International Scholarly Research Network
ISRN Biomathematics
Volume 2012, Article ID 384275, 8 pages
doi:10.5402/2012/384275

Research Article

The Ancestry of Genetic Segments

R. B. Campbell

Department of Mathematics, University of Northern Iowa, Cedar Falls, IA 50614-0506, USA

Correspondence should be addressed to R. B. Campbell, campbell@math.uni.edu

Received 21 November 2011; Accepted 4 January 2012

Academic Editor: O. François

Recombination within a DNA segment during the neutral fixation process is studied to determine the number of individuals in previous generations which carry genetic material ancestral to that region in the present generation. If $Nr \ll 1$, where N is the population size and r is the probability of a recombination event within that region per individual in a generation, the ancestors of all the base pairs in that segment were probably in the same individual in an arbitrary generation in the asymptotic past (prior to the most recent common ancestor) and all the base pairs in that segment share a common coalescent. If $Nr \gg 1$, the ancestors of the base pairs in a segment are probably spread among several individuals in asymptotic generations; hence, there is not an ancestral individual, but an ancestral pool, and the coalescents of base pairs do not coincide. The overlap of the ancestral pools of unlinked genetic segments is less than $2pq$ where p and q are the relative frequencies of the two ancestral pools, which provides that the size of the ancestral pool for the human genome is close to the .80 upper bound which ensues from the Poisson progeny distribution.

1. Introduction

Gene substitution is a foundation of evolution. Greater understanding of this process has been provided by the diffusion approximation of Kimura and Ohta [1] which yielded an estimate of the time until fixation of a new mutation and the coalescent process of Kingman [2, 3] which provided an estimate of the time since a common ancestor (which is essentially the same quantity). This is the basis of the time since the mitochondrial Eve [4] and the Y-chromosome Adam [5] which penetrated the popular press.

But these calculations for Eve and Adam are based on the fact that there is no recombination in the mitochondrial DNA or the Y-chromosome. Eve and Adam only contained the genes ancestral to all present genes in the mitochondria and Y-chromosome, and the present genetic material in the 22 autosomes and the X-chromosome had its ancestral material in many different contemporaries of Eve and Adam. There is not one genetic ancestor of the human population, but an ancestral pool, in each generation a set of individuals which contain genetic material ancestral to the present population. (The pool may contract to a single individual in some generations which provides a grand-most recent common ancestor [6] but will expand in previous generations.)

This paper studies how many base pairs (nucleotide sites) a genetic segment (a contiguous set of base pairs in DNA) can contain and have no recombination in that segment as a reasonable model for evolution; and how many individuals in a generation will contain material ancestral to the present population (base pairs identical by descent to base pairs in the present population) if recombination splits the genetic segment, hence the ancestral graph. The number of individuals in a given generation which contain material ancestral to the present population is the size of the ancestral genetic pool. Of course, recombination can split the ancestry of two adjacent base pairs, and there may be some generations where the genetic material ancestral to the present population is in a single individual no matter how long the genetic segment, but estimates for the expected size of the ancestral genetic pool are obtained. This paper helps delineate when recombination is an important factor in evolution.

There are two results which provide information on the size of the ancestral genetic pool. Chang [7] showed that asymptotically as time goes back, 80 percent of

Table 1: Bounds on identity probabilities for a genetic segment.

rN	$2N$	MRCA		Asymptotic ancestor		Asymptotic pool size		
		Lower	Upper	Lower	Upper	Lower	Upper	$1.28R/\ln(1+R)$
.001	200	.98	1.000	.996	1.00	1.00	1.004	1.28
.01	2000	.77	.998	.96	.98	1.02	1.04	1.29
.1	2×10^4	.03	.977	.6	.83	1.19	1.4	1.40
1	2×10^5	$<10^{-19}$	.795		.20	2.32	5	2.33
10	2×10^6	$<10^{-234}$	.100		.00018	6.59	41	8.41
100	2×10^7	"0"	1.048×10^{-10}		10^{-15}	20.25	401	48.27

The diploid population size is N, and r is the probability of recombination within a segment. The value $r = 10^{-5}$ is used for the columns which bound the probability that the MRCA of a base pair is the MRCA of the entire segment. The bounds on the probability that an asymptotic ancestor of a base pair is an asymptotic ancestor of the entire segment and the asymptotic expected size of the ancestral pool of a segment are functions of rN. The last column is the estimate from Wiuf and Hein [8] which was obtained for a limited range of parameter values ($R = 2rN$).

the population are pedigree ancestors of the present population, the others have no living descendants. This does not mean that entire 80 percent contains genetic material ancestral to the present population, rather that is an upper bound on the size of the ancestral pool for the entire genome.

Wiuf and Hein [8] obtained an estimate for the size of the ancestral pool of chromosome 20 using the model of Hudson and Kaplan [9] for incorporating recombination into the coalescent process. Their estimate is $1.28R/\ln(1 + R)$, where R is defined as the (effective) population size (N) times the length of the genetic material in morgans (r) (the number of morgans is the expected number of recombination events in an individual in one generation). This formula, which was obtained from curve fitting based on numerical simulations, produces the estimate that the ancestral pool for chromosome 20 is 13 percent of the diploid population size ($R = 20,000$). They employed the range of values $1000 \leq R \leq 20,000$ for their numerical simulations, which includes neither 1000 contiguous base pairs (unless $N > 10^8$) nor the entire genome (unless $N < 400$). The formula $1.28R/\ln(1 + R)$ is consistent with our results for 1000 contiguous base pairs but cannot be valid for the entire genome if $N < 10^{12}$ (because the size of the ancestral pool would exceed the size of the population). Since their formula is obtained from a diffusion approximation holding $N \times r$ constant as $N \to \infty$, it should not be expected to remain valid for large r.

We first calculate asymptotic bounds for the expected size of the ancestral pool, hence the probability that the ancestral pool is a single individual. This addresses the question: does a common ancestor exist (i.e, is there high probability that the ancestral pool is a single individual for most generations in the asymptotic past)? We use the word "common" in the sense of shared by all the individuals in the present generation (which is the standard usage), but also in the sense of shared by all the nucleotide sites in a segment. The results depend on the product of the (effective) population size (N) and the length of the genetic segment (r) in morgans. For concreteness, we identify the results with the product rN and also various population sizes for a segment of 1000 contiguous base pairs (i.e., $r = 10^{-5}$ morgans). This choice is motivated as a contiguous DNA sequence coding for a 333 amino acid protein.

We next calculate bounds for the probability that the most recent common ancestor (MRCA) of a nucleotide site in a DNA segment is indeed the MRCA of the entire segment (i.e., the MRCA of every base pair in the segment is in the same individual). These bounds are not functions of rN, so we employ the value $r = 10^{-5}$ above and various values for N. However, we have numerically confirmed that the results do not change much as r and N vary with rN constant. Results for the asymptotic pool size and for the MRCA are presented in Table 1.

Sets of base pairs which are not contiguous (i.e., multiple segments) are of interest but difficult to analyze because recombination between the segments will depend on the locations within the segments. But our last results provide information on multiple genetic segments by bounding the overlap of ancestral pools of unlinked genetic segments. This provides a loose bound for the size of genetic pools of multiple genetic segments. In particular, it is informative for the size of the ancestral pool of the entire genome if the sizes of the ancestral pools of chromosomes are known.

2. Results

2.1. The Model. The results are obtained using the coalescent [6, 10]. The population size is N diploid individuals (i.e., $2N$ haploid gametes); we are assuming this is also the effective population size. However, the analysis is haploid; hence, the word "individual" (when not preceded by "diploid") refers to a single copy of the genetic segment. The length of a segment (r) is measured in morgans, 1 morgan is the length over which the expected number of crossover events in one individual (in one generation) is 1. When we study the MRCA, we shall employ the length $r = 10^{-5}$, which is motivated by a segment of 1000 contiguous base pairs with the crossover probability between two adjacent nucleotides of 10^{-8}. The value 1000 corresponds to DNA coding for 333 amino acids, and 10^{-8} was used by Wiuf and Hein [8] (the recombination rate varies between species, and hotspots may

impact the recombination rate by a factor of 10; Wiuf and Hein [11] assumed the recombination rate 10^{-7}). This model is for a single contiguous segment.

By coalescent, we are always referring to the coalescent of the entire population which is the ancestral graph containing all of the ancestors of the individuals in the present generation. The coalescent process (merging of ancestral lineages) is essentially the inverse of the fixation process. Time (t) is measured in generations from the common ancestor hence increases with real time. Recombination (crossing over) within the segment is incorporated using the model of Hudson and Kaplan [9] as employed by Wiuf and Hein [8].

In computing bounds, some approximations are employed (such as rounding off to lowest-order terms or employing estimates for the coalescent size). Hence, the bounds could be interpreted as approximate bounds but, when paired, give a good indication of the measures of identity for various parameter values.

2.2. Asymptotic Ancestral Pools. The coalescent may not exist for a segment, different base pairs may have different ancestral pedigrees; but it does exist for every base pair. Before (i.e., after in negative time) the MRCA of a base pair, there is an ancestral lineage which extends back to the dawn of time. Such a lineage exists for each base pair. The ancestral pool of a segment is the union of the individuals (gametes) which contain the ancestral lineages of the base pairs in that segment in a given generation. By asymptotic, we mean the behavior of those pools as time goes backward to negative infinity. Two questions which are of interest are what is the probability that all the lineages coincide in a single gamete (i.e., a common ancestor exists) in a given generation, and what is the average size of the ancestral pool (averaged as time goes back to negative infinity)? It is possible to bound these two quantities.

A sequence [8] is defined as a segment which contains one or more ancestral base pairs, perhaps contiguous, perhaps with intervening nonancestral base pairs. For a given segment (region of DNA), denote the number of sequences in a generation in the past as k. At equilibrium, the number of coalescent events decreasing the number of sequences is equal to the number of crossing over events increasing the number of sequences. Unfortunately, we cannot characterize the latter exactly but have two inequalities:

$$r \le E\left[\frac{k(k-1)}{(4N)}\right] \le E[k \times r]. \tag{1}$$

The outer quantities are bounds on the number of crossing over events, and the middle quantity is the frequency of coalescent events. Equality on the left assumes all the ancestral base pairs in a sequence are contiguous so that only crossovers between adjacent ancestral base pair can increase the number of sequences. Equality on the right assumes that ancestral material is dispersed everywhere (within the segment region) in sequences carrying ancestral material so that crossovers anywhere within the segment region will generate an additional sequence. (Simulations by Wiuf and Hein [8] suggest that the former is closer to reality.)

From convexity and the right hand inequality,

$$(E[k])^2 - E[k] \le E[k^2] - E[k] \le 4NE[k] \times r. \tag{2}$$

Solving this quadratic inequality for $E[k]$ yields $E[k] \le 1 + 4N \times r$.

This provides $E[k] \le 1.004$ for $Nr = .001$, 1.04 for $Nr = .01$, 1.4 for $Nr = .1$, 5 for $Nr = 1$, 41 for $Nr = 10$, and 401 for $Nr = 100$ (the number of base pairs is always an upper bound, since each sequence contains at least one ancestral base pair). Because $k \ge 1$ (there is at least one ancestor), we can calculate $P(k = 1) > .996$ for $Nr = .001$, .96 for $Nr = .01$, and .6 for $Nr = .1$ (these bounds are based on the worst case scenario $k = 2$ if $k \ne 1$). These values are in Table 1.

An upper bound for the probability of there being a single sequence (a true coalescent common ancestor) and a lower bound for the expected number of sequences is obtained by using the lower bound for the frequency of crossover events generating new sequences r with the coalescent probability $k(k-1)/4N$ (i.e., the left hand inequality in (1)). Recall that increased frequency of crossing over increases the number of sequences and coalescence decreases the number of sequences (going backward in time). Hence, a model employing a smaller frequency of crossovers will generate fewer sequences than the actual crossover frequency would generate. This will provide a higher probability that there will be a single sequence in the asymptotic past and a smaller asymptotic expected number of sequences than the actual crossover rate would provide.

To calculate the bounds, the transitions r and $k(k-1)/4N$ can be put into an infinite stochastic matrix governing the distribution of the number of sequences with r on the subdiagonal increasing the number of sequences by recombination, $k(k-1)/4N$ on the superdiagonal decreasing the number of sequences due to coalescence, and $1-r-k(k-1)/4N$ on the diagonal manifesting no change in the number of sequences. (The coalescent probability $k(k-1)/4N$ is an approximation which is only valid for small k, but this does not affect our calculations which only employ small k.) The ith entry in the stochastic vector the matrix acts on is the probability that the ancestral pool contains i sequences. The upper left hand corner of this matrix is displayed below:

$$\begin{pmatrix}
1-r & \frac{2}{4N} & 0 & 0 & 0 & \cdots \\
r & 1-r-\frac{2}{4N} & \frac{6}{4N} & 0 & 0 & \cdots \\
0 & r & 1-r-\frac{6}{4N} & \frac{12}{4N} & 0 & \cdots \\
0 & 0 & r & 1-r-\frac{12}{4N} & \frac{20}{4N} & \cdots \\
0 & 0 & 0 & r & 1-r-\frac{20}{4N} & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{pmatrix} \tag{3}$$

Because (3) is a nondegenerate stochastic matrix, there is a unique stochastic eigenvector which is the equilibrium (asymptotic) distribution for the stochastic process governed

by (3), and repeated multiplication of any stochastic vector by (3) will converge to that equilibrium distribution. The first component of this eigenvector is the asymptotic probability that there is a single sequence, and the expected number of sequences is $\sum_{i=1}^{\infty} i \times e_i$ where e_i is the ith component of the eigenvector.

This eigenvector can be calculated iteratively using 1 as the first component, $2N \times r$ for the second component, and $((i-1)(i-2)e_{i-1} + (4N \times r)(e_{i-1} - e_{i-2}))/(i(i-1))$ for the ith component where e_i is the ith component, and then normalizing to a stochastic vector. Computations were performed truncating both at 10,000 components and at 50 components to make sure that error was not introduced by k being too large (the results were the same for both truncations) and normalizing. (Truncating is consistent with the direction of the bound.)

To show that the result is really a function of the product $r \times N$, note that the eigenvectors of a matrix are unchanged when the matrix is multiplied by a nonzero constant or has a multiple of the identity matrix added to it (excluding degenerate cases). Hence, the eigenvectors for (3) are the same as the eigenvectors for

$$\begin{matrix}
-Nr & \frac{2}{4} & 0 & 0 & 0 & \cdots \\
Nr & -Nr-\frac{2}{4} & \frac{6}{4} & 0 & 0 & \cdots \\
0 & Nr & -Nr-\frac{6}{4} & \frac{12}{4} & 0 & \cdots \\
0 & 0 & Nr & -Nr-\frac{12}{4} & \frac{20}{4} & \cdots \\
0 & 0 & 0 & Nr & -Nr-\frac{20}{4} & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{matrix}, \tag{4}$$

which is obtained by multiplying (3) by N, and then subtracting $N\mathbf{I}$ from it ($\mathbf{I}$ is the identity matrix). Since the matrix (4) is a function of rN, so are its eigenvectors, hence the bound for the asymptotic ancestral pool sizes associated with (3).

The result from calculating the eigenvectors is that for $rN = .001$, the probability of a single ancestral sequence was less than 1.00, the expected number of sequences was greater than 1.00; for $rN = .01$, the probability of a single ancestral sequence was less than .98, the expected number of sequences was greater than 1.02; for $rN = .1$, the probability of a single ancestral sequence was less than .83, the expected number of sequences was greater than 1.19; for $rN = 1$, the probability of a single ancestral sequence was less than .20, the expected number of sequences was greater than 2.32; for $rN = 10$, the probability of a single ancestral sequence was less than .00019, the expected number of sequences was greater than 6.59; for $rN = 100$, the probability of a single ancestral sequence was less than 10^{-15}, the expected number of sequences was greater than 20. Note that $rN = 1$ corresponds to $N = 10^5$ if $r = 10^{-5}$ which ensues from a segment length of 1000 base pairs. These values are in Table 1.

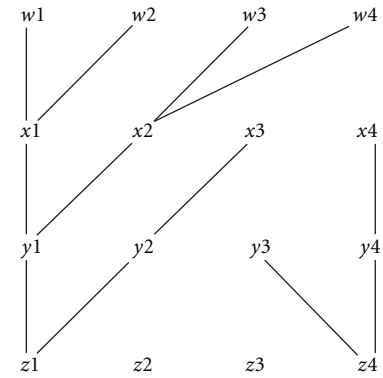

Figure 1: Schematic of coalescence. Lines connect individuals with their ancestors, with each generation a horizontal array of individuals (e.g., $x1$ $x2$ $x3$ $x4$). Time advances going up the page; hence, the parent of an individual is in the line below (e.g., $x2$ is the parent of $w4$). The coalescent is indicated with thick lines. Individuals $x1$ and $x2$ are in the coalescent; $x3$ is not in the coalescent but is descended from the MRCA of the coalescent; $x4$ is not in the coalescent and is not descended from the MRCA of the coalescent.

2.3. The Most Recent Common Ancestor. In addition to the asymptotic history, we can ask whether the MRCA really is an MRCA, that is, whether the MRCA of a single base pair (which must exist) is the MRCA of every base pair in the segment. This is not the requirement that the coalescents of all the base pairs in a segment coincide, merely that they terminate in the same individual. Crossing over during the coalescent process divides the genetic material in a single individual among two individuals, causing the ancestry of the gene to be contained in two different ancestral subgraphs; those graphs may terminate in the same MRCA or in different MRCAs. This is illustrated in Figure 1, where a crossover between individuals $x1$ and $x2$ or $x1$ and $x3$ would change the ancestral graph of the genetic material involved in the crossover but leave the same MRCA; a crossover between $x1$ and $x4$ would change the ancestral graph and change the MRCA to a more distant ancestor. The schematic of a coalescent in Figure 1 also illustrates that, during the process of coalescence or fixation, there are individuals not in the coalescent (ancestral pedigree) which share the common ancestor of the coalescent (e.g., $x3$) and individuals not in the coalescent which do not share the common ancestor of the coalescent (e.g., $x4$).

The probability of no crossing over involving individuals in the coalescent provides a lower bound for the probability of a common MRCA because that will assure a common MRCA, but allowing crossing over to individuals sharing the MRCA, whether inside or outside the coalescent, will also provide that MRCA. The probability of no crossing over involving individuals in the coalescent can be approximated employing the estimate for the cumulative number of individuals in the coalescent $4N(\ln(4N) - 0.5)$ ([12]; the cumulative size of the coalescent is the total number of individuals in the coalescent: in Figure 1, $z1$, $y1$, $x1$, $x2$, $w1$, $w2$, $w3$, and $w4$ are in the coalescent; hence, the cumulative

size is 8) and probability of a crossover in a single individual 10^{-5}, and assuming crossing over is a Poisson process. The result is that the probability of no crossover involving individuals in the coalescent is approximately $\exp(-10^{-5} \times 4N(\ln(4N) - 0.5))$. The quantity $4N(\ln(4N) - 0.5)$ is an estimate for the expected size of the coalescent based on the expected time between changes in the size of the coalescent; convexity of the exponential function provides that $\exp(E[X]) \leq E[\exp(X)]$ (in this case, X is the size of the coalescent), which is consistent with providing a lower bound.

A higher lower bound is obtained by calculating an upper bound for the probability that a recombination event involving a member of the coalescent resulted in at least one nucleotide base pair which did not share the MRCA of the coalescent being in the ancestry of that individual. To this end, we calculate the probability that a member of the coalescent crossed over with an individual outside the coalescent (e.g., $x1$ with $x3$ or $x4$); this overestimates the probability of recombination with an individual not sharing the MRCA because some individuals outside the coalescent (e.g., $x3$ in Figure 1) will share the same MRCA. The number of individuals in the coalescent at time t (t is the expected time from the MRCA until the coalescent has the specified size; this function is the inverse of the expected time to the coalescent size) is approximately $(1 + 1/2N - t/4N)^{-1}$ [12]. Because t is the expected time until the coalescent size, this is only valid until the expected time to fixation ($4N$) when the size of the coalescent becomes the population size ($2N$, which is N diploid individuals); hence, it is not relevant that the quantity becomes negative for $t > 4N + 2$. Because $(1 + 1/2N - t/4N)^{-1}$ is obtained from the coalescent process by employing the expected transition times for decreasing the number of individuals in the coalescent by one (i.e., manifests the expected time at each size), the summation (5) manifests the expected time at each coalescent size hence gives the expected number of crossing over events; variation in the timing of coalescent events does not introduce any error since expected times are used, any error results from the approximation $(1 + 1/2N - t/4N)^{-1}$ (and perhaps summing instead of integrating). The expected number of crossover events between individuals inside and outside the coalescent is

$$10^{-5} \times \sum_{t=0}^{4N} \left(1 + \frac{1}{2N} - \frac{t}{4N}\right)^{-1} \frac{2N - (1 + 1/2N - t/4N)^{-1}}{2N}, \quad (5)$$

where 10^{-5} is the probability that a crossover occurs in a single individual, $(1 + 1/2N - t/4N)^{-1}$ is the number of individuals in the coalescent at time t, and $1/2N \times (2N - (1 + 1/2N - t/4N)^{-1})$ is the probability that the crossover is with an individual outside the coalescent. This, assuming crossover events are a Poisson process, provides the probability of no such crossovers

$$e^{-10^{-5} \times \sum_{t=0}^{4N} (1+1/2N-t/4N)^{-1}(2N-(1+1/2N-t/4N)^{-1})/2N}, \quad (6)$$

(The variation in duration of the coalescent process will provide greater variation than a Poisson process; hence, the exponentiation in (5) underestimates the probability of no crossovers, which is consistent with providing a lower bound.)

For a population of 100 diploid individuals (i.e., 200 gametes, $2N = 200$), this provides the lower bound for the probability that all nucleotide sites in a segment have the same MRCA .98; for $2N = 2000$, .77; for $2N = 20,000$, .03; for $2N = 200,000$ or more, less than 10^{-19}. Thus, all the nucleotide sites in a segment probably have the same MRCA in populations smaller than 1000 but may not in larger populations (this is only a lower bound for all nucleotide sites having the same MRCA). This information is presented in Table 1.

In order to obtain an upper bound for the probability that the MRCA for a nucleotide base pair is indeed the MRCA for the entire 1000 base pairs in the segment, we shall use a lower bound for the probability that a crossover occurred between an individual in the coalescent and an individual not sharing the MRCA of the coalescent (e.g., $x1$ and $x4$ in Figure 1).

Heuristically, this can be obtained from the growth of the coalescent $(1 + 1/2N - t/4N)^{-1}$ and the rate of increase of the allele destined to fixation (which includes individuals such as $x3$ which are not in the coalescent). For the Poisson progenies distribution with $\lambda = 1$, the expected number of siblings of an individual is 1. Therefore, since all progeny are equally likely to become fixed, the expected increase in frequency, conditioned on fixation, is $1 - (k-1)/(2N-1) < 1$, where the 1 is the expected number of siblings of the progeny destined for fixation and the $(k - 1)/(2N - 1)$ reflects that the other $2N - 1$ individuals in the parental generation ($k - 1$ of which are of the same type as the progeny destined for fixation) must have on average $1 - 1/(2N - 1)$ progeny to maintain a constant population size. This provides that the expected number of copies of the allele destined for fixation is less than or equal to t at time t; hence, $r \sum_0^{2N} (1 + 1/2N - t/4N)^{-1}(2N - t)/2N$ should be a lower bound for the probability that the MRCA of a nucleotide pair is not the MRCA of all the nucleotide pairs (a crossover occurred with an individual not descended from the MRCA). Truncating the summation at $2N$ is consistent with calculating a lower bound, but because the factors in the summation are an expected value and a bound on an expected value, this may not be a lower bound.

Rigorously, a weaker bound can be obtained using Tchebychev's theorem. The variance of the change in allele frequency in a generation is $k(2N - k)/2N$ where k is the number of alleles of the designated type (the actual model is the binomial distribution, the Poisson progeny distribution is an approximation which is useful for many purposes, but the binomial variance is tractable here). Because the rate of increase of the designated allele is less than 1, the expected number of copies of the designated allele at time t is less than t (assuming one copy at time 1); hence, the variance of the change in allele frequencies at time t is less than t (i.e., $k \times (2N - k)/2N < t$; because of the convexity of $k(2N - k)$, the expected value of the variance is less than the variance calculated using the expected value). Independence between generations provides that the variance of the cumulative

change over t generations is less than $\sum_{i=1}^{t} i = t(t+1)/2 < t^2$; hence, the cumulative standard deviation is less than t.

This provides that $4t$ is three standard deviation units above the expected number of copies at time t; hence, by Tchebychev's theorem, there are at least $2N - 4t$ alleles not identical by descent with the designated allele at time t with probability 8/9. Because the argument t of the coalescent size $(1 + 1/2N - t/4N)^{-1}$ is the expected time to that size and $2N - 4t$ is linear, multiplying $(1 + 1/2N - t/4N)^{-1}$ by $2N - 4t$ entails an accurate pairing of coalescent and nondescendant sizes (i.e., for a given $E(t)$ which is the argument of $(1 + 1/2N - t/4N)^{-1}$, the actual value of t in $2N - 4t$ will vary, but conditioning on $E(t)$ as the argument for $(1+1/2N-t/4N)^{-1}$, averaging over all the associated values of $2N - 4t$ will be the same as using that $E(t)$ as the argument for $2N - 4t$. (The truncation of $2N - 4t$ is consistent with the direction of the bound.) This provides the upper bound for the probability that the MRCA of a nucleotide pair is the MRCA of all the nucleotide pairs in the segment:

$$e^{-10^{-5}\times\sum_{t=0}^{N/2}(1+1/2N-t/4N)^{-1}\times(2N-4t)/2N\times.88}, \tag{7}$$

where $r = 10^{-5}$ and .88 is the 8/9 from Tchebychev's theorem.

Numerical evaluation of this expression produces 1.000 for $2N = 200$, .998 for $2N = 2000$, .977 for $2N = 20,000$, .795 for $2N = 200,000$, .100 for $2N = 2,000,000$, and 10^{-10} for $2N = 20,000,000$. As noted above, this is a generous bound; hence, there is very low probability that all the nucleotide sites in a gene have the same MRCA for N greater than 1,000,000. These values are in Table 1.

2.4. Multiple Unlinked Segments. Genetics is seldom concerned with single contiguous segments of DNA, but often multiple segments with significant separation, hence recombination, between them. Although we should consider an arbitrary recombination frequency between segments, that frequency will depend on the locations within the segments (recombination within one segment will result in part, but not all, of that segment recombining with another segment), making it a difficult problem. Free recombination is the opposite extreme to no recombination and is appropriate for some cases including segments on different chromosomes or segments which are entire chromosomes. The specific question which we address is if the sizes of the ancestral pools of two unlinked segments are known, what is the size of the combined ancestral pool? It is at least the size of the larger of the two pools and at most the sum of the sizes of the pools. We provide a more precise bound. Calculations are based on lowest-order terms in power series.

First consider the case where the segment lengths and population size are small enough so that each ancestral pool is a single individual; hence, there are two ancestral lineages. This case lays a foundation for the following cases hence is of interest beyond the circumstances when its assumptions are met. The population size is N, hence $2N$ gametes. If the ancestral lineages of two unlinked segments are in the same gamete, then the previous generation they were in the same gamete half the time (because the zygote they came from was two gametes). If they are in different gametes, then $1/N$ of the time they came from the same zygote (this follows from Kingman's [3] observation that the Wright-Fisher model is equivalent to each individual choosing its parent independently from the previous generation), hence $1/2N$ of the time they came from the same gamete the previous generation. This defines a Markov process going backward in time with the two states that the lineages are or are not in the same gamete, and the matrix for this Markov process is

$$\begin{matrix} .5 & \dfrac{1}{2N}, \\ .5 & 1 - \dfrac{1}{2N}, \end{matrix} \tag{8}$$

which has the eigenvector (stable distribution) $\langle(1/(1 + N), N/(1 + N))\rangle$, hence the diploid structure provides that two independent lineages will coincide (be in the same gamete) approximately $1/N$ of the time rather than $1/2N$ which would occur from random association.

Next consider a single ancestral lineage (ancestral pool of size one) and the ancestral pool of size greater than one of an unlinked segment; u is the relative frequency (size/$2N$) of the ancestral pool at the gamete stage. In order to maintain an equilibrium size u of the ancestral pool, coalescence must be balanced by crossing over (recombination) going backward in time. Coalescence reduces the size of the ancestral pool from u to $1-e^{-u}$ in a generation, $u-(1-e^{-u}) = u^2/2$ to lowest order terms, hence crossing over must increase the number of ancestral lineages by that amount. Only crossing over in individuals in which exactly one of the alleles is ancestral to the ancestral pool will increase the size of the ancestral pool, the frequency of such individuals is $2e^{-u}(1 - e^{-u})$ (e^{-u} is the probability that a parental allele (half a zygote) is not an ancestor of the ancestral pool). Therefore, the frequency of crossing over, which we designate with ρ, satisfies $u^2/2 = \rho \times 2e^{-u}(1 - e^{-u})$ or $\rho = u/4$ to order u.

This provides that the probability that if the lineage was in a gamete with a part of the ancestral pool, it was in a gamete with part of the ancestral pool the previous generation is $.5 + .5(1 - e^{-u}) + .5\rho e^{-u}$, which is obtained by summing the probability the ancestral pool material was in the same gamete the previous generation (.5), the probability the gamete the previous generation contained the other copy of the allele in the zygote, but it was also ancestral ($.5(1 - e^{-u})$), and the probability the gamete the previous generation contained the other copy of the allele in the zygote which was not ancestral, but it was made ancestral by crossing over ($.5\rho e^{-u}$). To first-order terms in u, this is equal to $.5 + .625u$, hence the probability that if a lineage was in a gamete with part of the ancestral pool, it was in a gamete without part of the ancestral pool the previous generation is $.5 - .625u$. If the lineage was in a gamete without material from the ancestral pool, then its gamete the previous generation could have material from the ancestral pool if either its gamete the previous generation contained the ancestor of that nonancestral allele, but that allele had coalesced with an allele with ancestral material, or it contained the ancestor

of the other allele in the parent to the gamete and that allele contained ancestral material (crossing over produces higher-order terms), the respective probabilities are $.5(1 - e^{-u})$ and $.5(1 - e^{-u})$. To order u, summing these yields u. Hence, the probability that if the lineage was in a gamete without ancestral material, it was also in a gamete without ancestral material the previous generation is $1 - u$. This yields the Markov matrix governing cooccurrence of the lineage and ancestral pool

$$\begin{matrix} .5 + .625u & u, \\ .5 - .625u & 1 - u, \end{matrix} \tag{9}$$

which has the eigenvector (stable distribution) $\langle u/(.5 + .375u), (.5 - .625u)/(.5 + .375u)\rangle$; hence, the diploid structure provides that a lineage will coincide with part of an unlinked ancestral pool of size u approximately $u/(.5 + .375u)$ (i.e., approximately $2u$) of the time rather than u which would occur from random association.

Now consider two unlinked segments (or unlinked collections of genetic material) for which the sizes of the ancestral pools are known. Assume the asymptotic probabilities of gametes containing ancestral material for those segments are u and v, respectively (hence, we shall refer to them as "u" and "v" segments). Then, the ancestral lineage for each nucleotide pair in the "v" segment will be in a gamete with material in the "u" ancestral pool with probability $u/(.5 + .375u)$ (or $u/(.5 + .375u)$ of such lineages will be in "u" gametes). If all gametes containing "v" ancestral material had equal probability of containing "u" ancestral material, the probability that a gamete with "v" ancestral material contained "u" ancestral material would be $u/(.5 + .375u)$, the probability for a "v" lineage containing "u" ancestral material. Hence, the probability that a gamete contained ancestral material from both segments would be $vu/(.5 + .375u)$ (v is the probability of containing ancestral material from the second segment, and $u/(.5 + .375u)$ is the conditional probability of containing ancestral material from the first segment).

However, gametes containing many (as opposed to fewer) "v" ancestral lineages are likely to have recently coalesced (because coalescence combines ancestral lineages and crossing over separates them). The "u" segment (whether or not ancestral) in that gamete is also likely to have recently coalesced because the sexual reproduction process keeps independent segments together (with probability .5 each generation), and because it coalesced, it is more likely to contain ancestral material. Hence, gametes with many ancestral "v" lineages are more likely to contain ancestral "u" material than gametes with few ancestral "v" lineages. This provides that the probability that a gamete containing ancestral "v" material also contains ancestral "u" material will be less than the probability that an ancestral "v" lineage is in a gamete with ancestral "u" material. Thus, the probability that a gamete contains both "u" and "v" ancestral material is less than $vu/(.5 + .375u)$ (and less than $vu/(.5 + .375v)$ by symmetry). In particular, the probability that an individual contains ancestral material from both pools is less than twice the product of the probabilities of the two pools ($2uv$).

Therefore, the size of the combined ancestral pool is at least $u + v - 2uv$ (and at most $u + v$). This argument can be extended recursively to find a bound on the size of the ancestral pool of an arbitrary number of unlinked segments for which the ancestral pool size is known. In particular, it can be used to find a bound on the size of the ancestral pool of the entire genome if the size of the ancestral pool for each chromosome is known.

3. Discussion

The main result from Table 1 is that a segment will probably have a single ancestor (i.e., ancestral pool of size 1) if $rN \ll 1$ (the probability is greater than .6 if $rN = .1$, greater than .96 if $rN = .01$, and greater than .99 if $rN = .001$). Complementarily, the probability of a single ancestor is close to zero for $rN \gg 1$ (the probability is less than .00019 for $rN = 10$ and less than 10^{-15} for $rN = 100$). The bounds on the expected size of the asymptotic pool are of course close to 1 for $rN < 1$, but are not very useful for $rN > 1$ (numerical calculations provide that the lower bound approaches 51 as rN gets large while the upper bound is approximately $4rN$). For $rN = 1$, there is a rather tight bound on the expected size of the asymptotic pool size (between 2.3 and 5). However, $rN = 1$ is of limited interest. $rN = 1$ corresponds to a gene or a piece of a gene of 10^3 or 10^2 contiguous base pairs if the population size is 10^5 or 10^6. But it certainly does not correspond to an entire chromosome, a chromosome in man or *Drosophila* is about one morgan in size, which would require an effective population size close to 1. (This assumes a recombination rate of 10^{-8} between adjacent base pairs, there are other estimates for that rate, and variation in the rate (hotspots) further complicates the analysis [13].)

These results provide insight into the question: what is the integrity of the gene? Is the gene the atom of evolution or does evolution occur on a finer scale? In small populations ($N < 1000$), the gene (defined as 1000 contiguous base pairs) is indeed a meaningful entity, the most recent common ancestor (MRCA) is the same for all of its base pairs and that individual has an ancestral lineage which contains common ancestors for all the nucleotide pairs in that gene. Periods when the ancestral material is spread among multiple individuals are infrequent; hence, all the base pairs change their frequency as a unit. In larger populations ($N > 1,000,000$), the MRCAs for the various base pairs in the gene do not coincide, and it is rare that the ancestral lineages for all the base pairs coincide. There is not an ancestral individual, but an ancestral pool. Positive probability, no matter how small, provides that the lineages of all the base pairs will coincide at some time in the past (hence, there is a common ancestor), but, if $Nr \gg 1$, the base pairs will not all stay together and evolve (change frequency) as a unit. These conclusions are from the numerical bounds calculated in Table 1. Some of the bounds are quite loose, but they still support the conclusions.

These results are for neutral drift with no mutation (i.e., identity by descent). Selection will speed up the fixation process and increase identity by descent [14], hence increase

the likelihood that the MRCA for a base pair is the MRCA for all the base pairs in the gene, it might also eliminate aberrant forms of the gene, thereby further contributing to integrity. Mutation will decrease the physical identity of the genes. Since the mutation rate is comparable to the recombination rate (both are around 10^{-8} (per nucleotide site or between adjacent nucleotide sites; both have great variation)), probabilities of identity by type will be similar. But because much recombination will be with individuals which are identical by descent, identity by type is less likely than identity by descent.

The bounds in this paper on the size of the ancestral pool are most useful for a genetic segment of 1000 contiguous base pairs, and Wiuf and Hein [8] have presented an estimate for the size of the ancestral pool for a chromosome. Indeed, it would be nice to have tighter bounds for a genetic segment and an estimate for chromosomes which does not rely on simulation for the population size of interest. But it is also necessary to extend results for genetic segments to results for unions of genetic segments, whether a few separated contiguous segments or the entire genome. We have improved the bounds obtained by assuming that the genetic material in different segments (or chromosomes) is in the same individuals as much as possible, or in different individuals as much as possible (i.e., if the sizes of two genetic pools are u and v, the size of the combined pool is between $\max(u, v)$ and $u + v$); we have shown that the overlap of the two pools is less than $2uv$ if the genetic segments are unlinked. This enables us to show, based on the chromosomal pool size of Wiuf and Hein [8] and recursively applying the $2uv$ bound, that the size of the ancestral pool of the human genome is close to the 80 percent pedigree ancestor upper bound of Chang [7]. But tighter bounds should be sought in general, especially for the difficult problem of genetic segments which are linked.

Acknowledgment

This paper has been significantly improved due to suggestions from Joe Felsenstein and anonymous reviewers.

References

[1] M. Kimura and T. Ohta, "The average number of generations until fixation of a mutant gene in a finite population," *Genetics*, vol. 61, pp. 763–771, 1969.

[2] J. F. C. Kingman, "The coalescent," *Stochastic Processes and Their Applications*, vol. 13, pp. 235–248, 1982.

[3] J. F. C. Kingman, "On the genealogy of large populations," *Journal of Applied Probability*, vol. 19, pp. 27–43, 1982.

[4] C. Wills, "When did Eve live? An evolutionary detective story," *Evolution*, vol. 49, pp. 593–607, 1995.

[5] R. L. Dorit, H. Akashi, and W. Gilbert, "Absence of polymorphism at the ZFY locus on the human Y chromosome," *Science*, vol. 268, no. 5214, pp. 1183–1185, 1995.

[6] J. Hein, M. H. Schierup, and C. Wiuf, *Gene Genealogies, Variation, and Evolution: A Primer in Coalescent Theory*, Oxford University Press, New York, NY, USA, 2005.

[7] J. T. Chang, "Recent common ancestors of all present-day individuals," *Advances in Applied Probability*, vol. 31, no. 4, pp. 1002–1026, 1999.

[8] C. Wiuf and J. Hein, "On the number of ancestors to a DNA sequence," *Genetics*, vol. 147, no. 3, pp. 1459–1468, 1997.

[9] R. R. Hudson and N. L. Kaplan, "Statistical properties of the number of recombination events in the history of a sample of DNA sequences," *Genetics*, vol. 111, no. 1, pp. 147–164, 1985.

[10] J. Wakely, *Coalescent Theory: An Introduction*, and Company Publishers, Greenwood Village, Colo, USA, 2005.

[11] C. Wiuf and J. Hein, "The ancestry of a sample of sequences subject to recombination," *Genetics*, vol. 151, no. 3, pp. 1217–1228, 1999.

[12] R. B. Campbell, "A logistic branching process for population genetics," *Journal of Theoretical Biology*, vol. 225, no. 2, pp. 195–203, 2003.

[13] C. Wiuf and D. Posada, "A coalescent model of recombination hotspots," *Genetics*, vol. 164, no. 1, pp. 407–417, 2003.

[14] A. Albrechtsen, I. Moltke, and R. Nielsen, "Natural selection and the distribution of identity-by-descent in the human genome," *Genetics*, vol. 186, no. 1, pp. 295–308, 2010.