R. B. Campbell
Department of Mathematics
University of Northern Iowa
Cedar Falls, IA 50614-0506
http://www.math.uni.edu/∼campbell
campbell@math.uni.edu
(319)273-2447

July 2012
Ottawa
Canada

The Distribution of Ancestral Segments

Question: What are the roots of present day populations: How has our genetic material been transmitted through our forebears, how few individuals comprise the ultimate source of our genetic material, how was that genetic material distributed among them? Is the concept of most recent common ancestor meaningful for genetic regions larger than a single nucleotide base site? The specific question we shall address is the distribution of the lengths of genetic segments which are ancestral to the present generation, i.e., contiguous segments of DNA where each base pair is the common ancestor of that nucleotide site in the current generation, in a generation in the asymptotic past. This does not mean that those segments have been transmitted to the present generation intact, they may have bifurcated and subsequently reunited before their base pairs became fixed in the population.

Model: The standard Wright-Fisher model for a diploid population is assumed. DNA is modelled as a single long strand with a uniform recombination rate (r is the probability of a cross-over between two adjacent base pairs). (Results will not be significantly impacted by hotspots where the recombination rate varies, including free recombination between chromosomes.) The coalescent model of individuals randomly choosing their parents which follows from the Poisson progeny distribution is used (this entails selective neutrality). The size of the genome (number of base pairs), denoted L, is constant, and the base pairs are not rearranged. The (effective) population size N (haploid size $2N$, we shall census at the gamete phase) is also constant.

Analysis: The coalescent process results in an ancestral lineage for every base pair before (i.e., after going backward in time) coalescence has occurred. If the ancestral lineages of two base pair sites are in the same gamete, they will have been in the same gamete the previous generation unless recombination (crossing over) occurred. If they are in different gametes, they will have been in different gametes the previous generation unless coalescence or recombination occurred. The analysis is based on adjacent base pairs, for which case recombination joining adjacent lineages is less frequent than coalescence and can be ignored as a cause of adjacent base pairs becoming in the same gamete. The probability of crossing over between two adjacent base pairs r counters the probability of coalescence $1/2N$. This provides a Markov chain (going backward) for whether two adjacent ancestral lineages are in the same gamete, which will approach equilibrium in the asymptotic past. Assuming crossing over occurs at random locations and coalescence is random, adjacent ancestral lineages in the same gamete should occur uniformly throughout the genome.

Results: Pairs of adjacent ancestral lineages in different gametes demarcate the ends of ancestral genetic segments, hence the number of such pairs equals the number of ancestral segments. The asymptotic state of the Markov chain provides that the probability that two adjacent ancestral lineages are in different gametes is $\frac{2N}{2N+1/r}$. If $2Nr \gg 1$, this ratio is essentially one, hence most pairs of adjacent ancestal lineages are in different gametes (most ancestral segments have length one) and there are approximately L ancestal segments. If $2Nr \ll 1$, $\frac{2N}{2N+1/r}$ is almost zero, entailing that few adjacent ancestral lineages are in different individuals. The number of such pairs, which is the number of ancestral segments, is $L \times \frac{2N}{2N+1/r} \doteq 2LNr$. The segment lengths will follow a geometric distributon (actually, a geometric distribution shifted by 1) with mean $\frac{2Nr+1}{2Nr}$ and standard deviation $\frac{\sqrt{2Nr+1}}{2Nr}$.

Asymptotic Past: The asymptotic past can be defined as beginning (ending in real time) when the ancestral lineages are established and a crossover has occurred between each pair of adjacent ancestral lineages. The time until a common ancestor (establishment of ancestral lineage) is about 4N generations with standard deviation $2N$ generations, so the asymptotic past must begin some time before several N generations ago. The expected time (exponential waiting time) until a recombination event at a specific location is $1/r$ with standard deviation $1/r$, so the asymptotic past must begin several $1/r$ generations earlier (in real time). Therefore, the asymptotic past must begin max(several N, few $1/r$) generations ago, since one term will generally dominate the other. In particular, if $N > 1/r$ the asymptotic past begins when ancestral lineages are established for single nucleotide sites, and this work is concerned with whether common ancestry is meaningful for segments longer than one nucleotide site.

Discussion: This work provides a good characterization of the distribution of ancestral genetic material in populations in the asymptotic past under neutrality, but several caveats for the utility of the results should be mentioned. Identity by descent is studied, identity by type may be of greater interest for many applications. An estimate for the recombination rate between adjacent nucleotides is $r = 10^{-8}$, which provides that if the (effective) population size is $N < 10^6$, there will be many nontrivial ancestral segments, many the lengths of genes; but if $N > 10^{10}$, most ancestral segments will be trivial (of length 1). The nature of the results depends on the (effective) population size. Assuming $r = 10^{-8}$, the asymptotic past begins at least 10^9 generations ago, longer ago if the (effective) population size is greater than $N = 10^8$. The ancestral past may be before a species had its current population size, perhaps before the species was formed.