The Ancestry of a Gene

R. B. Campbell

Department of Mathematics

University of Northern Iowa

Cedar Falls, IA 50614-0506

USA

running head: Ancestry of a gene

phone: (319)273-2447

fax: (319)273-2546

e-mail: campbell@math.uni.edu

homepage: http://www.math.uni.edu/∼campbell

Abstract

Recombination within a gene during the neutral fixation process is studied to determine the number of individuals in previous generations which carry genetic material ancestral to the genes at a specified locus in the present generation. In populations smaller than 1000, the most recent common ancestor (MRCA) of a single nucleotide base pair in a gene is probably in the same individual as the MRCAs of all of the base pairs in that gene, hence there is indeed an MRCA for the entire gene. But in populations larger than 1,000,000, the MRCA of a single nucleotide base pair in a gene is probably not in the same individual as the MRCAs of other base pairs in that gene, hence there is not an MRCA of the entire gene, but an ancestral pool. Furthermore, in populations smaller than 1000 most of the generations prior to the MRCA will contain a single individual which is a common ancestor of all the base pairs in a gene, but in populations larger than 1,000,000 very few generations will contain a single individual which is a common ancestor of all the base pairs in a gene, rather an ancestral pool persists backward in time.

keywords: coalescent; MRCA; genetic ancestry; pedigree; recombination

INTRODUCTION

Gene substitution is a foundation of evolution. Greater understanding of this process has been provided by the diffusion approximation (Kimura and Ohta, 1969) which yielded an estimate of the time until fixation of a new mutation, and the coalescent process (Kingman, 1982a, 1982b) 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 (Wills, 1995) and the y-chromosome Adam (Dorit, Akashi, and Gilbert, 1995) 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, 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.

Is this situation at the level of the genome also manifested at the level of the gene? Is the concept of gene fixation meaningful (in the sense of identity by descent)? (I.e., is there a single ancestor of all the nucleotide pairs in all the genes at a specified locus?) Related questions include: What is the unit of evolution? How does recombination (crossing over) impact genetic ancestry including the results from diffusion and coalescent theory? When is it necessary to consider an ancestral pool rather than a common ancestor? How large are the ancestral pools?

There are two results which provide information on the size of the ancestral genetic pool. Chang (1999) showed that asymptotically as time goes back, 80

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

Wiuf and Hein (1997) obtained an estimate for the size of the ancestral pool of chromosome 20 using the model of Hudson and Kaplan (1985) for incorporating recombination into the coalescent process. Their estimate, $1.28R/\ln(1+R)$ (R is defined as the (effective) population size times the length of the genetic material in morgans (the number of morgans is the expected number of recombination events in an individual in one generation)), is obtained from curve fitting based on numerical simulations. Their formula 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 a single gene (unless $N > 10^8$) nor the entire genome (unless $N < 400$). The formula $1.28R/\ln(1+R)$ is consistent with our results for a gene, 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).

This paper is concerned with the ancestral pool of a single gene (a gene is defined as 1000 contiguous base pairs (or 10^{-5} morgans)). Two questions will be addressed: First, what is the probability that the most recent common ancestor (MRCA) of a nucleotide site in the gene is indeed the MRCA of the entire gene (i.e, of every base pair in the gene). Second, asymptotically as time goes backward, what is the probability that a single copy of the gene is a common ancestor for all the base pairs in the gene, i.e., what fraction of the time does

a common ancestor exist. These are two variations on the question: does a common ancestor exist? 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 gene. A variation of the second question is: if there is not a single common ancestor, what is the size of the ancestral pool? Our results are presented in Table 1.

RESULTS

The Model: The population size is N diploid individuals (i.e., $2N$ copies of the gene); 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 gene. The size of the gene is 1000 contiguous base pairs (i.e., 10^{-5} morgans, 1 morgan is is the length where the expected number of crossover events in one individual (in one generation) is 1). This corresponds to the cross over probability between two adjacent nucleotides of 10^{-8} which was used by Wiuf and Hein (1997) (hotspots may impact the recombination rate by a factor of 10, and it varies between species; Wiuf and Hein (1999) assumed the recombination rate 10^{-7}). This model is not appropriate for a gene containing more than one exon separated by significant distance, it may be appropriate for individual exons in such cases.

By coalescent, we are always referring to the coalescent of the entire population, so that the coalescent process is essentially the inverse of the fixation process. The schematic of a coalescent in Fig. 1 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., $x4$). Time (t) is measured in generations from

the common ancestor, hence increases with real time. Recombination (crossing over) within the gene is incorporated using the model of Hudson and Kaplan (1985) as employed by Wiuf and Hein (1997).

[FIGURE 1 NEAR HERE]

The Most Recent Common Ancestor: The first question is whether the MRCA really is an MRCA, i.e., whether the MRCA of a single nucleotide site (which must exist) is the MRCA of every nucleotide site in the gene. This is not the requirement that the coalescents of all the nucleotide sites in a gene 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 graphs; those graphs may terminate in the same MRCA, or in different MRCAs. For example, 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.

A concise lower bound for the probability that all the nucleotide sites in a gene have the same MRCA is obtained from the estimate for the cumulative number of individuals in the coalescent $4N(\ln(4N) - 0.5)$ (Campbell 2003); they will share the same MRCA if none of the individuals in the coalescent were involved in a crossover. Since the probability of a crossover in a single individual is 10^{-5}, assuming crossing over is a Poisson process, the probability of no crossover is approximately $\exp(-10^{-5} \times 4N(\ln(4N) - 0.5))$, which is a lower bound for the probability of all the nucleotide sites sharing the same MRCA.

A better lower bound is obtained by calculating an upper bound for the

probability that a recombination event occurred between a member of the coalescent and an individual not sharing the MRCA for that nucleotide site. 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 provides an upper bound because some individuals outside the coalescent (e.g., $x3$ in Fig. 1) will share the same MRCA. The number of individuals in the coalescent at time t (t measures time from the MRCA) is approximately $(1+1/2N-t/4N)^{-1}$ (Campbell, 2003). Because this 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 expected number of cross-over events is not impacted by the great variation in the timing of occurrence of coalescence events. This approximation 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 $t > 4N + 2$. The expected number of crossover events between individuals inside and outside the coalescent is:

$$10^{-5} \times \sum_{t=0}^{4N} (1 + \frac{1}{2N} - \frac{t}{4N})^{-1} \frac{2N - (1 + \frac{1}{2N} - \frac{t}{4N})^{-1}}{2N} \tag{1}$$

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

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

(The variation in duration of the coalescent process will provide greater variation than a Poisson process, hence the exponentiation in (2) underestimates the

probability of no cross-overs.)

For a population of 100 diploid individuals (i.e., 200 genes, $2N = 200$), this provides the lower bound for the probability that all nucleotide sites in a gene 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 gene 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 gene, 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 Fig. 1). To this end, we use $4t$ as an upper bound on the number of copies of a gene in the population t generations after there was a single copy, this includes individuals such as $x3$ which are not in the coalescent. Hence $2N - 4t$ (truncated to 0 at $t = N/2$) is a lower bound for the number of individuals not sharing the common ancestor. This is a very generous bound, but suffices to generate the desired results.

The bound $4t$ is obtained from the fact that the number of copies, k, of a base pair which will become fixed increases in expectation at the rate $1-(k-1)/(2N-1)$, with a standard deviation $\sqrt{1 + (k-1) \times ((2N-1) - (k-1))/(2N-1)}$. This is true because for the Poisson progeny distribution with $\lambda = 1$, the expected number of siblings of an individual is 1, but maintaining constant population size requires reducing the number of progeny of individuals which will not become fixed. It follows that the expected number of descendants t generations after the initial individual is less than $t + 1$, and the standard deviation of the

number of descendants is less than $\sqrt{t(t+1)/2} \leq t$ ($E[1+(k-1)\times((2N-1)-(k-1))/(2N-1)] \leq t$, and sum the variances). Therefore $4t$ is three standard deviation units above the expected number of descendants and Tchebycheff's theorem provides that the number of descendants will exceed $4t$ less than $1/9$ of the time. Multiplying the final result by $8/9 = .88$ compensates for that $1/9$.

The bound $2N - 4t$ is deterministic (except for the possible $1/9$ of the time which is corrected for) and linear. Because the coalescent size $(1 + 1/2N - t/4N)^{-1}$ is defined by 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. (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 gene:

$$e^{-10^{-5}\times\sum_{t=0}^{\frac{N}{2}}(1+\frac{1}{2N}-\frac{t}{4N})^{-1}\times\frac{2N-4t}{2N}\times.88}. \tag{3}$$

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.

Less recent common ancestors: The coalescent may not exist for a gene, different base pairs may have different ancestral pedigrees; but it does exist for every base pair. After (i.e., before in positive time) the MRCA of a base pair is reached, 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 gene is the union

of the individuals (genes) which contain the ancestral lineages of the base pairs in that gene. Two questions which are of interest are: what is the probability that all the lineages coincide in a single gene (i.e., a common ancestor exists), and what is the average size of the ancestral pool (in a single generation)? It is possible to bound these two quantities.

A sequence (Wiuf and Hein 1997) is defined as a gene which contains one or more ancestral base pairs, perhaps contiguous, perhaps as multiple segments. Denote the number of sequences 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:

$$10^{-5} \leq E[k(k-1)/(4N)] \leq E[k \times 10^{-5}]. \tag{4}$$

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 base pairs on a chromosome are contiguous so that only crossovers between adjacent base pair can increase the number of sequences. Equality on the right assumes that ancestral material is dispersed everywhere (within the 1000 base pair region) in genes carrying ancestral material so that crossovers anywhere within the 1000 base pair region will generate an additional sequence. (Simulations by Wiuf and Hein (1997) suggest the former is closer to reality).

From convexity and the right hand inequality,

$$(E[k])^2 - E[k] \leq E[k^2] - E[k] \leq 4NE[k] \times 10^{-5}. \tag{5}$$

Solving this quadratic inequality for $E[k]$ yields $E[k] \leq 1 + 4N \times 10^{-5}$.

This provides $E[k] \leq 1.004$ for $N = 100$, 1.04 for $N = 1000$, 1.4 for $N = 10,000$, 5 for $N = 100,000$, 41 for $N = 1,000,000$, and 401 for $N = 10,000,000$

(1000 is always an upper bound, since there are 1000 base pairs). Because $k \geq 1$ (there is at least one ancestor), we can calculate $P(k = 1) > .996$ for $N = 100$, .96 for $N = 1000$, and .6 for $N = 10,000$ (these bounds are based on the worst case scenario $k = 2$ if $k \neq 1$). 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 probability of recombination is 10^{-5} (a lower bound for recombination) and the coalescent probability $k(k-1)/4N$. Recall that recombination increases the number of sequences and coalescence decreases the number of sequences. These transitions can be put into an infinite stochastic matrix governing the number of sequences with 10^{-5} on the subdiagonal representing increasing the number of sequences by recombination, $k(k-1)/4N$ on the superdiagonal representing decreasing the number of sequences due to coalescence, and $1 - 10^{-5} - k(k-1)/4N$ on the diagonal representing no change in the number of sequences. The upper left hand corner of this matrix is displayed below:

$$\begin{array}{ccccccc}
1-10^{-5} & \frac{2}{4N} & 0 & 0 & 0 & \cdots \\
10^{-5} & 1-10^{-5}-\frac{2}{4N} & \frac{6}{4N} & 0 & 0 & \cdots \\
0 & 10^{-5} & 1-10^{-5}-\frac{6}{4N} & \frac{12}{4N} & 0 & \cdots \\
0 & 0 & 10^{-5} & 1-10^{-5}-\frac{12}{4N} & \frac{20}{4N} & \cdots \\
0 & 0 & 0 & 10^{-5} & 1-10^{-5}-\frac{20}{4N} & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{array} \tag{6}$$

Because a lower bound for recombination which generates sequences is used, this shifts the stable distribution to the left (toward 1, a single sequence), hence overestimates the probability of a single sequence and underestimates the expected number of sequences. The principal eigenvector (unnormalized stable distribu-

tion) can be calculated iteratively using 1 as the first component, $2N \times 10^{-5}$ for the second component, and $((i-1)(i-2)e_{i-1}+(4N\times 10^{-5})(e_{i-1}-e_{i-2}))/(i(i-1))$ for the i^{th} component where e_i is the i^{th} component. This eigenvector was truncated at 10,000 components (truncating is consistent with the direction of the bound) and normalized. The result is that for $N = 100$, the probability of a single ancestral sequence was less than 1.00, the expected number of sequences was greater than 1.00; for $N = 1000$ the probability of a single ancestral sequence was less than .98, the expected number of sequences was greater than 1.02; for $N = 10,000$ the probability of a single ancestral sequence was less than .83, the expected number of sequences was greater than 1.19; for $N = 100,000$ the probability of a single ancestral sequence was less than .20, the expected number of sequences was greater than 2.32; for $N = 1,000,000$ the probability of a single ancestral sequence was less than .00019, the expected number of sequences was greater than 6.59; for $N = 10,000,000$ the probability of a single ancestral sequence was less than 10^{-15}, the expected number of sequences was greater than 20. These values are in Table 1.

DISCUSSION

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 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. 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. 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, 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.

This paper presents bounds on the size of the ancestral pool for a gene defined as 1000 contiguous base pairs, and Wiuf and Hein (1997) 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 gene, and an estimate for chromosomes which does not rely on simulation for the population size of interest. But the most useful extension of these results would be a technique to extend them to genes with multiple exons which are separated by significant distance (perhaps on different chromosomes) or for multiple chromosomes (perhaps the entire genome). Bounds for multiple exons or chromosomes can be obtained by assuming that the genetic material in different exons (or chromosomes) is in the same individuals as much as possible, or in different individuals as much as possible (the 80 percent pedigree ancestors of Chang (1999) also provides a bound). But that will provide very loose bounds. The nature of the dependence of the location of genetic material in different exons or different chromosomes should be determined.

REFERENCES

Campbell, R. B., 2003. A logistic branching process for population genetics. J. Theoret. Biol. 225, 195-203.

Chang, J. T., 1999. Recent common ancestors of all present-day individuals. Adv. Appl. Prob. 31, 1002–1026.

Dorit, R. L., Akashi, H., Gilbert, W., 1995. Absence of polymorphism at the ZFY locus on the human Y chromosome. Science 268, 1183–1185.

Hudson, R. R., Kaplan N., 1985. Statistical properties of the number of recombination events in the history of DNA sequences. Genetics 111, 147–164.

Kimura, M., Ohta, T., 1969. The average number of generations until fixation of a mutant gene in a finite population. Genetics 61, 763–771.

Kingman, J. F. C., 1982a. The coalescent. Stoch. Proc. Appl. 13, 235–248.

Kingman, J. F. C., 1982b. On the genealogy of large populations. J. Appl. Prob. 19A, 27–43.

Wills, C., 1995. When did Eve live? An Evolutionary detective story. Evolution 49, 593-607.

Wiuf, C., Hein, J., 1997. On the number of ancestors to a DNA sequence. Genetics 147, 1459–1468.

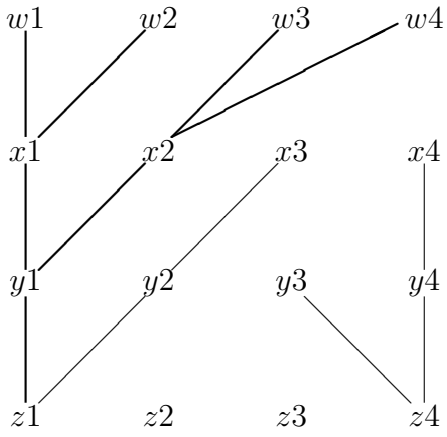
Wiuf, C., Hein J., 1999. The ancestry of a sample of sequences subject to recombination. Genetics 151, 1217–1228.

Table 1: BOUNDS ON IDENTITY PROBABILITIES FOR A GENE

2N	MRCA		Asymptotic ancestor		Asymptotic pool size		
	lower	upper	lower	upper	lower	upper	$\frac{1.28R}{\ln(1+R)}$
200	.98	1.000	.996	1.00	1.00	1.004	1.28
2000	.77	.998	.96	.98	1.02	1.04	1.29
2×10^4	.03	.977	.6	.83	1.19	1.4	1.34
2×10^5	$< 10^{-19}$	.795		.20	2.32	5	1.85
2×10^6	$< 10^{-234}$	.100		.00012	6.59	41	5.34
2×10^7	"0"	1.048×10^{-10}		10^{-15}	20.25	401	27.73

The diploid population size is N, the pairs of columns bound the probability that the MRCA of a base pair is the MRCA of the entire gene, the probability that an asymptotic ancestor of a base pair is an asymptotic ancestor of the entire gene, and the asymptotic expected size of the ancestral pool of a gene. The last column is the estimate from Wiuf and Hein (1997) which was obtained for a different range of parameter values.

Figure 1: Schematic of coalescence



Time advances going up the page. 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.