The Ancestry of a Gene

R. B. Campbell
Department of Mathematics
University of Northern Iowa
Cedar Falls, IA 50614-0506 USA
campbell@math.uni.edu
http://www.math.uni.edu/∼campbell

June 2009
Moscow, Idaho

Introduction Gene fixation in the sense that there is a single ancestor from which all the base pairs in all the copies of a gene in the population are descended only occurs in small ($N < 1000$) populations. In large populations ($N > 1\,000\,000$) crossing over (recombination) within the gene provides that there is an ancestral pool rather than a single ancestor of the gene. In the absence of recombination there is a common ancestor, such as the mitochondrial Eve or the y-chromosome Adam. Wiuf and Hein (1997) have provided an estimate for the size of the ancestral pool for chromosome 20, and Chang (1999) has provided an upper bound for the size of the ancestral pool for the entire genome. This paper presents upper and lower bounds for the probability of existence of a common ancestor, and the expected size of the ancestral pool, for a gene.

The Model We assume a diploid population of N individuals (haploid size $2N$, the analysis is haploid), the effective population size is the same as the actual population size. A gene is defined as 1000 contiguous base pairs. The probability of recombination between adjacent base pairs is 10^{-8}. Coalescence decreases the number of ancestors as modelled by Kingman (1982a,b); only coalescence of the entire population is considered. Crossing over increases the number of ancestors in accordance with the recombination model of Hudson and Kaplan (1985) as employed by Wiuf and Hein (1997).

Existence of a MRCA Whether all the base pairs share the same most recent common ancestor (MRCA) is measured by the probability that a cross over event occurred between a member of the coalescent and an individual not descended from the MRCA of the coalescent (in Fig. 1, $x1$ and $x2$ are in the coalescent, $x3$ is outside the coalescent but is descended from the MRCA, and $x4$ is outside the coalescent and is not descended from the MRCA. Calculations are based on the number of individuals in the coalescent t generations after the MRCA $(1 + 1/2N - t/4N)^{-1}$ and the rate of increase of the number of descendants of the MRCA $1-(k-1)/(2N-1)$ (k is the number of descendants). Results in Table 1 show that in populations smaller than 1000 all the base pairs in a gene share the same MRCA more than .77 of the time, but in populations larger than $1\,000\,000$ they share the same MRCA less than .10 of the time.

Size of Ancestral Pool An upper bound for the expected size of an ancestral pool (and a lower bound for the asymptotic probability of there being a single ancestor in a given generation) is obtained from the inequality $E[k(k-1)/(4N)] \leq E[k \times 10^{-5}]$ where k is the size of the ancestral pool, the left hand side is the probability a coalescent event occurs, and the right hand side is an upper bound for the probability a cross over event occurs. With convexity, this yields the inequality $E[k] \leq 1 + 4N \times 10^{-5}$. A lower bound for the expected size of the ancestral pool (and an upper bound for the asymptotic probability of there being a single ancestor in a given generation) is obtained from the infinite stochastic matrix

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

where the superdiagonal reflects coalescence, the subdiagonal is a lower bound for crossing over, and the diagonal reflects no change. This provides an asymptotic pool size of approximately 1 for $N < 1000$ and an asymptotic pool size greater than 6 for $N > 1\,000\,000$ as indicated in Table 1.

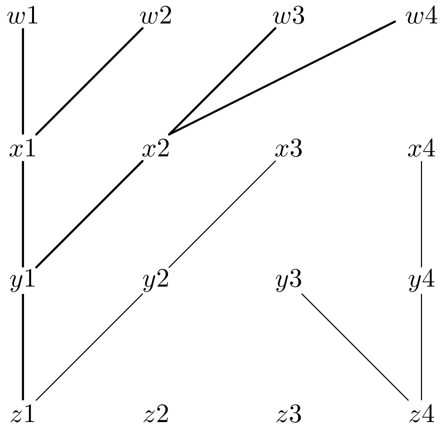
Discussion One conclusion of this study is that the gene is not the atom of evolution in large populations. A single gene does not become fixed in the population, rather crossing over during the fixation process entails that at every locus the genes have an ancestral pool rather than a common ancestor. If one wants to think of mutations becoming fixed, mutations must be viewed as the base pair which mutates, not the gene which contains the base pair.

This study assumed selective neutrality. Selection would shorten the substitution process, and might eliminate aberrant forms resulting from crossing over, hence make fixation of an intact gene more likely. These results should be generalized to incorporate selection.

In general, genes are not a sequence of contiguous base pairs, but a union of separated exons. such separation will of course increase the frequency of recombination, and these results should be extended to incorporate multiple exons. The results of Wiuf and Hein should also be extended to multiple chromosomes to provide an estimate of the size of the ancestral pool of the genome.

(over)

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.

Table 1: BOUNDS ON IDENTITY PROBABILITIES FOR A GENE

2N	MRCA		Asymptotic ancestor		Asymptotic pool size	
	lower	upper	lower	upper	lower	upper
200	.98	1.000	.996	1.00	1.00	1.004
2000	.77	.998	.96	.98	1.02	1.04
2×10^4	.03	.977	.6	.83	1.19	1.4
2×10^5	$< 10^{-19}$	.795		.20	2.32	5
2×10^6	$< 10^{-234}$	.100		.00012	6.59	41
2×10^7	"0"	1.048×10^{-10}		10^{-15}	20.25	401

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.

REFERENCES

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

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

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.

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