On the Number of Ancestors to a DNA Sequence

Carsten Wiuf and Jotun Hein

Institute of Biological Sciences, University of Aarhus, DK-8000 Aarhus, Denmark

Manuscript received March 3, 1997
Accepted for publication August 12, 1997

ABSTRACT

If homologous sequences in a population are not subject to recombination, they can all be traced back to one ancestral sequence. However, the rest of our genome is subject to recombination and will be spread out on a series of individuals. The distribution of ancestral material to an extant chromosome is here investigated by the coalescent with recombination, and the results are discussed relative to humans. In an ancestral population of actual size 1.3 million a minority of <6.4% will carry material ancestral to any present human. The estimated actual population size can be even higher, 5 million, reducing the percentage to 1.7%.

THE process of evolution of sequences subject to both coalescence and recombination in a population was first described by HUDSON (1983). In Hudson's setup a combined coalescence and recombination process is followed back in time until any nucleotide position in the extant sequences has only one ancestral nucleotide. The ancestral nucleotides can be located on different sequences. The process further back in time than this point has usually been of no interest, as it does not influence the relationship between the sequences in the sample.

However, it *does* determine the distribution of ancestral material to extant chromosomes on individuals, and raises some questions: (1) How many ancestors are there to a present human chromosome? An ancestor to an extent sequence is defined as a sequence carrying material ancestral to the extant sequence. The number of ancestors will thus vary with time and grow in mean until a steady state is reached. If the ancestral material to an extant chromosome are spread out on different individuals due to recombinations there will be more than one ancestor, but does this number of ancestors constitute a large or small part of humans that lived, say 300,000 years ago? (2) How many different sequences in an ancestral population can one sample by sequencing extant sequences? This is of importance when making assertions about species trees. A species tree can be inferred from a collection of gene trees of different genes. The number of ancestral chromosomes can be less than the number of genes, because ancestors to genes can be on the same ancestral chromosome.

The combined coalescence and recombination process is here studied beyond the point of most recent ancestral nucleotides by simulation and some analytical results are derived. The distribution of the process is determined only by the product $N_e r$, where N_e is the effective population size and r is the expected number of recombinations experienced by a sequence in one generation.

Main mathematical results for the process being in a steady state are (R denotes sequence length measured in expected number of recombinations per N_e generations) as follows:

1. The mean and variance of number of ancestral segments is $1 + R$ and approximately $2R$, respectively.
2. The mean length of a segment is 1.
3. The mean amount of ancestral material on the ancestor to the leftmost base on an extant sequence is $\log(1 + R)$.

A segment is a maximal interval of ancestral material on an ancestral sequence. Based on simulations it is conjectured that

4. The mean number of ancestor sequences is of the order $R/\log(1 + R)$.

The human chromosome *1* is ~263 Mb and 293 cM long (SCIENCE WEB PAGE 1997), and assuming the effective population of humans is $N_e = 20{,}000$, the mean segment length is ≈5000 (=263 Mb/(2.93·20,000)) bases long, and the length of ancestral material on the ancestor to the leftmost base is ≈50,000 [=263 Mb·log(1 + 2.93·20,000)/(2.93·20,000)] bases long.

If statement 4 holds true, the mean number of ancestors to the human chromosome *1* is ~6800 individuals.

This article is divided into four sections apart from the Introduction. The first section introduces the coalescent process with recombination and sequence evolution subject to recombination. In the second section mathematical results are derived from the point of view

Corresponding author: Carsten Wiuf, Department of Genetics and Ecology, Institute of Biological Sciences, University of Aarhus, Ny Munkegade, Building 540, DK-8000 Aarhus C, Denmark.
E-mail: wiuf@pop.bio.aau.dk

of discrete processes and results obtained by simulation studies are presented. The third section discusses the program used in the simulation studies, and finally the fourth section is a discussion. Here the results obtained in previous sections are applied to humans, and problems concerning the choice of model and factors included in the model are discussed relative to this application.

We estimate that the number of ancestors to the human genome is <86,000 chromosomes, which is <3.3% of the estimated entire ancestral chromosomes. This percentage is based on an actual ancestral population size of 1.3 million individuals 300,000 years ago (WEISS 1984).

EVOLUTION OF SEQUENCES SUBJECT TO RECOMBINATION

The model of a population of sequences subject to recombination is the following. Each sequence is L nucleotides long and recombination is assumed to occur to the right of a nucleotide. The population is constantly of size N and diploid, *i.e.*, there are $2N$ sequences. A new generation is obtained from the present by sampling $2N$ sequences in the previous generation with replacement, and forming random pairs of sequences and letting the pairs recombine between any two nucleotides with probability r. Time will start at the present and increase going backward in time.

This process is transformed to a continuous time and continuous sequence process by letting $N \rightarrow \infty$ and measuring time in $2N$ generations; letting $L \rightarrow \infty$ and $r \rightarrow 0$, such that $2rLN \rightarrow R$; and measuring length in expected number of recombinations per $2N$ generations. HUDSON (1983) showed that the waiting time to a sequence having been created by a recombination from two sequences is exponentially distributed with intensity parameter R_0, where R_0 is the rescaled length of the sequence. For the extant sequences, R_0 is simply the length of the sequences, *i.e.*, $R_0 = R$. For ancestral sequences, R_0 is the length of the interval spanned by regions that have ancestral material. These sequences can include regions with nonancestral material. The recombination point will be uniformly distributed within this material. The waiting time going backward in time until k sequences had only $(k - 1)$ ancestors in the population is exponentially distributed with intensity parameter $k(k - 1)/2$, and the two sequences having a common ancestor at that time are uniformly distributed among all different pairs. This was first realized by WAT-

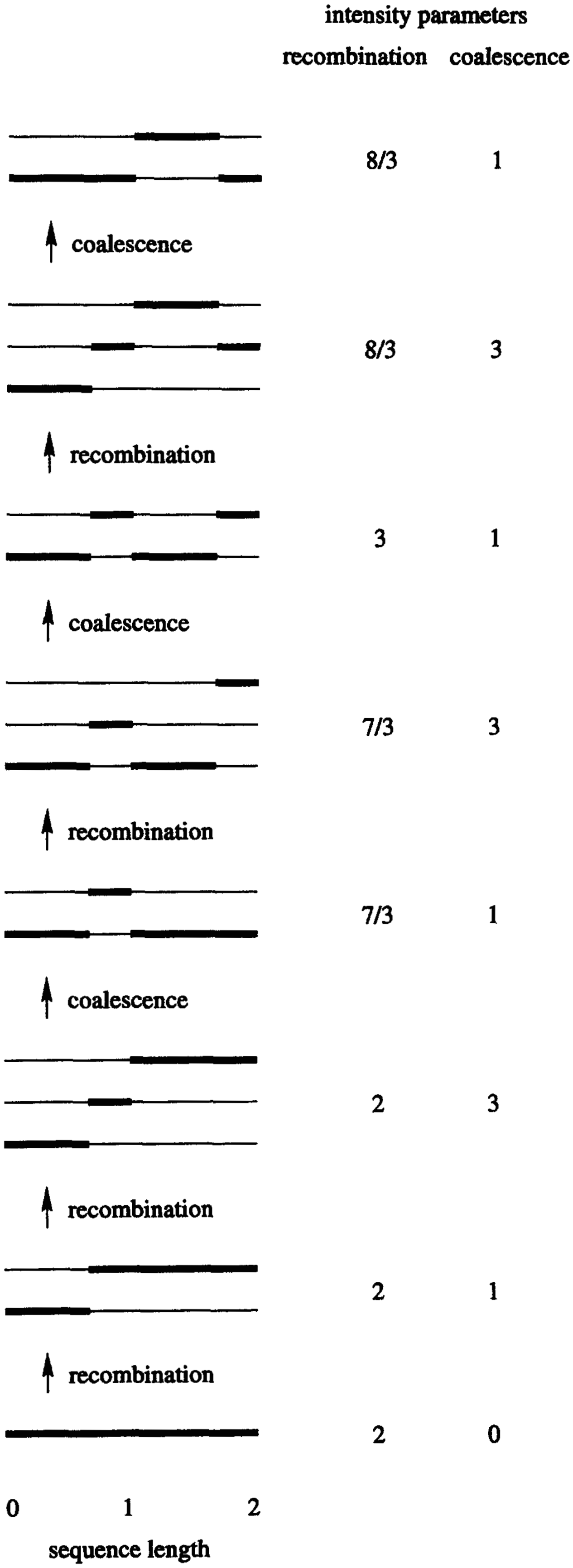


FIGURE 1.—Reshuffling of ancestral material. At time 0 one sequence is sampled. After the first recombination event, the sequences wait for either a recombination or a coalescent to occur. Each is associated an exponential waiting time and are independent of each other. The first coalescent traps some nonancestral material. The rate of recombination is the sum of the length of regions spanned by segments (including trapped material), and the rate of coalescence is $k(k - 1)/2$, where k denotes the number of sequences carrying ancestral material. For example, the rate of recombination after the second coalescent event is the length of segments plus the length of trapped material: $^2/_3 + {}^2/_3 + {}^1/_3 + {}^1/_3$ (segments) $+ {}^1/_3 + {}^2/_3$ (trapped material) $= 3$. Since the rate of coalescence is quadratic in k and rate of recombination is at most linear (less than $2k$ in this example), the process will reach a steady state and the number of segments and sequences will not increase indefinitely.

TERSON (1975) and later developed into the theory of the coalescent by KINGMAN (1982).

The coalescent with recombination has further been investigated by HUDSON and KAPLAN (1985), KAPLAN and HUDSON (1985), GRIFFITHS and MARJORAM (1996, 1997).

This process has some nice properties. First, the process is invariant under translations along the sequence, *i.e.*, the marginal distribution of a subsequence depends on the length of the subsequence only, not the position. Second, the process is symmetric, *i.e.*, the distribution is the same seen from either end points. Third, the distribution of m fixed points on the sequence is identical to the distribution of m loci in an m-locus model with recombination rates between loci given by the distances between points.

The history of a sequence can be simulated by going back in time, waiting for what occurs first, a recombination or a coalescence, and then performing the appropriate operation on the set of ancestral sequences. Recombination will increase the number of sequences carrying ancestral material by one, but will not increase the total amount of ancestral material. A coalescence will decrease the number of sequences with ancestral material by one. It can increase the amount of material, where recombination can occur, because coalescence can trap some nonancestral material. When any position on the extant sequences has found one ancestor, all segments with ancestral material spliced together will constitute one sequence. Above this point coalescence cannot reduce the amount of ancestral material and all that will occur is redistributions of ancestral material on different sequences by recombinations and coalescences. Since the rate of coalescences is quadratic in the number of sequences and the rate of recombinations is at most linear, the process will reach a steady state and not increase indefinitely. This is illustrated in Figure 1.

The distribution of ancestral material on different sequences can be classified and counted as follows. Figure 2 illustrates a simple situation. Start leftmost on the sequences. The sequence with the leftmost segment of ancestral material is labeled 1; the sequence with the second leftmost segment is labeled 2. If the third segment is not on the same segment as the first segment, it will be labeled 3. If it is on the same sequence as the first, it will be labeled 1. As the sequences carrying ancestral material are traversed and a segment on a new sequence is encountered, this sequence will be labeled by an integer one higher than any previously used. These numbers will in the sequel be referred to as sequence numbers.

The number of ways to distribute n segments on k sequences in this way will be called $C(n, k)$. This number is a reminiscent of the Stirling numbers of the second kind, $S(n, k)$, which is the number of ways to partition $\{1, \ldots, n\}$ into k subsets. The difference between configurations and partitions is that in a configuration two consecutive numbers cannot be in the same set, while partitions are not subject to this restriction. If a configuration did not obey this restriction, then two consecutive segments could be on the same sequence and then they would be one segment, not two. $S(n, k)$ and $C(n, k)$ obey very similar recursions:

$$S(n, k) = kS(n-1, k) + S(n-1, k-1),$$

$$S(n, n) = S(n, 1) = 1,$$

$$C(n, k) = (k-1)C(n-1, k) + C(n-1, k-1),$$

$$C(n, n) = C(n, 2) = 1.$$

From these recursions it is seen that $C(n+1, k+1) = S(n, k)$, *i.e.*, the number of ways to distribute $(n+1)$ segments on $(k+1)$ sequences is the same as the number of ways to partition n labeled elements into k sets.

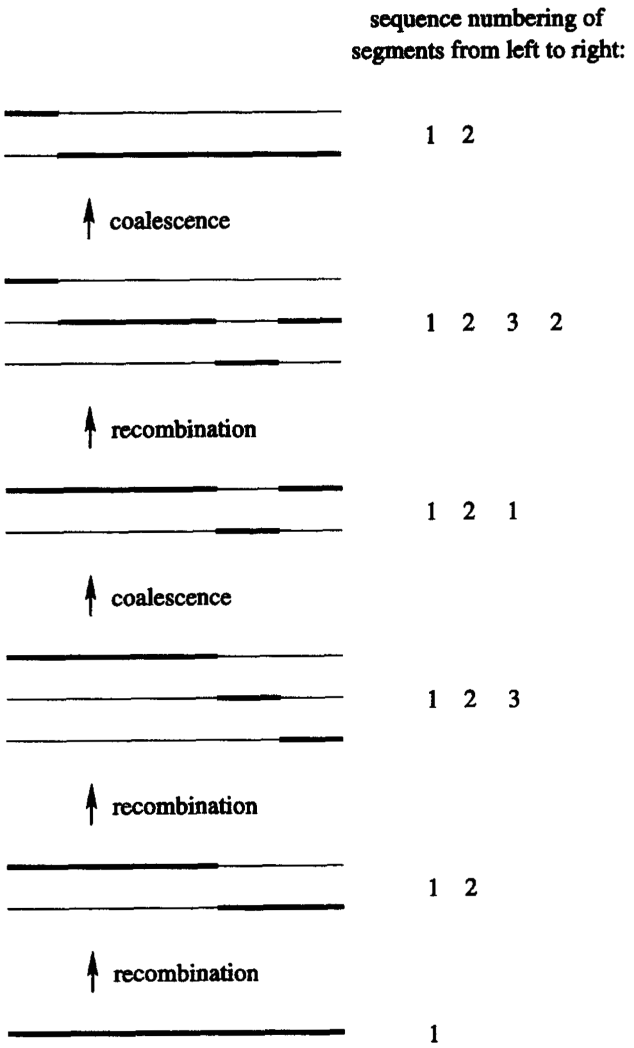


FIGURE 2.—Sequence numbering of segments. An example of the evolution of the distribution of material ancestral to a sequence. The first event (going back in time) is a recombination implying that just before this event material ancestral to the present sequence was located on two sequences, one sequence having trapped material. The amount of ancestral material remains constant.

RESULTS

In this section both results derived analytically as well as results obtained by simulations will be presented. These fall in two groups. The first group consists of results related to the sample of ancestor sequences, *e.g.*, the number of segments of ancestral material and number of ancestor sequences. The second group of results is related to the structure of a single ancestor, *e.g.*, length of ancestral material and number of segments on a single ancestor.

The coalescent process with recombination is a Markov process with state space given by all possible configurations of ancestral material and the multiplicity of sequences of each configuration. A strict mathematical formulation of the state space can be found in Griffiths and Marjoram (1996). Only results concerning the equilibrium distribution of the Markov chain will be presented and henceforth P will refer to this distribution.

No closed expression for the equilibrium distribution has been found and hence other approaches are required. The coalescent process with recombination can be embedded in a birth and death process with rates $\mu_k = k(k-1)/2$ and kR, $k = 1, 2, \ldots$, stressing once again that a steady state will be reached. The birth and death process ignores the fact that recombination events outside ancestral and trapped material will have no effect on the history of the sample. Trapped material is a segment located between two segments of ancestral material on a sequence (Figure 2).

The approach used here is to study the process through the distributional behavior of two, three and four points in the ancestral material. This limits the results to results on moments.

Define by A_t, $t \geq 0$, the sequence number at the point t (Figure 2), and let N_t be defined by $N_t = 1$ iff $\lim_{\epsilon \to 0}(A_{t-\epsilon} - A_{t+\epsilon}) \neq 0$ and $N_t = 0$ otherwise, *i.e.*, $N_t = 1$ iff t is a recombination point. Since the distribution of ancestral material in any interval of finite length is translation invariant, the process $\{N_t | t \geq 0\}$ is a stationary point process, whereas $\{A_t | t \geq 0\}$ is neither stationary nor a point process, since the value of A_t depends on all A_s, $0 \leq s \leq t$.

Let $[0, R]$ denote a sequence of length R, and let t_1 and t_2 be two points in $[0, R]$ with a distance $r = |t_1 - t_2|$ between them. Write A_i short for A_{t_i}. The two points will be on the same sequence with probability $1/(1 + r)$ and on different sequences with probability $r/(1 + r)$, *i.e.*,

$$P(A_1 = A_2) = 1 - P(A_1 \neq A_2) = \frac{1}{1+r}. \qquad (1)$$

Similarly one obtains for the case of three points $t_1 < t_2 < t_3$ in $[0, R]$ and distances $r_1 = t_2 - t_1$ and $r_2 = t_3 - t_2$ (Figure 3):

$$\begin{aligned}
P(A_1 = A_2 = A_3) &= \frac{1}{(1+r_1)(1+r_2)} + \frac{r_1 r_2}{K(r_1, r_2)}, \\
P(A_1 = A_2 \neq A_3) &= \frac{r_2}{(1+r_1)(1+r_2)} - \frac{r_1 r_2}{K(r_1, r_2)}, \\
P(A_1 \neq A_2 = A_3) &= \frac{r_1}{(1+r_1)(1+r_2)} - \frac{r_1 r_2}{K(r_1, r_2)}, \qquad (2) \\
P(A_1 = A_3 \neq A_2) &= \frac{r_1 r_2 (2 + r_1 + r_2)}{K(r_1, r_2)}, \\
P(A_i \neq A_j,\ i \neq j) &= \frac{r_1 r_2 (2 + r_1 + r_2)}{(1+r_1)(1+r_2)(3+r_1+r_2)},
\end{aligned}$$

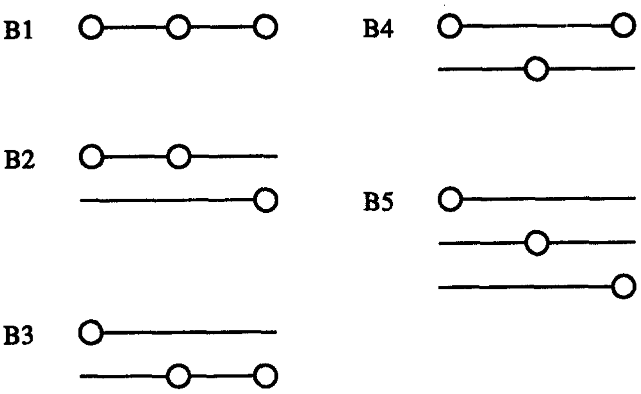


Figure 3.—The five different locations of three points. B1, $A_1 = A_2 = A_3$; B2, $A_1 = A_2 \neq A_3$; B3, $A_1 \neq A_2 = A_3$; B4, $A_1 = A_3 \neq A_2$; B5, $A_i \neq A_j$, $i \neq j$.

where

$$K(r_1, r_2) = (1 + r_1)(1 + r_2)(1 + r_1 + r_2)(3 + r_1 + r_2).$$

Both (1) and (2) can be obtained from two- and three-locus theory. For more than three points closed expressions are long and hard to derive even in special cases using programs like *Maple V* (Char *et al.* 1991). In the case of four points a single expression is required. Assume $t_1 < t_2 < t_3 < t_4$ and that the distance r_1 between t_1 and t_2 equals the distance between t_3 and t_4. Let r_2 be the distance between t_2 and t_3. For small values of r_1

$$\begin{aligned}
P(A_1 \neq A_2, A_3 \neq A_4) &= r_1^2\left(1 + \frac{2r_2^2 + 10r_2 + 9}{(2r_2^2 + 13r_2 + 9)(3 + r_2)(1 + r_2)}\right) \\
&\quad + O(r_1^3). \qquad (3)
\end{aligned}$$

In the sequel (1), (2) and (3) will be used several times to derive quantities related to the coalescent with recombination.

Put $\epsilon_n = 2^{-n}$, $n \in \mathbf{N}$, and define X_{in} and Z_{in}, $i \in \mathbf{N}$ by

$$X_{in} = 1_{\{A_{(i-1)\epsilon_n} \neq A_{i\epsilon_n}\}} \quad \text{and} \quad Z_{in} = 1_{\{A_0 = A_{i\epsilon_n}\}}.$$

The distributions of X_{in} and Z_{in} are found using (1). The number of segments S_R in $[0, R]$ can be defined through the X_{in}'s:

$$S_R = 1 + \lim_{n \to \infty} \sum_{i=1}^{R_n} X_{in},$$

with $R_n = 2^n R$, and similarly the length L_R of all ancestral material in $[0, R]$ located on the sequence containing the point $t = 0$ can be defined through the Z_{in}'s:

$$L_R = \lim_{n \to \infty} \epsilon_n \sum_{i=1}^{R_n} Z_{in}.$$

The above expression of S_R, (1) and (3) fairly easily lead to the following proposition:

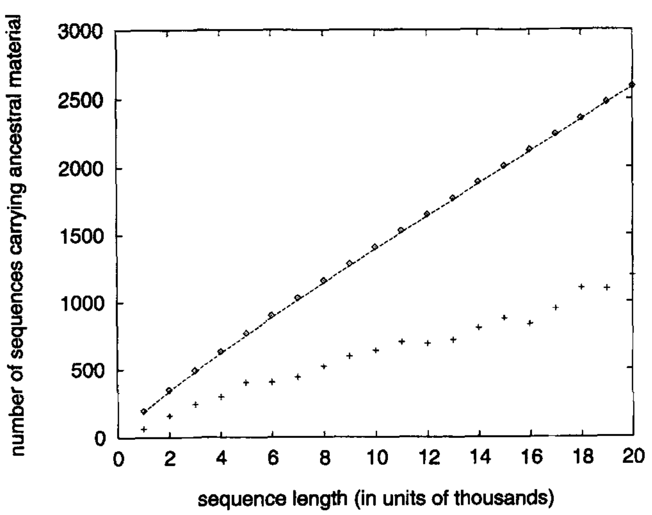


FIGURE 4.—Moments of the number of ancestor sequences C_R. The sequence length varies from 1000 to 20,000, and the number of simulations performed was 100. ◇ denotes the mean value, and +, the value of the variance. There is no theoretical justification for the choice of curve; this is only meant to show the order of magnitude. The curve fitted to the mean values is $f(x) = 1.28x/\log(1 + x)$ [1.28 is the third term in the expression of K (*Proposition 1*)].

Proposition 1: The mean and the variance of the number of segments S_R in $[0, R]$ *are given by*

$$E(S_R) = 1 + R,$$

$$V(S_R) = (1 + K)R + O(\log(1 + R)),$$

with K defined by

$$K = \tfrac{3}{8}\log(2) - \tfrac{1}{2}\log(3) + \frac{51}{4\sqrt{97}}\coth^{-1}\left(\frac{13}{\sqrt{97}}\right) \approx 0.993.$$

Proof: See APPENDIX A. □

The linear bound on $V(S_R)$ in R implies by the Markov inequality that the number of segments will tend to infinity with probability one as R becomes large.

In contrast to the above it is difficult to evaluate the number of sequences C_R carrying ancestral material. This number is given by $C_R = \max_{t\leq R} A_t$, hence it depends possibly on all values of A_t, which makes it impossible to come up with an expression of $E(C_R)$ based on quantities related to two, three and four points. Thus the mean and variance of C_R was simulated for different values of R (Figure 4), and in Figure 5 the number C_R is compared to the actual observed sequence number in the sequence end point $t = R$.

From Figure 4 it is seen that the increase in number of sequences as the length of the sequence increases one unit is of order $1/\log(R)$, hence the number of sequences grows slowly with length.

The variance of C_R is less than the mean. Intuitively, the difference in dispersion between the number of segments and the number of sequences can be explained with reference to possible fluctuations in these numbers. The number of segments is subject to higher fluctuations than the number of sequences, since a single coalescent event can reduce the number of segments by any number, even zero (Figure 1), whereas a recombination event will create one new segment, if the recombination happens with ancestral material, and zero new segments, if it happens within trapped material. The number of sequences will either be decreased by one (coalescent event), or increased by one (recombination event), and thus subject to less fluctuations.

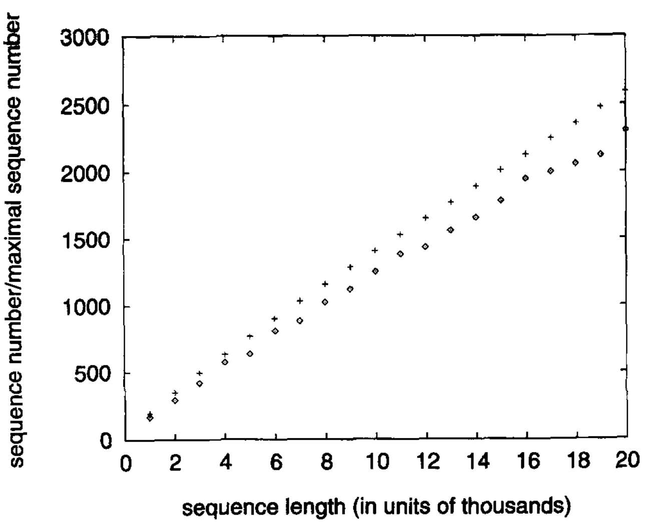


FIGURE 5.—Actual sequence number A_R *vs.* the total number of sequences C_R. The actual sequence number is the sequence number at the endpoint of the sequence, *i.e.*, the value of A_t in the point $t = R$. One hundred simulations were performed with sequence length varying from 1000 to 20,000. + denotes the mean number of sequences, and ◇ denotes the mean of the actual sequence number. The actual number is between 10–15% less than the total, which shows that a return to sequences with small sequence numbers rarely occurs.

Proposition 2: The mean and the variance of the length L_R of all ancestral material in $[0, R]$ *located on the sequence containing the point* $t = 0$ *are given by*

$$E(L_R) = \log(1 + R),$$

$$V(L_R) = (\log(1 + R))^2 + O(\log(1 + R)).$$

Proof: See APPENDIX A. □

The mean value of L_R suggests that the number of ancestral sequences C_R is not less than $R/\log(1 + R)$ (length of ancestral material divided by mean length of L_R that probably is larger than mean length of ancestral material on other ancestor sequences) and this order of magnitude is in fact confirmed in Figure 4.

Note that doubling the sequence length only increases the mean length by ≈ log(2), which is <1. Since mean segment length is 1 (*Proposition 5* below), the chance is low of a segment on sequence 1 in the interval $[R, 2R]$. Moreover *Proposition 3* can be interpreted in the following sense. If there is a segment there will be a whole battery of segments adjacent to it, similar to the battery of segments adjacent to the first segment labeled one.

Proposition 3: The conditional mean length L_R of all ancestral material in $[0, R]$ *on the sequence containing* $t = 0$ *given* $A_R = 1$ *(i.e., given* $t = 0$ *and* $t = R$ *are on the same sequence) is*

$$E(L_R \mid A_R = 1)$$
$$= 2\,\frac{1+R}{3+R}\log(1+R) + \frac{R}{3+R} \approx 2\log(1+R),$$

and for all $\epsilon > 0$

$$\lim_{R\to\infty} \frac{L_R}{R^\epsilon} = 0 \quad \text{in } L^2\text{-norm.}$$

Proof: See APPENDIX A. □

The second result in *Proposition 3* gives an upper bound on how fast the sequence length L_R can grow. It will grow less rapidly than any root of R, and hence segments cannot be uniformly spread with increasing sequence length. Simulation results (not shown) indicate that the distribution of $l_R = L_R/\log(1 + R)$ fulfilling $E(l_R) = 1$ and $V(l_R) \approx 1$ tends to an exponential distribution with parameter 1. If that is so, L_R grows like $\log(1 + R)$.

Denote by S^*_R the number of segments ending in $[0, R]$ on the sequence containing $t = 0$, *i.e.*,

$$S^*_R = 1 + \lim_{n\to\infty} \sum_{i=1}^{R_n} X_{in} Z_{in}.$$

Then

Proposition 4: The following are true for S^*_R*:*

$$P(S^*_{2R} - S^*_R = 0) > 1 - \log(2) \approx 0.307 \quad \text{for all } R,$$

$$P(\limsup_{R\to\infty} \{S^*_{2R} - S^*_R = 0\}) > 1 - \log(2),$$

and

$$P(\lim_{R\to\infty} S^*_R = \infty) > 0.$$

Proof: See APPENDIX A. □

Despite that the mean sequence length L_R presumably is growing logarithmically in R, there is a positive chance (> ≈0.307) that sequence 1 is not represented in an arbitrary large region of the ancestral material and that there will be infinitely many such regions (*Proposition 4*). The last statement gives a converse to the second statement: there is a positive chance that there will be an infinite number of segments labeled one.

If the chance of returning to the first sequence is significant, it should be reflected in the distribution of trapped material: the higher this chance, the higher is the probability of long pieces of trapped material. Figure 6 shows the simulated distribution of trapped material in the sequence labeled one.

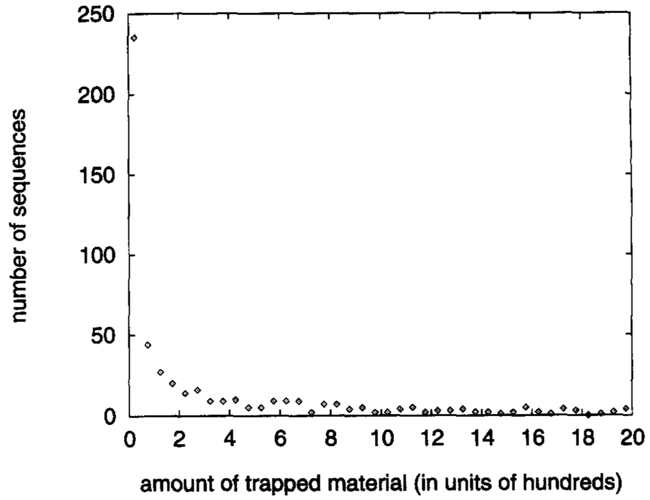


FIGURE 6.—Trapped material. The empirical distribution of trapped material located on the sequence containing the point $t = 0$. R was chosen to 2000, and 500 simulations were performed. Seventy-five of these simulated sequences consisted of only one segment, and ~50% (including the 75) had very little amount of trapped material, *i.e.*, the segments were very close to each other. The distribution has a long tail, and four sequences had trapped material of size 1975 or more.

Proposition 5: Let Λ *be the length of a segment measured from* $t = 0$*. Conditional on the event* $\{N_0 = 1\}$*, the mean value of* Λ *is*

$$E(\Lambda \mid N_0 = 1) = 1.$$

Proof: See APPENDIX A. □

Figure 7 sums up the structure of a single ancestor as described in *Propositions 2–5* and Figure 6 simply by showing an ancestor to an extant sequence. Segments tend to be placed in small batteries.

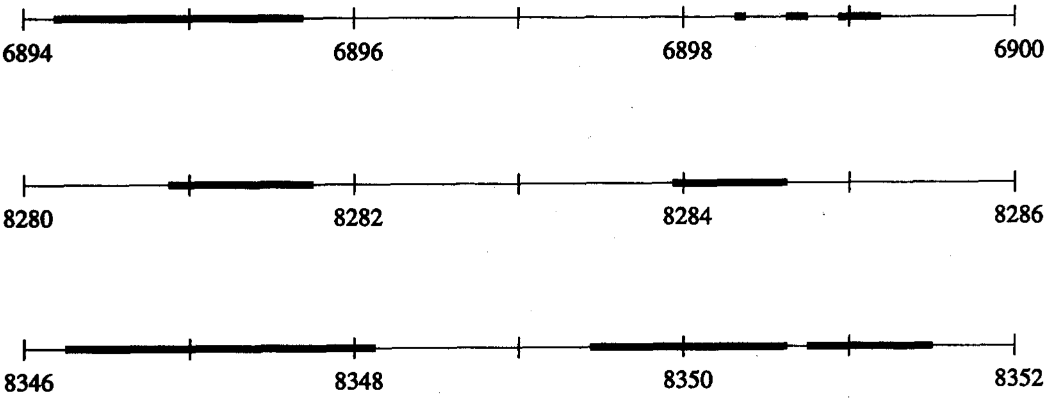
SIMULATIONS

A program was written in the computer language C that simulated the evolution of a sequence going backward in time. It tabulates the empirical distribution function of selected statistics of the process, and takes three input parameters:

1. the length of the sequence, R in expected number of recombinations per N generations,
2. the amount of time of evolution, T in units of N generations,
3. the number of simulations to perform.

A few criteria were chosen to measure the divergence from equilibrium:

1. The mean number of segments. Since waiting times until recombination events in ancestral material (not trapped) occur are independent and exponentially distributed with intensity R, the waiting time until



1. group	2. group	3. group
[6894.18, 6895.71]	[8280.88, 8281.74]	[8346.23, 8348.15]
[6898.32, 6898.36]	[8283.91, 8284.60]	[8349.42, 8350.67]
[6898.64, 6898.72]		[8350.73, 8351.46]
[6898.91, 6899.15]		

FIGURE 7.—Example of a sequence of length 20,000 with ancestral material. The sequence is divided into three groups of four, two and three segments, respectively. The groups are spread in the entire ancestral material with a distance between the two first groups of size ~1400. There are both very small segments (0.04 in the first group) and larger segments (1.92 in the third group). Compared to $\log(1 + R) \approx 9.90$, this is a normal sequence of length 7.34 of ancestral material, and the amount of nonancestral material is 1449.94.

the nth recombination event is distributed as $\Gamma(R, n)$. Moreover the waiting time until neighbor segments coalesce is exponential with intensity 1, hence the mean number of segments $S_R(t)$ present t time units back is

$$E(S_R(t)) = 1 + \sum_{n=1}^{\infty} \int_0^{\infty} \exp\{-(t - x) - Rx\} \frac{R^n x^{n-1}}{(n-1)!} dx$$
$$= E(S_R) - R\exp(-t),$$

i.e., the convergence to the equilibrium mean value is negative exponential.

2. The distribution at time t of three points. There are five possible configurations of three points (Figure 3), and if the process starts in the configuration with all points on the same sequence, the equilibrium distribution is approached with an error of order $\exp(-t(1 + r))$, where $r = \min(r_1, r_2)$, and r_1 and r_2 are distances between neighboring points (see *e.g.*, KEILSON 1979).

DISCUSSION

The questions addressed in this article are relevant for at least two reasons. First, it addresses on how many sequences ancestral material to present chromosomes were located. Obviously, there was not one ancestral sequence (Figure 4), but a series of ancestral sequences, carrying different amounts of ancestral material in different sized segments and with different number of segments. Second, it is of interest to know how many sequences one could sample in an ancestral population by sequencing extant sequences. This is of importance when making assertions about the dynamics of ancestral populations.

The model of the coalescent with recombination as discussed in this article is based on a number of assumptions, not all of which are realistic for natural populations including humans. Major assumptions are as follows: (1) constant population size, N, (2) no geographical subdivisions, (3) no selection. All results are derived under the further assumption of (4) the process being in steady state. In natural populations the violation of one or more of the above assumptions will often invalidate the coalescent with recombination as a reasonable description of a sample's history and genealogy. However, it is predictable what qualitative effects violations of the assumptions will have.

To answer the questions raised in the beginning of this section a discussion of the human ancestral population is required. Recently a debate of the origin and history of modern humans has flourished, initiated by the question of the homeland of mitochondrial common ancestor (CANN *et al.* 1987). A consensus seems to be reached that the species *Homo sapiens* originated somewhere in Africa and spread from there to the rest of the world (ROGERS 1995). It is here assumed the spread happened 100,000 years ago.

Before this date it is assumed that the effective population size was approximately constant for such a long time that the premodern human population can be assumed to have been in equilibrium. TAKAHATA *et al.* (1995) and ROGERS (1995) argue in favor of this and estimate the effective population size. Effective population sizes are in the range of 1500–7000 breeding females.

Based on the above it is assumed that assumptions 1, 2 and 4 are sufficiently fulfilled for the human popula-

TABLE 1

Percentage of ancestors in the ancestral population

Size of ancestral population (in millions)	All chromosomes: Percentage of ancestral chromosomes	All chromosomes: Percentage of ancestral individuals	Chromosome *20*: Individuals carrying two ancestral chromosomes	Chromosome *20*: *P*
0.5	8.3	15.9	3.4	0.03
1.3	3.3	6.4	1.3	0.27
5.0	0.9	1.7	0.3	0.71

The second and third columns show the numbers in percent of ancestral chromosomes and ancestral individuals, respectively, carrying material ancestral to one or more chromosome *1–22* of an extant individual. Column three shows the mean number of ancestral individuals with two chromosomes carrying material ancestral to an extant chromosome *20*. Column four shows the probability (*P*) that no ancestral individual carries more than one chromosome with ancestral material to an extant chromosome *20*. Algebraic expressions of the mean and the probability can be found in APPENDIX B.

tion for a sufficiently long period. Assumption 2 is possibly violated due to a rather small population density and a large populated area (Weiss 1984). Finally assumption 3 cannot be completely true, but this will be neglected here.

The ancestral material of an extant sequence was in previous generations distributed on a series of different sequences. One hundred thousand years ago this number is unknown and cannot be predicted by the coalescent with recombination due to violations of one or more of the assumptions 1–4, *e.g.*, the population is increasing. Since the population before 100,000 years ago fulfills assumption 1–4, the coalescent with recombination applies to the further history back in time to the ancestor sequences.

Using the information from Science Web Page (1997) on chromosomal length measured in crossovers, the first question raised can be answered. Counting the number of females as 5000, the effective sequence size in the diploid population is 20,000, and hence the rate of recombination of the human chromosome *20* is approximately $R = 2rLN = 20{,}000$ (length of chromosome *20* is 100 cM). Figure 4 supports the conclusion that the mean number of ancestor sequences is ~2600. This means that material ancestral to an extant human chromosome *20* was spread out on 2600 chromosomes, or the number of ancestors to chromosome *20* is ~2600.

The number of ancestors for human chromosomes will vary with the length of the chromosome. These range from ~58 to ~293 cM (Science Web Page 1997), and hence the number of ancestors varies from 1600 to 6800 chromosomes in the ancestral population.

Summing up, all ancestors on different chromosomes yield an upper bound of size 86,000, with the lower bound and far less probable being 6800. These numbers should be compared to the actual physical population size, say ~300,000 years ago, which is estimated to be ~1.3 million individuals (Weiss 1984). This number could both be too high and too low, but should be contrasted to the relatively low effective population size. The percentage of ancestors is ≈3.3% of the chromosomes in ancestral physical population. Even 5 millions individuals could be a realistic population size, which reduces the percentage of chromosomal ancestors to ~0.9% of the ancestral chromosomes.

Some ancestral individuals will carry material ancestral to an extant chromosome on both chromosomes and some on one chromosome only. If the ancestral population size is large, the chance will be low that an individual carries two chromosomes with ancestral material. Table 1 shows the percentage of chromosomes and individuals carrying material ancestral to an extant chromosome, and the mean number of individuals with two chromosomes with ancestral material.

Second, it is of interest to know how many different sequences one could sample in an ancestral population by sequencing extant sequences. This is of relevance when attempting to reconstruct a species phylogeny. The time of speciation can be determined more accurately if more extant loci are available and the loci are not linked. In the situation with totally unlinked loci there will be as many ancestral sequences as loci, and in the situation with linked loci, there will be one ancestral sequence only. But in between these extremes the number of ancestral sequences to extant sequences are not that easily deduced. The number of extant sequences sequenced is of little importance, since most loci will find a common ancestor long before the time of speciation.

A sequence of length $R = 100$ has ~30 ancestors in the ancestral population, and this is certainly sufficient in examples with only three species, *e.g.*, humans, chimpanzees and gorillas. The long time span involved (in the mentioned example several million years) make the assumptions less trustworthy and should be kept in mind.

The assumptions 1–4 were taken for granted in the discussion. These assumptions could partly be justified, but it is of interest to know the answer to questions like the above under less restricted models.

In most cases it will be difficult to choose a model that describes extant sequences appropriately. Often it is possible to describe the variation in extant sequences by several models, without there being reasonable criteria to choose among the different models. The effect of introducing more factors can in some cases easily be predicted: a bottleneck will decrease the number of ancestral sequences, since the effective population size in the bottleneck period will be drastically lowered. Subdivision with migration will have the opposite effect, because the size of the effective population will be larger than in a population without geographical structure.

However, in all cases the population structure needs to be modeled. The choice of model in this paper is from the point of simplicity. In the application to human chromosomes the assumptions made seem to be approximately fulfilled, so that the results derived about the number of ancestors gives the order of magnitude of the "true" value.

We thank OLE CAPRANI and MORTEN LAURITZEN for help with implementation of the simulation program and BERNT GULDBRANDTSEN and ANNE-METTE KRABBE PEDERSEN for reading and commenting on the manuscript. J.H. was supported by Danish Research Council grant SNF 94-0163-1.

LITERATURE CITED

CANN, R. L., M. STONEKING and A. C. WILSON, 1987 Mitochondrial DNA and human evolution. Nature **325:** 31–36.

CHAR, B. W., K. O. GEDDES, G. H. GONNET, B. L. LEONG, M. B. MONAGAN *et al.*, 1991 *Maple V Reference Manual.* Springer-Verlag, New York.

DALEY, D. J., and D. VERE-JONES, 1988 *An Introduction to the Theory of Point Processes.* Springer-Verlag, New York.

GRIFFITHS, R. C., and P. MARJORAM, 1996 Ancestral inference from samples of DNA sequences with recombination. J. Comp. Biol. **3/4:** 479–502.

GRIFFITHS, R. C., and P. MARJORAM, 1997 An ancestral recombination graph, pp. 257–270 in *Progress in Population Genetics and Human Evolution, IMA volumes in Mathematics and its Applications,* 87, edited by P. DONNELLY and S. TAVARÉ. Springer-Verlag, Berlin.

HUDSON, R. R., 1983 Properties of the neutral allele model with intergenic recombination. Theoret. Popul. Biol. **23:** 183–201.

HUDSON, R. R., and N. KAPLAN, 1985 The use of sample genealogies for studying a selectively neutral *m*-loci model with recombination. Theoret. Popul. Biol. **28:** 382–396.

KAPLAN, N., and R. R. HUDSON, 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics **111:** 147–164.

KEILSON, J., 1979 *Markov Chain Models—Rarity and Exponentiality.* Springer-Verlag, New York.

KINGMAN, J. F. C., 1982 The coalescent. Stoch. Process. Appl. **13:** 235–248.

ROGERS, A. R., 1995 Genetic evidence for a Pleistocene population explosion. Evolution **49:** 608–615.

SCIENCE WEB PAGE, 1997 *A Gene Map of Human Genome.* http://www.ncbi.nlm.nih.gov/SCIENCE96/

TAKAHATA, N., Y. SATTA and J. KLEIN, 1995 Divergence time and population size in the lineage leading to modern humans. Theoret. Popul. Biol. **48:** 198–221.

WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theoret. Popul. Biol. **7:** 256–276.

WEISS, K. M., 1984 On the number of members of the genus Homo, who have ever lived and some evolutionary implications. Hum. Biol. **56:** 637–649.

Communicating editor: A. G. CLARK

APPENDIX A

This appendix contains proofs of propositions given in "Ancestors to an extant sequence."

Proof of Proposition 1: Note that $X_{in} \leq X_{(2i-1)(n+1)} + X_{2i(n+1)}$, hence $\Sigma_{i=1}^{R_n} X_{in}$ is increasing in n. Consider $E(S_R)$,

$$E(S_R) = 1 + E\left(\lim_{n\to\infty} \sum_{i=1}^{R_n} X_{in}\right) = 1 + \lim_{n\to\infty} \sum_{i=1}^{R_n} E(X_{in}),$$

where the second equality follows from the monotone convergence theorem. Using (1) and the definition of ϵ_n,

$$E(S_R) = 1 + \lim_{n\to\infty} \frac{R_n\epsilon_n}{1+\epsilon_n} = 1 + R.$$

The proof of the expression for $V(S_R)$ is quite similar to the above. It follows easily that

$$V(S_R) = R - R^2 + 2 \lim_{n\to\infty} \sum_{i<j} E(X_{in}X_{jn}),$$

and using (3)

$$V(S_R) = R - R^2 + 2 \lim_{n\to\infty} \sum_{k=1}^{R_n} (R - x_{kn})$$

$$\times \left(1 + \frac{2x_{kn}^2 + 10x_{kn} + 9}{(2x_{kn}^2 + 13x_{kn} + 9)(3 + x_{kn})(1 + x_{kn})}\right)\epsilon_n,$$

where x_{kn} is short for $k\epsilon_n$. Taking the limit of the sum

$$V(S_R) = R + 2 \int_0^R \frac{(R - x)(2x^2 + 10x + 9)}{(2x^2 + 13x + 9)(3 + x)(1 + x)}\, dx.$$

This integral can be solved by standard decomposition methods (but is very cumbersome) and results in

$$V(S_R) = (1 + K)R + O(\log(1 + R)).$$

This completes the proof. □

Proof of Proposition 2: Note that $\epsilon_n \Sigma_1^{R_n} Z_{in} \leq \epsilon_n R_n = R$, and hence by Lebesgue's dominated convergence theorem (LDCT)

$$E(L_R) = \lim_{n\to\infty} \epsilon_n \sum_{i=1}^{n} E(Z_{in}).$$

Inserting (1) results in

$$E(L_R) = \lim_{n\to\infty} \epsilon_n \sum_{i=1}^{R_n} \frac{1}{1 + \epsilon_n i} = \int_0^R \frac{1}{1 + x}\, dx = \log(1 + R).$$

Concerning $V(L_R)$ note that

$$L_R^2 = 2 \lim_{n \to \infty} \epsilon_n^2 \sum_{i<j} Z_{in} Z_{jn} \le R^2,$$

and applying LDCT and (2) Equation 1

$$E(L_R^2) = 2 \int_0^R dx \int_0^{R-x} \frac{1}{(1+x)(1+y)}\, dy$$

$$+ \mathrm{O}(\log(1+R)) = 2 \log(1+R) \log(2+R)$$

$$+ O(\log(1+R)).$$

There is no exact solution to the integral. Subtracting $E(L_R)^2$ from $E(L_R^2)$ yields

$$V(L_R) = (\log(1+R))^2 + O(\log(1+R)),$$

and the proof is complete. □

Proof of Proposition 3: The proof of the first statement is similar to the proof of *Proposition 2.*

Note that for $\epsilon > 0$,

$$\lim_{R\to\infty} E\left(\frac{L_R}{R^\epsilon}\right)^2 = \lim_{R\to\infty} \frac{2 \log^2(1+R)}{R^{2\epsilon}} = 0$$

by *Proposition 2,* and hence L_R/R^ϵ converges in L^2 – *norm.* This completes the proof. □

Proof of Proposition 4: Similar to the first statement in *Proposition 1* it can be proved that

$$E(S_R^*)$$

$$= \log(1+R) + {}^3\!/\!_4 \log \frac{3+R}{1+R} + \frac{1}{2}\frac{2+R}{1+R} - {}^3\!/\!_4 \log(3).$$

Subtracting the means of S_{2R}^* and S_R^*,

$$E(S_{2R}^* - S_R^*) \le \log(1+2R) - \log(1+R) \le \log(2),$$

but

$$E(S_{2R}^* - S_R^*) \ge P(S_{2R}^* \ne S_R^*) = 1 - P(S_{2R}^* = S_R^*),$$

and hence the first statement is proved. The second follows from the first by Fatou's lemma.

Assume $P(\lim_{R\to\infty} S_R^* = \infty) = 0$, *i.e.,* $S_R^* < \infty$ *a.s.* From the remark below *Proposition 1:* $S_R = \infty$ *a.s.,* and hence if $S_R^* < \infty$ *a.s.* then $l_R = L_R/\log(1+R) \to 0$ *a.s.* (if $S_R < \infty$, then S_R^* could be finite with $l_R = \infty$). Both $E(l_R)$ and $V(l_R)$ are bounded in R therefore uniformly integrable and hence: $\lim_{R\to\infty} E(l_R) = E(\lim_{R\to\infty} l_R) = 0$, which contradicts $E(l_R) = 1$, and $P(\lim_{R\to\infty} S_R^* = \infty) > 0$ must be true. □

Proof of Proposition 5: The number of segments tends to infinity almost surely as R becomes large (the remark below *Proposition 1*) and the result follows from Daley and Vere-Jones (1988), theorem 3.4.II. □

APPENDIX B

Let M denote the size of the ancestral population, and K the number of chromosomes carrying material ancestral to an extant chromosome. The probability p that no ancestral individual carry two chromosomes with ancestral material is

$$p = \frac{2^K (M)_K}{(2M)_K}.$$

Divide the $2M$ chromosomes into M "first" chromosomes and M "second" chromosomes, such that each individual has exactly one first chromosome and one second. The probability $p(K_1, x)$ to pick K chromosomes of $2M$, such that K_1 are first chromosomes, K_2 are second, $K_1 + K_2 = K$, and such that in x cases the first and the second chromosome belong to the same individual is

$$p(K_1, x) = \frac{\binom{M}{K_1}\binom{M}{K_2}\binom{K_1}{x}\binom{M-K_1}{K_2-x}}{\binom{2M}{K}\binom{M}{K_2}},$$

$x \le K_1$, $x \le K_2$. The distribution of x, $x \le K$ is

$$p(x) = \frac{2^{K-2x} N!}{\binom{2M}{K} x!\,(K-2x)!\,(N-K+x)!}.$$

The variable x has mean value

$$E(x) = \frac{1}{2}\frac{K(K-1)}{2M-1}.$$

This number is less than one if $K \approx\, < 2\sqrt{M}$. □