Research Article

Genetic Modeling of Human History Part 2: A Unique Origin Algorithm

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

This paper presents a mathematical unique origin model of humanity. It suggests algorithms for testing different historical scenarios of the human population under the assumption that we all descend from one single couple. For each such scenario, DNA variation is repeatedly simulated from a sample of individuals of today in order to estimate statistics of DNA variation. Comparison of these statistics to real data makes model validation possible. Each simulation repeat is divided into three steps, where first the genealogy of the sampled individuals is simulated backwards in time until the founder generation is reached, then founder DNA is generated and thereafter spread forwards in time to the present, along the lineages of the ancestral tree. The model is applicable to predefined demographic scenarios that may include population expansions and bottlenecks. Colonization/range expansion and geographic migration is achieved by dividing the metapopulation into geographically separated, but more or less connected, subpopulations. Age structure is modeled in terms of overlapping generations, with various mating rules for males and females and reproduction rules of mating couples. On the genetic level, our model incorporates mitochondrial as well as nuclear (autosomal, X and Y chromosomal) DNA, ordinary (reciprocal) recombination events and gene conversion. The source of genetic variation is selectively neutral germline mutations, and for autosomal and X chromosomal DNA, a second source of variation is created diversity. An extension of the model allows for balancing selection. It combines forward and backward simulation of the genealogy. Our paper is a first step towards a future goal to compare a best fitting unique origin model with a common descent model where humans and other species have a shared ancestry.

Cite As: 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.

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.2016.4.c.

*Email: ola.hossjer@gmail.com

INTRODUCTION

The study of human history combines lines of evidence from several fields. Fossil records, archaeological findings and radiometric techniques are used to analyze and date morphological and cultural traits, and genetic data is used to trace common ancestry. The earliest genetic studies of human history relied on markers from blood groups and proteins [1,2]. The advent of new sequencing technologies in the 1980s initiated studies of mitochondrial DNA [3-9], the nonrecombining region of the Y chromosome [7,10-14] and nuclear autosomal DNA using microsatellite markers [15,16]. As the high throughput sequencing technology developed, millions of single nucleotide polymorphism markers could be analyzed at increasingly smaller costs, and DNA sequence variation was cataloged between and within human subpopulations [17-21]. This made it possible to analyze human ancestry with much larger autosomal DNA data sets, using increasingly sophisticated mathematical models [22-26]. Some of these methods are particularly well suited for studying the more recent human history [27]. Autosomal marker data has also been used to infer human ancestry and diversity from *Alu* insertion polymorphisms [28-30].

All proposed models of human evolution assume a common ancestry of man and other species. The most common one, the Out of Africa replacement model, asserts that modern humans arose in Africa more than 100,000 years (100 kyrs) ago. Then a very small subpopulation migrated to the Middle East around 50 kyrs ago, spread to Europe, Asia, Oceania and America, replacing Neanderthals, Denisovans and other existing local archaic populations [31-34].

Other models have also been proposed. The most well known of these is multiregional evolution. It posits that humans arose at several continents, but still with an African dominance through migration rather than replacement. It is believed that these human lineages, from different parts of the world, originate from Africa about two million years ago [35,36]. The last few years, technologies have been developed for sequencing of ancient DNA. They reveal a Neanderthal and Denisovan ancestry in many present-day human populations [37-39]. This 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 [40]. The common ancestry framework has also been used in cross-species studies in order to analyze autosomal DNA of humans, chimpanzees and gorillas jointly, with the purpose of estimating divergence times and ancestral population sizes of the assumed species tree [41-42].

The difficulty of inferring human population history from present genetic data is well known, since population size changes over time

and replacement of one population by another is confounded by migration and regional population size variation, see for instance [43-46] and Chapter 2 of [47]. Recent estimates of divergence times of gene trees in [48] confirm these difficulties, although the best fitting model for human history in this paper was a variant of the Out of Africa model, where different African archaic populations were connected by gene flow, although only one of them eventually colonized the other continents.

But it is not only difficult to choose between different common descent scenarios of human history. A number of evidences in favour of a different scenario have recently been pointed out in [49], whereby humans descend from one single couple that was created unique without further ancestry. We will refer to this as a unique origin model. In a previous paper (Part 1) we used a qualitative argument to compare the common descent and unique origin scenarios of humanity [50]. In this follow-up paper (Part 2) we build a quantitative framework by which a unique origin model can be tested.[1]

The methodological novelty of our paper is to put together a number of previously known methods into one model that incorporates many features, and our approach can briefly be summarized as follows: As in many other papers we first develop a mathematical model for human ancestry. Then based on this we devise an algorithm for simulating genetic data in terms of an ancestral tree for a sample of individuals for which DNA is supposedly collected. This tree aims to reflect human history, but, in contrast to previous work, it is built under the assumption that humanity originates from one single couple. In order to obtain a computationally feasible algorithm, the genealogy is built backwards in time, starting from the sampled individuals. The model is very flexible and allows for various demographic scenarios such as population expansions, bottlenecks and different geographic colonization and migration patterns. It also incorporates different mating and reproduction schemes, and age structure in terms of overlapping generations. The mechanisms of genetic change are applicable to autosomal, sex chromosome or mitochondrial segments of DNA. It also includes neutral mutations, reciprocal recombination events and gene conversion. An important parameter of the model is the created diversity of the founder generation, since it facilitates a higher degree of genetic diversity for a relatively young population within autosomal and X chromosomal regions, and possibly also for mitochondrial DNA.

The paper is organized as follows: In Section 1 we give a detailed overview of the model, including its input parameters and output statistics. We also propose methods for validating simulations with real data. Section 2 gives a more detailed description of how the genealogy is built backward in time, whereas Section 3 explains how ancestral mutations are generated and spread to the present. The model gets more complicated when mutations are not selectively neutral. In Section 4 we propose an extension for balancing selection based on a mixture of forward and backward simulation. Finally, in Section 5 we end with some concluding remarks. A list of notation can be found in Table 1.

1. MODEL

This section gives an overview of our proposed model for human history. Its demography in terms of population size variation, geographic subdivision, migration and colonization is specified in Section 1.1. In Sections 1.2 and 1.3 we describe the format for storing genetic data. We specify how chromosomes of individuals are represented as strings, divided into blocks within which no recombinations occur. Sections 1.4-1.7 contain a detailed description of the algorithm under a neutral model of microevolution. This is achieved by first simulating the genealogy backwards in time (Section 1.4) and then spreading DNA from the founder population to the present (Section 1.5). This second step is simplified considerably if double mutations are ignored (Section 1.6). The last two Sections 1.8 and 1.9 deal with methods of validating the model. This is accomplished by comparing how well the simulated output fits real data.

1.1 Demographics

We will first specify a demographic model of human history in terms of a world population of males and females, whose size varies over time, with geographic division into more or less isolated subpopulations. This is formalized by considering a two-sex population at a sequence of time points $T_0 = 0 < T_1 < \ldots < T_{\max}$, where $T_0 = 0$ represents the present, and $T_{\max}$ is the time point of the founding generation. If the population has non-overlapping generations, then $t = 0, \ldots, t_{\max}$ is a generation number, and $T_t - T_{t-1}$ is the time interval between generations t and $t-1$. The model is more general though, allowing for overlapping generations, so that $T_t - T_{t-1}$ represents a fraction of a generation. For this reason, we will mostly refer to t as a time point.

It is assumed that the world population has size $N_t = M_t + F_t$ at time t, of which M_t are males and F_t females. In particular $M_{t_{\max}} = F_{t_{\max}} = 1$ represents the founding couple. The members of time point t are numbered with the M_t males first, and then the F_t females, as $\{1, 2, \ldots, M_t, M_t + 1, \ldots, M_t + F_t\}$. This makes it possible to represent any individual by a pair (t, i) of numbers, a time index t when he or she lives and an order number i within that time point. Consequently, the whole human race is represented as a collection

$$\mathrm{I} = \{(t, i);\, 0 \leq t \leq t_{\max}, 1 \leq i \leq N_t\} \tag{1}$$

of individuals, males and females. For overlapping generations, some individuals will appear in (1) more than once, first as a newborn and then later on as an adult.

The next-generation sequencing data allows for the resolution, not only between continents, but also within continents and countries. The average genetic composition of individuals in a region will typically vary continuously with geographic location, see [51] and references therein. This is most likely a combination of ancestral colonization and founder events on one hand, and isolation by distance on the other. Here we propose a model which incorporates both of these two mechanisms. Its geographic substructure is discrete in terms of a number of subpopulations that either represent larger regions/continents (African, European, Middle East, East Asian, Polynesian, native American) or a finer division.

The island model of [52,53] is the first example of such an approach, where the metapopulation is divided into a fixed number of equally large and homogeneous islands, with the same migration rate between any pair of them. Because of its simplicity and analytical tractability, variants of this model have frequently been used for making inference about human history, see for instance [54-56]. With our simulation based approach, it is possible to consider more general models, with a metapopulation that is divided into a possibly time varying number D_t of demes of variable size. The migration rates can be chosen with great flexibility between any pair of demes in order to mimic geographic location (see Section 1.4), and the migration habits of males and females can be different

[1] It is possible, also within a common descent framework, that all humans descend from one single man and woman, if this couple had further ape-like ancestors. In Part 1 we argued that inbreeding depression is a severe problem for such a model. For this reason, we will assume that a single founding couple implies no common ancestry of humans and other species, so that any genetic diversity of the founding couple is created, not inherited from ancestors.

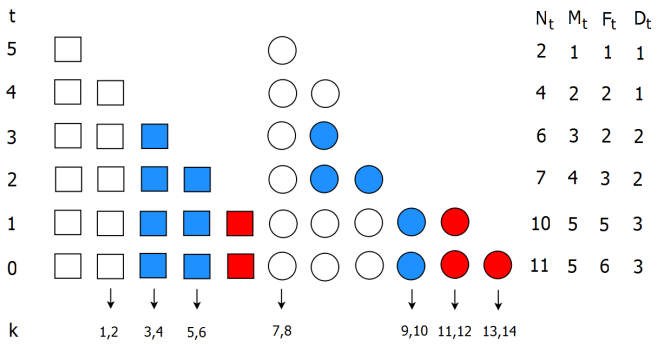

Figure 1: A population with three demes. It has six non-overlapping generations t, males are depicted as squares and females as circles, and the founder population ($t = t_{\max} = 5$) consists of one couple. The colour of each individual represents a deme number d, with white for $d = 1$, light blue for $d = 2$, and red for $d = 3$. Since generations are non-overlapping, the total number of individuals is $|I| = 40$, and the total number of chromosomes is $|C| = 80$. The genealogy is not shown, so it is not possible to tell from the figure which individuals and chromosomal regions are ancestral. A total of 7 individuals are sampled from the $t = 0$ generation, from all three demes ($n_{mm1} = 1, n_{mm2} = 2, n_{mm3} = 0, n_{ff1} = n_{ff2} = 1, n_{ff3} = 2$ in (16)). They contribute 14 chromosomes c_k, $k = 1, \ldots, 14$. **doi:**10.5048/BIO-C.2016.4.f1

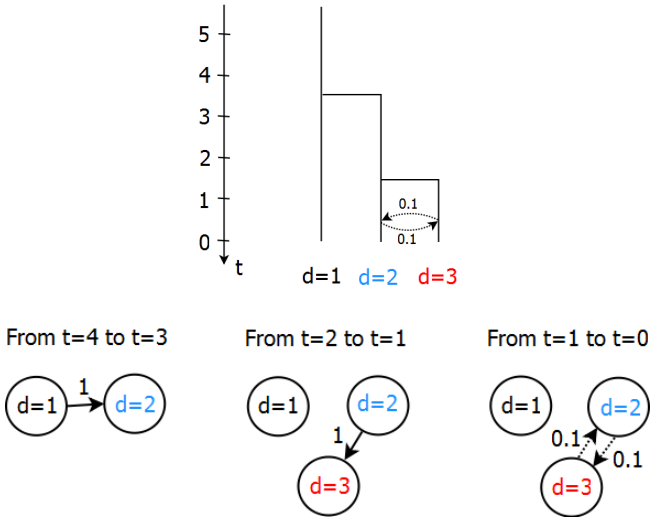

Figure 2: Colonization and migration history of the population in Figure 1. The upper plot shows the deme branching pattern. Each horizontal line corresponds to a branching event or colonization, when the daughter deme receives all its immigrants from the mother deme, such as $d = 2$ does from $d = 1$ between time points 3 and 4, and $d = 3$ does from $d = 2$ between time points 1 and 2. Between time points 0 and 1 there is a migration rate 0.1 back and forth between demes 2 and 3. This is the probability for a parent to originate from a deme other than the child. The lower plot shows the demes as circles with different colours. Those generation shifts are shown where either a colonization or migration occurs. Colonization events are illustrated with solid arrows that point to the newly populated deme. Since each newly colonized deme receives all its parents from the same deme, this corresponds to a migration rate of 1. Ordinary migration is illustrated with dashed lines, with the migration rate next to it. **doi:**10.5048/BIO-C.2016.4.f2

[57,58]. One option is to locate the demes on a two-dimensional lattice with different latitudes and longitudes, where migration from a deme is possible only to one of its four neighboring lattice points, each of which may or may not correspond to an existing deme. This is a version of the two-dimensional stepping stone model [59], and various extensions of it have been used in order to infer population history locally in Europe [60-62]. In these models, a deme may not only represent geographic location, but also ethnicity, for instance hunter/gatherers and farmers. Since the number of demes is very large, it is not feasible to specify their sizes in advance. The demography is rather simulated forwards in time, with parameters that regulate population growth and migration.

Formally, we write the male and female population sizes at time point t as sums

$$\begin{array}{rcl} M_t & = & \sum_d M_{td}, \\ F_t & = & \sum_d F_{td}, \end{array} \qquad (2)$$

over all demes $d = 1, \ldots, D_t$, with $N_{td} = M_{td} + F_{td}$ the total number of males and females in deme d. If one new population is colonized or founded at time point t, D_t increases by one, and N_{tD_t} is the size of the founding population of this deme. Figures 1-2 illustrate a six generation population with 40 individuals, 3 demes and possible migration between the most recently founded demes 2 and 3. Since generations are non-overlapping in this example, each male or female only appears once.

1.2 Chromosomes

We look at genetic inheritance in the population at a chromosome, or a segment of it. This segment could either represent nuclear DNA, as part of an autosome (non-sex chromosome), X-chromosome, Y chromosome, or it may represent mitochondrial DNA. For nuclear DNA, each individual has two homologous copies of this chromosome, inherited from the father and mother (although the two sex chromosomes X and Y within males are not truly homologous). The chosen segment of such a chromosome forms a haplotype, i.e. a sequence $\boldsymbol{h} = (a_1, \ldots, a_L)$ of DNA at L loci. Such a locus may represent a single nucleotide or site, with

$$a_l \in \mathcal{A}_{\text{sn}} = \{A, G, C, T\} \qquad (3)$$

the allele at site number $l \in \mathcal{L}$, where

$$\mathcal{L} = \{1, \ldots, L\} \qquad (4)$$

is the collection of all loci. Each such allele is either an adenine (A), guanine (G), cytosine (C) or thymine (T). An averaged sized whole chromosome has length $L = 1.5 \cdot 10^8$ nucleotides. A locus may also represent a codon, i.e. a triple of nucleotides. Then each allele

$$\begin{array}{rcl} a_l & \in & \mathcal{A}_{\text{cod}} \\ & = & \{AAA, AAG, AAC, AAT, ACA, \ldots, TTT\} \\ & \setminus & \{TAA, TAG, TGA\} \end{array} \qquad (5)$$

belongs to the set of $4^3 - 4 = 61$ non-stop codons, i.e. those triplets that code for amino acids. A third possibility is to include insertions and deletions, so that l more generally represents a locus which has been aligned between different copies of the chromosome in the population. At a short tandemly repeated or microsatellite locus, the allele

$$a_l \in \mathcal{A}_{\text{ms}} = \{1, 2, 3, \ldots\} \qquad (6)$$

refers to the number of repeats of a certain short sequence (tandem), typically of length 1-6 base pairs. Figure 3 shows a region with 9 single nucleotide or microsatellite loci, and how the alleles at each locus vary for a sample of 6 chromosomes.

It is possible to include copy number variation (CNV) as well, similar to microsatellite markers. An autosomal locus l may also represent an *Alu* insertion polymorphism. These are genetic elements that mobilize through a process called retroposition, with an allele

$$a_l \in \mathcal{A}_{\text{Alu}} = \{0, 1\}, \qquad (7)$$

that represents absence (0) or presence (1) of an *Alu*.

For nuclear DNA there are $2N_t = 2M_t + 2F_t$ chromosomal copies (or haplotypes of length L) at time point t, numbered as $c = 1, \ldots, 2M_t, 2M_t + 1, \ldots, 2M_t + 2F_t$, so that the $2M_t$ chromosomes within males come first, and then the $2F_t$ chromosomes within

Table 1: List of notation for some of the most important quantities. See also Table 3 for all input parameters of the algorithm.

t	Time point
N_t	Population size at time t
d	Deme or subpopulation number
N_{td}	Size of deme d at time t
i	Order number of individual
(t, i)	Identifier of individual number i at time point t
I	Collection of all individuals at all time points
I_t	All individuals at time t
c	Order number of a chromosomal copy
(t, c)	Identifier of chromosome number c at time point t
C	Collection of chromosomes at all time points
n	Number of sampled copies of a chromosome or chromosomal segment
k	Order number of a sampled chromosome
c_k	Order number of the kth sampled chromosome at time point 0
l	Order number of a locus
$\mathcal{A}(l)$	Set of possible alleles at locus l
$\mathcal{L}$	Set of all loci of the chromosomal segment
$\boldsymbol{h}_k$	Haplotype of the kth sampled chromosome at time point 0
a_{kl}	Allele of the the kth sampled chromosome at locus l
q	Order number of a haplotype block
hb_q	Haplotype block number q
AI	All individuals that are ancestral to at least one sampled chromosome, for at least one haplotype block
AI_t	All ancestral individuals at time t
AC	All chromosomes that are ancestral to at least one sampled chromosome, for at least one haplotype block
AC_t	All ancestral chromosomes at time t
ARG	Ancestral recombination graph
AHB	Ancestral haplotype block
AME	Ancestral mutational events

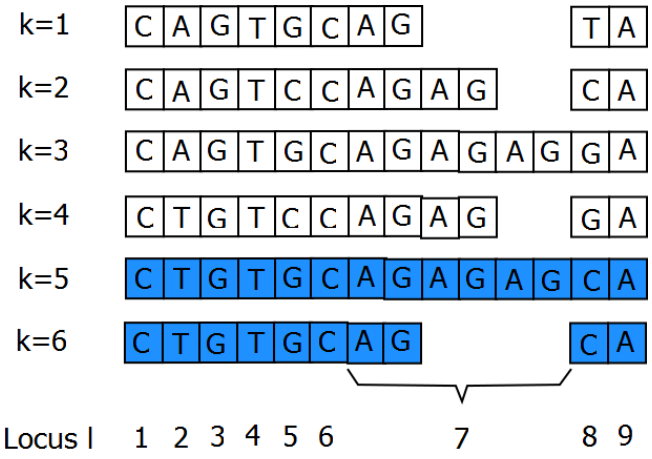

Figure 3: A sample of six aligned sequences. These are numbered c_k, $k = 1, \ldots, 6$, and they originate from a chromosomal region of $L = 9$ loci ($\mathcal{L} = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$), of which 8 consist of one single nucleotide ($\mathcal{L}_{\mathrm{sn}} = \{1, 2, 3, 4, 5, 6, 8, 9\}$) and one is a microsatellite locus ($\mathcal{L}_{\mathrm{ms}} = \{7\}$) with a two letter repeat sequence AG. The polymorphic loci, i.e. the ones with variation between sequences, are $\mathcal{L}_{\mathrm{pol}} = \{2, 5, 7, 8\}$, and of these $\mathcal{L}_{\mathrm{ba}} = \{2, 5\}$ are biallelic, with just two alleles or variants. Chromosomes $c_1, \ldots, c_4$ are from deme 1 (white), whereas chromosomes c_5, c_6 are from deme 2 (light blue). **doi:**10.5048/BIO-C.2016.4.f3

females. More specifically, individual i at time t contributes with two chromosomes $c = 2i - 1$ and $c = 2i$, inherited from the father and mother respectively. An arbitrary chromosome is referred to as (t, c), including its time point number t and its order number c at this time point. For autosomal nuclear DNA, the total set of chromosomes

$$\mathrm{C} = \{(t, c);\, 0 \le t \le t_{\max}, 1 \le c \le 2N_t\} \tag{8}$$

has size $|\mathrm{C}| = 2\sum_{t=0}^{t_{\max}} N_t$. Since Y chromosomes only reside within males, and are inherited from fathers, it corresponds to a subset

$$\mathrm{C}_Y = \{(t, c);\, 0 \le t \le t_{\max}, c = 2i - 1, 1 \le i \le M_t\} \tag{9}$$

of (8). The other sex chromosome, X, resides within males and females, but males always inherit their single copy from their mother, so that

$$\begin{aligned} \mathrm{C}_X \quad = \quad & \{(t, c);\, 0 \le t \le t_{\max}, c = 2i, 1 \le i \le M_t, \\ & \text{or } 2M_t + 1 \le c \le 2N_t\}. \end{aligned} \tag{10}$$

Since each individual i has only one copy of mitochondrial DNA (mtDNA), inherited from the mother, we represent it as $c = 2i$ (regarding $c = 2i - 1$ as empty or non-existing). The mtDNA of the female population then corresponds to

$$\mathrm{C}_{\mathrm{mit}} = \{(t, c);\, 0 \le t \le t_{\max}, c = 2i, M_t + 1 \le i \le N_t\}. \tag{11}$$

It is easy to extend the model to the whole genome, by letting $\boldsymbol{h}$ be a haplotype with alleles from different chromosomes. It is only for notational simplicity that we speak of a region within one chromosome. But (t, c) could equally well represent genome number c at time point t.

1.3 Haplotype Blocks

A large part of the genetic diversity in a population between different parts of an autosome or X-chromosome is due to (reciprocal) recombinations. Recombinations are caused by crossovers, i.e. switching the DNA that an individual receives along the chromosome between grandpaternal and grandmaternal modes of inheritance. It is well known that recombination rates vary along chromosomes on a coarse scale of order 10 Mb [63], and there is also increasing evidence [64] for a substantial recombination rate variation on a much finer 1-100 kb scale, as modeled for instance by [65]. A likely explanation of this local variation is either recombination hotspots (positions with a high crossover probability) or randomly located positions where ancestral crossover events occurred, or a combination of both. This is related to existence of so-called haplotype blocks. It has been argued that a large part of the genome can be divided into such blocks, with few recombinations within blocks, and many recombinations between them, see for instance [66-71]. Although the extent to which these blocks explain all genetic diversity is not settled [72,73], they nevertheless provide a very useful framework, particularly for a relatively young population with only two founders.

Haplotype blocks can be defined in different ways. We assume that recombinations in an autosomal chromosome or X-chromosome inherited from a mother only occur between haplotype blocks, similarly as in [74]. In more detail, we divide the L loci of each chromosome (t, c) into Q haplotype blocks

$$\mathrm{hb}_q = \{l_{q-1}, \ldots, l_q\}, \quad q = 1, \ldots, Q, \tag{12}$$

where $l_0 = 0$, and the rightmost end points of all blocks satisfy $0 < l_1 < l_2 < \ldots < l_Q = L$. Since recombinations only occur between blocks, the loci within each block will have the same genealogy, but loci of different blocks may have different genealogies. For Y-chromosome and mitochondrial DNA we assume there are no recombinations, and put $Q = 1$.

Early studies predicted that haplotypoe blocks vary in length between 5 kb and 200 kb [75,76], but more recent analyses with more genetic markers suggest that their average length could be as small as 5kb [77]. In addition, the blocks tend to be shorter in African populations than in non-African ones [78]. These results imply that a popoulation averaged Q is at least of the order 10^4 for a whole chromosome. It is possible though to have smaller haplotype blocks. Since we don't allow for recombinations within blocks, this is appropriate for models that incorporate gene conversion. In the extreme case, one could allow each locus to be a separate block ($Q = L$). If the haplotype blocks represent widely separated regions from the same chromosome, with gaps in between, a recombination between two neighboring blocks represents an odd number of crossovers. If, on the other hand, the haplotype blocks are adjacent chromosomal regions, a recombination between two neighboring haplotype blocks represents one single crossover event.

Our model incorporates two options, the first of which involves fixed or random (simulated) haplotype block boundaries that are specified before the genealogy is built. These boundaries may then be interpreted as recombination hot spots, where all crossovers are enforced to occur (at least if haplotype blocks are adjacent chromosomal regions, without any gaps between them). For the second option of our algorithm, the haplotype blocks are not specified in advance. Instead they are generated as the genealogy is built. Then haplotype block boundaries correspond to ancestral recombination events rather than recombinational hotspots, although many of these ancestral recombinations may still have occurred at hotspots, if a locally varying recombination rate is assumed.

1.4 Backward Simulation of Genealogy

In order to build a genealogy, we must not only set up parental relationships between the individuals (1) of a population, but also specify how DNA is inherited for the chromosomal segment that was defined in Sections 1.2-1.3, at all of its haplotype blocks. Since the DNA of our ancestors is not fully known, the genealogy is not known either. We therefore have to reconstruct it with some uncertainty. This is accomplished by viewing it as a random object, with a statistical distribution that quantifies our incomplete knowledge of the actual genealogy. From this distribution we simulate a number of plausible genealogies. The simulation algorithm that accomplishes this task should ideally include the following six major mechanisms of genetic change:

(i) Genetic drift, due to randomly varying reproductive success of mating couples, and Mendelian inheritance, with its randomness in choosing whose grandparental DNA to pass on to the grandchildren.

(ii) Recombination of DNA from homologous chromosomes.

(iii) Colonization of new demes, and then isolation or migration between them.

(iv) Mutations, i.e. changes of DNA.

(v) Natural selection. due to a systematic variation in reproductive fitness.

(vi) Founder diversity of autosomal, X-chromosome and possibly also mitochondrial DNA, i.e. allowing for different alleles of homologous founder chromosomes, or different founder mitochondria, at various loci.

Microevolutionary common descent models only include the first five mechanisms, but (vi) is important in order to generate enough diversity for a population with only one founding couple. The demography will only influence some of these six forces of genetic change; (iii) through migration, (i) through population size changes (since the genetic drift is much larger in a small deme that experiences a founding event or bottleneck, than in a large one) and to some extent (v), since the carrying capacity of a population may influence which characteristics that favour survival.

A number of computer programs have been developed for the purpose of simulating how demographics and the genetic composition of a population changes over time, see for instance [79] for a review. The most straightforward approach is to use a forward algorithm [80-86]. The name reflects that these programs simulate haplotypes of chromosomes by starting at the founder generation and ending at the present. They are often very flexible, allowing for many demographic scenarios, with all forces (i)-(v) of microevolution included. But since they require DNA of all individuals to be simulated for all haplotype blocks (with a complexity of the order $|\mathrm{C}|Q$), with current computer speed they seem to be limited to relatively small populations, not the whole human history ($10^{10} < |\mathrm{C}| < 10^{11}$).

To circumvent this difficulty we will mostly simulate the genealogy backward in time. The backward ancestry or genealogy of the sampled chromosomes at each haplotype block is a coalescence tree [87] with lineages that merge until the founding generation $t = t_{\max}$ is reached, when at most $2N_{t_{\max}} = 4$ lineages remain. Due to recombinations, different haplotype blocks may have different coalescence trees, and the whole collection of genealogies is referred to as an ancestral recombination graph [88-90]. Overviews of coalescence and ancestral recombination graph theory can be found [47] and [91-93].

The main advantage of backward simulation is that only a small subset of human history needs to be sampled. A price to pay is the

difficulty of incorporating natural selection. The algorithm that we describe in this and the next section will therefore only include the other five mechamisms (i)-(iv),(vi) of genetic change, whereas a discussion of natural selection is postponed to Section 4.

Several simulation programs generate ancestries in reverse time. Many of them [94-96] focus on relatively short chromosomal regions for monoecious and diploid populations. They use large time scale approximations, only simulating those generations where coalescence or recombination events occur. The algorithm in [97] simulates all generations, and is applicable for whole chromosomes, whereas those in [98,99] incorporate geographic substructure and two-sex models into the coalescence framework as well. We will use a similar approach and build a discrete time simulation model that includes not only geographic substructure and two sexes, but also overlapping generations.

With reversed time simulation, it is possible to sample only a subset of n chromosomes

$$\{(0,c_1),(0,c_2),\ldots,(0,c_n)\} \tag{13}$$

of generation 0, with $1 \le c_1 < c_2 < \ldots < c_n \le 2N_0$. This number n can be quite large (like $10^4 - 10^5$), but still much smaller than $2N_0 \sim 10^{10}$. We refer to a chromosome (t,c) as ancestral if it is an ancestor of at least one of the n sampled chromosomes for at least one haplotype block hb_q. The set of ancestral chromosomes of time point t is denoted as

$$\mathrm{AC}_t = \{c;\ (t,c) \in \mathrm{AC}\}. \tag{14}$$

The main idea of backward simulation is that the set AC of ancestral chromosomes only comprises a small fraction of all chromosomes in (8). In particular, AC_t is typically a small subset of all chromosomes $\{1,\ldots,2N_t\}$ at time t. Figure 4 displays a population with 10 individuals, as well as its genealogy forwards and backwards in time. It can be seen that the backward pedigree is a subset of the forward pedigree. At the current time point there are $N_0 = 4$ individuals, of which $n = 3$ are sampled, one adult female and one newborn cousin pair.
In order to build the genealogy backwards, we must first specify the demes and sex of the individuals from which DNA samples are taken. Since individual i at time 0 contributes with two chromosomes $c = 2i-1$ and $c = 2i$ of nuclear DNA, it suffices to specify four sample sizes

$$\begin{aligned}
n_{mmd} &= \text{number of sampled males from deme } d \text{ with both chromosomes in the sample,}\\
n_{md} &= \text{number of sampled males from deme } d \text{ with only one chromosome in the sample,}\\
n_{ffd} &= \text{number of sampled females from deme } d \text{ with both chromosomes in the sample,}\\
n_{fd} &= \text{number of sampled females from deme } d \text{ with only one chromosome in the sample,}
\end{aligned} \tag{15}$$

for each deme $d = 1,\ldots,D = D_0$, so that the total sample size can be written as

$$\begin{aligned}
n &= \textstyle\sum_{d=1}^{D}\left[n_{md} + n_{fd} + 2(n_{mmd} + n_{ffd})\right]\\
&= n_m + n_f + 2(n_{mm} + n_{ff}),
\end{aligned} \tag{16}$$

where n_m, n_f, n_{mm} and n_{ff} are the total numbers of sampled males and females that contribute with one or two haplotypes. If DNA-variation within a specific subpopulation is of interest, one may take $D = 1$, but in order to get a representative sample for the

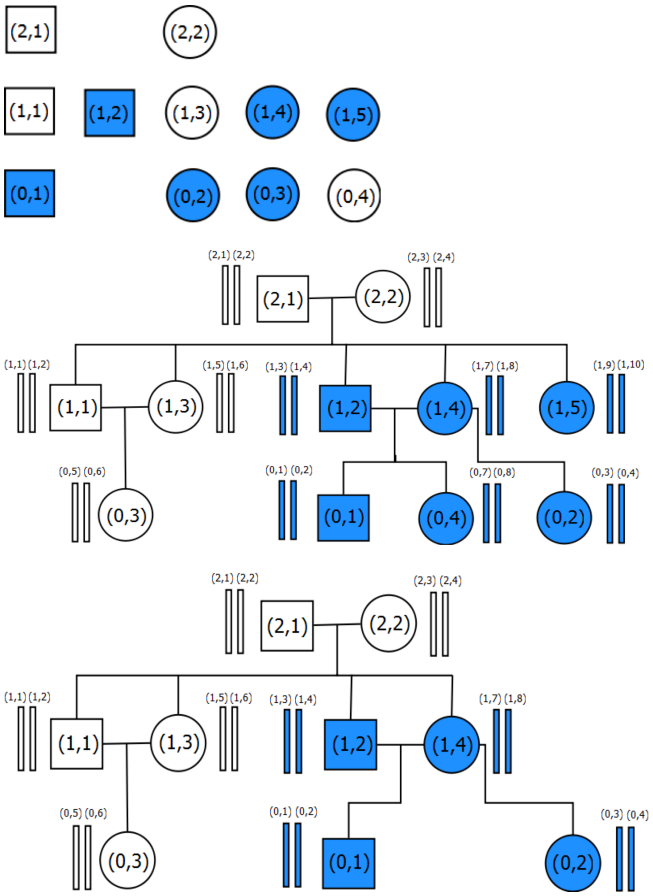

Figure 4: The history of a population during three time points. Upper graph: The population is displayed as in Figure 1, with squares and circles for males and females. The male and female population sizes at time points $0, 1, 2 = t_{\max}$ are $M_0 = 1, M_1 = 2, M_2 = 1$ and $F_0 = 1, F_1 = 3, F_2 = 3$. The population consists of two demes 1 (white) and 2 (light blue), with number of demes $D_0 = D_1 = 2$, $D_2 = 1$ at the three time points. Each individual that lives at time t is assigned a number (t,i), with males numbered first in at each time point. Middle graph: The pedigree of the population, generated forwards in time, with all mated pairs connected by horizontal lines, and a vertical line in between that connects to their children. Although $|I| = 11$, the population has only 10 individuals, since $(1,4)$ survives to the next time point. The two rectangles next to each non-founder individual (t,i) represent its two homologous chromosomes, with numbers $(t,2i-1)$ and $(t,2i)$, of which the first is inherited from its father and the second from its mother. There are $2 \cdot 11 = 22$ such rectangles $\mathsf{C} = \{(0,1),(0,2)\ldots,(2,4)\}$ in the graph, corresponding to $2 \cdot 10 = 20$ distinct chromosomal copies. For sex chromosome DNA, $\mathsf{C}_Y = \{(0,1),(1,1),(1,3),(2,1)\}$ and $\mathsf{C}_X = \mathsf{C} \setminus \mathsf{C}_Y$, whereas for mitochondrial DNA, $\mathsf{C}_{\mathrm{mit}} = \{(0,4),(0,6),(0,8),(1,6),(1,8),(1,10),(2,4)\}$. Lower graph: Subpedigree with $|\mathsf{AI}| = 9$ nodes, corresponding to 8 individuals. It is built backwards in time from the three sampled individuals $(0,1)$, $(0,2)$ and $(0,3)$, one male and two females ($n_m = n_f = 0, n_{mm} = 1, n_{ff} = 2$). The sampled chromosomes at time 0 are $(c_1,\ldots,c_6) = (1,\ldots,6)$. **doi:**10.5048/BIO-C.2016.4.f4

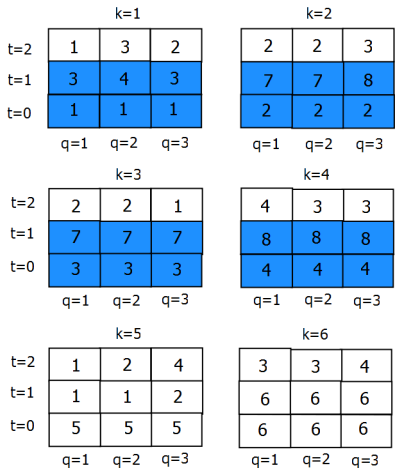

Figure 5: Ancestral recombination graph generated from the pedigree in the lower part of Figure 4. It corresponds to a non-sex chromosome segment, so that each individual carries two haplotypes. It starts from one male $(0,1)$ and two females $(0,2)$ and $(0,3)$, giving a total of $n = 6$ sampled chromosomes $(0,c_k)$ for $k = 1,\ldots,6$, with $c_k = k$. The number ARG_{tkq} within each rectangle shows ancestry of a sampled chromosome c_k at one of the $Q = 3$ haplotype blocks hb_q, $q = 1,2,3$ (columns), at one of three time points $t = 0,1,2$ (rows), as well as the deme d in which the ancestor lived ($d = 1$: white, $d = 2$: light blue). For instance, in the middle row of the upper left rectangle, the paternally inherited chromosome $(0,1)$ of individual $(0,1)$ experiences two recombination events, so that grandpaternal DNA from chromosome $(1,3)$ is inherited at haplotype blocks hb_1 and hb_3 ($\mathrm{ARG}_{111} = \mathrm{ARG}_{113} = 3$), whereas grandmaternal DNA from chromosome $(1,4)$ is inherited at hb_2 ($\mathrm{ARG}_{112} = 4$). **doi:**10.5048/BIO-C.2016.4.f5

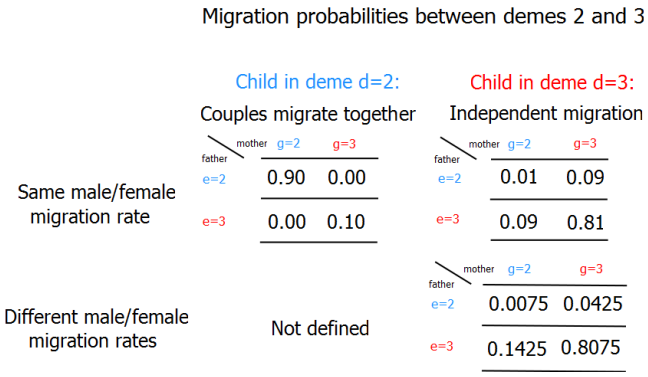

Figure 6: Mating probabilities between time points 0 and 1 for the population of Figure 2. This is the probability $B_{0d,eg}$ that the father comes from deme e and the mother from deme g, given that the child lives in deme d at time $t = 0$. Row (column) sums in each table correspond to the fraction of fathers (mothers) that migrate or stay in the deme in which they were born. In the upper part, the fraction of migrating fathers and mothers is the same, 0.1, whether the child lives in deme 2 or 3. In the lower part, the fraction migrating of fathers (mothers) is 0.05 (0.15) for children living in deme 3. Parents migrate together to deme 2, whereas they migrate independently to deme 3. When parents migrate together, the male and female migration rates must be the same, so therefore the lower left scenario is not possible. They overall migration rate equals the average of the fraction of males and females that lived in another deme than the child. In all cases it equals 0.1 for children living in deme 2 ($B_{023} = 1$ in equation (74)) and 0.1 for children living in deme 3 ($B_{032} = 1$). These two migration rates 0.1 are shown in the lower subplot of Figure 2. **doi:**10.5048/BIO-C.2016.4.f6

metapopulation, samples should be taken from all demes ($D = D_0$), at least if they represent larger regions or continents. In any case, it is either assumed that individuals and their genetic markers are sampled randomly within each subpopulation or according to some ascertainment procedure that mimics how real data was collected (see Section 3.3). The sample sizes in (15) and numbering of chromosomes in (13) will depend on DNA type, as described in more detail in Section 2.

The ancestral recombination graph (ARG) summarizes the backward genealogy of the sample, and it incorporates the first three forces of microevolution; genetic drift (i), recombination (ii) and migration (iii). It is stored in a three-dimensional array

$$\mathrm{ARG} = (\mathrm{ARG}_{tkq};\, 0 \le t \le t_{\max}, 1 \le k \le n, 1 \le q \le Q), \qquad (17)$$

where $\mathrm{ARG}_{tkq} \in \mathrm{AC}_t$ specifies which ancestral chromosome at time t in (14) that is an ancestor of chromosome $(0,c_k)$ at haplotype block hb_q. As an example, the ancestral recombination graph of the pedigree in the lower part of Figure 4, is shown in Figure 5.

The ARG is generated recursively back in time. For each time point we need to assign parents to all individuals that have ancestral DNA and specify which grandparent that passed on DNA, using Mendelian laws of inheritance and recombinations events. As the founder generation is reached, we get a pedigree with all its members' recombination events. The most complicated part is to construct the pedigree backwards in time. It requires a reproduction model in terms of mating preferences and how the number of offspring varies between couples. Additional complications arise when geographic substructure in terms of more or less isolated islands or demes is allowed for, or if age structure in terms of overlapping generations is assumed (see Section 2 for details).

We will generalize a model in [99-101], and assign parents to the newborns of each time point in two steps. The (possibly different) demes in which the parents live are determined first, according certain migration probabilities $B_{td,eg}$ that a child at time t from deme d has the father (at time $t+1$) from deme e and the mother (at time $t+1$) from deme g. As the name suggests, these probabilities determine how often males and females migrate between demes. They make it possible to incorporate, for instance, a higher migration rate for women than for men [57], see Figure 6.

In the second step, the children assign parents within the chosen demes, using the two key parameters

$$\begin{aligned} \alpha &: \text{ controls distribution of number of offspring among females,} \\ \beta &: \text{ tunes the degree of monogamy,} \end{aligned} \qquad (18)$$

to specify how reproductive success varies between individuals. Together with the population size, they determine genetic drift, i.e. the rate at which allele frequencies change over time. Some extreme choices of $0 \le \alpha, \beta \le \infty$ are listed in Table 2. The parametrization in (18) favors polyandry (women having several men) when β is large. But it is straightforward to exchange the role of the two sexes, so that the model that favors polygymy (men having several women) instead when β is large. It has been argued in [57] that this is a more reasonable assumption. On the other hand, recent studies suggest that the degree of monogamy is quite high in human populations. These conclusions are based on comparing estimates of effective population sizes or of historical recombination rates between autosomal and X chromosome regions [102].

Each row of the ancestral recombination graph of Figure 5 represents a time point. For autosomal and X-chromosome DNA, the inherited haplotype blocks of such a row is a mosaic from different ancestral chromosomes at this time point. The breakpoints of the mosaic are caused by historical recombination events between haplotype blocks, so that the more recombinations there are, the finer is the mosaic. In order to model recombinations, we specify

Table 2: Some parameter choices for the pedigree building algorithm of Section 2.1, and their interpretation.

(α, β)	Mating scenario
(∞, ∞)	Children have parents independently according to a two-sex Wright Fisher model, within the demes that have first been assigned to the parents.
$(\infty, 0)$	Children have mothers independently within the deme that has first been assigned to the mother, and each mother has only one husband per deme.
$(0, \infty)$	All children with mother from the same deme, have the same mother, but the mother chooses fathers independently for each mating, given that demes have been assigned to her spouses.
$(0, 0)$	All children with parents from a particular pair of demes, have the same mother and father.

two vectors

$$\begin{array}{rcl} \boldsymbol{r}_m & = & (r_{m1}, \ldots, r_{m,Q-1}), \\ \boldsymbol{r}_f & = & (r_{f1}, \ldots, r_{f,Q-1}) \end{array} \tag{19}$$

of recombination probabilities for males and females. If haplotype blocks are not specified in advance, so that the number of haplotype blocks is not an input parameter, we replace Q by L in (19). For mitochondrial or Y-chromosome DNA there is no need to specify the probabilities in (19), since these sequences only have one haplotype block ($Q = 1$).

If two neighboring haplotype blocks hb_q and hb_{q+1} of autosomal or X-chromosome DNA are adjacent chromosomal regions, with no gap in between, then r_{mq} (r_{fq}) is the probability of one crossover between hb_q and hb_{q+1} when a sperm (ovum) cell is formed. Neighboring haplotype blocks may also, more generally, be separated by a stretch of DNA, and then r_{mq} (r_{fq}) is the probability of an odd number of crossovers between hb_q and hb_{q+1}. The recombination probabilities vary over regions [63] and are typically larger for females than for males (see for instance [103], Section 1.3 of [104], and Section 3.12 of [105]). Recent research indicates that recombination rates may vary between individuals of the same sex, with a higher average rate for Africans than for people with European ancestry [78].

A common simplification is to start with a model (19), and then use sex-averaged recombination probabilities

$$r_q = \frac{1}{2}(r_{mq} + r_{fq}) \tag{20}$$

between all pairs of haplotype blocks hb_q and hb_{q+1}. The sex- and region-averaged crossover probability is roughly 10^{-8} for one single nucleotide site. Consequently, if there are no gaps between any of the $Q - 1$ pairs of neighboring haplotype blocks, we get a sex and region-averaged recombination probability

$$\frac{1}{L-1} \sum_{q=1}^{Q-1} r_q \sim 10^{-8}$$

between a randomly chosen pair of neighboring blocks. In a simple model where recombinations occur uniformly over the chromosome,

$$r_q = r = \frac{(L-1) \cdot 10^{-8}}{Q-1}.$$

The averaging in (20) is not necessary though, but it is sometimes preferable in order to reduce the number of input parameters.

When LD patterns over shorter ranges are of interest, it is important to include gene conversion [106,107] as well. Gene conversions are closely spaced double crossovers (see Section 2.2) that tend to break up LD. We specify how often gene conversions occur for autosomal or X-chromosome DNA by the two vectors

$$\begin{array}{rcl} \boldsymbol{\gamma}_m & = & (\gamma_{m1}, \ldots, \gamma_{m,Q-1}), \\ \boldsymbol{\gamma}_f & = & (\gamma_{f1}, \ldots, \gamma_{f,Q-1}), \end{array} \tag{21}$$

where γ_{mq} (γ_{fq}) is the probability that the leftmost of the two crossovers occurs between haplotype blocks hb_q and hb_{q+1}, for males and females. As for single recombinations, we may take a sex average $\gamma_q = (\gamma_{mq} + \gamma_{fq})/2$, and assume $Q = L$ for a version of the algorithm where the haplotype block structure is not specified in advance. The gene conversion ratio GCratio $= \gamma_q / r_q$ reveals how much more likely gene conversions are compared to single recombination events. We also need a tract length distribution (TLD) for the distance between the two crossovers. The sizes of the haplotype blocks have to be at least as small as the expected tract length.

Until recently, quite little was known about gene conversion in mammals. The first studies focused on yeast and fruit flies, with a mean tract length ranging between 300 and 2000 bp [106,108,109]. It has been established since then that gene conversion is also the most important determinant of LD patterns over short distances for humans [110]. For instance, recombination models were fitted in [109] to patterns of LD in three populations. A mean tract length of 500 bp for the African population gave an estimate GCratio = 7.3. A model with GCratio = 2 was found in [111] to fit their data set better than one with single recombinations alone. A fixed tract length of 500 bp was assumed in [23], and $\gamma_q = 4.5 \cdot 10^{-9}$ per base pair was obtained for their best fitting model, which corresponds to a GCratio less than 0.5. More recent estimates in [112] suggest an average tract length of 75 bp. Combining this with estimates of the average number of sites affected by gene conversions per generation [113], one obtains a gene conversion rate of $\gamma_g = 8 \cdot 10^{-8}$, and a GCratio slightly less than 10.

1.5 Founder Diversity, Mutations and Gene Dropping

In this section we describe how to incorporate into our model two other mechanisms of genetic change; mutations (iv) and created founder diversity (vi). It is well known that assumptions concerning the "molecular clock", mutations, and the amount of founder diversity, are both crucial for the timing of human history [114,115]. Since mutations give rise to genetic differences over time, and founder diversity generates differences in the first generation, it is helpful to first assign a reference haplotype

$$\boldsymbol{h}^{\text{ref}} = (a_1^{\text{ref}}, \ldots, a_L^{\text{ref}}) \tag{22}$$

of the first generation as a yardstick to which other haplotypes are compared. For nuclear autosomal DNA, we regard the haplotype of the first chromosome $(t_{\max}, 1)$ of the founder generation as this reference haplotype. Its alleles a_l^{ref} have to be chosen in some way at all loci l. A simple option is to assume that a_l^{ref} are drawn independently between loci, whereas more advanced models take dependency between neighbouring loci into account [116,117]. In the first case, it suffices to specify probabilities of picking different alleles at each loci l. We denote[2] these probabilities as $\text{Prob}(a_l^{\text{ref}} = a) = \pi_a$ for any a that belongs to the set $\mathcal{A}(l)$ of possible alleles at l. This set could be any of (3), (5) or (6), depending on the type of locus l represents. For a single nucleotide marker (3),

$$\boldsymbol{\pi} = (\pi_A, \pi_G, \pi_C, \pi_T) \tag{23}$$

contains the frequencies of all four nucleotides in the population. The model in [118], for instance, asserts that all four types of

[2] A more accurate notation would be π_{al}, to account for that allele frequencies depend on the type of marker at locus l. This may also take into account that allele frequencies for the same type of marker varies over the genome, for instance between coding and non-coding regions.

nucleotides are equally frequent ($\pi_A = \pi_G = \pi_C = \pi_T = 1/4$), but we could also use sequence data to estimate (23). For a codon locus (5),

$$\boldsymbol{\pi} = (\pi_{AAA}, \pi_{AAG}, \ldots, \pi_{TTT}) \tag{24}$$

consists of the frequencies of all 61 non-stop codons, and for a microsatellite marker (6) it lists the frequencies

$$\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3, \ldots) \tag{25}$$

of the number of tandem repeats.

In order to allow for some variability between the four haplotypes of the founder generation, we assume that the haplotypes of the other three chromosomes, $(t_{\max}, 2)$, $(t_{\max}, 3)$ and $(t_{\max}, 4)$, are generated by copying the reference haplotype $\boldsymbol{h}^{\text{ref}}$, but this copying mechanism is interrupted when markers experience "founder mutations" independently with probabilities

$$\boldsymbol{\nu} = (\nu_1, \ldots, \nu_L). \tag{26}$$

Such a change from the reference allele a_l^{ref} at l to some other allele is not a germline mutation, where DNA in one of a child's two homologous chromosomes is changed compared to a parental chromosome. It rather contributes to created diversity, a way of generating genetic difference between the four founder haplotypes. The frequency of these founder mutations will depend heavily on the assumed time point when the founding couple lived. An estimate of $\nu_l = 10^{-3}$ is given in [119] for a recent founding event (of the order 10,000 yrs), by extrapolating heterozygosity within individuals from the HapMap dataset. But at some chromosomal regions ν_l could be lower. For instance, there are several genes at the end of chromosome 6 for which only about 200 base pairs are polymorphic. If these genes are a few Mbp long, this suggests a region averaged value of the order $\nu_l = 10^{-4}$.

For sex chromosome DNA, the founder generation consists of one Y-chromosome $(t_{\max}, 1)$ as well as three X-chromosomes $(t_{\max}, 2)$, $(t_{\max}, 3)$ and $(t_{\max}, 4)$. We take $(t_{\max}, 1)$ as the reference haplotype of the Y-chromosome population, and $(t_{\max}, 2)$ as the reference haplotype of the X chromosome population. The alleles of these two haplotypes are chosen independently, according the specified allele frequencies at all loci. Founder mutations are needed for the X-chromosomes, in order to generate $(t_{\max}, 3)$ and $(t_{\max}, 4)$ from $(t_{\max}, 2)$, whereas no created diversity is required for the founder generation of Y-chromosome DNA. For mitochondrial DNA within females, $(t_{\max}, 4)$ carries the only haplotype of the founder generation. We choose it as a reference, and need no founder mutations[3].

When all four chromosomes of the founding generation $t = t_{\max}$ have been assigned alleles at all loci, the ARG is used to spread it to the present ($t = 0$). This so called gene dropping will be interrupted by some additional and selectively neutral germline mutations at various loci for time points $t_{\max} - 1, \ldots, 0$. For each chromosome (t, c) with $0 \leq t < t_{\max}$, the mutation probabilities at all loci are contained in the two vectors

$$\begin{aligned} \boldsymbol{\mu}_m &= (\mu_{m1}, \ldots, \mu_{mL}), \\ \boldsymbol{\mu}_f &= (\mu_{f1}, \ldots, \mu_{fL}), \end{aligned} \tag{27}$$

where μ_{ml} (μ_{fl}) is the germline mutation probability at locus l when sperm (ova) are formed. Equation (27) takes into account that the mutation probability not only depends on the type of marker, but also on the chromosomal position, with regions of low mutation rate surrounded by mutational hotspots. The mutation probabilities also seem to vary between individuals [120], but generally they are higher for males than females, and for males they also increase with age [121,122]. The age-dependency can be accounted for when age structure is included in the model, but a simplified option is to regard μ_{ml} as an age-averaged mutation rate at locus l for all males that reproduce. Although not necessary, we will mostly consider sex-averaged mutation probabilities $\mu_l = (\mu_{ml} + \mu_{fl})/2$ at all loci, as summarized by

$$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_L). \tag{28}$$

This will simplify notation and reduce the number of parameters.

It has recently been shown that a majority of DNA has function in terms of affecting phenotypes or molecular activity [123,124]. Since most mutations are neutral or slightly deleterious, and there is a repairing mechanism for copying errors during cell division, it is reasonable to believe that mutation rates should be low for most parts of the genome. Previous estimates were often based on sequence comparisons between species and various assumptions on the existence and timing of divergence events. For instance, [23] used a value of $\mu_l = 1.5 \cdot 10^{-8}$ per base pair per generation at single nucleotides in their best fitting model for human history. Numerous recent studies (see [122]) from family data make it possible to detect and count *de novo* mutation events, so that the mutation probability for single nucleotides can be estimated directly. These studies indicate an autosomal mutation rate that is lower than previously believed, a higher mutation rate for males than for females, and a rate that increases with age for men. Combining the results of these studies, the sex and genome-averaged mutation rate is found to be in the range $1 \cdot 10^{-8} - 1.2 \cdot 10^{-8}$ per base pair per generation, and it is slightly larger (range $1.3 \cdot 10^{-8} - 2.2 \cdot 10^{-8}$) within exons. In [125], the authors estimated mutation rates for single nucleotides of Y-chromosomes. They accounted for its dependency on paternal age, and found a de novo rate of $1.9 \cdot 10^{-8}$ per base pair per generation (if the generation time is 30 years), whereas [126] used large Icelandic pedigrees to infer a mutation rate of $3 \cdot 10^{-8}$, which is close to estimates derived from a Chinese family in [127] and for a small Central Asian population in [128]. For mitochondrial single nucleotides, the mutation rate was previously believed to be about ten times as high as for nuclear sites. For instance, a value of $3.5 \cdot 10^{-7}$ per base pair per generation is used in [7], but recent direct estimates based on family data suggest that the rate is a lot larger, around 10^{-5} [129-132], although it varies considerably between different regions of the mitochondrion.

The higher SNP mutation rate of mtDNA makes it more suitable for infering the more recent human history, whereas nuclear DNA is better suited for studying the more distant past. The authors of [133] used this observation (and the fact that the mitochondrion population is four times as small) to suggest a bottleneck followed by a population expansion as part of the Out of Africa scenario. This would reconciliate the seemingly contradictory observations of an expanding population from mtDNA, and a bottleneck for nuclear DNA. With a unique origin approach, a similar argument applies, but the bottleneck is replaced by created diversity from one founding couple.

Mutations rates for microsatellites are several orders of magnitudes larger than for single nucleotides. For non-sex chromosomes, the mutation rate for males was found to be higher than for females [134], with a sex-averaged estimate of the mutation probability of the order $2 \cdot 10^{-4} - 3 \cdot 10^{-4}$ per locus and generation for tandem repeats of length 2bp, and 10^{-3} for tandems of length 4bp. In [135], the mutation rate for Y-chromosomes was estimated to be an order of magnitude larger.

Given that a mutation occurs at some locus l with probabilities (26) or (28), at a founder or non-founder chromosome, we need

[3] It is possible though to incorporate created diversity of mitochondrial DNA, if the woman of the founding generation had diverse mitochondria that she passed on differently to her children.

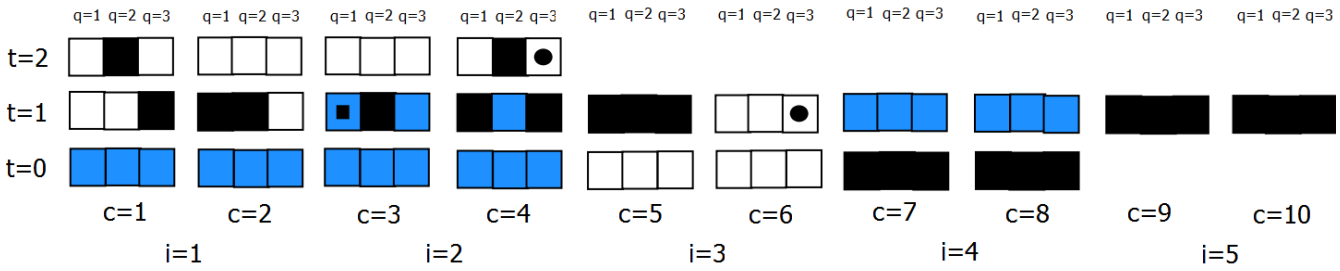

Figure 7: Illustration of which genetic material is ancestral from the ARG of Figure 5. It is shown for each chromosome (t, c) which of its haplotype blocks hb_q are ancestral by depicting those from deme 1 (2) as white (light blue), whereas and non-ancestral blocks are shown in black, so that for instance $\text{AHB}_{01} = \{1, 2, 3\}$ and $\text{AHB}_{13} = \{1, 3\}$. Two of all individuals (t, i) in the pedigree of Figure 4 are not ancestral; $(0, 4)$, who is not part of the sample, and $(1, 5)$, who has no children (and in particular no children in the sample). Five of all 22 chromosomes (t, c) are not ancestral, i.e. have no ancestral haplotype blocks. This includes the four chromosomes of the two non-ancestral individuals, and in addition chromosome $(1, 5)$, since individual $(0, 2)$ received her maternal chromosome from the grandmother (chromosome $(1, 6)$) at all haplotype blocks. Mutations occur at two loci: A single mutation (filled square) occurs at a locus within haplotype block hb_1 for chromosome $(1, 4)$, so that $|\text{AME}_{141}| = 1$. A double mutation (filled circles) occurs at a locus within haplotype block hb_3, one for chromosome $(2, 4)$ and one for chromosome $(1, 6)$, so that $|\text{AME}_{243}| = |\text{AME}_{163}| = 1$. For all other t, c, q we have $\text{AME}_{tcq} = \emptyset$. **doi:**10.5048/BIO-C.2016.4.f7

to define the probabilities by which the allele before the mutation changes to other alleles. At single nucleotide loci we have point mutations, i.e. changes between two different nucleotides, such as $A \to T$ or $C \to G$. We collect all probabilities by which new nucleotides are generated, for all $4 \times 3 = 12$ possible mutations, into a matrix

$$\boldsymbol{P} = \begin{pmatrix} 0 & P_{AG} & P_{AC} & P_{AT} \\ P_{GA} & 0 & P_{GC} & P_{GT} \\ P_{CA} & P_{CG} & 0 & P_{CT} \\ P_{TA} & P_{TG} & P_{TC} & 0 \end{pmatrix} \tag{29}$$

with four rows and four columns. For instance, $P_{AT} = \text{Prob}(A \to T|A)$ is the probability that a mutation from A to T takes place, given that we know it has happened and that the nucleotide before the mutation was A. Since any nucleotide gets mutated to some other nucleotide, the sum of the elements in each row of (29) is 1. The model in [118], for instance, sets all mutation probabilities in (29) to 1/3.

More elaborate models can be defined by dividing nucleotides into purines (A, G) and pyrimidines (C, T). Point mutations within the same group (purines or pyrimidines) are called transitions, and changes between the two groups are called transversions. For single nucleotide loci l, it is less likely that a transition is a nonsynonymous mutation that changes the codon, i.e. the triplet of nucleotides to which l belongs, to one that codes for another amino acid. The transition/transversion ratio R quantifies how more frequent transitions are compared to transversions, and it is believed that R is close to 2 [122,136]. The transition and transversion probabilities in (29) can sometimes be chosen so that they conform not only with the nucleotide probabilities in (23), but also with a certain value of R. It will be seen in Section 3.1.2 that this is not always possible though, and since the mutation probability also depends on the allele that mutates, we will discuss more flexible and general mutation models in Section 3.2.

At loci that represent a codon (5), transition probabilities of a mutational model are specified between all 61 codons that code for amino acids. For instance, the authors of [137] define a model that makes it possible to distinguish between nonsynonymous mutations (that change the amino acid) and synonymous mutations (that don't change the amino acid), making the latter much more likely.

For a microsatellite locus we get a matrix $\boldsymbol{P}$ with rows and columns as indexed by the number of tandem repeats (6). The simplest stepwise mutation model in [138] asserts that the number of repeats only changes one unit down or up with equal probabilities, so that

$$P_{ab} = \begin{cases} 1, & a = 1, b = 2, \\ 0.5, & a > 1, b = a - 1 \text{ or } a + 1, \\ 0, & \text{otherwise.} \end{cases} \tag{30}$$

This model is quite accurate for longer (4-6 bps) tandems, although a drawback is that the number of repeats may increase beyond all limits over time. In addition, for short repeats of size 2, (30) does not account for the fact that repeat numbers sometimes change by more than one. Several mutation models for AC repeat data were analyzed in [139], including one with

$$P_{ab} = \begin{cases} \exp\left[-\lambda_d(a - b)\right], & b < a, \\ \exp\left[-\lambda_u(b - a)\right], & a > b, \end{cases} \tag{31}$$

where λ_d and λ_u determine how much the repeat number may change downwards or upwards respectively. If $\lambda_d > \lambda_u$, we can tune the two parameters λ_d and λ_u so that the equilibrium distribution $\{\pi_a\}_{a=1}^{\infty}$ has a mean and standard deviation that conform with observed data. It is well known however that the mutation rate also increases with number of repeats, and this requires the extended model of Section 3.2.2.

We only need to create mutations (at single nucleotides, codons or microsatellites) for the ancestral chromosomes, since all non-ancestral mutations in $\text{C} \setminus \text{AC}$ are censored, that is, they never reach the sampled chromosomes. More specifically, the ancestral mutational events

$$\text{AME} = (\text{AME}_{tcq};\, 0 \le t \le t_{\max}, c \in \text{AC}_t, q \in \text{AHB}_{ct}) \tag{32}$$

is a three-dimensional array of sets, where $\text{AME}_{tcq} \subset \text{hb}_q$ consists of those loci of haplotype block hb_q within ancestral chromosome (t, c) where a mutation occurs. The mutational events in (32) include founder mutations when $t = t_{\max}$ and germline mutations when $t = 0, \ldots, t_{\max} - 1$. The founder mutations occur with probabilities (26), and the germline mutations probabilities (28). It is only needed to compute mutational events AME_{tcq} for those haplotype blocks of chromosome (t, c) that contain ancestral material. These ancestral haplotype blocks of (t, c) are denoted as

$$\begin{aligned} \text{AHB}_{tc} \quad = \quad & \{q;\, (t, c) \text{ ancestral to at least one of} \\ & (0, c_1), \ldots, (0, c_n) \in \text{AC}_t \text{ at } \text{hb}_q\}. \end{aligned} \tag{33}$$

All other mutations are silent, and not visible in the sampled set of chromosomes. Figure 7 depicts the ancestral mutational events and ancestral haplotype blocks for the genealogy of Figure 5.
By gene dropping we mean that the ancestral recombination graph ARG and the ancestral mutational events AME can be used to

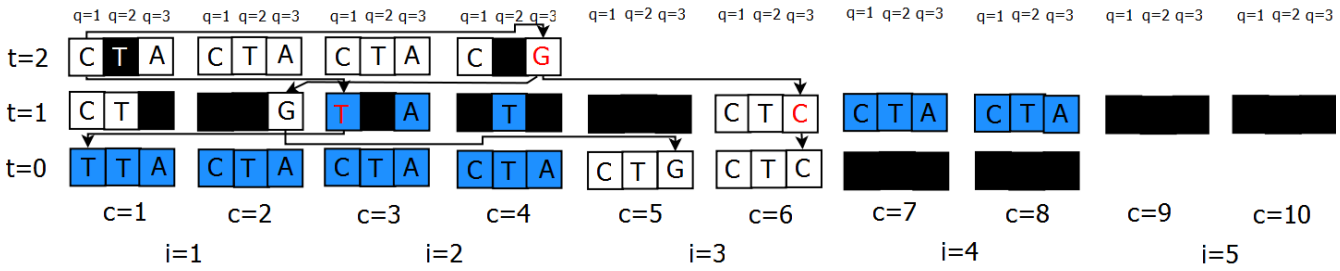

Figure 8: Illustration of gene dropping for nuclear autosomal DNA. The genealogy is as in Figure 5, with ancestral haplotype blocks and mutational events as in Figure 7. It is assumed that each haplotype block consists of one single nucleotide, with haplotype $\boldsymbol{h}^{\text{ref}} = (C, T, A)$ of the reference chromosome $(t, c) = (2, 1)$. Only the ancestral nucleotides (with white or light blue background) are shown for non-reference chromosomes. Mutational events are marked by red, and gene dropping is shown for all sampled nucleotides that have been mutated somewhere along their ancestral lineages. The sampled haplotypes are $\boldsymbol{h}_1 = (T, T, A)$, $\boldsymbol{h}_2 = (C, T, A)$, $\boldsymbol{h}_3 = (C, T, A)$, $\boldsymbol{h}_4 = (C, T, A)$, $\boldsymbol{h}_5 = (C, T, G)$ and $\boldsymbol{h}_6 = (C, T, C)$. **doi:**10.5048/BIO-C.2016.4.f8

spread the alleles of the reference chromosome $\boldsymbol{h}^{\text{ref}}$ to the whole sample. In other words, $\boldsymbol{h}^{\text{ref}}$, ARG and AME determine the output

$$\boldsymbol{H} = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_n) \tag{34}$$

of the algorithm uniquely for each simulation round, where

$$\boldsymbol{h}_k = (a_{k1}, a_{k2}, \ldots, a_{kL}) \tag{35}$$

is the haplotype of the k^{th} sampled chromosome $(0, c_k)$, see Figure 8 for an illustration.

1.6 Ignoring Double Mutations

If the infinite sites model [140] is assumed, each new mutation occurs at a locus where there have not been any previous mutations. Then there is no locus with double mutations, and all polymorphic loci are biallelic. This makes it possible to speed up the algorithm, since there is no need to generate a reference haplotype $\boldsymbol{h}^{\text{ref}}$, nor to gene drop from the founder generation to find the haplotypes of the sampled chromosomes. The reason is that in order to compute the output statistics of Section 1.8, it suffices to keep track of which chromosomes are descendants of the mutated chromosome at each polymorphic locus.

We formalize this by introducing the set

$$\text{SMC}_l \subset \{(0, c_1), \ldots, (0, c_n)\} \tag{36}$$

of sampled mutated chromosomes that experience a mutation at locus l. Obviously $\text{SMC}_l = \emptyset$ for loci without mutations, and for loci with a single historical mutation, SMC_l is the set of descendants of the mutated ancestral chromosome (see Figure 22 of Section 3.1.2). The absence of double mutations allows us to introduce a simplified set of two alleles for any locus l ($\mathcal{A}(l) = \{0, 1\}$), with 0 and 1 corresponding to an unmutated and mutated chromosome at this locus. Since the reference haplotype $\boldsymbol{h}^{\text{ref}} = (0, \ldots, 0)$ has no mutated sites, it is fixed and need not be evaluated. With "ism" an acronym for infinite sites model, the output of the algorithm is a collection

$$\boldsymbol{H}^{\text{ism}} = (\boldsymbol{h}_1^{\text{ism}}, \ldots, \boldsymbol{h}_n^{\text{ism}}) \tag{37}$$

of haplotypes

$$\boldsymbol{h}_k^{\text{ism}} = (a_{k1}^{\text{ism}}, \ldots, a_{kL}^{\text{ism}}) \tag{38}$$

from all sampled chromosomes $(0, c_k)$. The allele

$$a_{kl}^{\text{ism}} = 1\left((0, c_k) \in \text{SMC}_l\right) \tag{39}$$

at locus l equals 1 if and only if a mutation has occurred at this locus for some ancestor of $(0, c_k)$. Since it is only needed to store (37) at loci where a mutation occurs, it has a much simpler format than the output (34) of a model with double mutations, see Figure 9.

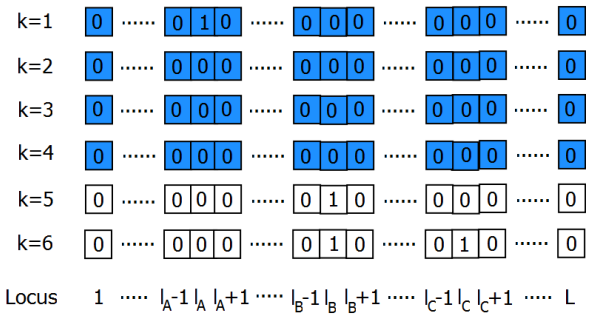

Figure 9: Illustration of haplotypes for the infinite sites model. The genealogy is the same as in Figures 5-7, and all three mutations are forced to occur at distinct loci $l_A < l_B < l_C$. The haplotypes $\boldsymbol{h}_k^{\text{ism}}$ in (38) correspond to rows for $k = 1, \ldots, 6$, whereas each column represents a locus. It is assumed that the single mutation of Figure 7 within chromosome $(1, 3)$ occurs at l_A, whereas the double mutation of Figure 7 is split into two distinct loci; l_B for chromosome $(2, 4)$, and l_C for chromosome $(1, 6)$. Mutated alleles are depicted as 1, and unmutated ones as 0. It is only needed to store l_A, l_B, l_C and alleles at these three loci, since all other alleles are 0. The sets of mutated chromosomes (36) at loci l_A, l_B, l_C are $\text{SMC}_{l_A} = \{(0, 1)\}$, $\text{SMC}_{l_B} = \{(0, 5), (0, 6)\}$, $\text{SMC}_{l_C} = \{(0, 6)\}$, and at all other loci l we have $\text{SMC}_l = \emptyset$. **doi:**10.5048/BIO-C.2016.4.f9

It is often reasonable to assume an infinite sites model for single nucleotide polymorphisms of nuclear DNA, apart from those mutations that occur at hotspot regions. But the mutation rate of microsatellites is too high to ignore that several repeat changes will occur at the same locus. On the other hand, for *Alu* polymorphisms in (7), it is usually assumed that each insertion only occurs in one ancestral chromosome [29]. This makes the infinite sites model very appropriate, with a reference haplotype without any polymorphic *Alu* insertions.

1.7 Summary of Algorithm and Input Parameters

A summary of the combined backward simulation, gene dropping and mutation generating algorithm is shown in Figure 10 when double mutations are allowed, with the accompanying input parameters listed in Table 3. The corresponding simplified algorithm for the infinite sites model is shown in Figure 11. It only requires a subset of the input parameters from Table 3.

1.8 Output Parameters

In this section we will look in more detail at the output of the algorithm. This output consists of haplotypes (34) for all sampled individuals, and it needs to be summarized in some convenient way. Table 4 lists a number of statistics that can be computed from (34),

Table 3: A list of input parameters for the backward simulation algorithm.

Parameter	Description
$t_{\max}$	Number of time points back to founder generation. The same as number of generations back to the founders, if generations are non-overlapping.
$\boldsymbol{T} = (T_1, \ldots, T_{t_{\max}})$	A list of all time points.
$\boldsymbol{D} = (D_0, \ldots, D_{t_{\max}})$	Number of demes/subpopulations at all time points.
$\boldsymbol{M} = \big((M_{td})_{d=1,\ldots,D_t}\big)_{t=0,\ldots,t_{\max}}$	Total male deme sizes at all time points.
$\boldsymbol{F} = \big((F_{td})_{d=1,\ldots,D_t}\big)_{t=0,\ldots,t_{\max}}$	Total female deme sizes at all time points.
$\boldsymbol{B} = \Big((B_{td,eg})_{d=1,\ldots,D_t,e,g=1,\ldots,D_{t+1}}\Big)_{t=0,\ldots,t_{\max}-1}$	Mating rule, i.e. joint backward migration probabilities of male and female parents between pairs of demes at all time points.
L	Number of loci of the chromosome.
Q	Number of haplotype blocks (optional).
α	Mating parameter controlling distribution of number of offspring of females.
β	Mating parameter controlling degree of monogamy.
$\boldsymbol{l} = (l_1, \ldots, l_{Q-1})$	Locations of rightmost loci of all haplotype blocks but the last (optional).
$\boldsymbol{r}_m = (r_{m1}, \ldots, r_{m,Q-1})$	Male recombination probabilities.
$\boldsymbol{r}_f = (r_{f1}, \ldots, r_{f,Q-1})$	Female recombination probabilities.
$\boldsymbol{\gamma}_m = (\gamma_{m1}, \ldots, \gamma_{m,Q-1})$	Male gene conversion probabilities.
$\boldsymbol{\gamma}_f = (\gamma_{f1}, \ldots, \gamma_{f,Q-1})$	Female gene conversion probabilities.
TLD	Tract length distribution for gene conversion.
$\boldsymbol{\nu} = (\nu_1, \ldots, \nu_L)$	Mutation probabilities for non-reference chromosomes of the founder generation, at all loci.
$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_L)$	Mutation probabilities per generation during gamete formation, at all loci.
$\boldsymbol{\pi} = (\pi_A, \pi_T, \pi_C, \pi_G)$ or $(\pi_1, \pi_2, \ldots)$	Allele frequencies of nucleotides or number of tandem repeats.
$\boldsymbol{P} = (P_{a,a'})$	Matrix of transition probabilities between all possible different pairs a, a' of alleles of a given marker.
$\boldsymbol{n} = \big(n_{md}, n_{mmd}, n_{fd}, n_{ffd}\big)_{d=1,\ldots,D}$	The number of sampled males and females that contribute with one or two chromosomes, from D demes ($1 \leq D \leq D_0$).

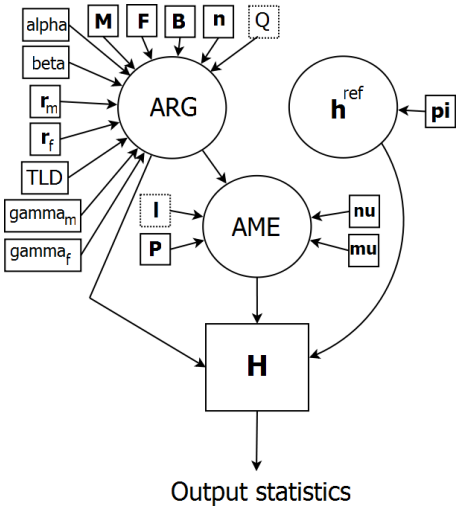

Figure 10: Directed acyclic graph illustrating the backwards-in-time simulation algorithm. Fixed or observed quantities are shown in boxes, including the input parameters from Table 3, the n sampled haplotypes $\boldsymbol{H}$ at present time 0, and the output statistics of Table 4. Since the number of haplotype blocks Q and their boundaries $\boldsymbol{l}$ are optional input parameters, we frame them with dashed boxes. Random quantities are shown in circles, and arrows indicate causal (deterministic or random) relationships. **doi:**10.5048/BIO-C.2016.4.f10

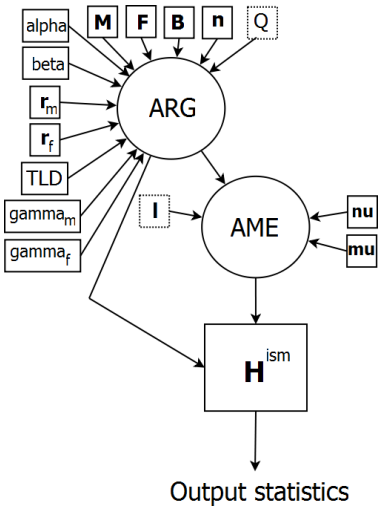

Figure 11: Directed acyclic graph that illustrates the backwards-in-time simulation algorithm for an infinite sites model, where double mutations are ignored. Compared to Figure 10, there is no need to generate a reference haplotype or do any gene dropping. Fixed or observed quantities are shown in boxes, including the input parameters from Table 3, the n sampled haplotypes $\boldsymbol{H}^{\text{ism}}$ at time 0, and the output statistics of Table 4. Since the number of haplotype blocks Q and their boundaries $\boldsymbol{l}$ are optional input parameter, we mark them as dashed boxes. Random quantities are shown in circles, and arrows indicate causal (deterministic or random) relationships. **doi:**10.5048/BIO-C.2016.4.f11

several of which are deme-specific. This is important for model validation, since real data sets reveal that subpopulations don't have the same patterns of DNA variation. A well fitting model should capture these differences; with separate population histories for each region in terms of how and when it was colonized, how its size varied in the past and how much interchange with other regions it had.

1.8.1 Single Locus Statistics

The single locus statistics summarize information about genetic diversity, without taking covariation between different loci into account. The amount of diversity is roughly proportional to the number of polymporphic loci. Whereas mutations (iv) and created founder variation (vi) tend to increase diversity, genetic drift (i) has the opposite effect of decreasing it. The smaller and younger a population is, the smaller is the amount of genetic diversity. The frequencies of different alleles at the polymorphic loci form a so called allele frequency spectrum. It tells how many of the alleles are common or rare, and it gives additional demographic information in terms of population size variations, and geographic division into demes. There are also single locus statistics that quantify how much variation there is between known subpopulations or demes.

Analytical results for the allele frequency spectrum have been obtained in [141] for homogeneous population of constant size, and in [142-145] for populations of time varying size. We will use simulations instead, since no analytical results have been obtained for models with two sexes, geographic substructure and time-varying sizes of the subpopulations. This will require simulation of separate statistics for each deme. In addition, we also quantify how genetically different the various demes are.

In order to describe the single locus statistics in more detail, we assume for simplicity that codon loci are disregarded, so that the chromosomal region (4) can be divided into two disjoint sets $\mathcal{L}_{\text{sn}}$ and $\mathcal{L}_{\text{ms}}$ of single nucleotides (3) and microsatellite loci (6), of sizes $L_{\text{sn}} = |\mathcal{L}_{\text{sn}}|$ and $L_{\text{ms}} = |\mathcal{L}_{\text{ms}}|$ (see Figure 3). Let also $\mathcal{D}_d \subset \{1, \ldots, n\}$ refer to the set of those k for which the k^{th} sampled chromosome c_k is from an individual of deme d. The number sampled chromosomes for which allele $a \in \mathcal{A}(l)$ appears at locus l in the whole sample, or in subpopulation d, are denoted as

$$\begin{aligned} n_{la} &= |\{k;\, 1 \le k \le n, a_{kl} = a\}|, \\ n_{lad} &= |\{k;\, k \in \mathcal{D}_d, a_{kl} = a\}|, \end{aligned}$$

see Table 5 for an illustration.

It is appropriate to treat single nucleotide and microsatellite loci separately. For single nucleotide loci, we let

$$\begin{aligned} \hat{\pi}_a &= \tfrac{1}{L_{\text{sn}} n} \textstyle\sum_{l \in \mathcal{L}_{\text{sn}}} n_{la}, \\ \hat{\pi}_{ad} &= \tfrac{1}{L_{\text{sn}} n_d} \textstyle\sum_{l \in \mathcal{L}_{\text{sn}}} n_{lad}, \end{aligned}$$

be the estimated fraction of allele a in the whole sample and deme d, with $n_d = n_{md} + n_{mmd} + n_{fd} + n_{ffd} = |\mathcal{D}_d|$ the number of sampled individuals from deme d. Let also

$$\begin{aligned} \text{nralleles}_l &= |\{a \in \mathcal{A}(l);\, n_{la} > 0\}|, \\ \text{nralleles}_{ld} &= |\{a \in \mathcal{A}(l);\, n_{lad} > 0\}|, \end{aligned} \tag{40}$$

be the number of different alleles that appear at a single nucleotide locus l, in the whole sample and deme d.

The amount of genetic variation varies quite a lot within different people groups. For instance, several Afriacan populations tend to have more diversity than non-African ones, see for instance [146] and references therein. In order to quantify how much variation

Table 4: A summary of output parameters.

Parameter	Description
$\hat{\boldsymbol{\pi}} = \{\hat{\pi}_a,\, a \in \mathcal{A}_{\rm sn}\}, \hat{\boldsymbol{\pi}}_d = \{\hat{\pi}_{ad};\, a \in \mathcal{A}_{\rm sn}\}$	Estimated frequencies for all single nucleotides in the whole sample, and within deme d.
S, S_d	Number of segregating sites in the whole sample and within deme d.
$\hat{\Pi}$, $\hat{\Pi}_d$	Estimated nucleotide diversity in the whole sample and within deme d.
$\hat{\boldsymbol{\Phi}} = \{\hat{\Phi}_j\}_{j=1,\ldots,M}$, $\hat{\boldsymbol{\Phi}}_d = \{\hat{\Phi}_{jd}\}_{j=1,\ldots,M}$	Allele frequency spectrum at biallelic loci in the whole sample and within deme d, for M frequency classes.
$\boldsymbol{F}_{ST} = \{F_{ST,de}\}_{1\leq d<e\leq D}$	Fixation indeces between all pairs of D demes.
$\boldsymbol{r}^2 = \{r^2_j\}_{j=1,\ldots,K}$, $\boldsymbol{r}^2_d = \{r^2_{jd}\}_{j=1,\ldots,K}$	Average squared correlation LD measure between pairs of biallelic loci at K different distances, for the whole sample and within deme d.
$\boldsymbol{CLDP} = \{\mathrm{CLDP}_j\}_{j=1}^{K}$, $\boldsymbol{CLDP}_d = \{\mathrm{CLDP}_{jd}\}_{j=1}^{K}$	The complete LD proportion of all pairs of biallelic loci at distance within each of K different classes, for the whole sample and within deme d.

Table 5: Allele counts for the six chromosomal region copies of Figure 3. An allele a either refers to a nucleotide (A, C, G, T) or a number of tandem repeats ($1, 2, 3, \ldots$). The upper part of the table gives the allele counts n_{la} for the two demes $d = 1, 2$ combined. The middle and lower parts of the table give the allele counts n_{lad} for each of the two demes.

		Locus l								
Sample	Allele	1	2	3	4	5	6	7	8	9
Comb	A	0	3	0	0	0	0	0	0	6
	C	6	0	0	0	2	6	0	3	0
	G	0	0	6	0	4	0	0	2	0
	T	0	3	0	6	0	0	0	1	0
	1	0	0	0	0	0	0	2	0	0
	2	0	0	0	0	0	0	2	0	0
	3	0	0	0	0	0	0	2	0	0
	Sum	6	6	6	6	6	6	6	6	6
Deme 1	Allele	1	2	3	4	5	6	7	8	9
	A	0	3	0	0	0	0	0	0	4
	C	4	0	0	0	2	4	0	1	0
	G	0	0	4	0	2	0	0	2	0
	T	0	1	0	4	0	0	0	1	0
	1	0	0	0	0	0	0	1	0	0
	2	0	0	0	0	0	0	2	0	0
	3	0	0	0	0	0	0	1	0	0
	Sum	4	4	4	4	4	4	4	4	4
Deme 2	Allele	1	2	3	4	5	6	7	8	9
	A	0	0	0	0	0	0	0	0	2
	C	2	0	0	0	0	2	0	2	0
	G	0	0	2	0	2	0	0	0	0
	T	0	2	0	2	0	0	0	0	0
	1	0	0	0	0	0	0	1	0	0
	2	0	0	0	0	0	0	0	0	0
	3	0	0	0	0	0	0	1	0	0
	Sum	2	2	2	2	2	2	2	2	2

there is in the whole human population, or within a subregtion (deme) d, we can use the number of segregating sites

$$\begin{aligned} S &= |\{l \in \mathcal{L}_{\rm sn};\, {\rm nralleles}_l > 1\}|, \\ S_d &= |\{l \in \mathcal{L}_{\rm sn};\, {\rm nralleles}_{ld} > 1\}|, \end{aligned} \tag{41}$$

in the whole sample and in deme d. This is the same thing as the number of single nucleotide polymorphisms (SNPs). Two related quantities are the nucleotide diversities Π and Π_d [147-148]. In our context they can be interpreted as the probabilities that two randomly chosen *different* chromosomes, from the whole population or from deme d, have different alleles at a randomly selected locus. The observed human nucleotide diversity $\hat{\Pi}$ of nuclear DNA is around 0.08% [149-151], although it is highest for non-sex chromosomes, lower for X chromosomes, and smallest for Y chromosomes. For mitochondrial DNA it is higher, about 0.25% [29].

For autosomal DNA, the nucleotide diversity can be estimated as

$$\begin{aligned} \hat{\Pi} &= \frac{1}{L_{\rm sn}} \sum_{l \in \mathcal{L}_{\rm sn}} \left[1 - \sum_{a \in \mathcal{A}_{\rm sn}} \hat{\pi}_{la} \frac{2\hat{\pi}_{la} N_0 - 1}{2N_0 - 1} \right], \\ \hat{\Pi}_d &= \frac{1}{L_{\rm sn}} \sum_{l \in \mathcal{L}_{\rm sn}} \left[1 - \sum_{a \in \mathcal{A}_{\rm sn}} \hat{\pi}_{lad} \frac{2\hat{\pi}_{lad} N_{0d} - 1}{2N_{0d} - 1} \right], \end{aligned} \tag{42}$$

for the whole population or within deme d. Here $\hat{\pi}_{la} = n_{la}/n$ and $\hat{\pi}_{lad} = n_{lad}/n$ are the estimated frequencies of allele a at locus l, globally or locally within deme d respectively. The rationale of (42) is that autosomal DNA has $2N_0$ chromosomes in the total population at present, of which $2N_{0d}$ belong to deme d. A fraction π_{la} (π_{lad}) of these chromosomes carry allele a at locus l, and $\hat{\pi}_{la}$ ($\hat{\pi}_{lad}$) are estimates of these quantities derived from the sample. Since there is only a single copy of Y-chromosome and one kind of mitochondrial DNA per individual, the 2 factors of (42) are removed.

Let

$$\begin{aligned} \mathcal{L}_{\rm ba} &= \{l \in \mathcal{L}_{\rm sn};\, {\rm nralleles}_l = 2\}, \\ \mathcal{L}_{{\rm ba},d} &= \{l \in \mathcal{L}_{\rm sn};\, {\rm nralleles}_{ld} = 2\} \end{aligned}$$

be the set of biallelic loci in the whole sample and in deme d, and write $S_{\rm ba} = |\mathcal{L}_{\rm ba}|$ and $S_{{\rm ba},d} = |\mathcal{L}_{{\rm ba},d}|$ for the corresponding number of segregating biallelic sites. At such loci it is of interest to compute estimates of the minor allele frequencies

$$\begin{aligned} \widehat{\rm MAF}_l &= \min\{n_{la};\, a \in \mathcal{A}(l), \hat{\pi}_{la} > 0\}/n, \\ \widehat{\rm MAF}_{ld} &= \min\{n_{lad};\, a \in \mathcal{A}(l), \hat{\pi}_{la} > 0\}/n_d. \end{aligned} \tag{43}$$

In order to quantify the allele frequency spectrum, we will summarize the minor allele frequencies at all biallelic loci. This is achieved by dividing the frequency range $\pi \in [0, 0.5]$ into M equispaced intervals $\mathcal{P}_j = [(j-1)/(2M), j/(2M)]$ of length $1/(2M)$. With ϕ and ϕ_d the minor allele frequency densities in the whole population and in deme d, we estimate the fraction of biallelic loci

$$\begin{aligned} \Phi_j &= \int_{\mathcal{P}_j} \phi(\pi) d\pi, \\ \Phi_{jd} &= \int_{\mathcal{P}_j} \phi_d(\pi) d\pi \end{aligned}$$

that belong to interval j, by

$$\begin{aligned} \hat{\Phi}_j &= |\{l \in \mathcal{L}_{\rm ba};\, \widehat{\rm MAF}_l \in \mathcal{P}_j\}|/S_{\rm ba}, \\ \hat{\Phi}_{jd} &= |\{l \in \mathcal{L}_{\rm ba};\, \widehat{\rm MAF}_{ld} \in \mathcal{P}_j\}|/S_{{\rm ba},d}, \end{aligned} \tag{44}$$

for $j = 1, \ldots, M$. Table 6 shows the frequency spectrum with $M = 5$ intervals, for a standard model with constant population size and no geographic subdivision [149]. Population increase and subdivision will both have the effect of increasing the frequency of rare variants even more, whereas a recent bottleneck has the opposite effect of reducing their number. It is well known that many human subpopulations have an excess of rare variants compared to the standard model. It is likely that recent population increase is the most important explanation [152,153]. Since population histories vary between regions d, their allele frequency spectra will not be the same. For instance, many African populations tend to have a slightly smaller fraction of common variants [21,23,154] compared to non-African ones.

Table 6: Allele frequency spectrum under a standard neutral model of evolution. There is no subpopulation division and population size is constant over time. The interval with the smallest frequency range has been truncated, so that only SNPs with a minor allele frequency of at least 0.01 are included.

MAF interval $\mathcal{P}_j$	Frequency Φ_j
(0.01,0.10)	0.522
(0.10,0.20)	0.177
(0.20,0.30)	0.117
(0.30,0.40)	0.096
(0.40,0.50)	0.088

It is also important to assess how genetically different the various demes are. To this end, we define a fixation index $F_{ST,de}$ for each pair d, e of demes from which samples are taken ($1 \le d < e \le D$). This is a number between 0 and 1 that quantifies genetic difference between the two demes [52,155]. Its multilocus and multiallelic version is defined as

$$F_{ST,de} = \frac{\sum_l \sum_a \frac{1}{2} \left[(\hat{\pi}_{lad} - \bar{\pi}_{lade})^2 + (\hat{\pi}_{lae} - \bar{\pi}_{lade})^2 \right]}{\sum_l \left[1 - \sum_a \bar{\pi}^2_{lade} \right]}, \tag{45}$$

where the outer and inner sums are taken over all single nucleotide loci ($l \in \mathcal{L}_{\rm sn}$) and all alleles at each locus ($a \in \mathcal{A}(l)$) respectively, and $\bar{\pi}_{lade} = (\hat{\pi}_{lad} + \hat{\pi}_{lae})/2$ is the average frequency of allele a at locus l in the two demes d and e.

The diversity statistics in (41) and (42) are defined in the same way for microsatellite loci, but the allele frequency spectrum is more complicated, so that (44) is not used. Instead, a number of other statistics are defined in [15,16] and references therein from the observed distribution of number of tandem repeats. The subpopulation differentiation statistic, F_{ST}, is usually replaced by another quantity R_{ST} defined in [156] for microsatellites, or by a closely related distance measure between subpopulations [157].

Any of the output statistics in (41)-(45) are appropriate for the infinite sites model (including *Alu*s), since all of its mutated loci are biallelic.

1.8.2 Statistics for Pairs of Loci

Covariation of alleles in a population between different loci (linkage disequilibrium, LD) is summarized by various statistics. LD is caused by a number of different factors [158], but among the forces of change introduced in Section 1.4, it basically occurs as a balance between genetic drift (i), admixture of/migration between subpopulations (iii) and founder diversity (vi) on one hand, and recombinations (ii) on the other. Whereas genetic drift, migration and founder diversity tend to increase LD, recombinations tend to decrease it. Statistics that quantify allelic covariation at several loci therefore complement those in Section 1.8.1 for single loci, which mainly quantify another balance - between genetic drift and mutations. The level of LD over shorter or longer distances is highly influenced by demographic history, such as the age of a population, rapid size expansions, severe bottlenecks or geographic substructure. This is known both from simulations [151,159] and theoretical investigations [160]. The amount of LD between closely located markers, on one hand, give information about human demography in the more distant past [161], whereas the degree of LD between distant markers reflect more recent demographic events [162].

Table 7: A table with two rows and two columns showing counts for all four haplotypes from two biallelic loci. These loci are denoted l and l'.

	a'	b'	Sum
a	$n_{ll'aa'}$	$n_{ll'ab'}$	n_{la}
b	$n_{ll'ba'}$	$n_{ll'bb'}$	n_{lb}
Sum	$n_{l'a'}$	$n_{l'b'}$	n

Table 8: Measures of association between two of the loci of Figure 3. These are the only two loci $l = 2$ and $l' = 5$ that correspond to a single nucleotide that is biallelic. The upper part of the table gives twolocus haplotype counts $n_{25aa'}$, $n_{25aa',1}$ and $n_{25aa',2}$ for the whole sample, deme $d = 1$ and deme $d = 2$ respectively. The haplotypes (a, a') consist of two alleles, one from each of l and l'. The lower part quantifies dependency between the two loci in terms of the unstandardized and standardized measures of linkage disequilibrium. The two standardized measures (r^2, D') are not well-defined for $d = 2$, since locus $l' = 5$ is not biallelic within this deme.

Hapl counts	Haplotype (a, a')	Comb $n_{25aa'}$	Deme 1 $n_{25aa',1}$	Deme 2 $n_{25aa',2}$
	(A, C)	1	1	0
	(A, G)	2	2	0
	(T, C)	1	1	0
	(T, G)	2	0	2
	Sum	6	4	2
Dependency	LD measure	Comb	Deme 1	Deme 2
	$\delta_{25,\cdot}$	0	-2	0
	$r^2_{25,\cdot}$	0	2/3	-
	$D'_{25,\cdot}$	0	1	-

We will focus on statistics for pairs of biallelic loci (typically a single nucleotide), which is appropriate, for instance, for the infinite sites model. Several measures of LD for pairs of biallelic loci exist, see for instance [163] and Chapter 8 of [164]. We concentrate on two of the most commonly used, the squared correlation coefficient r^2 and Lewontin's D'. They are computed between loci at various distances, for the whole population and within each deme. It is also possible to use the population recombination rate. This measure of LD is more sophisticated than r^2 and D', but also more complicated to estimate, and in particular it requires some assumptions about the population history [110].

In order to define statistics for pairs of loci, let $n_{ll'aa'}$ be a two-locus haplotype count, i.e. the number of sampled chromosomes $(0, c_k)$ that have haplotype (a, a') at two biallelic loci $l, l' \in \mathcal{L}_{\rm ba}$ (that is, $a_{kl} = a$ and $a_{kl'} = a'$). If a, b and a', b' are the two alleles present at l and l', we may summarize all haplotype counts from these two loci as in Table 7. Haplotype counts $n_{ll'aa',d}$ for the subsample of deme $d = 1, \ldots, D$ are defined in the same way.

The coefficients of linkage disequilibrium

$$\begin{array}{rcl} \delta_{ll'} & = & n_{ll'aa'}n_{ll'bb'} - n_{ll'ba'}n_{ll'ab'}, \\ \delta_{ll',d} & = & n_{ll'aa',d}n_{ll'bb',d} - n_{ll'ba',d}n_{ll'ab',d}, \end{array} \tag{46}$$

are unstandardized measures of dependency of the allelic variation at l and l', for the whole sample and deme d, with values of 0 corresponding to independence (see Table 8 for an illustration).

The squared correlation coefficient

$$\begin{array}{rcl} r^2_{ll'} & = & \delta^2_{ll'}/(n_{la}n_{lb}n_{l'a'}n_{l'b'}), \\ r^2_{ll',d} & = & \delta^2_{ll',d}/(n_{la,d}n_{lb,d}n_{l'a',d}n_{l'b',d}), \end{array}$$

and Lewontin's

$$\begin{array}{rcl} D'_{ll'} & = & |\delta_{ll'}|/\delta_{\max,ll'}, \\ D'_{ll',d} & = & |\delta_{ll',d}|/\delta_{\max,ll',d}, \end{array} \tag{47}$$

are different standardizations of (46), with values between 0 and 1, for the whole population and for each deme d. Both quantities are 0 for allelic independence between loci l and l', and larger values indicate a stronger association. The denominator of (47) is

$$\delta_{\max,ll'} = \begin{cases} \min(n_{la}n_{l'b'}, n_{lb}n_{l'a'}), & \text{if } \delta_{ll'} > 0, \\ \min(n_{la}n_{l'a'}, n_{lb}n_{l'b'}), & \text{if } \delta_{ll'} < 0, \end{cases} \tag{48}$$

for the whole sample, and $\delta_{\max,ll',d}$ is defined analogously within deme d. Complete LD means that one of the four haplotype counts of Table 7 are absent. It commonly occurs between two nearby loci l and l', when one of them has recently mutated. The normalization in (48) implies that $D'_{ll'}$ equals 1 for complete LD. This is not the case for $r^2_{ll'}$, which is smaller than 1 even for complete LD, unless the allele frequencies at the two loci l and l' are the same.

Since recombinations tend to break up LD, and the expected number of recombinations is larger between more widely separated pairs l, l' of loci, $r^2_{ll'}$ and $D'_{ll'}$ are both, *on average*, decreasing functions of the distance between l and l', although the variation is large. This decay of average LD with distance can be estimated by dividing all pairs of loci within a certain range or maximal distance (md) into K distance classes

$$\begin{array}{rcl} \mathcal{L}^{\rm pairs}_j & = & \{l, l' \in \mathcal{L}_{\rm ba};\ \frac{j-1}{K} \cdot {\rm md} < |l - l'| \le \frac{j}{K} \cdot {\rm md}\}, \\ \mathcal{L}^{\rm pairs}_{jd} & = & \{l, l' \in \mathcal{L}_{{\rm ba},d};\ \frac{j-1}{K} \cdot {\rm md} < |l - l'| \le \frac{j}{K} \cdot {\rm md}\}, \end{array}$$

for $j = 1, \ldots, K$, with typical values md = 100 kb and $K = 40$. Then

$$\begin{array}{rcl} r^2_j & = & \frac{1}{|\mathcal{L}^{\rm pairs}_j|} \sum_{l,l' \in \mathcal{L}^{\rm pairs}_j} r^2_{ll'}, \\ r^2_{jd} & = & \frac{1}{|\mathcal{L}^{\rm pairs}_{jd}|} \sum_{l,l' \in \mathcal{L}^{\rm pairs}_{jd}} r^2_{ll',d}, \end{array} \tag{49}$$

is the average value of r^2 for all pairs of loci at a distance within class j, for the whole sample and deme d. In order to get a measure of LD that is more complementary to r^2, we follow [23] and define the Complete LD Proportion

$$\begin{array}{rcl} {\rm CLDP}_j & = & \frac{1}{|\mathcal{L}^{\rm pairs}_j|} \sum_{l,l' \in \mathcal{L}^{\rm pairs}_j} 1(D'_{ll'} = 1), \\ {\rm CLDP}_{jd} & = & \frac{1}{|\mathcal{L}^{\rm pairs}_{jd}|} \sum_{l,l' \in \mathcal{L}^{\rm pairs}_{jd}} 1(D'_{ll',d} = 1), \end{array} \tag{50}$$

i.e. the proportion of pairs of loci within each distance class that are in complete LD ($D'_{ll'} = 1$), with $1(A)$ equal to 1 if A holds and 0 otherwise.

The deme-specific LD measures in (49) and (50) will reflect differences in population histories. For instance, it is well known that linkage disequlibrium extends over longer distances in non-African populations compared to African ones [78,109,149].

The average LD measures in (49) and (50) are simple, but have a disadvantage of clumping chromosomal regions with short and long range LD into one statistic. This can be avoided by focusing on recombination based genetic map distance classes rather than physical distance classes that are defined in terms of number of base pairs [23,159]. The expected r^2_j values in (49), for instance, are inversely related to the probability of recombination between the two loci [165]. However, since recombination probabilities are more difficult to estimate over short distances, we use physical distance classes instead.

When haplotype blocks are not specified in advance, we can use the number of haplotype blocks Q as output statistic [115]. Although Q is simpler than the LD measures (49) and (50), is roughly conveys the same information, since it is inversely proportional to how far away linkage disequilibrium extends.

1.9 Model Fitting and Validation

In order to fit and validate our model, one training data set and one validation data set is needed. The training data set

$$\boldsymbol{H}^{\text{train}} = (\boldsymbol{h}_1^{\text{train}}, \ldots, \boldsymbol{h}_n^{\text{train}}) \tag{51}$$

consists of haplotypes from a number of individuals. Since our model contains a large number of parameters $\boldsymbol{\theta}$, it is very challenging to estimate $\boldsymbol{\theta}$ from training data by simulation-based maximum likelihood or Bayesian techniques, see for instance [25,166-168] and references therein.

We will use a simpler approach that is similar to the one in [23], where $\boldsymbol{\theta}$ is tuned so that a number of output statistics

$$\boldsymbol{X}_u^{\text{train}} = (X_{u1}^{\text{train}}, \ldots, X_{uV_u}^{\text{train}}), \quad u = 1, \ldots, U, \tag{52}$$

of Table 4 for the training data set (51) get as close as possible to the corresponding output statistics

$$\boldsymbol{X}_u = (X_{u1}, \ldots, X_{uV_u}), \quad u = 1, \ldots, U, \tag{53}$$

computed from R simulated sets. These simulated data sets are either generated from (34), if double mutations are accounted for, or from (37), if double mutations are ignored. Scalar statistics such as the deme-specific number of segregating sites S_d and nucleotide diversity $\hat{\Pi}_d$ have only one component ($V_u = 1$), the fixation index $\boldsymbol{F}_{ST}$ has one component for each pair of demes ($V_u = D(D-1)/2$), the deme-specific allele frequency spectrum $\hat{\boldsymbol{\Phi}}_d$ has one component for each allele frequency class ($V_u = M$), and the deme-specific LD statistics $\boldsymbol{r}_d^2$ and $\boldsymbol{CLD}_d$ have one component for each distance class ($V_u = K$).

One option is to choose $U = 5D + 1$ statistics in (53), including $\boldsymbol{F}_{ST}$, and S_d, $\hat{\Pi}_d$, $\hat{\boldsymbol{\Phi}}_d$, $\boldsymbol{r}_d^2$, $\boldsymbol{CLD}_d$ for all demes $d = 1, \ldots, D$. A simpler option is to use only $\boldsymbol{F}_{ST}$, and all $\hat{\boldsymbol{\Phi}}_d$, $\boldsymbol{r}_d^2$. The reason is that the two LD measures are correlated, and the extent of LD is inversely related to the nucleotide diversity and the number of segregating sites [151].

Let

$$\boldsymbol{H}^{(r)} = (\boldsymbol{h}_1^{(r)}, \ldots, \boldsymbol{h}_n^{(r)})$$

be the simulated haplotypes (34) or (37) for all sampled chromosomes in repeat r, and $X_{uv}^{(r)}$ the value of the vth component of the uth output statistic for the same repeat. We summarize these R output statistics, for each u, v, by their sample mean and sample variance

$$\begin{aligned} X_{uv} &= \textstyle\sum_{r=1}^{R} X_{uv}^{(r)}/R, \\ \sigma_{uv}^2 &= \textstyle\sum_{r=1}^{R} (X_{uv}^{(r)} - X_{uv})^2/(R-1). \end{aligned}$$

It is also possible to estimate σ_{uv}^2 from the empirical data set by resampling [23]. In any case, the goodness of fit is defined as

$$\Delta_u = \sqrt{\frac{1}{V_u} \sum_{v=1}^{V_u} \frac{(X_{uv}^{\text{train}} - X_{uv})^2}{\sigma_{uv}^2}},$$

for statistic number u, and

$$\Delta = \sqrt{\sum_{u=1}^{U} \Delta_u^2} \tag{54}$$

for all statistics combined. A scenario is chosen by adjusting the parameters in $\boldsymbol{\theta}$ so that Δ gets as small as possible. A value around $\Delta = 1$ for the chosen parameters signifies a perfect fit, and the larger $\Delta > 1$ is, the poorer is the fit.

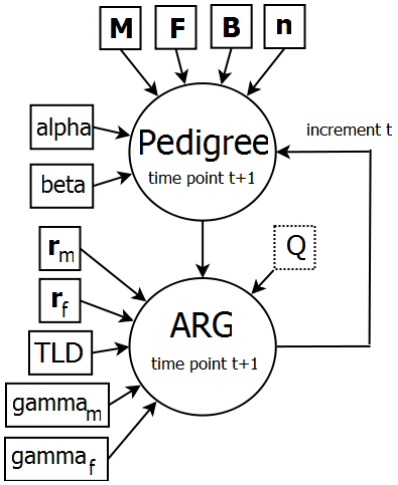

Figure 12: Illustration of how a directed acyclic graph is built recursively in time. The graph is a subpart of Figure 10, and in each step $t = 0, \ldots, t_{\text{max}} - 1$, members of the pedigree at time point t are assigned parents at time $t + 1$. Then Mendelian inheritance and recombination events generate the ancestral recombination graph ARG for time point $t + 1$, before updating time and assigning parents or self numbers to this time point. Fixed input parameters (see Table 3) appear in boxes, random quantities are shown in circles, and arrows indicate causal (deterministic or random) relationships. The number of haplotype blocks Q is an optional input parameter and appears in a dashed box. **doi:**10.5048/BIO-C.2016.4.f12

In order to validate the model, we compute a quantity Δ^{val} in the same way as in (54), using the output statistics of a validation data set

$$\boldsymbol{H}^{\text{val}} = (\boldsymbol{h}_1^{\text{val}}, \ldots, \boldsymbol{h}_n^{\text{val}})$$

from another chromosome, and R simulated data sets. These simulations are based on the model that was fitted to the training data set, possibly changing some chromosome-specific parameters, such as recombination or mutation rates.

If data is not sampled randomly, it is important for model fitting and model validation that the simulated and real data sets are ascertained in the same way, both in terms of which individuals and loci that are sampled. If the ascertainment mechanism for the real data set is known, some polymorphic loci of the simulated data sets can be removed, as described in Section 3.3. A common non-random sampling mechanism (for instance in the HapMap data sets) is that more polymorphic loci have a higher sampling probability than the less polymorphic ones.

2. GENERATING ANCESTRY

In this section we give a detailed description of how the ancestral recombination graph ARG is generated. It requires a pedigree of individuals and how DNA is inherited in this pedigree, in particular recombination events between the haplotype blocks of the ancestral chromosomes, see Figure 12.

The ARG in (17) starts at time point $t = 0$, with

$$\text{ARG}_{0kq} = c_k$$

for the kth sampled chromosome c_k at haplotype block hb_q. In order to initiate the ARG, we thus need to specify all c_k. This numbering of sampled chromosomes will depend on DNA type as well as the sample size numbers in (15). For nuclear autosomal DNA it seems most natural, in order to use all available data, to include both chromosomes of within a sampled male or female.

Then there is no need to specify, for any chromosome that did not have its homologous copy in the sample, whether it was inherited from a grandfather or grandmother. For these reasons, we assume $n_{md} = n_{fd} = 0$ for all demes d, so that the total sample size (16) simplifies to $n = 2\sum_{d=1}^{D}(n_{mmd} + n_{ffd})$. If the sampled males and females are assigned the lowest numbers at time 0 within their sex, we get

$$c_k = \begin{cases} k, & k = 1, \ldots, 2n_{mm}, \\ 2M_0 + k - 2n_{mm}, & k = 2n_{mm} + 1, \ldots, n. \end{cases} \quad (55)$$

The numbering of chromosomes is similar for other types of DNA. Since Y chromosomes are not sampled within females and only one is sampled per male, we put $n_{mmd} = n_{ffd} = n_{fd} = 0$, so that the total number of sampled haplotypes $n = \sum_{d=1}^{D} n_{md}$ are taken from chromosomes $c_k = 2k - 1$, $k = 1, \ldots, n$. Mitochondrial DNA is only sampled one per female, so that $n_{mmd} = n_{md} = n_{ffd} = 0$, $n = \sum_{d=1}^{D} n_{fd}$, and $c_k = 2M_0 + 2k$ for $k = 1, \ldots, n$. X chromosomes, finally, are sampled one per male and two per female, so that $n_{mmd} = n_{fd} = 0$, $n = \sum_{d=1}^{D}(n_{md} + 2n_{ffd})$, $c_k = 2k$ for $k = 1, \ldots, n_m$, and $c_k = 2M_0 + (k - n_m)$ for $k = n_m + 1, \ldots, n$, where $n_m = \sum_{d=1}^{D} n_{md}$ is the total number of sampled males.

Once the ancestral recombination graph has been generated for time point $t = 0$, we need to specify individuals of the pedigree and inheritance of their DNA, recursively back in time $t = 0, 1, \ldots, t_{\max}$. In order to describe the pedigree, we introduce for each t the set

$$\mathrm{AI}_t = \{1, \ldots, m_t, M_t + 1, \ldots, M_t + f_t\} \quad (56)$$

of ancestral individuals that are alive at this time point t. It consists of $m_t \leq M_t$ males and $f_t \leq F_t$ females, assuming without loss of generality that males and females of each time point are numbered so that the ancestral ones come first. An individual $i \in \mathrm{AI}_t$ if at least one of its two chromosomes is ancestral. That is, $i \in \mathrm{AI}_t$ if either the chromosome $c = 2i - 1$ that i inherited from its father belongs to AC_t, or if the chromosome $c = 2i$ that i inherited from its mother does.

The pedigree is a directed acyclic graph on AI_t. The details of how it is built can be found in Section 2.1. Briefly, as we proceed backwards in time $t = 0, 1, \ldots, t_{\max}$, in the recursive step from t to $t + 1$, we need to specify for each ancestral individual $i \in \mathrm{AI}_t$ that is a newborn, at a time point before the founder generation $t < t_{\max}$, its two male and female parents

$$\begin{aligned} m_t(i) &\in \{1, \ldots, M_{t+1}\}, \\ f_t(i) &\in \{1, \ldots, F_{t+1}\}, \end{aligned} \quad (57)$$

see Table 9 for an example. All non-founders in this table have ancestral parents. This is not always the case though. A necessary and sufficient condition for an individual at time $t+1$ to be ancestral is that he or she is the parent of at least one ancestral individual at time t, *and* transmitted at least one haplotype block to a child for which this block is ancestral. In view of (56), this implies

$$\begin{aligned} m_t(i) &\in \{1, \ldots, m_{t+1}\}, \text{ iff } 2i - 1 \in \mathrm{AC}_t \\ &\quad \text{for at least one his children } i, \\ f_t(i) &\in \{1, \ldots, f_{t+1}\}, \text{ iff } 2i \in \mathrm{AC}_t \\ &\quad \text{for at least one her children } i, \end{aligned} \quad (58)$$

where iff is short for "if and only if". If the population size $N_t = M_t + F_t$ is large for many time points, it is very important to prune the pedigree by removing non-ancestral individuals. Otherwise the computational advantage of backward simulation diminishes, as the pedigree size quickly increases when time proceeds backwards [169,170]. Pruning ensures that only those parents (57) that satisfy (58) are part of the pedigree. This is described in detail in Section 2.1, and here we assume that the pedigree has been pruned.

Table 9: Numbers associated with individuals for the pedigree in the lower part of Figure 4. This includes order number $m_t(i)$ of the father, order number $f_t(i)$ of the mother, and deme number $d_t(i)$ for all newborn non-founders (t, i) of the pedigree. For the adult female ($(t, i) = (0, 2)$), the self number $s_t(i)$ is displayed instead of the parental order numbers.

(t,i)	$m_t(i)$	$f_t(i)$	$s_t(i)$	$d_t(i)$
(0,1)	2	2	-	2
(0,2)	-	-	2	2
(0,3)	1	1	-	1
(1,1)	1	1	-	1
(1,2)	1	1	-	2
(1,3)	1	1	-	1
(1,4)	1	1	-	2

All individuals are newborns for a population with non-overlapping generations. But if generations are overlapping there are adults, whose parents were born more than one time step earlier. Since each adult $i \in \mathrm{AI}_t$ existed at time T_{t+1}, it is assigned a "self number"

$$\begin{aligned} s_t(i) &\in \{1, \ldots, m_{t+1}, M_{t+1} + 1, \ldots, M_{t+1} + f_{t+1}\} \\ &= \mathrm{AI}_{t+1}, \end{aligned} \quad (59)$$

rather than a father $m_t(i)$ and mother $f_t(i)$, regardless of whether the parents of (t, i) were still alive at time T_{t+1}.

The set of ancestral individuals AI_{t+1} at time T_{t+1} is completely specified, once we have defined parents (57) of newborns or self numbers (59) for adults, for all $i \in \mathrm{AI}_t$, and then removed all of these individuals that are non-ancestral. Before generating parents and self numbers for the next time point T_{t+2}, we must first define how parental DNA from time point T_{t+1} was transmitted from parents to newborns. If (t, c) is an ancestral chromosome ($c \in \mathrm{AC}_t$) of a non-founder time point $t < t_{\max}$, it has at least one ancestral haplotype block. Such a chromosome (t, c) resides within individual $i = [(c + 1)/2] \in \mathrm{AI}_t$, and its haplotype block hb_q (whether ancestral or not) was inherited from chromosome $(t + 1, p_{tq}(c))$ at time $t + 1$, where

$$\begin{aligned} p_{tq}(c) &= \text{number of the chromosome in } \mathrm{AC}_{t+1} \\ &\quad \text{that is parental to } (t, c) \text{ at } \mathrm{hb}_q. \end{aligned}$$

For DNA of an ancestral chromosome (t, c) within a newborn $i = [(c+1)/2]$ that is either mitochondrial, from an Y chromosome, or from an X chromosome within a female and inherited from a father, we know not only whether i received this DNA from the father $m_t(i)$ or the mother $f_t(i)$, but also which grandparent that transmitted DNA through this parent to i. If (t, c) is inherited from a father (c odd), we can formalize this as

$$p_{tq}(c) \overset{c \text{ odd}}{=} \begin{cases} 2m_t\left([\frac{c+1}{2}]\right) - 1, \\ 2m_t\left([\frac{c+1}{2}]\right), \end{cases} \quad (60)$$

where the upper and lower rows on the right hand side of (60) are for Y DNA and X DNA of a female respectively, since a father always passes on its Y chromosome to the child from the grandfather and its X chromosome from the grandmother. For a chromosome (t, c) inherited from a mother (c even), we have

$$p_{tq}(c) \overset{c \text{ even}}{=} 2M_t + 2f_t\left([\frac{c+1}{2}]\right), \quad (61)$$

for mt DNA, since mitochondria are always inherited through the mother from her mother. On the other hand, for autosomal DNA, X chromosome DNA within males, or X chromosome DNA that a female inherited from her mother, we need to specify which of its two chromosomes each of the two parents in (57) passed on to the child at each haplotype block hb_q. In order to determine

these parental chromosome numbers we use Mendelian laws of inheritance. For a chromosome (t,c) inherited from a father (c odd), these laws imply that the parental chromosome has number

$$\text{Prob}(p_{tq}(c)=c') \overset{c \text{ odd}}{=} \begin{cases} 0.5, & c'=2m_t\left(\left[\frac{c+1}{2}\right]\right)-1, \\ 0.5, & c'=2m_t\left(\left[\frac{c+1}{2}\right]\right), \end{cases} \tag{62}$$

for autosomal DNA, depending on whether a grandpaternal or grandmaternal chromosome was transmitted at haplotype block hb_q from time $t+1$ to t. In the same way, a parental chromosome that (t,c) inherited from a mother (c even) has number

$$\text{Prob}(p_{tq}(c)=c') \overset{c \text{ even}}{=} \begin{cases} 0.5, & c'=2M_t+2f_t\left(\left[\frac{c+1}{2}\right]\right)-1, \\ 0.5, & c'=2M_t+2f_t\left(\left[\frac{c+1}{2}\right]\right), \end{cases} \tag{63}$$

for autosomal or X DNA. It is much easier to assign parental chromosomes if (t,c) resides with an adult $i=[(c+1)/2]$. We only need to know the self number in (59), and whether c was transmitted from a father (c odd) or mother (c even). Formally we write this as

$$p_{tq}(c)=2s_t\left(\left[\frac{c+1}{2}\right]\right)-1+1(c \text{ even}), \tag{64}$$

for $q=1,\ldots,Q$. As a next step, for autosomal or X chromosome DNA we generate recombination events between haplotype blocks. To this end, we first assume that all haplotype blocks are known in advance, and write the haplotype block boundaries as

$$\text{hbb}=\{\text{hbb}_1,\ldots,\text{hbb}_{Q-1}\}.$$

The qth haplotype block boundary hbb_q is located between the adjacent pair hb_q and hb_{q+1} of haplotype blocks, that is, between loci l_q and l_q+1. For each ancestral chromosome $(t,c)\in\text{AC}$ of a non-founder time point, we let

$$\text{REC}_{tc}\subset\text{hbb} \tag{65}$$

refer to the set of haplotype block boundaries at which a recombination event (that is, an odd number of crossovers) occurred when the first copy of (t,c) was formed in a germ cell. There are no recombination events ($\text{REC}_{tc}=\emptyset$) for adults, or for DNA of newborns that is located within mitochondria, the non-recombining parts of Y chromosomes (NRY), or X chromosomes from fathers. On the other hand, for DNA within newborns that is located within an autosome or within an X chromosome that is inherited from a mother, REC_{tc} describes how the parental chromosome in (62)-(63) switches between grandpaternal and grandmaternal modes of inheritance, including both single recombination and gene conversion events. Section 2.2 describes in more detail how to generate REC_{tc}.

Thus, in order to define $p_{tq}(c)$ at all haplotype blocks hb_q we may first use Mendelian laws and specify $p_{t1}(c)$ for the leftmost block hb_1 according to any of (60)-(64), depending on whether (t,c) is inherited from a father or a mother. Then we use (65) to specify $p_{tq}(c)$ recursively at all other haplotype blocks hb_q according to

$$\begin{array}{rcll} p_{tq}(c) & = & p_{t,q-1}(c), & \text{if } \text{hbb}_{q-1}\notin\text{REC}_{tc}, \\ p_{tq}(c) & \neq & p_{t,q-1}(c), & \text{if } \text{hbb}_{q-1}\in\text{REC}_{tc}, \end{array} \tag{66}$$

for $q=2,\ldots,Q$. Once $p_{tq}(c)$ has been specified for all ancestral chromosomes (t,c) at all haplotype blocks hb_q, we have enough information to generate the ancestral recombination graph (17) and all ancestral haplotypes blocks (33) at time T_{t+1}. When this is repeated backwards in time we get the algorithm of Figure 13. See also Table 10 for an illustration of concepts, for the genealogy of Figures 4-7.

After this general overview of how to generate the ancestral recombination graph, it remains to describe in more detail how to build the pedigree (assigning parents and selfing numbers, and removing non-ancestral parents), how to define recombination events, and how the algorithm is modified when the haplotype blocks are not known in advance. These are the topics of Sections 2.1-2.3.

2.1 Building the Pedigree

We will generalize the backward simulation method for pedigrees in [99], allowing the population to have a time-varying size and different types (geographic or age) of substructure. This will be done in steps, starting with a homogeneous population.

2.1.1 Homogeneous Population

The nodes of the pedigree are the ancestral individuals

$$\text{AI}=\{(t,i);\,0\le t\le t_{\max}, i\in\text{AI}_t\}$$

at all time points, of which those in AI_t live at time t, cf. (56). For a population with non-overlapping generations without geographic substructure, each individual $(t,i)\in\text{AI}$ of a non-founder generation ($t<t_{\max}$) has two upward edges, to the father $m_t(i)$ and mother $f_t(i)$, cf. (57), and there is no need to determine in which deme (t,i) lives. In order to build the pedigree we proceed recursively and start with those individuals AI_0 that belong to the current generation ($t=0$). They are

$$\text{AI}_0=\begin{cases} \{1,\ldots,n_{mm},M_0+1,\ldots,M_0+n_{ff}\}, \\ \{1,\ldots,n_m\}, \\ \{1,\ldots,n_m,M_0+1,\ldots,M_0+n_{ff}\}, \\ \{M_0+1,\ldots,M_0+n_f\}, \end{cases} \tag{67}$$

for autosomal, Y-, X- and mitochondrial DNA respectively, with n_m, n_f, n_{mm} and n_{ff} defined as in (15). Given that AI_t has been specified for $0\le t<t_{\max}$, we will generate AI_{t+1} (which, in view of (56), is completely specified by m_{t+1} and f_{t+1}), the edges between AI_t and AI_{t+1}, and the number of children

$$\begin{array}{rcl} C_{t+1,f} & = & |\{i\in\text{AI}_t;\,f_t(i)=f\}|, \\ & & \text{for } f=1,\ldots,f_t, \\ C_{t+1,mf} & = & |\{i\in\text{AI}_t;\,m_t(i)=m,f_t(i)=f\}|, \\ & & \text{for } m=1,\ldots,m_t,f=1,\ldots,f_t, \end{array}$$

of all mothers and couples of generation $t+1$. All these quantities will be computed recursively, by selecting parents for all $i\in\text{AI}_t$. Suppose $i\in\text{AI}_t$ is preceded by $0\le j=j(i)<|\text{AI}_t|$ other individuals of AI_t, which have already been assigned parents. Then choose parents $(m_t(i),f_t(i))$ of (t,i) according to a Polya urn scheme [171], with probabilities

$$\begin{aligned} &\text{Prob}(m_t(i)=m,f_t(i)=f) \\ &=\begin{cases} \frac{\alpha+C_{t+1,f}}{F_{t+1}\alpha+j}\cdot\frac{\beta+C_{t+1,mf}}{M_{t+1}\beta+C_{t+1,f}}, & \text{(I)} \\ \frac{\alpha+C_{t+1,f}}{F_{t+1}\alpha+j}\cdot\frac{\beta(M_{t+1}-m_{t+1})}{M_{t+1}\beta+C_{t+1,f}}, & \text{(II)} \\ \frac{\alpha(F_{t+1}-f_{t+1})}{F_{t+1}\alpha+j}\cdot\frac{1}{M_{t+1}}, & \text{(III)} \\ \frac{\alpha(F_{t+1}-f_{t+1})}{F_{t+1}\alpha+j}\cdot\frac{M_{t+1}-m_{t+1}}{M_{t+1}}, & \text{(IV)} \end{cases} \end{aligned} \tag{68}$$

where $\sum_f C_{t+1,f}=j$ and $\sum_m C_{t+1,mf}=C_{t+1,f}$. The rows on the right hand side of (68) represent the four possible combinations

$$\begin{array}{rl} \text{(I)}: & m\le m_{t+1},f\le f_{t+1}, \\ \text{(II)}: & m=m_{t+1}+1,f\le f_{t+1}, \\ \text{(III)}: & m\le m_{t+1},f=f_{t+1}+1, \\ \text{(IV)}: & m=m_{t+1}+1,f=f_{t+1}+1, \end{array}$$

as to whether the father m and mother f of (t,i) have previously been assigned children or not. Figure 14 illustrates how (68) is used to add parents in generation $t+1$ to all members of the pedigree in generation t.

After these steps we have a set

$$\text{AIC}_{t+1}=\{1,\ldots,m_{t+1},M_{t+1}+1,\ldots,M_{t+1}+f_{t+1}\} \tag{69}$$

INITIALIZIZATION $t = 0$
  Define AI_0 as in (67)
  Assign deme numbers $d_0(i)$ to all $i \in \mathrm{AI}_0$ from (75)
  $\mathrm{AC}_0 = \{c_1, \ldots, c_n\}$
  FOR $k = 1$ TO n
    $\mathrm{AHB}_{0c_k} = \{1, \ldots, Q\}$
    FOR $q = 1$ TO Q
      $\mathrm{ARG}_{0kq} = c_k$
    END
  END
END
FOR $t = 0, \ldots, t_{\max} - 1$
  Generate parents $m_t(i), f_t(i)$ or selfing numbers $s_t(i)$ for all $i \in \mathrm{AI}_t$ (*)
  Define AI_{t+1} as all $s_t(i)$ and all parents in previous step that are ancestral (**)
  Record the deme number $d_{t+1}(i)$ of all $i \in \mathrm{AI}_{t+1}$
  $\mathrm{AC}_{t+1} = \emptyset$
  FOR all $c \in \mathrm{AC}_t$ DO
    Compute $p_{t1}(c)$ according to (60), (61), (62), (63) or (64)
    Generate recombination events REC_{tc} (***)
    Compute $\{p_{tq}(c)\}_{q=2}^{Q}$ according to (66)
    FOR all $c' \in \{p_{tq}(c); q \in \mathrm{AHB}_{tc}\}$ DO
      $\mathrm{AC}_{t+1} \leftarrow \mathrm{AC}_{t+1} \cup \{c'\}$
    END
  END
  FOR all $c \in \mathrm{AC}_{t+1}$ DO
    $\mathrm{AHB}_{t+1,c} = \emptyset$
  END
  FOR $k = 1$ TO n
    FOR $q = 1$ TO Q
      $c = p_{tq}(\mathrm{ARG}_{tkq})$
      $\mathrm{ARG}_{t+1,kq} = c$
      $\mathrm{AHB}_{t+1,c} \leftarrow \mathrm{AHB}_{t+1,c} \cup \{q\}$
    END
  END
END

Figure 13: Algorithm for generating the ancestral recombination graph (ARG) and ancestral haplotype blocks (AHB), when haplotype blocks are specified in advance. The initiation of ancestral chromosomes AC_0 will depend on DNA type, such as (55) for autosomes, and in the text below this formula for other types of DNA. The lines marked (*) and (**) are explained in more detail in Section 2.1, and the line marked (***) in Section 2.2. **doi:**10.5048/BIO-C.2016.4.f13

Table 10: Inference of recombination events and inheritance patterns from the ancestral recombination graph (ARG). For each non-founder chromosome (t, c) of the pedigree in the lower part of Figure 4, the ARG of Figure 5 is used to obtain partial information about its recombination events REC_{tc} and paternal chromosomes $p_{tq}(c)$ at all haplotype blocks hb_q. The last three columns specify which haplotype blocks of (t, c) that are ancestral ($\in \mathrm{AHB}_{tc}$), precisely those that are colored as white or light blue in Figure 7. We notice that $p_{tq}(c)$ is known for all ancestral haplotype blocks.

(t, c)	REC_{tc}	$p_{tq}(c)$			$q \in \mathrm{AHB}_{tc}$?		
		$q = 1$	$q = 2$	$q = 3$	$q = 1$	$q = 2$	$q = 3$
(0,1)	$\{1, 2\}$	3	4	3	Yes	Yes	Yes
(0,2)	$\{2\}$	7	7	8	Yes	Yes	Yes
(0,3)	$\emptyset$	7	7	7	Yes	Yes	Yes
(0,4)	$\emptyset$	8	8	8	Yes	Yes	Yes
(0,5)	$\{2\}$	1	1	2	Yes	Yes	Yes
(0,6)	$\emptyset$	6	6	6	Yes	Yes	Yes
(1,1)	$\{1\}$	1	2	?	Yes	Yes	No
	$\{1, 2\}$						
(1,2)	?	?	?	4	No	No	Yes
(1,3)	$\{1\}$	1	?	2	Yes	No	Yes
	$\{2\}$						
(1,4)	?	?	3	?	No	Yes	No
(1,5)	?	?	?	?	No	No	No
(1,6)	$\{2\}$	3	3	4	Yes	Yes	Yes
(1,7)	$\{2\}$	2	2	1	Yes	Yes	Yes
(1,8)	$\{1\}$	4	3	3	Yes	Yes	Yes

```
INITIATE
  j = 0
  m_{t+1} = f_{t+1} = 0
  C_{t+1,f} = C_{t+1,mf} = 0 for all m, f
END
FOR all i ∈ AI_t DO
  SELECT (m_t(i), f_t(i)) with probabilities as in (68)
  j ← j + 1
  m_{t+1} ← m_{t+1} + 1 for cases II and IV
  f_{t+1} ← f_{t+1} + 1 for cases III and IV
  C_{t+1,f_t(i)} ← C_{t+1,f_t(i)} + 1
  C_{t+1,m_t(i)f_t(i)} ← C_{t+1,m_t(i)f_t(i)} + 1
END
```

Figure 14: Algorithm for adding parents to a pedigree. The population is homogeneous with non-overlapping generations, and it is shown how all individuals in generation t are assigned parents. The algorithm corresponds to line (*) of Figure 13. **doi:**10.5048/BIO-C.2016.4.f14

```
INITIALIZE
  Define m_t and f_t from the algorithm of Figure 14
  Define the ancestral individual candidates AIC_{t+1} according
    to (69)
  Initiate ancestral function: A_{t+1}(i) = 0 for all i ∈ AIC_{t+1}
END
FOR all i ∈ AI_t DO
  IF 2i − 1 ∈ AC_t
    A_{t+1}(m_t(i)) = 1
  END
  IF 2i ∈ AC_t
    A_{t+1}(M_t + f_t(i)) = 1
  END
END
Compute m'_{t+1} and f'_{t+1} from (70)
Compute permutations τ_m and τ_f as in (71)
Update all m_t(i) and f_t(i) as in (72)
Update m_t and f_t as in (72)
Compute AI_t from (56)
```

Figure 15: Pruning of non-ancestral individuals. The algorithm above removes the non-ancestral parents of individuals from generation t. It corresponds to line (**) of Figure 13. **doi:**10.5048/BIO-C.2016.4.f15

of Ancestral Individual Candidates of generation $t+1$. We still have to prune the pedigree by removing those parents that are not ancestral. To this end, we define the ancestral function

$$A_{t+1}(i) = 1(i \in \mathrm{AI}_{t+1}),$$

for all candidate ancestral parents $i \in \mathrm{AIC}_{t+1}$. We compute this function recursively by checking for all children whether or not the chromosomes they inherited from their two parents are ancestral (see Figure 15). Once the ancestry function A_{t+1} has been computed, it enables us to find the number of males and females

$$\begin{array}{rcl} m'_{t+1} & = & |\{i;\, 1 \le i \le m_{t+1}, A_{t+1}(i) = 1\}|, \\ f'_{t+1} & = & |\{i;\, 1 \le i \le f_{t+1}, A_{t+1}(i) = 1\}|, \end{array} \tag{70}$$

of generation $t+1$ that are ancestral. The next step is to update the number $m_t(i)$ of the father and $f_t(i)$ of the mother for all ancestral individuals of generation t, by introducing two permutations τ_m and τ_f of $\{1, \ldots, m_{t+1}\}$ and $\{1, \ldots, f_{t+1}\}$ that move the ancestral males and females first within each group. In order achieve this, the permutations should satisfy

$$\begin{array}{rcl} \tau_m(i) & = & |\{i';\, 1 \le i' \le i, A_{t+1}(i') = 1\}|, \\ & & \text{if } A_{t+1}(i) = 1, \\ \tau_m(i) & \in & \{m'_{t+1} + 1, \ldots, m_{t+1}\}, \\ & & \text{if } A_{t+1}(i) = 0, \\ \tau_f(i) & = & |\{i';\, 1 \le i' \le i, A_{t+1}(M_t + i') = 1\}|, \\ & & \text{if } A_{t+1}(M_t + i) = 1, \\ \tau_f(i) & \in & \{m'_{t+1} + 1, \ldots, m_{t+1}\}, \\ & & \text{if } A_{t+1}(M_t + i) = 0. \end{array} \tag{71}$$

Finally, we update all relevant variables as

$$\begin{array}{rcll} m_t(i) & \leftarrow & \tau_m(m_t(i)), & i \in \mathrm{AI}_t, \\ f_t(i) & \leftarrow & \tau_f(f_t(i)), & i \in \mathrm{AI}_t, \\ m_{t+1} & \leftarrow & m'_{t+1}, & \\ f_{t+1} & \leftarrow & f'_{t+1}, & \end{array} \tag{72}$$

and define AI_{t+1} as in (56). A summary of the pruning algorithm can be found in Figure 15.

2.1.2 Population with Geographic Substructure

In order to extend the algorithm of the previous subsection to a population with geographic substructure, we assume it divides into a number of islands or demes. A similar model was proposed in [99], but here we allow the number of demes and the number of males and females within each deme to vary with time.

It is assumed that generations are non-overlapping, with D_t the number of demes of generation t. Recall that the number of males and females M_{td} and F_{td} of all demes d satisfy (2). We will use the backward migration probabilities

$$\begin{array}{l} \{B^m_{tde};\, 1 \le d \le D_t, 1 \le e \le D_{t+1}\}, \\ \{B^f_{tde};\, 1 \le d \le D_t, 1 \le e \le D_{t+1}\}, \end{array} \tag{73}$$

that an individual of deme d in generation t has its male or female parent from deme e in generation $t+1$ (see Figure 6). The overall backward migration rate

$$B_{tde} = \frac{1}{2}(B^m_{tde} + B^f_{tde}) \tag{74}$$

from deme e to deme d, is the probability that a randomly chosen parent of a child in d comes from e.

With geographic substructure, we must not only record the two parents $(m_t(i), f_t(i))$, but also the deme $d_t(i)$ in which a non-founder (t, i) lives. We will build the pedigree backwards in time $t = 0, 1, \ldots, t_{\max} - 1$, as in Subsection 2.1.1. Starting with $t = 0$, we first define the pedigree members AI_0 of generation 0 as in (67). Then we assign deme number

$$d_0(i) = d \tag{75}$$

for $i = \sum_{e=1}^{d-1}(n_{me} + n_{mme}) + 1, \ldots, \sum_{e=1}^{d}(n_{me} + n_{mme})$ and $i = M_0 + \sum_{e=1}^{d-1}(n_{fe} + n_{ffe}) + 1, \ldots, M_0 + \sum_{e=1}^{d}(n_{fe} + n_{ffe})$. Here. some of the numbers n_{md}, n_{mmd}, n_{fd} and n_{ffd} are put to zero, depending on type of DNA, as explain above and below (55).

Next we need to assign parents from AI_{t+1} to all individuals in AI_t, recursively for $t = 0, \ldots, t_{\max} - 1$. To this end, we will divide each recursive step into two parts. In the first part, assuming that deme membership $d_t(i)$ has already been defined for all individuals of AI_t, we also assign them paternal and maternal demes, with probabilities

$$\begin{array}{rcl} B_{td,eg} & = & \mathrm{Prob}(d_{t+1}(m_t(i)) = e, \\ & & \quad d_{t+1}(M_t + f_t(i)) = g | d_t(i) = d). \end{array} \tag{76}$$

The probabilities in (76) define a mating rule [172], which has to be consistent with the backward migration probabilities in (73). By this we mean that

$$\begin{array}{rcl} \sum_g B_{td,eg} & = & B^m_{tde}, \\ \sum_e B_{td,eg} & = & B^f_{tdg}, \end{array} \tag{77}$$

when summing over $g = 1, \ldots, D_{t+1}$ for the possible demes of the female spouse or likewise for the possible demes e of the male spouse. Because of (77), we only need to specify the mating probabilities $B_{td,eg}$ as input parameters, since the paternal and maternal

backward probabilities are functions of them. Two possible mating schemes are

$$B_{td,eg} = \begin{cases} B^m_{tde} B^f_{tdg}, \\ B_{tde} 1(\{e = g\}), \end{cases} \quad (78)$$

where in the first parents meet independently of their geographic origin, and in the second they come from the same deme. This second mating rule can only be achieved when the backward probabilities for fathers and mothers are the same ($B^m_{tde} = B^f_{tde} = B_{tde}$), see Figure 6.

For the second part of the recursive step, we generalize the parental assignment algorithm of Subsection 2.1.1 by keeping track of how many parents have been selected so far within each deme, and how many children they have. In more detail, suppose we are to assign parents to $i \in \mathrm{AI}_t$, and that

$$j = \sum_{g=1}^{D_{t+1}} j_g$$

parents have already been assigned to other members of AI_t, of which j_g mothers live in deme g. Let $m_t = \sum_e m_{t+1,e}$ and $f_t = \sum_e f_{t+1,e}$ be the total number of fathers and mothers that have these j children, of which $m_{t+1,e}$ and $f_{t+1,e}$ belong to deme e. As before, $C_{t+1,f}$ and $C_{t+1,mf}$ refer to the number of children that have been assigned so far to mother f and couple (m, f) respectively. This implies in particular that

$$j = \sum_{f=1}^{f_t} C_{t+1,f},$$

$$j_g = \sum_{\substack{f;1\le f\le f_t \\ d_{t+1}(M_t+f)=g}} C_{t+1,f}.$$

We also introduce

$$C^e_{t+1,f} = \sum_{\substack{m;1\le m\le m_t \\ d_{t+1}(m)=e}} C_{t+1,mf}$$

as the number of children mother f had so far with a spouse from deme e. With these definitions, we can generalize the parental selection scheme (68) to a population with geographic substructure, as

$$\begin{aligned} &\mathrm{Prob}(m_t(i) = m, f_t(i) = f| \\ &\quad d_{t+1}(m_t(i)) = e, d_{t+1}(f_t(i)) = g) \\ &= \begin{cases} \frac{\alpha+C_{t+1,f}}{F_{t+1,g}\alpha+j_g} \cdot \frac{\beta+C_{t+1,mf}}{M_{t+1,e}\beta+C^e_{t+1,f}}, & \text{(I)} \\ \frac{\alpha+C_{t+1,f}}{F_{t+1,g}\alpha+j_g} \cdot \frac{\beta(M_{t+1,e}-m_{t+1,e})}{M_{t+1,e}\beta+C^e_{t+1,f}}, & \text{(II)} \\ \frac{\alpha(F_{t+1,g}-f_{t+1,g})}{F_{t+1,g}\alpha+j_g} \cdot \frac{1}{M_{t+1,e}}, & \text{(III)} \\ \frac{\alpha(F_{t+1,g}-f_{t+1,g})}{F_{t+1,g}\alpha+j_g} \cdot \frac{M_{t+1,e}-m_{t+1,e}}{M_{t+1,e}}, & \text{(IV)} \\ 0, & \text{(V)} \end{cases} \end{aligned}$$

where the five numbered rows on the right hand side represent different choices

$$\begin{aligned} \text{(I)}: &\quad m \le m_{t+1}, f \le f_{t+1}, d_{t+1}(m) = e, d_{t+1}(f) = g, \\ \text{(II)}: &\quad m = m_{t+1} + 1, f \le f_{t+1}, d_{t+1}(f) = g, \\ \text{(III)}: &\quad m \le m_{t+1}, f = f_{t+1} + 1, d_{t+1}(m) = e, \\ \text{(IV)}: &\quad m = m_{t+1} + 1, f = f_{t+1} + 1, \\ \text{(V)}: &\quad \text{otherwise,} \end{aligned}$$

```
INITIATE
  j = 0, j_g = 0 for g = 1, ..., D_{t+1}
  m_{t+1} = f_{t+1} = 0, m_{t+1,e} = f_{t+1,e} = 0 for e = 1, ..., D_{t+1}
  C_{t+1,f} = C^e_{t+1,f} = C_{t+1,mf} = 0 for all m, f, e
END
FOR i ∈ AI_t DO
  Assign demes e and g to the father and mother of i,
    according to (76)
  Select (m_t(i), f_t(i)) with probabilities as in the above
    displayed formula
  j ← j + 1, j_g ← j_g + 1
  IF cases II or IV
    m_{t+1} ← m_{t+1} + 1, m_{t+1,e} ← m_{t+1,e} + 1
    d_{t+1}(m_{t+1}) = e
  END
  IF cases III or IV
    f_{t+1} ← f_{t+1} + 1, f_{t+1,g} ← f_{t+1,g} + 1
    d_{t+1}(M_{t+1} + f_{t+1}) = g
  END
  C_{t+1,f_t(i)} ← C_{t+1,f_t(i)} + 1
  C^e_{t+1,f_t(i)} ← C^e_{t+1,f_t(i)} + 1
  C_{t+1,m_t(i)f_t(i)} ← C_{t+1,m_t(i)f_t(i)} + 1
END
```

Figure 16: Algorithm for adding parents to a pedigree. The population has geographic substructure and non-overlapping generations. It corresponds to line (*) of Figure 13, where all individuals in generation t are assigned parents.
doi:10.5048/BIO-C.2016.4.f16

of father m, mother f and their deme numbers. The algorithm to find the parents of all individuals in the pedigree in generation t is summarized in Figure 16. After this a final pruning of the pedigree takes place, where all non-ancestral parents of generation $t+1$ are removed. This is done in the same way as in Figure 15.

2.1.3 Population with Age Structure

For a geographically homogeneous population with age structure, we think more generally of t as a time index rather than a generation number, with D_t the number of age classes at time T_t. The age classes at this time point range from newborns $d = 1$ up to the oldest age class $d = D_t$. If the age difference between two adjacent classes is constant, and equal to the difference $T_{t+1} - T_t$ between two consecutive time points, it follows that an individual who survives to the next time point always moves up one age class. If we look backwards in time, an adult at time T_t with $t < t_{\max}$ belonged to the population at time T_{t+1} as well, but in the nearest lower age class.

Let $d_t(i)$ be the age class to which a non-founder (t, i) belongs. Age structured populations differ from geographically structured ones in that newborns ($d_t(i) = 1$) and adults ($d_t(i) > 1$) are treated differently. Newborns are handled in the same way as in Section 2.1.2 by specifying a mating rule

$$B_{t1,eg} = \mathrm{Prob}(d_{t+1}(m_t(i)) = e, d_{t+1}(f_t(i)) = g | d_t(i) = 1) \quad (79)$$

for the two age classes e and g of the father and mother. In analogy with (77), we let

$$\begin{aligned} B^m_{t1e} &= \textstyle\sum_g B_{t1,eg}, \\ B^f_{t1g} &= \textstyle\sum_e B_{t1,eg} \end{aligned} \quad (80)$$

refer to the age distribution of the father and mother by summing over all possible age classes $g, e = 1, \ldots, D_{t+1}$ of the spouse, that is, the mother and father respectively. Assume for instance an age span $T_{t+1} - T_t$ between two consecutive time points of ten years, and that age classes correspond to $0, 10, 20, \ldots, 100$ years. If the age range fertility is 20-70 years for fathers ($B^m_{t1e} > 0$ for $e = 2, 3, 4, 5, 6, 7$) and 20-40 years for females ($B^f_{t1e} > 0$ for $e = 2, 3, 4$), and couples mate independently of age, we get a mating rule as in the upper part of (78). Since the fertility ranges

```
INITIALIZE
  m_{t+1} = f_{t+1} = 0, m_{t+1,e} = f_{t+1,e} = 0 for e = 1, ..., D_{t+1}
END
FOR i ∈ AI_t DO
  d = d_t(i)
  IF d > 1 DO
    IF i ≤ M_t DO
      m_{t+1} ← m_{t+1} + 1
      m_{t+1,d-1} ← m_{t+1,d-1} + 1
      d_{t+1}(m_{t+1}) = d − 1
      s_t(i) = m_{t+1}
    ELSEIF i > M_t DO
      f_{t+1} ← f_{t+1} + 1
      f_{t+1,d-1} ← f_{t+1,d-1} + 1
      d_{t+1}(M_{t+1} + f_{t+1}) = d − 1
      s_t(i) = M_{t+1} + f_{t+1}
    END
  END
END
```

Figure 17: Part of the pedigree building algorithm for a homogeneous population with overlapping generations. It corresponds to the second step of line (*) in Figure 13, where adults of time step t assign themselves as parents at time step $t+1$. **doi:**10.5048/BIO-C.2016.4.f17

of males and females are different, it is not possible to require that parents have the same age, as in the lower part of (78). One may define a mating rule though where couples are as close in age as possible, given the different male and female fertility ranges.

Adults (t,i), on the other hand, have only one parent (itself) the previous time point if $t < t_{\max}$. Let $s_t(i) \in \mathrm{AI}_{t+1}$ be the self number (59) that (t,i) had the previous time point. The backward migration probabilities are easily assigned to $s_t(i)$, since adults move down one age class in reversed time, i.e.

$$\begin{aligned} B_{tde} &= \mathrm{Prob}(d_{t+1}(s_t(i)) = e | d_t(i) = d > 1) \\ &= 1(\{e = d-1\}). \end{aligned} \tag{81}$$

We will build AI_{t+1} from AI_t in three steps. In the first step, all adults in AI_t will have themselves assigned as single parents, according to (81). To this end, we let $m_{t+1} = \sum_{e=1}^{D_{t+1}} m_{t+1,e}$ refer to the total number of males in AI_{t+1} that have so far either been assigned at least one child, or appear in AI_t as an adult. Among these males, $m_{t+1,e}$ belong to age class e. The corresponding variables for females are f_{t+1} and $f_{t+1,e}$. See Figure 17 for a summary of the first step. After its completion we have

$$\begin{aligned} m_{t+1,e} &= |\{i \in \mathrm{AI}_t;\, i \le M_t, d_t(i) = e+1\}|, \\ f_{t+1,e} &= |\{i \in \mathrm{AI}_t;\, i > M_t, d_t(i) = e+1\}|, \end{aligned}$$

for $e = 1, \ldots, D_{t+1}$.

In the second step, we assign parents to all newborns of AI_t. To this end, we use the algorithm of Section 2.1.2 in order to choose mother and father of all newborns in AI_t. This requires not only that we keep track of m_{t+1}, $m_{t+1,e}$, f_{t+1} and $f_{t+1,e}$, but also of j, the number of other newborns that have been assigned parents so far, of which j_g had a mother in age class g. We also need to know $C_{t+1,f}$, the number of newborn children of AI_t that have so far been assigned to female parent $f \in \mathrm{AI}_{t+1}$, and how many $(C^e_{t+1,f})$ of these children that had a male parent from age class e. Finally, $C_{t+1,mf}$ is the number of children that have so far been assigned to couple (m,f). Assignment of parents $m_t(i)$ and $f_t(i)$ to all newborns $i \in \mathrm{AI}_t$ starts with initial conditions

$$\begin{aligned} j &= 0, \\ j_g &= 0, \\ C_{t+1,f} = C^e_{t+1,f} = C_{t+1,mf} &= 0, \quad \text{for all } m, f, e. \end{aligned}$$

Together with the values of m_{t+1}, $m_{t+1,e}$, f_{t+1} and $f_{t+1,e}$ from the first step, they are used as input parameters for the algorithm of Figure 16, with the same updating rules for all variables.

In the third step, we prune the pedigree by removing those parents from generation $t+1$ that are not ancestral, as described in Figure15.

2.1.4 Combined Geographic and Age Structure

The theory of the last two sections can be combined, assuming that the population at each time point T_t has geographic *and* age substructure. Then each subpopulation $d = 1, \ldots, D_t$ represents a combined deme and age class. It is convenient to decompose all subpopulations

$$\{1, \ldots, D_t\} = \mathrm{NS}_t \cup \mathrm{AS}_t$$

at time T_t into newborn and adult subpopulations. Suppose $(t,i) \in$ AI lives at a time point T_t with $t < t_{\max}$. For a newborn $(d_t(i) \in \mathrm{NS}_t)$, we need to assign his or her male and female parents $m_t(i)$ and $f_t(i)$ $(\in \mathrm{AI}_{t+1})$, with probabilities specified by a mating rule

$$B_{td,eg} = \mathrm{Prob}(d_{t+1}(m_t(i)) = e, d_{t+1}(f_t(i)) = g | d_t(i) = d). \tag{82}$$

For an adult $(d_t(i) \in \mathrm{AS}_t)$, we need to specify from which subpopulation he or she migrated, with probabilities

$$B_{tde} = \mathrm{Prob}(d_{t+1}(s_t(i)) = e | d_t(i) = d), \tag{83}$$

where $s_t(i) \in \mathrm{AI}_{t+1}$ is his or her self index from the previous time point.

The recursive step of the pedigree algorithm, where AI_{t+1} is built from AI_t, is basically the same as in Section 2.1.3. In a first step we assign from which subpopulation all adults have migrated, according to (83). The major difference compared to Section 2.1.3 is that the subpopulation of time T_{t+1} can be chosen in more than one way. Then, in a second step, all newborns choose father and mother according to (82), and finally the non-ancestral parents at time T_{t+1} are removed.

2.2 Generating Recombination Events

The random collection of recombination events REC_{tc} in (65) for autosomal DNA, or for X chromosome DNA inherited from mothers, will be generated independently for each ancestral chromosome (t,c) of a non-founder time point. We write it as a disjoint union

$$\mathrm{REC}_{tc} = \left(\mathrm{REC}^{\mathrm{ord}}_{tc} \cup \mathrm{REC}^{\mathrm{gc},l}_{tc} \cup \mathrm{REC}^{\mathrm{gc},r}_{tc}\right) \cap \{1, \ldots, Q-1\}$$

of ordinary (or reciprocal) recombinations events, each one of which represents an odd number of crossovers, and those single crossover events generated by gene conversion, one to the left and one to the right of the tract. Since one of the two crossovers of a gene conversion may fall outside the studied chromosomal region, the intersection with $\{1, \ldots, Q-1\}$ assures that REC_{tc} is delimited to this region. Whenever gene conversion is included, it is assumed that all neighboring haplotype blocks hb_q and hb_{q+1} are neighboring chromosomal regions, with no gap in between, so that each recombination events represents one single crossover.

In the following two subsections we describe ordinary recombinations and gene conversions separately.

2.2.1 Ordinary Recombinations

In order to generate ordinary recombination events, we assume no chiasma interference, or Haldane's map function, see for instance [104,105]. It implies that crossovers occur independently within REC_{tc} at all haplotype block boundaries hbb_q, $q = 1, \ldots, Q-1$, with probabilities

$$\mathrm{Prob}(\mathrm{hbb}_q \in \mathrm{REC}^{\mathrm{ord}}_{tc}) = \begin{cases} r_{mq}, & \text{if } c \text{ is odd}, \\ r_{fq}, & \text{if } c \text{ is even}, \end{cases} \tag{84}$$

as in (19), since an odd (even) c corresponds to a chromosome inherited from a father (mother).

If Q is large and all recombination probabilities are small, it is computationally more efficient to use an approximation of (84) where first the total number $|\mathrm{REC}_{tc}^{\mathrm{ord}}|$ of ordinary recombination events are generated, and then these are randomly distributed to haplotype block boundaries. Let

$$\begin{array}{rcll} r_m & = & \sum_{q=1}^{Q-1} r_{mq}, & \text{if } c \text{ is odd,} \\ r_f & = & \sum_{q=1}^{Q-1} r_{fq}, & \text{if } c \text{ is even,} \end{array} \tag{85}$$

be the expected number of ordinary recombination events along chromosome c when a sperm or ovum cell is formed. We draw the number of recombination events from a Poisson distribution with mean (19), i.e.

$$\mathrm{Prob}(|\mathrm{REC}_{tc}^{\mathrm{ord}}| = j) = \begin{cases} e^{-r_m} r_m^j/j!, & \text{if } c \text{ is odd,} \\ e^{-r_f} r_f^j/j!, & \text{if } c \text{ is even,} \end{cases}$$

for $j = 0, 1, 2, \ldots$. These $|\mathrm{REC}_{tc}^{\mathrm{ord}}|$ events are then allocated independently between haplotype block boundaries hbb_q with probabilities r_{mq}/r_m and r_{fq}/r_f for male and female germ cells respectively. Any crossover is discarded if it appears at a haplotype boundary where another crossover has already been allocated.

2.2.2 Gene Conversion

By gene conversion we mean a process during meiosis where two homologous strands intersect, and the resulting Holliday junction is resolved by two nearby crossing over events. Only a short tract of DNA is transferred between the two strands, whereas the flanking regions are not. We use the mathematical model in [173,174] to generate gene conversion, or double crossover, events. It is mainly of interest when the simulated region comprises a small segment of a chromosome.

Each randomly generated gene conversion can be represented in terms of its left and right haplotype block boundaries $\{\mathrm{hbb}_{Q_1}, \mathrm{hbb}_{Q_2}\}$, and tract length $\mathrm{TL} = Q_2 - Q_1 > 0$, counted in units of haplotype blocks. If both end points are visible within the chromosomal segment we have $1 \le Q_1 < Q_2 \le Q - 1$, if only the left end point is visible we have $1 \le Q_1 \le Q - 1 < Q_2$, and finally, if only the right end point is visible we have $Q_1 < 1 \le Q_2 \le Q - 1$. Let $\mathrm{TL}_{\max}$ denote the maximal possible value of TL. In a first step, we generate left end points of gene conversion tracts independently for all haplotype boundaries hbb_q, $q = 1 - \mathrm{TL}_{\max}, \ldots, Q - 1$, with probabilities

$$\mathrm{Prob}(\mathrm{hbb}_q \in \mathrm{REC}_{tc}^{\mathrm{gc},l}) = \begin{cases} \gamma_{mq}, & \text{if } c \text{ is odd,} \\ \gamma_{fq}, & \text{if } c \text{ is even,} \end{cases} \tag{86}$$

as in (21). Typically, we will not distinguish these probabilities for meioses within males and females, and write $\gamma_{mq} = \gamma_{fq} = \gamma_q$. When all left end points have been defined, we write the right end points as

$$\mathrm{REC}_{tc}^{\mathrm{gc},r} = \bigcup_{q;\mathrm{hbb}_q \in \mathrm{REC}_{tc}^{\mathrm{gc},l}} \mathrm{hbb}_{q+\mathrm{TL}_q}, \tag{87}$$

where $1 \le \mathrm{TL}_q \le \mathrm{TL}_{\max}$ are independent random variables having a pre-specified tract length distribution TLD. This distribution may differ or be the same for males and females meioses.

There is a small probability that a ordinary single crossover event overlaps with a gene conversion tract, or that two gene conversion tracts overlap. In the first case we disregard the single crossover, and in the second case we remove the rightmost tract.

2.3 Recursive Computation of Haplotype Blocks

Instead of pre-specifying haplotype blocks, they can be computed as part of the algorithm by allowing recombination events to happen anywhere among the boundaries $\{1, \ldots, L-1\}$ between adjacent loci. It is then convenient to define haplotype boundaries as a subset these $L-1$ locus boundaries, rather than $\{1, \ldots, Q-1\}$. Similarly, for any ancestral chromosome $(t, c) \in \mathrm{AC}$ of a non-founder time point, we define the recombination events

$$\mathrm{REC}_{tc} = \left(\mathrm{REC}^{\mathrm{ord}} \cup \mathrm{REC}^{\mathrm{gc},l} \cup \mathrm{REC}^{\mathrm{gc},r}\right) \cap \{1, \ldots, L-1\}$$

as subset of the locus boundaries rather than the $Q-1$ haplotype boundaries. We can still use the model of Section 2.2, with ordinary recombination events and gene conversion to the left or right of the tract, provided we replace haplotype boundaries by locus boundaries. These are denoted lb_l between loci l and $l+1$, for $l = 1, \ldots, L-1$. For ordinary recombination events, we modify (84) to

$$\mathrm{Prob}(\mathrm{lb}_l \in \mathrm{REC}_{tc}^{\mathrm{ord}}) = \begin{cases} r_{ml}, & \text{if } c \text{ is odd,} \\ r_{fl}, & \text{if } c \text{ is even,} \end{cases} \tag{88}$$

where $(r_{ml})_{l=1}^{L-1}$ and $(r_{fl})_{l=1}^{L-1}$ contain recombination probabilities at all locus boundaries for meioses within males and females. For gene conversion, we similarly we replace (86) and (87) by

$$\mathrm{Prob}(\mathrm{lb}_l \in \mathrm{REC}_{tc}^{\mathrm{gc},l}) = \begin{cases} \gamma_{ml}, & \text{if } c \text{ is odd,} \\ \gamma_{fl}, & \text{if } c \text{ is even,} \end{cases} \tag{89}$$

and

$$\mathrm{REC}_{tc}^{\mathrm{gc},r} = \bigcup_{l;\mathrm{lb}_l \in \mathrm{REC}_{tc}^{\mathrm{gc},l}} \mathrm{lb}_{l+\mathrm{TL}_l}, \tag{90}$$

where $(\gamma_{ml})_{l=1}^{L-1}$ and $(\gamma_{fl})_{l=1}^{L-1}$ contain recombination probabilities for the left part of the tract at all locus boundaries within males and females, and TL_l is the length (in units of base pairs) for a tract starting at locus l.

The set of haplotype block boundaries consists of those locus boundaries where recombinations occurred, either in ancestral or trapped material, see for instance Section 5.5 in [175]. The trapped material of an ancestral chromosome $(t, c) \in \mathrm{AC}$ is the set of non-ancestral loci that are surrounded by ancestral regions along the chromosome, to the right *and* left. It is important to account for crossovers also in trapped material, as it will affect the ancestral recombination graph. Since the trapped material 'fills up' gaps between ancestral regions, it follows that the set of ancestral *or* trapped loci

$$\begin{array}{rcl} \mathrm{ATL}_{tc} & = & \{l \in \mathcal{L};\ l \text{ is ancestral or trapped in } (t, c)\} \\ & = & \{l_{tc}^1, l_{tc}^1 + 1, \ldots, l_{tc}^2\} \end{array}$$

of an ancestral chromosome $(t, c) \in \mathrm{AC}$ is a discrete interval, with both end point loci l_{tc}^1 and l_{tc}^2 containing ancestral material. The haplotype boundaries are then defined as

$$\mathrm{hbb} = \cup_{(t,c) \in \mathrm{AC}} \left[\mathrm{REC}_{tc} \cap \mathrm{lb}(\mathrm{ATL}_{tc})\right], \tag{91}$$

where $\mathrm{lb}(\mathrm{ATL}_{tc})$ consists of all locus boundaries in ATL_{tc}.

We use the notation $\mathrm{cl}(\mathcal{L}')$ for the closure of any subset $\mathcal{L}' \subset \mathcal{L}$ of loci. Figure 18 describes how the haplotype block boundaries (91) are computed recursively backwards in time, together with the genealogy.

3. GENERATING MUTATIONS AND GENE DROPPING

Once the genealogy is defined, the alleles of the reference chromosome and those that get mutated in other chromosomes, are spread forwards in time to all ancestral DNA. This will be done differently depending on whether mutation probabilities depend on the actual alleles that get mutated or not.

INITIALIZE
  $Q = 1$
  $\text{hbb} = \emptyset$
  Define AI_0 and AC_0
  FOR $k = 1, \ldots, n$ DO
    $\text{ATL}_{0,c_k} = \mathcal{L}$
  END
END
FOR $t = 0, \ldots, t_{\max} - 1$
  Generate parents $m_t(i)$, $f_t(i)$ or selfing numbers $s_t(i)$ for all $i \in \text{AI}_t$
  Define AI_{t+1} as all $s_t(i)$ and all ancestral parents in previous step
  $\text{hbb}_{\text{old}} = \text{hbb}$
  FOR all $c \in \text{AC}_t$ DO
    Compute $p_{t1}(c)$ according to (60), (61), (62), (63) or (64)
    Generate REC_{tc}
    $\text{hbb}_{tc} = \text{REC}_{tc} \cap \text{lb}(\text{ATL}_{tc})$
    $\text{newhbb}_{tc} = [\text{hbb} \cup \text{hbb}_{tc}] \setminus \text{hbb}$
    $Q \leftarrow Q + |\text{newhbb}_{tc}|$
    $\text{hbb} \leftarrow \text{hbb} \cup \text{newhbb}_{tc}$
  END
  IF $t = 0$
    FOR $k = 1, \ldots, n$ DO
      $\text{AHB}_{0c_k} = \{1, \ldots, Q\}$
      FOR $q = 1$ TO Q
        $\text{ARG}_{0kq} = c_k$
      END
    END
  ELSE
    $\text{newhbb}_t = \text{hbb} \setminus \text{hbb}_{\text{old}}$
    Use newhbb_t to update all $\text{ARG}_{t'kq}$ and $\text{AHB}_{t'c}$ with $t' \leq t$
  END
  $\text{AC}_{t+1} = \emptyset$
  FOR all $c \in \text{AC}_t$ DO
    Compute $\{p_{tq}(c)\}_{q=2}^{Q}$ as in (66)
    FOR all $c' \in \{p_{tq}(c); q \in \text{AHB}_{tc}\}$ DO
      $\text{AC}_{t+1} \leftarrow \text{AC}_{t+1} \cup \{c'\}$
      $\text{ATL}_{t+1,c'} \leftarrow \text{cl}\left[\text{ATL}_{t+1,c'} \cup (\text{ATL}_{tc} \cap \{\text{hb}_q; q \in \text{AHB}_{tc}, p_{tq}(c) = c'\})\right]$
    END
  END
  FOR all $c \in \text{AC}_{t+1}$ DO
    $\text{AHB}_{t+1,c} = \emptyset$
  END
  FOR $k = 1$ TO n
    FOR $q = 1$ TO Q
      $c = p_{tq}(\text{ARG}_{tkq})$
      $\text{ARG}_{t+1,kq} = c$
      $\text{AHB}_{t+1,c} \leftarrow \text{AHB}_{t+1,c} \cup \{q\}$
    END
  END
END

Figure 18: Algorithm for generating the ancestral recombination graph (ARG) and ancestral haplotype blocks (AHB), when haplotype blocks are not specified in advance. **doi:**10.5048/BIO-C.2016.4.f18

3.1 Mutation probabilities independent of nucleotides

We will mainly consider the case when the founder diversity probabilities ν_l and germline mutation probabilities μ_l can be written as in (26) and (28), regardless of which alleles that mutate. This is much simpler, since it is possible to first generate all founder/germline mutations (without knowing what kind of nucleotides get mutated), and then spread the alleles of the reference chromosome, as well as mutations, through gene dropping.

3.1.1 Generating ancestral mutations

Define the subset

$$\mathrm{AME}_{tc} = \cup_{q\in \mathrm{AHB}_{tc}} \mathrm{AME}_{tcq} \subset \mathcal{L} \tag{92}$$

of loci at which mutations occur for the ancestral part of chromosome (t, c), cf. (32). We only include ancestral haplotype blocks in (92), and disregard all silent mutations of (t, c).

When (t, c) belongs to a non-founder time point ($t < t_{\max}$), we will use the mutation probabilities in (28). Let $\mu = \sum_{l=1}^{L} \mu_l$ be the expected total number of mutations along (t, c). We assume that the total number nmut_{tc} of mutations along (t, c) has a Poisson distribution with expected value μ, i.e.

$$\mathrm{Prob}(\mathrm{nmut}_{tc} = j) = e^{-\mu}\frac{\mu^j}{j!}, \quad j = 0, 1, 2, \ldots. \tag{93}$$

Let

$$P_l = \frac{\mu_l}{\mu} \tag{94}$$

be the probability that a specific mutation occurs at locus l for all $l \in \mathcal{L}$. Given that $\mathrm{nmut}_{tc} = j$, we generate mutational loci independently with probabilities (94), until j different loci have been obtained. (There is a very small probability some locus may be obtained more than once. Then only its first occurrence is retained.) Finally we only keep the ancestral mutational loci l. That is, we only keep $l \in \mathrm{hb}_q$ if $q \in \mathrm{AHB}_{tc}$.

For the founder generation, we need to generate ancestral mutational events for its non-reference copies of the chromosome region of interest. Depending on type of DNA, these are

$$(t, c) \in \begin{array}{ll} \{(t_{\max}, 2), (t_{\max}, 3), (t_{\max}, 4)\}, & \text{aut-DNA,} \\ \{(t_{\max}, 3), (t_{\max}, 4)\}, & X\text{-DNA,} \\ \emptyset, & Y\text{- or mtDNA.} \end{array} \tag{95}$$

For each such chromosome, AME_{tc} is obtained similarly as for non-founder chromosomes, replacing the mutational probabilities μ_l in (28) by the founder diversity probabilities ν_l in (26). That is, we first generate the total number of mutations in (t, c) from

$$\mathrm{Prob}(\mathrm{nmut}_{tc} = j) = e^{-\nu}\frac{\nu^j}{j!}, \quad j = 0, 1, 2, \ldots, \tag{96}$$

where $\nu = \sum_{l=1}^{L} \nu_l$, then select mutated loci l with probabilities

$$P_l = \frac{\nu_l}{\nu}, \tag{97}$$

and keep those that are ancestral. Figure 19 summarizes how ancestral mutations are generated.

3.1.2 Building mutational trees and gene dropping

Once all ancestral mutational events are defined, we generate the alleles $a_l^{\mathrm{ref}} \in \mathcal{A}(l)$ of the reference haplotype $\boldsymbol{h}^{\mathrm{ref}}$ in (22) independently at all loci l, with probabilities π_a as in (23), (24) or (25). As mentioned in Section 1.5, this reference haplotype belongs to chromosome $(t_{\max}, c^{\mathrm{ref}})$, where

$$c^{\mathrm{ref}} = \begin{cases} 1, & \text{for aut- or Y-DNA,} \\ 2, & \text{for X-DNA,} \\ 4, & \text{for mtDNA.} \end{cases} \tag{98}$$

```
FOR all (t, c) ∈ AC DO
  AME_tc = ∅
  IF t < t_max
    Generate nmut_tc from distribution (93)
    FOR i = 1 TO nmut_tc DO
      Generate mutated locus l from distribution (94)
      Define q by l ∈ hb_q
      IF q ∈ AHB_tc
        AME_tc ← AME_tc ∪ {l}
      END
    END
  ELSEIF t = t_max AND (t, c) belongs to the set in (95)
    Generate nmut_tc from distribution (96)
    FOR i = 1 TO nmut_tc DO
      Generate mutated locus l from distribution (97)
      Define q by l ∈ hb_q
      IF q ∈ AHB_tc
        AME_tc ← AME_tc ∪ {l}
      END
    END
  END
END
```

Figure 19: Algorithm for generating ancestral mutation events (AME).
doi:10.5048/BIO-C.2016.4.f19

The alleles of $\boldsymbol{h}^{\mathrm{ref}}$ are then spread for each l by gene dropping along all ancestral lineages, forwards in time, down to present time $t = 0$. This continues until the alleles a_{kl} of locus l are defined for all sampled haplotypes $\boldsymbol{h}_k$ in (35), i.e. for $k = 1, \ldots, n$. Whenever an ancestral mutational event along a lineage occurs, a change of alleles takes places.

The set of polymorphic loci $\mathcal{L}_{\mathrm{pol}}$ are those loci with more than one allele in the sample, cf. (40). We may compute this set from the ancestral mutational events (92) as

$$\begin{aligned} \mathcal{L}_{\mathrm{pol}} &= \{l \in \mathcal{L};\ \mathrm{nralleles}_l > 1\} \\ &= \cup_{(t,c)\in \mathrm{AC}} \mathrm{AME}_{tc}. \end{aligned} \tag{99}$$

It is trivial to do the gene dropping at non-polymorphic loci $\mathcal{L}_{\mathrm{nonpol}} = \mathcal{L} \setminus \mathcal{L}_{\mathrm{pol}}$, since the sampled haplotypes will have alleles

$$a_{kl} = a_l^{\mathrm{ref}} \text{ for } l \in \mathcal{L}_{\mathrm{nonpol}} \text{ and } k = 1, \ldots, n,$$

identical to the reference haplotype at all such loci. It is more complicated to define the gene dropping process at polymorphic loci $l \in \mathcal{L}_{\mathrm{pol}}$. For each such locus, we first build a mutational tree

$$\mathrm{MT}_l \subset \mathrm{AC},$$

whose root is the reference chromosome $(t_{\max}, c^{\mathrm{ref}})$, and whose leaves are the n sampled chromosomes $(0, c_1), \ldots,$ $(0, c_n)$ of time point 0. All internal nodes belong to the set

$$\mathrm{AMC}_l = \{(t, c) \in \mathrm{AC};\ l \in \mathrm{AME}_{tc}\} \tag{100}$$

of ancestral mutated chromosomes at l, i.e. those ancestral chromosomes that experienced a mutation at l. For each node or vertex (t, c) of MT_l we introduce

$$\begin{aligned} \mathrm{pa}_l(t, c) &= \text{parent of } (t, c), \\ \mathrm{ch}_l(t, c) &= \text{set of children of } (t, c), \\ k_l(t, c) &= \text{index } k \text{ of one (arbitrary)} \\ &\quad \text{descendant } (0, c_k) \text{ of } (t, c), \end{aligned}$$

with $\mathrm{pa}_l(t_{\max}, c^{\mathrm{ref}}) = \emptyset$ for the reference chromosome and $\mathrm{ch}_l(0, c_k) = \emptyset$ for all leaves. The mutational tree is similar to an allelic genealogy [176], in that its internal nodes correspond to mutations rather than coalescence events. The allelic genealogy is in fact a subset of the mutational tree, where only those internal nodes that branch

```
INITIATE
  (t_max, c^ref) ∈ MT_l, where c^ref is defined in (98)
  pa_l(t_max, c^ref) = ch_l(t_max, c^ref) = ∅
  k_l(t_max, c^ref) = 1
  FOR k = 1 TO n
    (0, c_k) ∈ MT_l
    pa_l(0, c_k) = ch_l(0, c_k) = ∅
    k_l(0, c_k) = k
  END
  Define q by l ∈ hb_q
END
WHILE pa_l(t, c) = ∅ for some (t, c) ∈ MT_l \ {(t_max, c^ref)} DO
  k = k_l(t, c)
  t' = min{t''; t < t'' ≤ t_max, ARG_{t'',kq} ∈ AMC_l},
    where min ∅ = t_max + 1
  IF t' > t_max THEN
    pa_l(t, c) = (t_max, c^ref)
    ch_l(t_max, c^ref) ← ch_l(t_max, c^ref) ∪ {(t, c)}
  ELSEIF t' ≤ t_max THEN
    pa_l(t, c) = (t', c'), where c' = ARG_{t',kq}
    IF (t', c') ∉ MT_l
      MT_l ← MT_l ∪ {(t', c')}
      ch_l(t', c') = (t, c)
      k_l(t', c') = k
    ELSE
      ch_l(t', c') ← ch_l(t', c') ∪ {(t, c)}
    END
  END
END
```

Figure 20: Algorithm for building a mutational tree at a polymorphic locus. The tree is denoted MT_l at locus l. **doi:**10.5048/BIO-C.2016.4.f20

(i.e. have at least two children) are retained. The algorithm for building the mutational tree is summarized in Figure 20.

When MT_l has been defined, it remains to gene drop at l. That is, we need to spread the reference allele a_{kl}^{ref} from the root of the tree to all descendants, and generate new alleles randomly whenever a mutational node is found, with probabilities P_{ab} obtained either from (29), (30) or (31). To this end, we define an allele function

$$a_l : \mathrm{MT}_l \to \mathcal{A}(l)$$

for all nodes of MT_l, with $a_l(t, c) = \emptyset$ designating a node whose allele has not yet been assigned. The gene dropping algorithm is summarized in Figure 21. It is applicable both for single nucleotide, codon and microsatellite markers.

The mutational trees, and the subsequent gene dropping are illustrated in Figure 22 for two single nucleotide loci. The leftmost tree has only one internal node other than the root. This is always the case for mutational trees at loci without double mutations.

3.1.3 Example: Single nucleotide markers

For single nucleotide loci, we want the transition and transversion probabilities of matrix $\boldsymbol{P}$ in (29) to conform with the transition/transversion ratio R (cf. Section 1.5), as well as the probabilities π_A, π_G, π_C and π_T of all four nucleotides[4]. It can be shown that

$$\boldsymbol{P} = \begin{pmatrix} 0 & 1-u & ux & u(1-x) \\ 1-u & 0 & ux & u(1-x) \\ vy & v(1-y) & 0 & 1-v \\ vy & v(1-y) & 1-v & 0 \end{pmatrix} \qquad (101)$$

[4]Technically, $(\pi_A, \pi_G, \pi_C, \pi_T)$ should equal the stationary distribution of a Markov chain with transition matrix $\boldsymbol{P}$.

```
INITIATE
  a_l(t_max, c^ref) = a_l^ref
  FOR all (t, c) ∈ MT_l \ {(t_max, c^ref)}
    a_l(t, c) = ∅
  END
END
WHILE some internal node (t, c) has been assigned allele but
  not its children DO
  FOR all (t', c') ∈ ch_l(t, c) DO independently
    IF (t', c') ∈ AMC_l
      GENERATE a_l(t', c') randomly with
        Prob(a_l(t', c') = a' | a_l(t, c) = a) = P_{aa'}
    ELSE
      a_l(t', c') = a_l(t, c)
    END
  END
END
FOR k = 1 TO n
  a_kl = a_l(0, c_k)
END
```

Figure 21: Gene dropping of alleles at one locus. The locus number is l. **doi:**10.5048/BIO-C.2016.4.f21

indeed satisfies these constraints when

$$\begin{aligned} u &= 1/\left[2(R+1)(\pi_A+\pi_G)\right], \\ v &= 1/\left[2(R+1)(\pi_C+\pi_T)\right], \\ x &= \left[(2-v)\pi_C/(\pi_C+\pi_T)-(1-v)\right]/v, \\ y &= \left[(2-u)\pi_A/(\pi_A+\pi_G)-(1-u)\right]/u, \end{aligned}$$

and $0 \le x, y \le 1$. However, the last two inequalities are only satisfied when

$$\frac{1-u}{2-u} \le \frac{\pi_A}{\pi_A+\pi_G} \le \frac{1}{2-u},$$

and

$$\frac{1-v}{2-v} \le \frac{\pi_C}{\pi_C+\pi_T} \le \frac{1}{2-v}.$$

The matrix

$$\boldsymbol{P} = \begin{pmatrix} 0 & \frac{R}{R+1} & \frac{1}{2(R+1)} & \frac{1}{2(R+1)} \\ \frac{R}{R+1} & 0 & \frac{1}{2(R+1)} & \frac{1}{2(R+1)} \\ \frac{1}{2(R+1)} & \frac{1}{2(R+1)} & 0 & \frac{R}{R+1} \\ \frac{1}{2(R+1)} & \frac{1}{2(R+1)} & \frac{R}{R+1} & 0 \end{pmatrix} \qquad (102)$$

was suggested in [178]. It is a special case of (101) when $\pi_A = \pi_C = \pi_G = \pi_T = 1/4$, and it generalizes the symmetric matrix proposed in [118] when there are twice as many transversions as transitions ($R = 1/2$). See also Chapter X in [178].

3.1.4 Double mutations ignored

When all loci with double mutations are discarded, the algorithm of Section 3.1.2 can be simplified. The mutational tree at l can be computed more rapidly, since any of its internal nodes is either the reference chromosome, or has the reference chromosome as a parent. This amounts to adding two new rows to the algorithm of Figure 20, after its last IF statement:

```
  ⋮
IF (t', c') ∉ MT_l
  MT_l ← MT_l ∪ {(t', c')}
  ch_l(t', c') = (t, c)
  k_l(t', c') = k
  pa_l(t', c') = (t_max, c^ref)
  ch_l(t_max, c^ref) ← ch_l(t_max, c^ref) ∪ {(t', c')}
ELSE
  ⋮
```

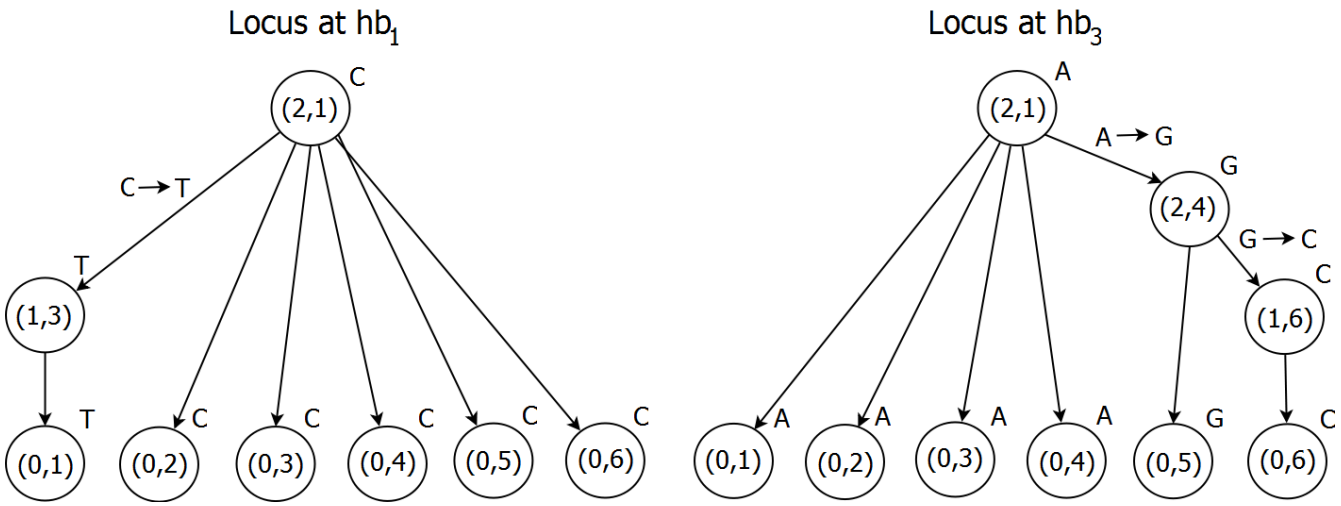

Figure 22: Illustration of mutational trees for nuclear autosomal DNA. Trees MT_l are shown at the two single nucleotide loci $l \in \mathcal{L}_{\mathrm{sn}}$ for which mutations occurred in Figure 7. At the leftmost locus within haplotype block hb_1, the gene dropping (see Figure 8) starts from nucleotide C of reference chromosome $(2,1)$ of the founder generation. The spread of this nucleotide is interrupted by one ancestral mutation $C \to T$ that occurs at chromosome $(1,3)$ at time point 1. At the second locus within hb_3, gene dropping starts from nucleotide A of the reference chromosome, and the spread of this allele is interrupted by two ancestral mutations. The first mutation $(A \to G)$ occurs in chromosome $(2,4)$ in order to generate diversity in the founder generation. The second mutation $(G \to C)$ occurs at time point 1, for an offspring of the mutated founder chromosome. Notice that the first locus is biallelic (C and T), whereas the second locus has three alleles A, G and C. The set of mutated sampled chromosomes SMC_l in (36) is $\{(0,1)\}$ for locus l of the leftmost graph, and $\{(0,5),(0,6)\}$ for locus l of the rightmost graph. **doi:**10.5048/BIO-C.2016.4.f22

Once the mutational tree at l is built, there is no need to gene drop alleles from the reference haplotype. In view of (39), it suffices to know which sampled chromosomes $(0, c_k)$ that descend from a mutated allele. This is equivalent to checking whether the parent of $(0, c_k)$ in the mutational tree is the reference chromosome $(t_{\max}, c^{\mathrm{ref}})$ or not:

```
FOR k = 1 to n DO
  IF pa_l((0, c_k)) = (t_max, c^ref)
    a^ism_kl = 0
  ELSE
    a^ism_kl = 1
  END
END
```

3.2 Mutation probabilities allele dependent

We will generalize the previous model of Section 1.5, and write the mutation probability between a pair of distinct alleles $a \neq b \in \mathcal{A}(l)$ at locus $l \in \mathcal{L}$ as

$$\begin{array}{rcl} \nu_{la} P_{ab} & = & \mathrm{Prob}(a \to b \text{ in founder chr. at } l), \\ \mu_{la} P_{ab} & = & \mathrm{Prob}(a \to b \text{ in non-founder chr. at } l). \end{array} \tag{103}$$

The mutation model of Section 1.5 is a special case of (103). The interpretation of P_{ab} in (103) is still the same, the probability that a mutation is from nucleotide a to b, given that we know it has happened and that the allele before the mutation was a. The main novelty of (103) is that the mutation probabilities $\nu_{la} = \nu_l$ and $\mu_{la} = \mu_l$ depend on which allele $a \in \mathcal{A}(l)$ that is changed. We have that

$$\begin{array}{rcl} \sum_{a \in \mathcal{A}(l)} \nu_{la} \pi_a & = & \nu_l, \\ \sum_{a \in \mathcal{A}(l)} \mu_{la} \pi_a & = & \mu_l, \end{array} \tag{104}$$

so that ν_l and μ_l are still the overall founder and germline mutation rates, when averaged over all possible alleles.

When the mutation probability μ_{la} depends on the allele a, we have to know this allele before generating the mutation. This has implications for the simulation algorithm, since mutations have to be defined simultaneously with gene dropping. This is more time consuming, as we have to apply gene dropping for lots of loci at which mutations never occur.

3.2.1 Example: Singe nucleotide markers

At single nucleotide markers the set of alleles is the four nucleotides A, G, C and T (cf. (3)). It is known for instance that the G and C alleles have at least a tenfold higher mutation rate than the A and T alleles for germline mutations [122].

For simplicity of notation we write $\mu_{la} = \mu_a$, not indicating that mutation probabilities depend on the locus l. The general time-reversible (GTR) model in [179] has the form

$$\begin{array}{rcl} (\mu_a P_{ab}) & = & \begin{pmatrix} -- & \mu_A P_{AG} & \mu_A P_{AC} & \mu_A P_{AT} \\ \mu_G P_{GA} & -- & \mu_G P_{GC} & \mu_G P_{GT} \\ \mu_C P_{CA} & \mu_C P_{CG} & -- & \mu_C P_{CT} \\ \mu_T P_{TA} & \mu_T P_{TG} & \mu_T P_{TC} & -- \end{pmatrix} \\ & = & \begin{pmatrix} -- & u_1 \pi_G & u_2 \pi_C & u_3 \pi_T \\ u_1 \pi_A & -- & u_4 \pi_C & u_5 \pi_T \\ u_2 \pi_A & u_4 \pi_G & -- & u_6 \pi_T \\ u_3 \pi_A & u_5 \pi_G & u_6 \pi_C & -- \end{pmatrix}, \end{array}$$

where $a, b \in \mathcal{A}_{\mathrm{sn}}$ range over all single nucleotide alleles, and $u_1, \ldots, u_6$ are free parameters that can the varied. This model automatically provides the correct nucleotide frequencies $\pi_A, \pi_G, \pi_C, \pi_T$. The other parameters can be varied as to fit the observed mutation probabilities μ_a and transition/transversion ratio R.

3.2.2 Example: Microsatellite markers

It is well known that the germline mutation probability $\mu_a = \mu_{la}$ for microsatellite markers depends on the number of repeats a. A data set with AC repeats was analysed in [139]. The authors found a best fitting model of the form

$$\mu_a P_{ab} = \begin{cases} \gamma_u \exp\left[\alpha_u a - \lambda(b-a)\right], & b > a, \\ \gamma_d \exp\left[\alpha_d a - \lambda(a-b)\right], & b < a, \\ 1 - \sum_{c; c \neq a} \mu_a P_{ac}, & b = a, \end{cases}$$

where λ controls how much the repeat length may change, γ_u (γ_d) determine the overall mutation probability for upward (downward) jumps, and α_u (α_d) controls how the upward (downward)

probabilities depend on length. Their parameter estimates where

$$\begin{aligned} \lambda &= 1.06, \\ \gamma_u &= 3.1 \cdot 10^{-6}, \\ \gamma_d &= 4.0 \cdot 10^{-7}, \\ \alpha_u &= 0.200, \\ \alpha_d &= 0.302. \end{aligned} \tag{105}$$

Since $\alpha_d > \alpha_u$, the stationary distribution $\{\pi_a\}_{a=1}^{\infty}$ for the number of tandem repeats exists. The values in (105) give a distribution that fits the empirically observed ones, with a mean around 20 repeats and standard deviation of order 5.

3.3 Ascertainment correction

When validating simulated data, it is important to mimic the sampling procedure of real data in order to avoid ascertainment bias. It is often the case that the set of markers has been found from a discovery panel with few individuals. As a consequence, highly polymorphic markers with at least two common alleles tend to be overrepresented. This implies that the observed allele frequency spectrum in the sample gets biased in comparison to that of the whole population.

A number of possible ascertainment schemes are possible, see for instance [180,181] and references therein. In our context we allow for ascertainment by retaining only a subset of polymorphic loci $\mathcal{L}_{\rm pol}$, whether they were generated as in Section 3.1 or 3.2. Let Asc_l be the event that a polymorphic locus l is retained after ascertainment. As a first approximation, we may assume that ascertainment events are independent between loci, with probabilities

$$\Pr_l = \text{Prob}(\text{Asc}_l). \tag{106}$$

The sparser $\mathcal{L}_{\rm pol}$ is, the more accurate is this approximation.

We mentioned above that most ascertainment schemes tend to select for loci that are polymorphic, i.e. have at least two common alleles. In order to illustrate this with a simple example, we let π_{la} be the frequency of allele a at locus l in the whole population. It equals $\hat{\pi}_{la} = n_{la}/n$ in (42) only if the whole population is sampled, i.e. if $n = 2N_0$. Suppose, for instance, that a reference sample of $n^{\rm ref}$ chromosomes from the population is used, where typically $n^{\rm ref}$ is much smaller than n. A simple ascertainment scheme retains l only if it is polymorphic in the reference panel. For a biallelic locus, both alleles must be present in the reference sample. Assuming that the $n^{\rm ref}$ chromosomes are drawn randomly without replacement from the whole population, we get an ascertainment probability

$$\Pr_l = \frac{\binom{2N_0 \text{MAF}_l}{n^{\rm ref}} + \binom{2N_0(1-\text{MAF}_l)}{n^{\rm ref}}}{\binom{2N_0}{n^{\rm ref}}} \tag{107}$$

in terms of the minor allele frequency

$$\text{MAF}_l = \min\{\pi_{al};\, a \in \mathcal{A}(l), \pi_{al} > 0\}$$

of locus l. The minor allele frequency in (43) can be viewed as an estimate of MAF_l for the sampled set of n chromosomes. A more refined ascertainment scheme would take into account that the reference sample is drawn from several different geographic regions.

Typically $\Pr_l$ is a function of the allele frequencies of the marker at l, as in (107). If these are not known, we replace these frequencies by appropriate estimates and plug them into (106), in order to obtain an an estimate $\widehat{\Pr}_l$ of $\Pr_l$. For instance, if the minor allele frequency MAF_l in (107) is not known, we replace it by an estimate. Figure 23 summarizes the ascertainment correction algorithm.

```
Define the set L_pol of polymorphic loci from the gene dropping
  algorithm
FOR l ∈ L_pol DO
  Compute estimate Pr_l-hat of ascertainment probability in (106)
  With probability 1 − Pr_l-hat DO L_pol ← L_pol \ {l}
  END
END
```

Figure 23: Removal of all non-ascertained polymorphic loci from the sampled haplotypes. **doi:**10.5048/BIO-C.2016.4.f23

4. NATURAL SELECTION

The algorithm of Sections 1.4-1.6 only includes four forces of microevolution; genetic drift, recombinations, migration and mutations. In this section we will briefly discuss the fifth one, natural selection. It occurs when the genetic composition of individuals influence their reproductive fitness. For the public, directional selection has often been described as the major force of evolution [182,183]. This is the most well known type of natural selection where one allele is believed to have a selective advantage over another at the same locs. But an increasing number of authors have emphasized its limited power to generate new genetic patterns [184-186], although it seems important for some specific genes [187]. Selection does first of all not act on DNA but on phenotypes, our observable characteristics like body weight and functioning organs. DNA only affects fitness indirectly, and it needs mutations to operate on, of which most are neutral or slightly deleterious. It seems that the major task of directional selection is to select against those individuals that have highly deleterious mutations. And since selection operates on phenotypes, its power is limited to favor beneficial alleles and/or to remove slightly deleterious alleles at many loci simultaneously [188-190].

But selection may still have an important role to play in order to generate diversity. Some genomic regions, such as the HLA-DRB1 gene of the major histocompability complex of chromosome 6, which is important for the human immune system, or a region from the ABO blood group system on chromosome 9, are known to be very polymorphic [191,192]. This is partially due to a higher mutation rate in these regions, but it is likely that selection is required as well in order to explain such a high diversity. Another example is the mouflon population from an isolated island of the Sub-Antarctive archipelago. It was founded in 1957 by two individuals, and its current size oscillates around 500 [193]. The population has been monitored since the 1970s, and analysis of genetic data revealed a diversity much higher than neutral models of microevolution would predict. Simulations indicate that natural selection could be a major explanation of this unexpectedly high genomic diversity.

A number of different types of natural selection are described in Chapter 6 of [47]. Of these balancing selection is the one most suitable to explain the occurrence of genetic diversity. It operates in such a way that heterozygous individuals, with different homologous alleles at one or several parts of the genome, have a selective advantage in terms of a higher survival probability [194-196]. Balancing selection was used in [193] for the microsatellite loci distributed throughout the genome in their simulations. With such an approach we cannot build a pedigree backwards in time without knowing DNA at the loci that are under selection. It is therefore not possible to use the backward simulation algorithm of Sections 2.1.1-2.1.4, since the pedigree is generated before DNA is assigned through gene dropping and mutations. Instead a forward simulation algorithm (see Section 1.4) is required, and this is very time consuming for the worldwide human population.

For this reason we confine ourselves to describe a hybrid model

that combines forward and backward simulation. It is applicable when balancing selection only operates on one single haplotype block hb_{q_0} $(1 \le q_0 \le Q)$, located in an autosomal region of DNA. This region has to be small enough so that ordinary recombinations and gene conversions within it can be neglected. This could be of interest, for instance, in order to generate diversity at parts of the HLA or blood group system regions referred to above. In more detail, the hybrid simulation algorithm looks as follows:

1. In the first step we build the pedigree, drop DNA and generate new mutations at all loci $l \in \mathrm{hb}_{q_0}$ within the selected haplotype block. This is done forwards in time for the whole population (1) of individuals.

2. In the second step we select a sample of n chromosomes out of the $2N_0$ simulated individuals of time point 0 from Step 1.

3. In the third step, we build the ancestral recombination graph of the n sampled chromosomes from Step 2, using the fact that a pedigree already exists, as well as genetic inheritance at hb_{q_0}. This means that each chromosome $(t,c) \in \mathrm{C}$ in (8) from a non-founder time point $t < t_{\max}$ already has a parental chromosome $p_{tq_0}(c)$ from which its DNA at haplotype block hb_{q_0} was inherited, and therefore the coalescence tree of this haplotype block is completely specified from Step 2.

   In order to build the rest of the ARG, crossovers are generated as in Subsection 2.2 for all ancestral chromosomes $(t,c) \in \mathrm{AC}$ from non-founder time points. This makes it possible to define inheritance of (t,c) recursively at all haplotype blocks hb_q other than hb_{q_0}. For the blocks to the right of hb_q we use (66) for $q = q_0+1,\dots,Q$, whereas for those to the left we use

$$\begin{array}{rcll} p_{tq}(c) & = & p_{t,q+1}(c), & \text{if } \mathrm{hbb}_q \notin \mathrm{REC}_{tc}, \\ p_{tq}(c) & \neq & p_{t,q+1}(c), & \text{if } \mathrm{hbb}_q \in \mathrm{REC}_{tc}, \end{array}$$

   for $q = q_0-1,\dots,1$. Once $p_{tq}(c)$ has been defined for all ancestral chromosomes and all haplotype blocks, we build the ARG as in Figure 13.

   When the haplotype block structure (apart from the selected block hbb_{q_0}) is not known in advance, it has to be generated simultaneously with the ARG, as described in Section 2.3.

4. Ancestral mutations are defined at all haplotype blocks other than hb_{q_0}. DNA is then dropped from the founder generation as described in Section 3, at all these haplotype blocks.

It remains to define the algorithm of Step 1 in more detail (see Figure 24 for a summary). For simplicity, we will assume a demographically homogeneous population without demes that has non-overlapping generations. This is the framework of the pedigree algorithm of Section 2.1.1, which we modify in three ways: First, the children of generation t that will have parents from generation $t+1$ assigned to them, is not the set of ancestral individuals AI_t, but the much larger set

$$\mathrm{I}_t = \{i;\ (t,i) \in \mathrm{I}\} = \{1,\dots,2N_t\}$$

of all individuals of generation t. Second, time proceeds forwards $(t = t_{\max}-1, t_{\max}-2,\dots,0)$ rather than backwards $(t = 0,1,\dots,t_{\max}-1)$. Third, only some children survive birth, with probabilities depending on their genotypes at hb_{q_0}. More specifically, we let

$$G_{ti} = (h_{t,2i-1}^{(q_0)}, h_{t,2i}^{(q_0)}) \tag{108}$$

be the genotype of individual $(t,i) \in \mathrm{I}$ at the selected haplotype block hb_{q_0}. It consists two haplotypes

$$h_{tc}^{(q_0)} = (a_{tc,l_{q_0-1}}+1,\dots,a_{tc,l_{q_0}})$$

of nucleotides a_{tcl} from the chromosome c that (t,i) inherited at this haplotype block from its father $(c = 2i-1)$ or mother $(c = 2i)$. Selection enters the algorithm of Figure 24 in terms of the probability $S(G_{ti})$ that an individual with genotype G_{ti} survives birth, grows up and becomes an adult. If not, we need to repeat the procedure and assign parents and a new genotype to (t,i), until we eventually obtain a child that reaches the adult stage.

For simplicity of notation we will write $h_{t,2i-1}^{(q_0)} = h_1$ and $h_{t,2i-1}^{(q_0)} = h_2$ for the two haplotypes of (t,i) at haplotype block hb_{q_0}. That is, we consider a fertilized egg whose genotype $G = (h_1,h_2)$ is formed at the selected locus, with $h_c = (a_{cl};\ l \in \mathrm{hb}_{q_0})$ the haplotype derived from the sperm cell $(c = 1)$ or ova cell $(c = 2)$ at haplotype block hb_{q_0}. The simplest kind of balancing selection has a survival probability

$$S(G) = \begin{cases} 1, & h_1 \neq h_2, \\ 1-s, & h_1 = h_2, \end{cases} \tag{109}$$

for some selection coefficient s. This means that heterozygous individuals always survive, whereas homozygous individuals have a probability s of not becoming adults. A more sophisticated survival function takes into account how different the two haplotypes h_1 and h_2 are in terms of their diversity (or Hamming distance)

$$\mathrm{div}(h_1,h_2) = |\{l;\ l_{q_0-1} < l < l_{q_0}, a_{1l} \neq a_{2l}\}|,$$

i.e. the number of loci at which they differ. Suppose the optimal diversity between h_1 and h_2 is $\mathrm{div}_{\mathrm{opt}}$ in terms of maximizing selective advantage. If the diversity increases beyond that, gene function or gene regulation gets interrupted, and when it exceeds $\mathrm{div}_{\max}$, the fertilized cell will die with certainty. We formalize this as

$$S(G) = \begin{cases} 1 - s\cdot\frac{(\mathrm{div}(h_1,h_2)-\mathrm{div}_{\mathrm{opt}})^2}{\mathrm{div}_{\mathrm{opt}}^2}, \\ 1 - \frac{(\mathrm{div}(h_1,h_2)-\mathrm{div}_{\mathrm{opt}})^2}{(\mathrm{div}_{\max}-\mathrm{div}_{\mathrm{opt}})^2}, \\ 0, \end{cases}$$

as to whether the diversity satisfies $0 \le \mathrm{div}(h_1,h_2) \le \mathrm{div}_{\mathrm{opt}}$, $\mathrm{div}_{\mathrm{opt}} \le \mathrm{div}(h_1,h_2) \le \mathrm{div}_{\max}$ or $\mathrm{div}(h_1,h_2) \ge \mathrm{div}_{\max}$. In particular, homozygots $(\mathrm{div}(h_1,h_2) = 0)$ have a survival probability of $1-s$, as in (109). Then the survival probability increases quadratically up to a maximum of 1 at $\mathrm{div}(h_1,h_2) = \mathrm{div}_{\mathrm{opt}}$, and after that it decreases quadratically down to 0 at $\mathrm{div}(h_1,h_2) = \mathrm{div}_{\max}$. Possible parameter values could be $\mathrm{div}_{\mathrm{opt}} = 10$ and $\mathrm{div}_{\max} = 100$. But this depends heavily on how the different amino acids interact in the protein that gene(s) in hb_{q_0} code for.
Another selection model used in [193], is

$$S(G) = \begin{cases} 0, & 0 \le \mathrm{div}(h_1,h_2) < \varepsilon\mathrm{div}_{\max}, \\ 1, & \varepsilon\mathrm{div}_{\max} \le \mathrm{div}(h_1,h_2) \le \mathrm{div}_{\max}. \end{cases}$$

This could be of interest, for instance, when all loci are single nucleotides, and it is known that the haplotype block hb_{q_0} harbours at most $\mathrm{div}_{\max}$ single nucleotides polymorphisms. For survival it is required that an individuals is heterozygous for at least a fraction ε of these SNPs.

5. CONCLUSIONS

In this paper we proposed a mathematical model for simulation of human genetic data based on the assumption that the worldwide human population originates from one single couple. The main idea is to build an ancestral recombination graph backwards in time for all sampled individuals. The model is very flexible and

FOR $t = t_{\max} - 1, t_{\max} - 2, \ldots 0$ DO
  INITIATE
    $f_{t+1} = m_{t+1} = C_{t+1,m} = C_{t+1,mf} = 0$ for all m, f
  END
  FOR $i = 1, \ldots, 2N_t$ DO
    WHILE parents for (t, i) and its DNA at hb_{q_0} have not yet been selected
      Parents (m, f) of (t, i) are proposed with distribution (68), where j=i-1
      Choose grandpaternal modes of inheritance of (t, i) for its DNA at hb_{q_0} as in (62)-(63)
      Pass on paternal DNA $h^{(q_0)}_{t,2i-1} = h^{(q_0)}_{t+1,p_{t,q_0}(2i-1)}$
        and maternal DNA $h^{(q_0)}_{t,2i} = h^{(q_0)}_{t+1,p_{t,q_0}(2i)}$ at hb_{q_0}
      Update $h^{(q_0)}_{t,2i-1}$ and $h^{(q_0)}_{t,2i}$ with mutations
      Define G_{ti} as in (108)
      Select parents (m, f) and DNA G_{ti} of (t, i) at hb_{q_0} with probability $S(G_{ti})$
    END
    Keep the selected parents $(m_t(i), f_t(i)) = (m, f)$ of (t, i) and its DNA G_{ti} at hb_{tq_0}
    Update the relevant $f_{t+1}, m_{t+1}, C_{t+1,m}, C_{t+1,mf}$
  END
END

Figure 24: Step 1 of the balancing selection algorithm, where the genealogy is generated forwards in time at an autosomal haplotype block. This block is denoted hb_{q_0}, and it is assumed that generations are non-overlapping with no geographic substructure. **doi:**10.5048/BIO-C.2016.4.f24

allows for different demographic scenarios, with time varying population sizes and possible migration between geographic subregions. Reproduction is based on a dioecious and diploid framework where males and females are treated separately, so that different mating scenarios are possible. The model also incorporates ordinary recombination events, gene conversion, neutral mutations, and age structure in terms of overlapping generations. An extension of the model with mixed forward and backward simulation allows for balancing selection as well. One particularly important parameter is the created diversity, which makes it possible to obtain a substantial amount of genetic diversity for nuclear autosomal and X-chromosome DNA, during a relatively short period of time.

In subsequent papers, we plan to simulate human DNA data from our proposed model in order to assess how well it fits real data. The description of the mathematics is quite detailed. The rationale for this is that other research groups may implement the model as well, in order to test a wide range of different population history scenarios. A challenging continuation of this project is first to develop a more automated inference procedure, in order to find the best fitting population history within a unique origin framework, and then to compare it with a best fitting common ancestry model.

ACKNOWLEDGMENTS

The authors wish to thank the editors, three anonymous reviewers, Peter Loose and Geoff Barnhard for valuable comments on the work.

1. Mourant AE (1954) The Distribution of Human Blood Groups. Blackwell Scientific Publicationis (Oxford).

2. Cavalli-Sforza L, Bodmer W (1971) The Genetics of Human Populations. Freeman (San Francisco, CA).

3. Cann RL, Stoneking M, Wilson AC (1987) Mitochondrial DNA and human variation. Nature 325:31-36. doi:10.1038/325031a0

4. Vigilant L, Stoneking M, Harpending H, Hawkeds K, Wilson AC (1991) African populations and the evolution of mitochondrial DNA. Science 253:1503-1507. doi:10.1126/science.1840702

5. Templeton AR (1993) The "Eve" hypothesis: A genetic critique and reanalysis. Amer Anthropol 95(1):51-72. doi:10.1525/aa.1993.95.1.02a00030

6. Ingman M, Kaessman H, Pääbo S, Gyllensten U (2000) Mitochondrial genome variation and the origin of modern humans. Nature 408:708-713. doi:10.1038/35047064

7. Tang H, Siegmund DO, Shen P, Oefner PJ, Feldman MW (2002) Frequentist estimates of coalescence times from nucleotide sequence data using a tree-based partition. Genetics 161:447-459.

8. Olivieri A, et al (2006) The mtDNA legacy of the Levantine early upper Palaeolithic in Africa. Science 314:1767-1770. doi:10.1126/science.1135566

9. 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

10. Underhill PA, et al (2000) Y chromosome sequence variation and the history of human populations. Nat Genet 26:358-361. doi:10.1038/81685

11. Zhivotovsky LA, et al (2011). The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. Am J Hum Genet 74:50-61. doi:10.1086/380911

12. Francallaci P, et al (2013) Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny. Science 342:565-569. doi:10.1126/science.1237947

13. Poznik GD, 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

14. Poznik GD, et al (2016) Punctuated bursts in human male demography inferred from 1 244 worldwide Y-chromosome sequences. Nat Genet 48(6):593-599. doi:10.1038/ng.3559

15. Zhivotovsky LA, Bennett L, Bowcock AM, Feldman MW (2000) Human population expansion and microsatellite variation. Mol Biol Evol 17(5):757-767. doi:10.1093/oxfordjournals.molbev.a026354

16. Zhivotovsky LA, Rosenberg NA, Feldman MW (2003) Features of evolution and expansion of modern humans, inferred from microsatellite markers. Am J Hum Genet 72:1171-1186. doi:10.1086/375120

17. The International HapMap Consortium (2003) The International HapMap Project. Nature 426:789-796. doi:10.1038/nature02168

18. The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437(27): 1299-1320. doi:10.1038/nature04226

19. The International HapMap Consortium (2007) A second generation human haplotype of over 3.1 million SNPs. Nature 449(18):851-862. doi:10.1038/nature06258

20. The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061-1073. doi:10.1038/nature09534

21. The 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. Nature 526:68-87. doi:10.1038/nature15393

22. Zhao Z, et al (2000) Worldwide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22. Proc Natl Acad Sci USA 97(21):11354-11358. doi:10.1073/pnas.200348197

23. 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

24. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachadran S, et al (2008) Worldwide human relationships from genome-wide patterns of variation. Science 319:1100-1104. doi:10.1126/science.1153717

25. Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. Nature 475:493-496. doi:10.1038/nature10231

26. Sheehan S, Harris K, Song YS (2013) Estimating variable effective population sizes from multiple genomes: A sequentially Markov conditional sampling distribution approach. Genetics 194:647-662. doi:10.1534/genetics.112.149096

27. Palmara PF, Lencz T, Darvasi A, Pe′er I (2012) Length distributions of identity by descent reveal fine-scale demography history. Am J Hum Genet 91:809-822. doi:10.1016/j.ajhg.2012.08.030

28. Sherry ST, Harpending HC, Batzer MA, Stoneking M (1997) Alu evolution in human populations: Using the coalescent to estimate effective population size. Genetics 147:1977-1982.

29. Nasidze I, et al (2001) Alu insertion polymorphisms and the genetic structure of human populations from the Caucasus. Eur J Hum Genet 9:267-272. doi:10.1038/sj.ejhg.5200615

30. Romualdi C, et al (2002) Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms. Gen Res 12:602-612. doi.10.1101/gr.214902

31. Stringer C (2002) Modern human origins: progress and prospects. Phil Trans R Soc Lond 357:563-579. doi:10.1098/rstb.2001.1057

32. Cavalli-Sforza L, Feldman MW (2003) The application of molecular genetic approaches to the study of human evolution. Nat Genet Suppl 33:266-275. doi:10.1038/ng1113

33. 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

34. Schiffels S, Durbin R (2014) Inferring human population size and history from multiple genome sequences. Nat Genet 46:919-925. doi:10.1038/ng.3015

35. Relethford JH (1998) Genetics of modern human origins and diversity. Ann Rev Anthropol 27:1-23. doi:10.1146/annurev.anthro.27.1.1

36. 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

37. 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

38. 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

39. Prüfer 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(43):9. doi:10.1038/nature12886

40. 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

41. 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

42. Yang Z (2002) Likelihood and Bayes estimation of ancestral population size in hominoids using data from multiple loci. Genetics 162:1811-1823.

43. Felsenstein J (1982) How do we infer geography and history form gene frequencies? J Theor Biol 96:9-20. doi:10.1016/0022-5193(82)90152-7

44. Relethford JH, Harpending HC (1995) Ancient differences in population in population size can mimic recent African origin of modern humans. Curr Anthropol 36(4):667-674. doi:10.1086/204415

45. Meyers S, Fefferman C, Patterson N (2008) Can one learn history from the allelic frequency spectrum? Theor Popul Biol 73:342-348. doi:10.1016/j.tpb.2008.01.001

46. Mazet O, Rodriquez W, Grusea S, Boitard S, Chikhi L (2016) On the importance of being structured: instantaenous coalescence rates and human evolution - lessons for ancestral population size inference? Heredity 116:362-371. doi:10.1038/hdy.2015.104

47. Durrett R (2008) Probability Models for DNA Sequence Evolution, 2nd ed. Springer (New York). doi:10.1007/978-0-387-78168-6

48. Blum MGB, Jakobsson M (2011) Deep divergences of human gene trees and models of human origins. Mol Biol Evol 28(2):889-898. doi:10.1093/molbev/msq265

49. Sanford JC, Carter R (2014) In light of genetics ... Adam, Eve and the Creation/Fall. Christ Apol J 12(2):51-98.

50. 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

51. Henn BM, Gravel S, Moreno-Estrada A, Acevedo-Acevedo S, Bustamante CD (2010) Fine-scale population structure and the era of next-generation sequencing. Hum Mol Genet 19(Review Issue 2):R221-R226. doi:10.1093/hmg/ddq403

52. Wright S (1931) Evolution in Mendelian populations. Genetics 16:97-159.

53. Wright S (1951) The general structure of populations. Ann Eugenic 15:323-354. doi:10.1111/j.1469-1809.1949.tb02451.x

54. Takahata N (1993) Allelic genealogy and human evolution. Mol Biol Evol 10:2-22.

55. Takahata N (1995) A genetic perspective of origins and history of humans. Ann Rev Ecol Syst 26:343-372. doi:10.1146/annurev.es.26.110195.002015

56. Excoffier L (2004) Patterns of DNA sequence diversity and genetic structure after a range expansion: lesson from the infinite-island model. Mol Ecol 13:853-864. doi:10.1046/j.1365-294X.2003.02004.x

57. Seielstad M, Minch E, Cavalli-Sforza LL (1998) Genetic evidence for a higher female migration rate in humans. Nat Genet 20:278-280. doi:10.1038/3088

58. Stoneking M (1998) Women on the move. Nat Genet 20:219-220. doi:10.1038/3012

59. Kimura M, Weiss WH (1964) The stepping stone model of genetic structure and the decrease of genetic correlation with distance. Genetics 49:561-576.

60. Barbujani G, Sokal RR, Oden NL (1995) Indo-European origins: A computer-simulation test of five hypotheses. Am J Phys Anthropol 96:109-132. doi:10.1002/ajpa.1330960202

61. Ray N, Currat M, Excoffier L (2003) Intra-deme molecular diversity in spatially expanding populations. Mol Biol Evol 20(1):76-86. doi:10.1093/molbev/msg009

62. Currant M, Excoffier L (2004) The effect of Neolithic expansion on European molecular diversity. Proc R Soc B 272, 679-688. doi:10.1098/rspb.2004.2999

63. 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

64. McVean GAT, et al (2004) The fine scale structure of recombination rate variation in the human genome. Science 304:581-584. doi:10.1126/science.1092500

65. Wiuf C, Posada D (2003) A coalescent model with recombination hotspots. Genetics 164:407-417.

66. 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

67. Goldstein DB (2001) Islands of linkage disequilibrium. Nat Genet 29:109-111. doi:10.1038/ng1001-109

68. Patil N, et al (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science 294:1719-1723. doi:10.1126/science.1065573

69. Gabriel SB, et al (2002) The structure of haplotype blocks in the human genome. Science 296:2225-2229. doi.10.1126/science.1069424

70. Phillips MS, et al (2003) Chromosome-wide distribution of haplotype block and the role of recombination hot spots. Nat Genet 33:382-387. doi:10.1038/ng1100

71. Rinaldo A, Bacanu S-A, Devlin B, Sompar V, Wasserman L, Roeder K (2005) Characterization of multilocus linkage disequilibrium. Genet Epidemiol 28:193-206. doi:10.1002/gepi.20056

72. Cardon LR, Abecasis GR (2003) Using haplotype blocks to map human complex trait loci. Trends Genet 19(3):135-140. doi:10.1016/S0168-9525(03)00022-2

73. Wall JD, Pritchard JK (2003) Assessing performance of the haplotype block model of linkage disequilibrium. Am J Hum Genet 73:502-515. doi:10.1086/378099

74. Wang N, Akey JM, Zhang K, Chakraborty R, Jin L (2002) Distribution of recombination crossovers and the origin of haplotype blocks: The interplay of population history, recombination and mutation. Am J Hum Genet 71:1227-1234. doi:10.1086/344398

75. Pääbo S (2003) The mosaic that is our genome. Nature 421:409-412. doi:10.1038/nature01400

76. 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

77. Rosenfeld JA, Mason CE, Smith TM (2012) Limitations of the human reference genome for personalized genomics. PLoS ONE 7(7):e40294. doi:10.1371/journal.pone.0040294

78. Hinch AG, et al (2011) The landscape of recombination in African Americans. Nature 476, 170-177. doi:10.1038/nature10336

79. Excoffier L, Heckel G (2006) Computer programs for population genetics data analysis: a survival guide. Nat Rev Genet 7:745-758. doi:10.1038/nrg1904

80. Terwilliger JD, Speer M, Ott, O (1993) Chromosome-based method for rapid computer simulation in human genetic linkage analysis. Genet Epidemiol 10:217-224. doi:10.1002/gepi.1370100402

81. Hambe J, Wienker T, Schreiber S, Nurberg P (1998) POPSIM: a general population simulation program. Bioinformatics 14:458-464. doi:10.1093/bioinformatics/14.5.458

82. Balloux F (2001) EASYPOP (version 1.7): A computer program for population genetics simulation. J Hered 92:301-302. doi:10.1093/jhered/92.3.301

83. Peng B, Kimmel M (2005) A forward-time population genetics simulation environment. Bioinformatics 21(18): 3686-3687. doi:10.1093/bioinformatics/bti584

84. Hoggart CJ, Chadeau M, Clark TG, Lampariello R, De IM, Whittaker JC, et al (2007) Sequence level population simulations over large genomic regions. Genetics 177(3):1725-1731. doi:10.1534/genetics.106.069088

85. 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.

86. Edwards TL, Bush WS, Turner SD, Dudek SM, Torstensson ES, Schmidt M, Martin E, Ritchie MD (2008) In: Marchiori E, Moore JH, eds. Generating linkage disequilibrium patterns in data simulations using genome SIMLA. EvoBIO 2008, LNCS 4973, Springer-Verlag (Berlin):24-35.

87. Kingman JFC (1982) The coalescent. Stoch Proc Appl 13:235-248. doi:10.1016/0304-4149(82)90011-4

88. 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

89. Hudson RR (1990) Gene genealogies and the coalescent process. Oxford Surv Evol Biol 7:1-44.

90. 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

91. Nordborg M, Tavaré S (2002) Linkage disequilibrium: what history has to tell us. Trends Genet 18(2):83-90. doi:10.1016/S0168-9525(02)02557-X

92. Wakeley J (2004) Recent trends in population genetics: More data! More math! Simple models? J Hered 95(5):397-405. doi:10.1093/jhered/esh062

93. Wakeley J (2009) Coalescence Theory: An Introduction, Roberts & Co. Publishers (Greenwood Village, Colorado).

94. Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation Bioinformatics 18(2):337-338. doi:10.1093/bioinformatics/18.2.337

95. Posada D, Wiuf C (2003) Simulating haplotype blocks in the human genome Bioinformatics 19(2):289-290. doi:10.1093/bioinformatics/19.2.289

96. Mailund T, Schierup MH, Pedersen CNS, Mechlenborg PJM, Madsen JN, Schauser L (2005) CoaSim: A flexible environment for simulating genetic data under coalescent models. BMC Bioinformatics 6:252, 6 pages. doi:10.1186/1471-2105-6-252

97. Liang L, Zöllner S, Abecasis R (2007) GENOME: a rapid coalescent based whole genome simulator. Bioinformatics 23:1565-1567. doi:10.1093/bioinformatics/btm138

98. Currant M, Ray N, Excoffier L (2004) SPLATCHE: a program to simulate genetic diversity taking into account environmental heterogeneity Mol Ecol Notes 4:139-142. doi:10.1046/j.1471-8286.2003.00582.x

99. 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

100. Gasbarra D, Pirinen M, Sillanpää MJ, Salmela E, Arjas E (2007) Backward simulation of ancestors of sampled individuals. Theor Popul Biol 72:305-322. doi:10.1016/j.tpb.2007.06.004

101. Gasbarra D, Pirinen M, Sillanpää MJ, Arjas E (2007) Estimating genealogies from linked marker data: a Bayesian approach. BMC Bioinformatics 8:411. doi:10.1186/1471-2105-8-411

102. Labuda D, Lefebvre J-F, Nadeau P, Roy-Ganon M-H (2010) Female-to-male breeding ratio in modern humans - an analysis

based in historical recombinations. Am J Hum Genet 86:353-363. doi:10.1016/j.ajhg.2010.05.009

103. 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

104. Ott J (1999) Analysis of Human Genetic Linkage, 3rd ed. The John Hopkins University Press (Baltimore).

105. Sham P (1998) Statistics in Human Genetics. Arnold Applications of Statistics (London).

106. Hilliker AJ, Harauz G, Reaume G, Gray M, Clark SH, et al (1994) Meiotic tract length distribution with the rosy locus of *Drosophila melanogaster*. Genetics 116:153-160.

107. Andolfatto P, Nordborg M (1997) The effect of gene conversion on intralocus associations. Genetics 148:1397-1399.

108. Hilliker AJ, Chovnick A (1981) Further observations of intragenic recombination in *Drosophila melano-goster*. Gen Res 38:281-296. doi:10.1017/S0016672300020619

109. Frisse L, Hudson RR, Bartoszewicz A, Wall JD, Donfack J, Di Rienzo A (2001) Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am J Hum Genet 69:831-843. doi:10.1086/323612

110. Wall JD (2001) Insights from lined single nucleotide polymorphisms: what we can learn from linkage disequilibrium. Curr Opin Genet Dev 11:647-651. doi:10.1016/S0959-437X(00)00248-3

111. Przeworski M, Wall JD (2001) Why is there so little intragenic linkage disequilibrium in humans? Gen Res 77:143-151. doi:10.1017/S0016672301004967

112. Cole F, Baudat F, Grey C, Keeney S, de Massy B, Jasin M (2014) Mouse tetrad analysis provides insights into recombination mechanisms and hotspot evolutionary dynamics. Nat Genet 46:1072-1080. doi:10.1038/ng.3068

113. Williams AL, Genovese G, Dyer T, Altemose N, et al (2015) Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. Elife 4:e04637. doi:10.7554/elife.04637

114. Parker G (1980) Creation, mutation and variation. Acts and Facts 9:11.

115. 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 J 9:81-122.

116. Reinert G, Schbath S, Waterman MS (2000) Probabilistic and statistical properties of words: an overview. J Comp Biol 7:1-46. doi:10.1089/10665270050081360

117. Arndt PF, Burge CB, Hwa T (2003) DNA sequence evolution with neighbour-dependent mutation. J Comp Biol 10:313-322. doi:10.1089/10665270360688039

118. Jukes TH, Cantor C (1969) Evolution of protein molecules. In: Munro MN, ed. Mammalian Protein Metabolism, Academic Press (New York): pp. 245-279. doi:10.1016/b978-1-4832-3211-9.50009-7

119. Carter RW (2014) The non-mythical Adam and Eve! Historical Adam biologos, `http://creation.com/historical-adam-biologos`.

120. Conrad, DF et al (2011) Variation in genome-wide mutation rates within and between human families. Nat Genet 43(7):712-714. doi:10.1038/ng.862

121. Crow JF (2000) The origins, patterns and implications of human spontaneous mutation. Nat Rev Genet 1:40-47. doi:10.1038/35036060

122. Campbell CD, Eichler EE (2013) Properties and rates of germline mutations in humans. Trends Genet 29(10):575-584. doi:10.1016/j.tig.2013.04.005

123. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489(7414):57-74. doi:10.1038/nature11247

124. Kellis M, Wold B, Snyder MP, Bertstein BE, Kundaje A, Marinov GK, et al (2014) Defining functional DNA elements in the human genome. Proc Natl Acad Sci USA 111(17):6131-6138. doi:10.1073/pnas.1318948111

125. Mendez FL, et al (2013) An African American paternal lineage adds an extremely ancient root to the human Y chromosome phylogenetic tree. Am J Hum Genet 92:454-459. doi:10.1016/j.ajhg.2013.02.002

126. Helgason A, et al (2015) The Y-chromosome point mutation rate in humans. Nat Genet 47:453-457. doi:10.1038/ng.3171

127. 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

128. Balanovsky O, 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

129. Heyer E, Zietkiewicz E, Rochowski A, Yotova V, Puymirat J, Labuda D (2001) Phylogenetic and familial estimates of mitochondrial substitution rates: Study of the control region mutations in deep-rooting pedigrees. Am J Hum Genet 69:1113- 1126. doi:10.1086/324024

130. Howell N, Bogolin Smejkal C, Makey DA, Chinnery PF, Turnbull DM, Herrnstadt C (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

131. 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

132. 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.

133. Fay JC, Wu C-J (1999) A human population bottleneck can account for the discordance between patterns of mitochondrial variation versus nuclear DNA variation. Mol Biol Evol 16:1003-1005. doi:10.1093/oxfordjournals.molbev.a026175

134. Sun JX, Helgasson A, Masson G, Ebenersersdóttir SS, Li H, Mallick S, Gnerre S, Patterson N, Kong A, Reich D, Stefansson K (2012) A direct characterization of human mutation based on microsatellites. Nat Genet 44(10):1161-1167. doi:10.1038/ng.2398

135. Kayser, et al (2000) Characteristics and frequency of germline mutations at microsatellite loci from the human *Y* chromosome, as revealed by direct observation in father/son pairs. Am J Hum Genet 66:1580-1588. doi:10.1086/302905

136. Lynch M (2010) Rate, molecular spectrum and consequences of human mutation. Proc Natl Acad Sci USA 107:961-968. doi:10.1073/pnas.0912629107

137. Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11(5):725-736.

138. Ohta T, Kimura M (1973) The model of mutation appropriate to calculate number of electrophoretically detectable alleles in a

genetic population. Gen Res 22:201-204. doi:10.1017/S0016672300012994

139. Whittaker JC, Harbord RM, Boxall N, Mackay I, Dawson G, Sibly RM (2003) Likelihood-based estimation of microsatellite mutation rates. Genetics 164:781-787.

140. 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.

141. Fu YX (1995) Statistical properties of segregating sites. Theor Popul Biol 48:172-197. doi:10.1006/tpbi.1995.1025

142. Wooding S, Rogers A (2002) The matrix coalescent and an application to human single-nucleotide polymorphisms. Genetics 161:1641-1650.

143. Marth GT, Czabarka E, Murvai J, Sherry ST (2004) The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. Genetics 166:351-372. doi:10.1534/genetics.166.1.351

144. Nawa N, Tajima F (2008) Simple method for analyzing the pattern of DNA polymorphism and its application to SNP data of human. Genes Genet Syst 83:353-360. doi:10.1266/ggs.83.353

145. Rafajlovic M, Klassman A, Eriksson A, Wiehe T, Mehlig B (2014) Demography-adjusted tests of neutrality based on genome-wide SNP data. Theor Popul Biol 95:1-12. doi:10.1016/j.tpb.2014.05.002

146. Kim HL, Ratan A, Perry GH, Montenegro A, Miller W (2014) Khoisan hunter-gatherers have been the largest population throughout most of modern-human demographic history. Nat Commun 5:5692. doi:10.1038/ncomms6692

147. Nei M, Li W-H (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci USA 76(10):5269-73. doi:10.1073/pnas.76.10.5269

148. Tajima F (1989) Statistical method for testing neutral mutation hypothesis by DNA polymorphisms. Genetics 123:585-595.

149. Przeworski M, Hudson RR, Rienzo AD (2000) Adjusting the focus on human variation. Trends Genet 16(7):296-302. doi:10.1016/S0168-9525(00)02030-8

150. 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

151. Ardlie KG, Kryglyak L, Seielstad M (2002) Patterns of linkage disequilibrium in the human genome. Nat Rev Genet 3:299-309. doi:10.1038/nrg777

152. Keinan A, Clark AG (2012) Recent explosive human population growth has resulted in an excess of rare genetic variants. Science 336:740-743. doi:10.1126/science.1217283

153. Nelson MR, et al (2012) An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science 337:100-104. doi.10.1126/science.1217876

154. Rafajlovic M (2012) Genetic variation in structured populations. Licentiate thesis, Department of Physcis, Gothenburg University.

155. Nei M (1973) Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci USA 70:3321-3323. doi:10.1073/pnas.70.12.3321

156. Slatkin M (1995) A measure of population subdivision based on microsatellite allele frequencies. Genetics 139:457-462.

157. Goldstein DB, Linares AR, Cavalli-Sforza LL, Feldman MW (1995) An evaluation of genetic distances for use with microsatellite loci. Genetics 139:463-471.

158. Slatkin M (2008) Linkage disequilibrium - understanding the evolutionary past and mapping the medical future. Nat Rev Genet 9:477-485. doi:10.1038/nrg2361

159. Pritchard JK, Przeworski M (2001) Linkage disequilibrium in humans: Models and data. Am J Hum Genet 69:1-14. doi:10.1086/321275

160. Schaper E, Eriksson A, Rafajlovic M, Sagitov S, Mehlig B (2012) Linkage disequilibrium under recurrent bottlenecks. Genetics 190:217-229. doi:10.1534/genetics.111.134437

161. Tenesa A, Navarro P, Hayes BJ, Duffy DL, Clarke GM, Goddard ME, Visscher PM (2007) Recent human effective population size estimated from linkage disequilibrium. Gen Res 17:520-526. doi:10.1101/gr.6023607

162. Park L (2011) Effective population size of current human population. Genet Res Cambr 93:105-114. doi:10.1017/S0016672310000558

163. Devlin B, Risch N (1995) A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 29:311-322. doi::10.1006/geno.1995.9003

164. Thomas D (2004) Statistical Methods in Genetic Epidemiology. Oxford University Press (New York).

165. Hudson RR (2001) Linkage disequilibrium and recombination. In: Balding D, Bishop M, Cannings C, eds. Handbook of Statistical Genetics. Wiley and Sons (New York):pp. 309-322.

166. Stephens M, Donnelly D (2000) Inference in molecular population genetics. J R Stat Soc Ser B 62(4):605-655. doi:10.1111/1467-9868.00254

167. Li N, Stephens M (2003) Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide data. Genetics 165:2213-2233.

168. Kuhner MK (2008) Coalescent genealogy samplers: windows into population history. Trends Ecol Evol 24(2):86-93. doi:10.1016/j.tree.2008.09.007

169. Chang JT (1999) Recent common ancestors of all present-day individuals. Adv Appl Probab 31:1002-1026. doi:10.1017/S0001867800009587

170. Derrida B, Manrubia SC, Zannette DH (2000) On the genealogy of a population of biparental individuals. J Theor Biol 203:303-315. doi:10.1006/jtbi.2000.1095

171. Blackwell D, MacQueen JB (1973) Ferguson distributions via Polya urn schemes. Ann Statist 1:353-355. doi:10.1214/aos/1176342372

172. Hössjer O, Olsson F, Laikre L, Ryman N (2015) Metapopulation inbreeding dynamics, effective size and subpopulation differentiation - a general analytical approach for diploid organisms. Theor Popul Biol 102:40-59. doi:10.1016/j.tpb.2015.03.006

173. Wiuf C (2000) A coalescence approach to gene conversion. Theor Popul Biol 57:357-367. doi:10.1006/tpbi.2000.1462

174. Wiuf C, Hein J (2000) The coalescent with gene conversion. Genetics 155: 451-462.

175. Hein J, Schierup MH, Wiuf C (2005) Gene Genealogies, Variation and Evolution. Oxford University Press (Oxford).

176. Takahata N, Nei M (1990) Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. Genetics 124:967-978.

177. Kimura M (1980) A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111-120. doi:10.1007/BF01731581

178. Felsenstein J (2009) Theoretical Evolutionary Genetics. Genome 562, Dept. of Genome Sciences and Dept. of Biology, University of Washington (Seattle).

179. Lanave C, Preparata G, Saccone C, Serio G (1984) A new method for calculating evolutionary substitution rates. J Mol Evol 20:86-93. doi:10.1007/BF02101990

180. Nielsen R, Hubisz MJ, Clark AG (2004) Reconstructing the frequency spectrum of ascertained single-nucleotide polymorphism data. Genetics 168:2373-2382. doi:10.1534/genetics.104.031039

181. Albrechtsen A, Nielsen FC, Nielsen R (2008) Ascertainment biases in SNP chips affect measures of population divergence. Mol Biol Evol 24(1):1-20. doi:10.1093/molbev/msl161

182. Dawkins R (1973) The Selfish Gene. Oxford University Press (New York).

183. Dawkins R (1996) Climbing Mount Improbable. Norton (New York).

184. Kimura M (1983) Neutral Theory of Molecular Evolution. Cambridge University Press (New York). doi:10.1017/CBO9780511623486

185. 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

186. Sanford JC, Brewer W, Smith F, Baumgardner J (2015) The waiting time problem in a model hominin population. Theor Biol Med Model 12(18):28 pages. doi:10.1186/s12976-015-0016-z

187. Sabeti PC, et al (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449:913-919. doi:10.1038/nature06250

188. ReMine WJ (1993) The Biotic Message. Evolution Versus Message Theory. St. Paul Science Publishers (Saint Paul, Minnesota).

189. Lynch M, Conery J, Bürger R (1995) Mutation accumulation and the extinction of small populations. Am Nat 146(4):489-518. doi:10.1086/285812

190. Sanford JC (2008) Genetic Entropy and the Mystery of the Genome, 3rd edition. FMS Publications (Waterloo, New York).

191. 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

192. Gauger A, Axe D, Luskin C (2012) Science and Human Origins: Discovery Institute Press (Seattle).

193. Keuffer R, et al (2007) Unexpected heteroygosity in an island mouflon population founded by single pair of individuals. Proc R Soc B 274:527-533. doi:10.1098/rspb.2006.3743

194. Hudson RR, Kaplan NL (1988) The coalescent process in models with selection and recombination. Genetics 120:831-840.

195. Kaplan NL, Darden T, Hudson RB (1988) The coalescent process in models with selection. Genetics 120:819-829.

196. Barton NH, Etheridge AM (2004) The effect of selection on genealogies. Genetics 166:1115-1131. doi:10.1534/genetics.166.2.1115