

The Coancestry of Unlinked Genetic Segments

June 2010
Portland, Oregon

R. B. Campbell
Department of Mathematics/University of Northern Iowa/Cedar Falls, IA 50614-0506

campbell@math.uni.edu
<http://www.math.uni.edu/~campbell>

The Problem: Whether studying a single gene or the entire genome, the question “How many individuals in previous generations contained genetic material ancestral to the present generation?” arises. Both genes and the entire genome are composed of multiple segments, often unlinked, which is the case we consider here. We assume the size of the ancestral pool (individuals in a previous generation with genetic material ancestral to the present generation) is known for contiguous segments of DNA (this is known in some cases, including an entire chromosome), and study the intersection of the ancestral pools of multiple DNA segments. The question is how much genetic association is maintained by the diploid segregation mechanism. The standard model of random mating with a Poisson progeny distribution is assumed, and we census at the gamete phase. Only the case of free recombination (no linkage) between segments is considered here.

Single Individual Skeletons: The heuristic answer that because unlinked genetic material remains together half the time in the segregation process, unlinked genetic material will be together twice as often as would randomly occur is essentially correct. A Markov process going back in time with the coalescent process for whether the skeletons (ancestral lineages) of two unlinked nucleotide sites (or segments for which $Nn < 1\,000\,000$, where N is the population size and n is the number of base pairs in a segment, since recombination is not significant then) were in the same gamete is governed by the matrix:

$$\begin{matrix} .5 & 1/2N \\ .5 & 1 - 1/2N \end{matrix}$$

which has the eigenvector (stable distribution) $\langle 1/(1+N), N/(1+N) \rangle$, hence the two skeletons will be in the same gamete approximately $1/N$ of the time instead of $1/2N$ as would occur by chance (the haploid size is $2N$).

Larger Ancestral Pools: If there is an ancestral pool rather than a single individual in the skeleton for one of the segments, the matrix governing the the Markov process for identity of a single individual skeleton with that pool of relative size u is approximately

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

which has eigenvector $\langle u/(.5 + .375u), (.5 - .625u)/(.5 + .375u) \rangle$, hence the skeleton will share a gamete with the ancestral pool of the other site approximately $u/(.5 + .375u)$ of the time (this is $2u$ for small u). This provides an upper bound on the probability a gamete contains material from two pools of size greater than $1/2N$, because gametes with skeletons of multiple base pair sites from one segment will tend to have multiple skeletons from the other segment.

Recursively Cumulating Ancestral Pools: The above bound on the intersection of ancestral pools can be used recursively to find a lower bound on the total percentage of gametes containing ancestral material of at least one segment, if the sizes of the ancestral pools of the (unlinked) segments are known:

$$u' = v + u - uv/(.5 + .375u)$$

where u is the size of the cumulative pool, v is the size of the pool being cumulated in, and $'$ designates the next iteration. Numerical calculations show that the cumulative pool will be at least 50% if the number of unlinked segments is $1/u$ where u is the relative frequency of each segment, at least 75% if there are $3/u$ segments, and at least 80% if there are $40/u$ segments. An example of interest is that 22 unlinked ancestral pools, each of which is .15 of population size will provide a cumulative ancestral pool of about 77% of the population size. This calculation is motivated by Wiuf and Hein’s estimate that the ancestral pool for chromosome 22 is approximately .13 of the population size, the observation that other chromosomes should have larger ancestral pools since they are larger, and the fact that there are 22 paired chromosomes in humans (excluding the sex chromosomes). Thus we see that for humans the size of the ancestral pool is close to Chang’s bound of 80%.

Summary: The diploid segregation structure approximately doubles the number of gametes which contain genetic material from unlinked loci from what would occur under random association. This is a small effect, in organisms such as humans the size of the cumulative ancestral pool will be close to the 80% bound.

REFERENCES

- Chang, J. T., 1999. Recent common ancestors of all present-day individuals. *Adv. Appl. Prob.* 31, 1002–1026.
Wiuf, C., Hein, J., 1997. On the number of ancestors to a DNA sequence. *Genetics* 147, 1459–1468.