens,* our species, came about:

1. **Out of Africa replacement model**: According to this theory [31], modern humans evolved from *Homo erectus* in Africa more than 100,000 years ago. Then they went through a severe bottleneck that reduced the population size to an order of 10,000 individuals or smaller. A large part of this group emigrated from Africa about 50,000 years ago to the Middle East, Europe, East Asia and America, gradually replacing existing archaic species. After leaving Africa, all non-African populations have experienced much more recent and severe bottlenecks before they started to grow.[12]
2. **Multiregional evolution** posits that our ancestors evolved in parallel from archaic species in several parts of the world, possibly with an African dominance [32]. As a consequence, we have to trace human lineages up to 2 million years before they all end up in Africa.

The replacement model has been the most popular common descent model of human history for several decades, but there is no distinct boundary between it and multiregional evolution. An Out of Africa scenario with some interbreeding with archaic populations is not too different from a multiregional model with an African dominance. The last few years, ancient DNA (aDNA) has been retrieved from Neanderthal and Denisovan bones in different parts of the world[13] and compared with that of present day humans. These studies reveal that all human populations except sub-Saharan Africans' have about 1–2% of Neanderthal DNA. In fact, it seems that as much as 40% of Neanderthal DNA is found in at least some individuals alive today [37]. Lower levels (fractions of 1%) of Denisovan ancestry can be found mainly in South East Asia, Oceania and among Native Americans. This has caused many researchers to adopt a hybrid of the replacement and multiregional models, according to which our ancestors originated from Africa, but still had some interbreeding with archaic populations [25].

2C. Unique Origin

A unique origin model does not exclude per se the possibility that humans arose from more than two individuals. A designer could choose to start the human race with a larger number of people than one single pair. But we will argue below that from a scientific point of view it is not even necessary to start with a larger group of individuals. By a unique origin model we therefore mean one in which humanity originates from a single couple. We also have to address when the first man and woman lived and their geographic origin. It turns out that these two questions are intermixed. There are at least two versions of the unique origin model, with different age *and* geographic ancestry of the first couple:

1. **African ancestry.** This is a scenario by which the first couple lived in Africa. It has many similarities with the Out of Africa model, except for the unique origin assumption. The subsequent migration scenario out of Africa could be similar for both models. In the next section we will argue that unique origin with an African ancestry typically gives old estimates for the age of humanity.
2. **Middle East ancestry** posits that the most recent common ancestors (MRCA) of all humans lived in the Middle East. The subsequent migration from Middle East to Europe, Asia, America and Oceania could be similar to that of the Out of Africa and the African unique origin models. The crucial difference is that Africa was colonized from the Middle East rather than the opposite. In the next section we will argue that in light of genetics this gives a much younger age of humanity.

3. COMPARISON AND EVALUATION

The crucial question is which of the common descent or unique origin scenarios actual genetic data supports the most. Since there are many types of evidence, we will divide the argument into several parts.

3A. Differences with Other Species

As briefly mentioned in Section 1C, a major drawback of the common descent models is their difficulty with handling larger genetic differences between humans and other species. There are indeed research papers that try to estimate a common genealogy of humans, chimps and gorillas, using those parts of their genomes that show more similarity, with gene trees within species trees and so called trans-species polymorphisms [38,39]. But comparisons of this kind have limited scope, since they focus too little on the regions where the species differ [40]. Models that compare human DNA with that of other species should incorporate the difficulty for mutations and other

[12] According to [3], these bottlenecks reduced effective population sizes to less than 1,500 individuals, lasted at least ten thousand years, and may have ended as recently as 15,000-20,000 years ago.

[13] Four of the first papers on sequencing of archaic DNA are references [33]–[36].

genome arrangements to build up such genomic interspecies divergence, as well as anatomical and physiological differences. Other studies that take this into account typically reveal that the time it takes for these mutations to appear is much longer than what macroevolution requires [41].

3B. Variability in Human Genetic Data

The main argument against a unique origin is that the nucleotide diversity of human DNA data seems too high in order make a single founding couple possible. Genetic data also indicates that all non-African populations (and some north-African ones) are quite closely related. According to the Out of Africa-model, this is explained by a severe and very recent and long-lasting bottleneck of a few thousand individuals (somewhere in the time period 10,000 to 50,000 years ago) that all non-Africans ancestors supposedly experienced after they left Africa [3]. At some time, after the departure out of Africa, these ancestors started to expand, diverge and spread to the rest of the world, possibly with some migration back to northern Africa [42]. This is perhaps the main reason why this model is more popular today than multiregional evolution. The evidence for such a recent ancestry between non-Africans includes nucleotide diversity estimates between their people groups, their allele frequency spectra and LD plots. Sub-Saharan African populations, on the other hand, look older, at least at first sight. Their nucleotide diversity is higher, their allele frequency spectra have a larger fraction of rare variants, and there is more variation along their chromosomes (a shorter range of LD). There are also considerable genetic differences between African groups, indicating that their ancestors lived in small and relatively isolated tribes [43,44].

In order to trace the roots of the non-African and African branches, one uses diversity estimates between these groups. For autosomal and X-chromosome DNA, this leads to a common ancestry of all humans of the order one million years ago, but the uncertainty of these estimates is large. Analysis of Y-chromosome and mitochondrial DNA leads to a male and female tree whose roots (Y-chromosome Adam and mitochondrial Eve) are dated between 100,000 to 200,000 years ago, see [45–48] and references therein.

These arguments for the Out of Africa model seem convincing at first. But they could also be used for a unique origin model with an old African ancestry. The only difference is that the ancestral population of all people alive today (supposed to have lived in Africa no later than the roots of the mitochondrial and Y-chromosome trees) is only a single couple. From a common descent point of view one may argue that such a founding couple is impossible, since the ancestral population at this time should have a certain amount of diversity in order to explain the diversity we see today. But this is indeed possible if they were *created* with genetic diversity in their autosomal and X-chromosome DNA. Although the first male had no created diversity in his Y-chromosome, this may not contradict the diversity among Y-chromosomes that we see today, since it is actually smaller than previously believed [49].

But what about a founding couple from the Middle East? Is this unique origin version incompatible with data? Not necessarily. This model requires that the age of humanity is much more recent. The reason is that the first unique couple replaces the long-lasting bottleneck that supposedly occurred after immigration to the Middle East from Africa [3]. And the genetic diversity that remains after this bottleneck population is replaced by created diversity in this founding pair. This would explain the relatively large genetic variability we find among humans of today for non-sex and X-chromosome DNA (recall that the diversity of Y-chromosomes is much smaller).

The main challenge of the unique origin Middle East ancestry model is to explain why African populations look older than the non-Africans ones, and how the variation of Y-chromosome DNA came about without any founder diversity.[14] There are some tentative explanations, and future research will tell which of them are most credible. First, recent studies indicate that the nucleotide diversity of African autosomal, X- and mitochondrial DNA is only moderately larger (less than twice) than for non-African DNA [49]. This relatively small difference need not only be explained by an older African population. Since Africans lived in small, more or less isolated tribes, this would increase nucleotide diversity among them, and the fraction of rare variants in the allele frequency spectrum, both in the whole African population as well as for African regions. Even though there will be less diversity in each small tribe (because of a high amount of genetic drift in a small population), different alleles tend to be fixed in different tribes, creating an overall larger variability in African regions as well as in the whole African population. The consequence of this is that African populations may look older than they are, as long as they are not analyzed at the tribe level. Second, it has recently been found that for Y-chromsomes it is the other way around. The diversity of Y-DNA is actually slightly smaller among Africans that among non-Africans [49]. This is more in line what one would expect if the African population is not older. Third, as mentioned above, nucleotide diversity of Y-chromosomes seems to be an order of magnitude smaller than for other types of DNA, across different people groups around the world [49]. Various explanations have been given, for instance that the male populations experienced much more severe bottlenecks that the female ones [42]. But these findings also fit a unique origin explanation whereby autosomal and X-DNA were created with diversity among the founding couple, whereas Y-DNA had no such created diversity. Mitochondrial DNA may or may not have had created diversity, but its higher mutation rate may in any case explain part of its larger diversity. Fourth, since African non-sex and X-chromosomes look more scrambled (shorter LD range), a possible explanation is that recombination rates among many Africans are higher due to different alleles of the PRDM9 gene, as a recent study indicates [50]. Fifth, the dating of the Y-chromosome male tree and the mitochondrial female tree depends crucially on where the roots of these trees are put. While the topologies of the two trees are unambiguous, whether one uses a common descent

[14] As mentioned in Section 1D, the first female carried multiple mitochondria, possibly with different DNA, and she may have passed that diversity on to her daughters.

or unique origin approach, their roots may be chosen in many ways. A Middle East ancestry of humanity is consistent with a root along the branch that unites all non-African lineages. One may argue that this is not reasonable. There are more mutations along the African branches, and this seems to indicate an older common ancestry of African groups. Therefore, the roots of the Y-chromosome [46,47] and mitochondrial [48] trees should be placed between those African branches that look oldest. While this is a legitimate objection, it is still too early to rule out a recent African and worldwide human ancestry. It is only the last few years that mutation rates are being estimated *de novo* with a fairly high precision for nuclear [29,51], mitochondrial [13,52] and Y-DNA [53–55], and we still have an incomplete knowledge of how it varies between chromosomal regions and people groups. These improved estimates suggest that the mtDNA mutation rate is much higher than previously believed [56,57], and since Y-chromosomes are male-specific, previous predictions that its mutation rate is higher than for autosomes [58] have recently been confirmed [54]. This makes a younger dating of its root at least more reasonable. The tribe isolation of African ancestors may additionally complicate the dating of the mitochondrial and Y-chromosome trees, and some African subpopulations may have had shorter generation times [59].

3C. Block Structure of DNA

There is another very interesting feature of human genetic data, discovered more than fifteen years ago [60,61]. It was seen from LD plots that a large part of our non-sex and X-chromosomes can be divided into blocks. Early studies [62,63] predicted that block lengths vary a lot, and that most of them are between 5,000 to 200,000 nucleotides. With larger amounts of genetic data it is possible to make more accurate predictions. A recent paper suggests that the block lengths are smaller, with an average size of about 5,000 nucleotides [64]. In any case, the blocks are several thousand nucleotides long, and in spite of this there is very little variation within them. Each block comes in a few variants, four for many parts of the genome. The non-sex and X-chromosomes of human sperm or ova cells are different mosaics of these block variants.

This DNA block structure is remarkably consistent with a unique origin hypothesis. If the first human couple was created with DNA diversity, there are four different copies of each non-sex chromosome; two in the male founder and two in the female one. Their four chromosomes have since then been scrambled by ancestral recombinations, and today each of us has inherited one mosaic of the four founder chromosomes from our father, and another one from our mother.

The DNA blocks can be seen in all human populations, but they tend to be longer for non-Africans than for Africans. This may indicate that African populations are older, but it is also possible that recombinations happen more often among Africans, as some recent research indicates [50].

The existence of these blocks could be problematic from a common descent point of view, or for a unique origin approach that dates humanity far back in time. If recombinations happen randomly throughout the chromosomes, and if our most recent common ancestor lived a long time ago, we should see many of them in DNA. This would require much shorter blocks than ten or hundred thousand nucleotides. The solution to this problem is to assume that recombination doesn't happen randomly along the chromosomes, but at certain hotspots. If so, many recombinations of the past happened at the same hotspot. And if we cannot distinguish them, the number of ancestral recombinations could have been much larger than the number of blocks. Then the argument goes that we cannot rule out an old age of humanity from the DNA blocks.

It seems too early to say to which extent there are recombination hotspots. There are first of all biological reasons for some recombination rate variation. It has been known for many decades that female recombination rates are higher than for males [65], and that the rates for both sexes vary along the chromosomes on a course scale of several million nucleotides [66]. It is also reasonable to assume that recombination rates vary between coding DNA, non-coding DNA within genes, and intergenic DNA. Since there are so many potential recombination hotspots, and the recombination probability is very small (on average about 10^{-8} per nucleotide of each germ cell, corresponding to a recombination probability of the order 10^{-4} or smaller per block), large amounts of data are required to test whether they exist or not. Still, observed de novo recombinations (for instance through sperm typing in males) reveal that some hotspots do exist [50]. It may be the case that some of the blocks we see are caused by hotspots with many recombinations, whereas others are caused by single ancestral recombination events.

3D. Inbreeding Depression and Genetic Entropy

It is well known that many alleles will be lost due to genetic drift when a population experiences a severe bottleneck. The long term consequence is a decreased ability for long term adjustment to environmental changes, but there is also a more acute risk of inbreeding depression when the frequency of recessive disorders increases, as more offspring receive from both of their parents the harmful variant of the disease-causing gene. The smaller the population is during the bottleneck, and the longer time it takes before its size starts to increase, the more severe are the consequences for the population's viability, so that ultimately it may die out [67]. Conservation biologists have devised rules for minimal population sizes in order to ensure short and long term protection of animal species [68]. For humans there are several well known examples of the drastic effects of continued inbreeding, such as the extinction of the Spanish Habsburg dynasty around 1700 [69] and the high occurrence of severe form of color blindness on the Micronesian atoll of Pingelap, after a typhoon hit the island in 1775, and 90% of the inhabitants died. All present-day inhabitants can trace their ancestry to one of the survivors that carried the harmful variant of the gene that codes of this type of color blindness [70].

Inbreeding depression is potentially a difficulty for the Out of Africa model. Recent calculations reveal that the model predicts a very small bottleneck of the non-African ancestors after they left Africa, of the order a few thousand individuals, and that it lasted for at least 1,000 generations [3]. In spite of this, the survivors of the bottleneck are believed to have conquered the rest of the world. A unique origin model with an old age of humanity faces a similar challenge of inbreeding depression, since some kind of bottleneck seems necessary in order to explain the relatively small diversity we see among humans today. But this problem is still smaller than for the common descent scenario, if the created diversity of the first pair had only neutral variants, whereas the harmful ones occurred later through germline mutations.

But what about a more recent unique origin model that starts with two people? Isn't such a bottleneck much more severe than a few thousand individuals, especially if sons and daughters of the first man and woman had children together? Not necessarily. Again, if the founding couple was created with diversity, with no harmful alleles, and had many children and grandchildren, the population would have expanded quickly without any risk of inbreeding depression. Such a short-lasting bottleneck is not associated with any appreciable loss of genetic variants. The same is true if there was a subsequent very short bottleneck followed by a rapid expansion.[15]

A model with a young age of humanity has another advantage. It can handle the problem of increased genetic entropy.[16] If all present-day harmful variants arrived through germline mutations in descendants of the first pair, there has not been enough time to accumulate them into large numbers [71]. In contrast, any model with an old age of humanity faces a problem, since either the number of slightly deleterious or deleterious alleles increases, or a bottleneck occurred that removed some of these harmful variants, but at the cost of spreading others so that inbreeding depression might occur.[17]

3E. Archaic Populations, Humans or Not?

Recall that significant fragments of Neanderthal and Denisovan DNA have been found among present day humans. This made researchers suggest that some interbreeding took place between archaic populations and the ancient humans that supposedly emigrated out of Africa.[18] This admixture is believed to have happened at least 50,000 years ago, and probably later on as well. It is in fact well known that gene flow between closely related populations is helpful in order to increase genetic variability and to avoid inbreeding, and indeed, the archaic introgression is believed to have had positive effects, like helping the Tibetans to adapt to high altitudes, and the non-Africans in general to adapt to colder temperature [28,72] and to ward off infections [73,74]. But Out of Africa replacement adherents also use various common descent assumptions (such as the divergence time of humans and chimps) and genetic diversity estimates between humans and archaic hominins, to predict a split between them about 500,000 years ago or earlier [28,75,76,77]. If this is true, it is remarkable that two populations, after such a long time of separation, were still able to get fertile offspring [78]. But even if this would be possible, because of the long separation, it is reasonable to believe that the offspring had low fitness, since our archaic ancestors had, most likely, accumulated many alleles which are deleterious for humans, before the admixture took place.

The large fraction of archaic DNA among present-day humans seems in view of this more reconcilable with a unique origin model in which Neanderthals and Denisovans are descendants of the first founding couple, and hence our fully human ancestors. Indeed, sequencing of mitochondrial DNA suggests that the diversity among Neanderthals is much smaller than among humans [79]. As a possible explanation, they could have been quite early descendants of the first man and woman. And the close genetic resemblance between Neanderthals, Denisovans and people of today suggests that the morphological differences are mostly explained by changed gene expression due either to mutations of regulatory DNA or to epigenetic changes [80].

3F. Conclusions

It is now time to summarize and answer the question that was posed in the introduction, whether a unique origin or common descent scenario for humanity is most consistent with DNA of humans, chimps and other apes. We have argued that a unique origin model with created diversity and an old African ancestry or a more recent Middle East ancestry should have at least the same explanatory power for human genetic data as the most popular common descent scenario of today, a variant of the Out of Africa replacement model with some interbreeding with archaic populations. Any common descent model faces a challenge to explain the genetic differences rather than the similarities with other species, the consequences of inbreeding depression and increased genetic entropy, human DNA mixture with archaic populations, and that our DNA resembles a mosaic of about four founder genomes. The provisional conclusion is that a unique origin model seems more plausible.

Among the unique origin scenarios, the one with Middle East ancestry and a young dating of humanity has on one hand some advantages over an African ancestry model, at least if the latter has its founding couple far back in time. If humanity is young, inbreeding depression is not a problem, and the block-like DNA structure could possibly be explained as a combination of single historical recombination events and recombination hotspots. On the other hand, the Middle East ancestry model faces some challenges, in particular to explain why African DNA looks older than non-African DNA. Future research will tell which of the two unique origin models fits data the best. Such a comparison should not only involve genetics, but the interpretation of fossils data is crucial as well.

In principle, it is possible to combine common descent with a single man and woman that are the ancestors of all humans. This couple would represent an extreme bottleneck of two individuals in the lineage that connects humans and chimps.

[15] This includes, for instance, a very severe bottleneck, with only eight persons surviving. But it can actually be shown that if the population prior to this bottleneck was fairly large, and after that quickly expanded, then eight persons are enough to retain most of the created genetic variants of the first couple.

[16] See Section 2D.

[17] See [17], and references therein, for estimates of the fractions of mutations that are neutral, slightly deleterious or deleterious. If the total number of slightly deleterious alleles is large, their cumulative effect may be large as well.

[18] See Section 2B.

However, we argue that such a scenario is less likely than a unique origin model, since the bottlenecked couple would have inherited many deleterious mutations from their ancestors, so that inbreeding depression is a major issue.

4. IMPLEMENTATION AND VALIDATION OF UNIQUE ORIGIN MODELS

The qualitative arguments of Section 3 lead us to conclude that a unique origin scenario of human history seems plausible. It is of interest to follow up this with a more formal way of testing the unique origin scenario. The basic idea is to simulate genetic data from each proposed model many times, and then compare how well the simulated output fits real data. This comparison should include autosomal, sex chromosome and mitochondrial DNA, using nucleotide diversity, allele frequency spectra, LD plots and other statistics. There are at least two different ways to proceed with these simulations.

4A. Forward simulation

The most straightforward way of simulating genetic data is to start at the founder generation and then proceed forward in time [81]. For each simulation round one first assigns genomes to the founding couple. Demographic and genetic data are then simulated one generation at a time, using all mechanisms I–VI of change. This includes rules for how couples mate, how their number of children varies between families, how often new geographic areas are colonized, and how often people move between regions. The hereditary principles of Section 1A are used to pass on DNA from one generation to the next, with Mendelian principles of inheritance and randomly located recombinations. Selection is included by allowing the survival probability of a fertilized egg to depend on its genome, according to some fitness function.

The main advantage of forward simulation is its great flexibility. Virtually any type of model for human history can be simulated and validated with real data. But the method requires that DNA of all humans is simulated. In view of the size of the worldwide human population, this is very time consuming.

4B. Backward simulation

There is another much faster simulation algorithm. A more detailed description of it can be found in our accompanying paper (Part 2) [82]. In each simulation round only a small subset of genetic data is generated. The main idea is to first select a small sample of humans alive today (for instance a few thousand individuals), then simulate their genealogy backwards in time (using a method called coalescence theory, see for instance [10]) until the founder generation is reached. When the genealogy has been generated, DNA is assigned, first to the founding couple, and then spread forwards in time to all their descendants along the branches of the simulated genealogy. The genealogy will only be a small subset of the ancestral human population. It only includes those individuals that passed on DNA to at least one of the individuals of today that were included in the sample (see Fig. 11).

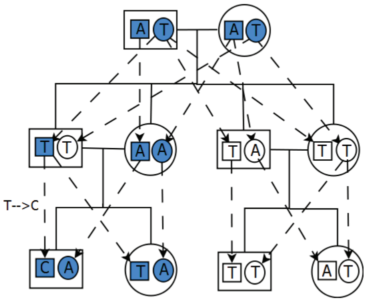

Figure 11: A three generation population with males and females shown as boxes and circles, respectively. The pedigree is drawn with solid lines and the gene tree at one autosomal nucleotide with dashed lines. (Other parts of the genome will have different gene trees due to recombinations.) Spouses are connected by horizontal lines, with their children vertically below them. If only the leftmost sibling pair of the last generation is sampled, and its gene tree is built backwards in time, only the shaded (light blue) part of the gene tree is needed. The nodes of the gene tree are variants (A,C,G,T) of the nucleotide. There is diversity of the founder generation, since the male and female both have an AT genotype. Then a germline mutation T⟶C occurs in the leftmost male of the third generation. **doi:**10.5048/BIO-C.2016.3.11

We refer to this method as backward simulation, since the genealogy is simulated backwards in time. While it is a lot faster than forward simulation, it is not as general. In order to build the genealogy, it is necessary to use a neutral model without natural selection. Although this is a limitation, we argued above[19] that selection has a limited role to play in order to explain most microevolutionary changes. It may be important for certain chromosomal regions, but a neutral model is likely to be accurate when genome-wide measures like nucleotide diversity, allele frequency spectra and LD plots are generated.

We are currently working on implementing a model based on backward simulation. The intent is to validate it with real data. This is a long-term project, whose outcome we hope to publish later.[20] Using this approach, it may be possible to demonstrate that a unique origin model is able to replicate current human diversity as well or better than the common descent model. That is the purpose of the model—to test this possibility. Therefore, if more than one plausible account of human origins can explain the data, the common descent model of our origin from ape-like ancestors can no longer be claimed as conclusive proof that there could not have been a single first pair.

[19] See Section 1D.

[20] Updates on the project can be found at the website uniqueoriginresearch.org.

1. Hartl D (2000) A Primer of Population Genetics, 3rd edition. Sinauer Associates (Sunderland, MA).
2. Haines JL and Pericak Vance MA eds. (1998) Approaches to Gene Mapping in Complex Human Diseases. Wiley-Loss (New York).
3. The 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. Nature 526:68–87. doi:10.1038/nature15393
4. Crow JF and Kimura M (1970) An Introduction to Population Genetics Theory. The Blackburn Press (Caldwell, NJ).
5. Huxley JS (1942) Evolution: The Modern Synthesis. Allen and Unwin (London).
6. Mayr E (1982) The Growth of Biological Thought. Harvard University Press (Cambridge, MA).
7. Allendorf FW, Luikart GH (2007) Conservation and the Genetics of Populations. Blackwell Publishing (Malden, MA).
8. Behe MJ (1996) Darwin's Black Box. Simon & Schuster (New York).
9. Ewens WJ (2004) Mathematical Population Genetics. 2nd ed. Springer Verlag (New York). doi:10.1007/978-0-387-21822-9
10. Durrett R (2008) Probability Models for DNA Sequence Evolution. 2nd ed. Springer Science (New York). doi:10.1007/978-0-387-78168-6
11. Shapiro JS (2011) Evolution: a View from the 21st Century. Pearson Education Inc (Upper Saddle River, NJ).
12. Noble D (2013) Physiology is rocking the foundations of evolutionary biology. Exp Physiol 98:1235–1243. doi:10.1113/expphysiol.2012.071134
13. Howell N (2003) The pedigree rate of sequence divergence in the human mitochondrial genome: There is a difference between phylogenetic and pedigree rates. Am J Hum Genet 72: 659–670. doi:10.1086/368264
14. Kimura M (1983) Neutral Theory of Molecular Evolution. Cambridge University Press (New York). doi:10.1017/CBO9780511623486
15. Lynch M (2007) The frailty of adaptive hypotheses for the origins of organismal complexity. Proc Natl Acad Sci USA 104 (suppl 1):8597–8604. doi:10.1073/pnas.0702207104
16. ReMine WJ (1993) The Biotic Message. Evolutions Versus Message Theory. St. Paul Science Publishers (Saint Paul, Minnesota).
17. Sanford J (2008) Genetic Entropy and the Mystery of the Genome, 3rd ed. FMS Publications (Waterloo, New York).
18. Gauger A, Axe D, Luskin C (2012) Science and Human Origins. Discovery Institute Press (Seattle).
19. Parker G (1980) Creation, mutation and variation. Acts and Facts 9:11.
20. 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 mechanism. Answ Res 9:81–122.
21. Ardlie KG, Kryglyak L and Seielstad M (2002) Patterns of linkage disequilibrium in the human genome. Nat Rev Genet 3:299–309. doi:10.1038/nrg777
22. Sachidanandam R et al (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409:928–933. doi:10.1038/35057149
23. Kim HL, Schuster SC (2013) Poor man's 1000 genome project: recent human population expansion confounds the detection of disease alleles in 7,098 complete mitochondrial genomes. Front Genet 4:Article 13. doi:10.3389/fgene.2013.00013
24. Keinan A, Clark A (2012) Recent explosive human population growth has resulted in an excess of rare genetic variants. Science 336:740–743. doi:10.1126/science.1217283
25. Thomas D (2004) Statistical Methods in Genetic Epidemiology. Oxford University Press (New York).
26. Li JZ, Absher DM, Tang H, Southwick AM et al (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319: 1000–1004. doi:10.1126/science.1153717
27. Pritchard JK et al (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959.
28. 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. doi:10.1186/s13059-015-0866-z
29. Campbell CD, Eichler EE (2013) Properties and rates of germline mutations in humans. Trends Genet 29:575–584. doi:10.1016/j.tig.2013.04.005
30. Mazet O, Rodriquez W, Grusea S, Boitard S, Chikhi L (2016) On the importance of being structured: Instantaneous coalescence rates and human evolution — lessons for ancestral population size inference? Heredity 116: 362–371. doi:10.1038/hdy.2015.104
31. Mellars P (2006) Going east: New genetic and archaeological perspectives on the modern human colonization of Eurasia. Science 313:796–800. doi:10.1126/science.1128402
32. Wolpoff MH, Hawks J, Caspari R (2000) Multiregional, not multiple origins. Am J Phys Anthropol 112:129–136. doi:10.1002/(SICI)1096-8644(200005)112:1<129::AID-AJPA11>3.0.CO;2-K
33. Green RE, Krause I, Biggs AW, Maricic T, Stenzel U, Kircher M et al (2010) A draft sequence of the Neandertal genome. Science 328:710–722. doi:10.1126/science.1188021
34. Reich D, Green RE, Kircher M, Krause I, Patterson N, Durand EY et al (2010) Genetic history of an archaic hominin group from Denisova cave in Siberia. Nature 468:1053–1060. doi:10.1038/nature09710
35. Prufer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S et al (2014) The complete sequence of a Neanderthal from the Altai mountains. Nature 505:9. doi:10.1038/nature12886
36. Meyer M, Kircher M, Gansauge M-T et al (2012) A high coverage genome sequence from an archaic Denisovan individual. Science 338:212–216. doi:10.1126/science.1224344
37. Pääbo S (2015) The contribution of ancient hominin genomes from Siberia to our understanding of human evolution. Herald of the Russian Academy of Sciences 85:392–396. doi:10.1134/S1019331615050081
38. Chen F-C, Li W-H (2001) Genetic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am J Hum Genet 68:444–456. doi:10.1086/318206
39. Yang Z (2002) Likelihood and Bayes estimation of ancestral population size in hominoids using data from multiple loci. Genetics 162:1811–1823.
40. Kehrer-Sawatzki H, Cooper D (2007) Understanding the recent evolution of the human genome: Insights from human-chimpanzee genome comparisons. Hum Mutat 28:99–130. doi:10.1002/humu.20420
41. Sanford JC, Brewer W, Smith F, Baumgardner J (2015) The waiting time problem in a model hominin population. Theor Biol Med Model 12:28 pages.
42. Poznik GD Xue Y, Mendez FL, Williams TH, Massaia A, Sayres MAW et al (2016) Punctuated bursts in human male demography inferred from 1244 worldwide Y-chromosome sequences. Nat Genet 48:593-599. doi:10.1038/ng.3559
43. Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D (2005) Calibrating coalescent simulation of human genome simulation. Gen Res 15:1576–1583. doi:10.1101/gr.3709305
44. Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. Nature 475:493–496 doi:10.1038/nature10231
45. Blum MGB, Jakobsson, M (2011) Deep divergence of human gene trees and models of human origin. Mol Biol Evol 28(2):889–898. doi:10.1093/molbev/msq265

46. Poznik GD, Henn BM, Yee M-C et al (2013) Sequencing Y-chromosomes resolves discrepancy in time to common ancestor of males versus females. Science 341:562–565. doi:10.1126/science.1237619
47. Fracalacci P, Morelli L, Anguis A et al (2013) Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny. Science 341:565–569. doi:10.1126/science.1237947
48. Olivieri A, Achilli A, Pala M, Battaglia V et al (2006) The mtDNA legacy of the Leventine early upper Palaeolithic in Africa. Science 314:1767–1770. doi:10.1126/science.1135566
49. Wilson Sayres MA, Lohmuller KE, Nielsen, R (2014) Natural selection reduced diversity on human Y chromosomes. PLoS Genet 10 (2014):e1004064. doi:10.1371/journal.pgen.1004064
50. Hinch AG et al (2011) The landscape of recombination in African Americans. Nature 476:170–177. doi:10.1038/nature10336
51. Kondrashov AS (2003) Direct estimate of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. Hum Mut 21(1):12–27. doi:10.1002/humu.10147
52. Ding, JC, Li, C-I et al (2015) Assessing mitochondrial DNA variation and copy number in lymphocytes of ~2000 Sardinians using tailored sequencing analysis tools. PLoS Genet 11(7):e1005306. doi:10.1371/journal.pgen.1005306
53. Xue Y et al (2009) Human Y-chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. Curr Biol 19:1453–1457. doi:10.1016/j.cub.2009.07.032
54. Helgason A et al (2015) The Y-chromosome point mutation rate in humans. Nat Genet 47:453–457. doi:10.1038/ng.3171
55. Balanovsky O, Zhabagin M, Agdzhoyan A, Chukhryaeva M, Zaporozhchenko V et al (2015) Deep phylogenetic analysis of haplogroup G1 provides estimates of SNP and STR mutation rates on the human Y chromosome and reveals migrations of Iranic speakers. PLoS ONE 10(4):e0122968. doi:10.1371/journal.pone.0122968
56. Jeanson NT (2013) Recent, functionally diverge origin for mitochondrial genes from ~2700 metazoan species. Answ Res J 6:467–501.
57. Jeanson NT (2015) A young-earth creation human mitochondrial DNA ‘‘clock’′: Whole mitochondrial genome mutation rate confirms D-loop results. Answ Res J 8:375–378.
58. Jobling MA, Tyler-Smith C (2003) The human Y-chromosome: An evolutionary marker comes of age. Nat Rev Genet 4:598–612. doi:10.1038/nrg1124
59. Jeanson NT (2016) On the origin of human mitochondrial DNA differences, new generation time data both suggest a unified young-earth creation model and challenge the evolutionary out-of-Africa model. Answ Res J 9:123–130.
60. Daly MJ, Rioux JD, Schaffner SF, Hudson TF, Lander EL (2001) High-resolution haplotype structure in the human genome. Nat Genet 29:229–232. doi:10.1038/ng1001-229
61. Gabriel SB et al (2002) The structure of haplotype blocks in the human genome. Science 296:2225–2229. doi:10.1126/science.1069424
62. Pääbo S (2003) The mosaic that is our genome. Nature 421:409–412. doi:10.1038/nature01400
63. 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
64. 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
65. 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
66. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudbjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G et al (2002) A high-resolution recombination map of the human genome. Nat Genet 31:241–247. doi:10.1038/ng917
67. Lynch M, Conery J, Burger R (1995) Mutation accumulation and extinction of natural populations. Am Nat 146:489–518. doi:10.1086/285812
68. 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).
69. Alvarez G, Ceballos FC, Quinteiro, C (2009) The role of inbreeding in the extinction of a royal dynasty. PLoS ONE 4:e5175.
70. Norton NE, Lew R, Hussels IE, Little GF (1972) Pingelap and Mokil atolls: Historical genetics. Am J Hum Genet 24:277–289.
71. Sanford JC, Carter R (2008) In light of genetics ... Adam, Eve and the Creation/Fall. Christ Apol J 12(2):51–98.
72. Sankararaman S et al (2014) The genomic landscape of Neanderthal ancestry in present-day humans. Nature 505:43–49. doi:10.1038/nature12961
73. 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
74. Danneman M, Andrés AM, Kelso J (2016) Introgression of Neandertal- and Denisovan-like 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
75. Green RE, Malaspinas A-S, Krause J, Briggs AW, Johnson PLF et al (2008) A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. Cell 134:416–426. doi:10.1016/j.cell.2008.06.021
76. Mendez FL, Poznik DG, Castellano S, Bustamante, CD (2016) The divergence of Neandertal and modern human Y-chromosomes. Am J Hum Genet 98:728–734. doi:10.1016/j.ajhg.2016.02.023
77. Stringer C (2016) The origin and evolution of Homo Sapiens. Phil Trans R Soc B 371:20150237 doi:10.1098/rstb.2015.0237
78. Burgoyne PS, Mehadevala, SK, Turner, JMA (2009) The consequences of asynapsis for mammalian meiosis. Nat Rev Genet 10:207–216. doi:10.1038/nrg2505
79. Briggs, AW, Good JM, Green RE, Krause J, Maricic T et al (2009) Targeted retrieval and analysis of five Neandertal mtDNA genomes. Science 325:318–321. doi:10.1126/science.1174462
80. Gockman D, Lavi E, Prüfer K, Fraga MF et al (2014) Reconstructing the DNA methylation maps of the Neandertal and Denisovan. Science 344: 523–527. doi:10.1126/science.1250368
81. 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.
82. Hössjer O, Gauger A, 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