CS 262 Lecture 14 Notes
Human Genome Diversity, Coalescence and Haplotypes

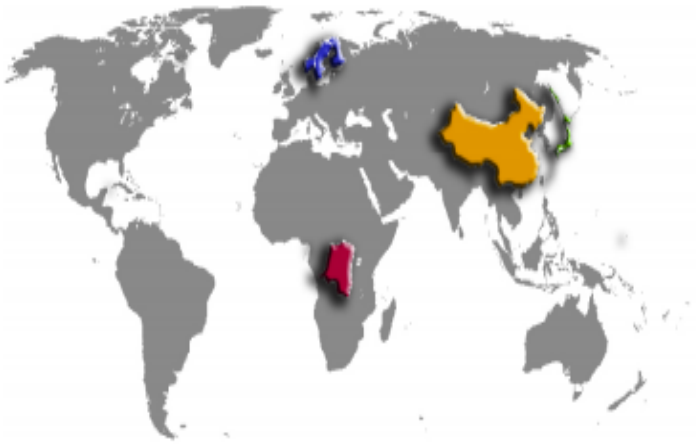
Scribe: Alex Wells
2/18/16

Coalescence

Whenever you observe two sequences that are similar, there is actually a single individual who is the ancestor of everyone sharing those sequences today. This phenomenon is known as coalescence. Coalescence can be within the same species, or to an ancestor that is not within the same species as its descendants. For example, consider a homologous gene that is shared by humans and mice. This suggests that there is some species that had this gene which is an ancestor of both humans and mice.

The HapMap Project

The HapMap project was the first attempt to catalog the diversity of the human genome, and was stared soon after the genome was sequenced (~2003). The project began by collecting sequences from 3 populations in Asia, Africa, and Europe (later expanded to 10 populations).



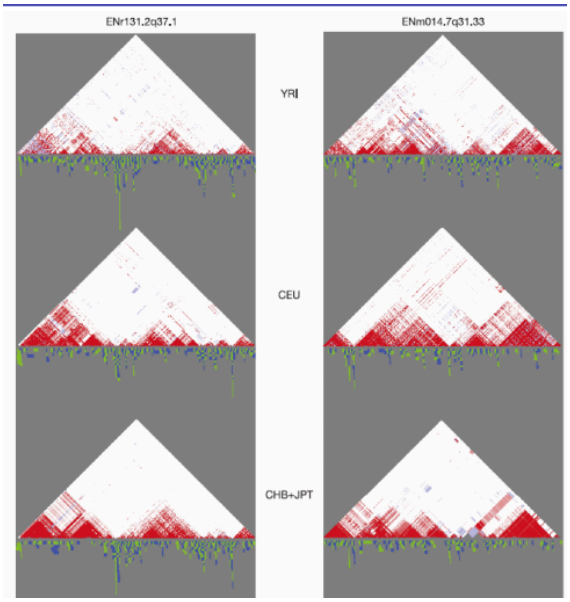
The image to the left shows the initial populations sequenced by the HapMap project. Populations from Asia, Europe and Africa were sequenced.

The overall goal was to find positions in the genome that had a high minor allele frequency. Once the locations of these alleles are known, then it becomes easy to tell if an individual has two major alleles, two minor alleles, or one of each at a given position. To accomplish this, we first determine the complementary strand around the DNA position of interest for each of the possible alleles that someone can have at that position. Then we can put these oligos on a microarray, and wash human DNA over the chip. The DNA will stick to the appropriate complementary locations in the microarray and fluoresce a different color depending on the genotype present. Each microarray can be used to query many different positions of

interest in the human genome. 23andMe and other genetic companies use this method.

Linkage Disequilibrium

Definition: the non-random association of alleles at different loci.
To calculate linkage disequilibrium, we look at two positions that vary in the human genome and find the allele frequency at each position, for example P_A and P_G. Then linkage disequilibrium = P(A and G)- P_A P_G. Linkage disequilibrium exists because recombination happens very infrequently per genomic position, so across a number of generations there is not enough time to break allele combinations for nearby alleles.



For each genomic positions i and j, we plot (at a right angle) the degree of linkage disequilibrium between the two positions. Therefore, in the image to the left, a solid red triangle means that all position between i and j has very high linkage disequilibrium, so the region is inherited as a block.

Recombination hotspots: locations that are more likely than others to recombine during meiosis.

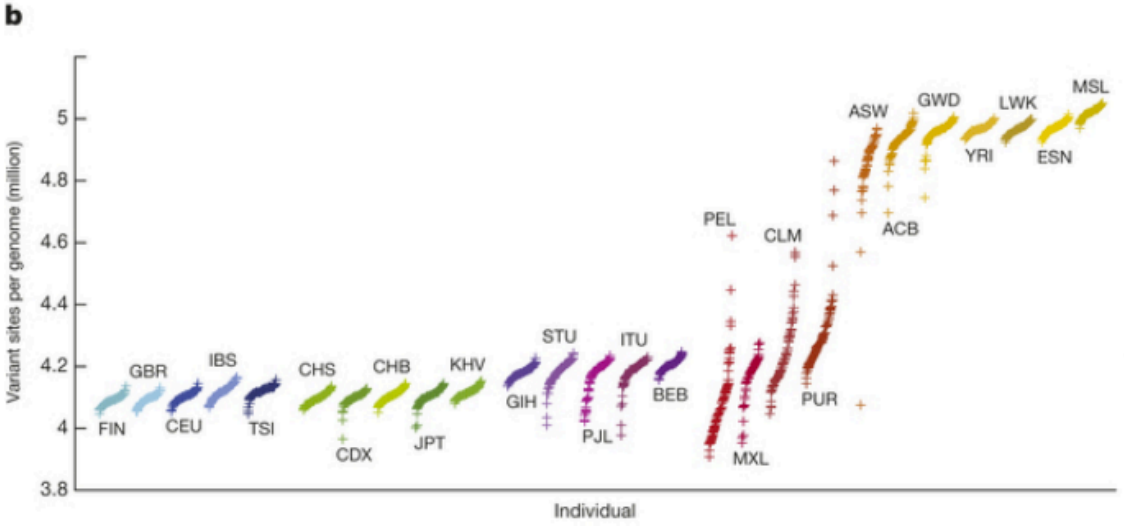
If we look at the human population, we have very specific hotspots. The chimpanzee also has very specific recombination hotspots, therefore since human and chimpanzee are 99% identical we would expect hotspots to be highly related. However, this turns out to be false, as hotspots between humans and chimpanzees are completely uncorrelated.

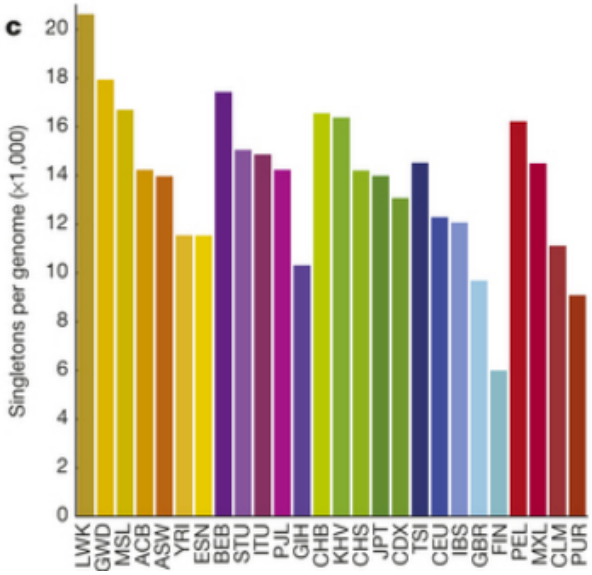
Once a hotspot is "used" in a generation, it is not a hotspot anymore (won't be used again).

1000 Genome Project

The goal of the 1000 genome project is to find the most genetic variants that have frequency of at least 1% in the populations studied (rare alleles). Even though common alleles occur more than rare alleles, most of the variation in the human genome comes from rare alleles.

If we take any two single haplotype human genomes, they are roughly (on a nucleotide level) 99.9% identical (3 million differences). If we take two diploid genomes and put them together, there will be more differences. The graph below shows the number of diploid differences between paired individuals. On the very left portion of the graph, we can see that Finland has the fewest differences per paired diploid individuals (4.1 million differences), which makes sense because the population is very confined with lower heterozygosity. The African populations studied have close to 5 million differences (more heterozygosity).





This graph shows the number of singletons per genome across different populations. Singletons are variants that only occur once. African populations have between 11,000 and 20,000 singletons, while European populations have the fewest singletons per individual (~6,000 to 10,000).

UK10K

The goal of this project is to better understand the link between low frequency and rare genetic changes, and human disease caused by harmful changes to proteins the body makes. This project studied the genetic code of 10,000 individuals.

Purifying selection is when a given allele is harmful for fitness. Then the allele frequency tends to be lower, either because they are very new or they are getting eliminated. Singletons have a high proportion of deleterious alleles compared to non-singletons.

Derived allele frequency: the allele frequency of an allele that is not an ancient allele. Not the same as a minor allele frequency. If we look at a given allele, it can be a major or minor allele, or a new or ancient allele. For example, if chimpanzees and gorillas have an A in a specific position, and humans have an A and C at that position, then the A is the ancient allele and the C is the derived allele.

Population Sequencing

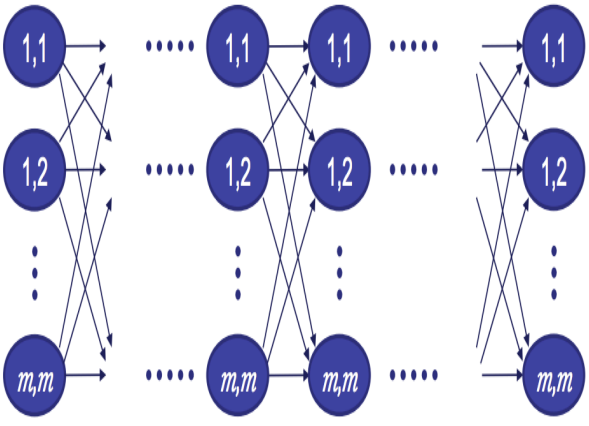
Given a specific budget, how many genomes do we want to sequence. If we had an unlimited budget, we could sequence every person very deeply, so at each position in the genome, we will have enough reads lining up so we can confidently call each individual. However, there are usually budget restrictions, so instead we can use the fact that individuals tend to be similar to each other. Therefore we can use the information of some individuals to ascertain information about another individual. Using this method we can get by with low sequence read data (5-7x coverage).

Imagine we are given a set of individuals $G_1,G_2,\ldots,G_N$ where $Gi = g_{i1},\ldots,g_{in}$ (haplotype) and $P_1,P_2,\ldots,P_N$ where $P_i = [p_{ijg} = Pr(g_{ij} = g|data)]$.

When coverage is high (> 30x), then it is easy to calculate $Pr(g_{ij} = g|data)$ since this is approximately equal to $Pr(g_{ij} = g|reads\ mapping\ on\ (i,j))$. However, in the presence of little data (low coverage) it is difficult to separate sequencing errors from rare alleles. Therefore in this scenario, we want to leverage the LD of the population that is being sequenced. This can be done using an HMM.

Li and Stephens Model:

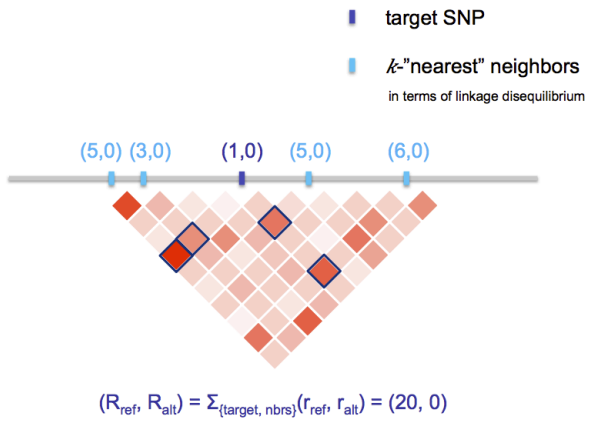
Models the haplotypes and variation of a population. A population is modeled as a collection of m halplotypes. A haplotype is modeled as a chain of states. Therefore a population is a collection of m chains (as shown below). An individual is a path through a markov chain, where at every position we have m^2 states and m^4 positions.



Reveel/Informative Neighbors:

The Li and Stephens model is problematic since LD does not follow an HMM architecture. Instead, in order to differentiate between errors in reads and rare alleles, we will use haplotype and linkage disequilibrium information.
First, given a target SNP, we must find the k-'nearest' neighbors in terms of linkage disequilibrium. To do this, we will use a graph similar to the one mentioned in the "linkage disequilibrium" section.

The diagram to the left shows a visual representation of k-nearest neighbors with respect to linkage disequilibrium. Note that the 'target SNP' is aligned at a right angle to other locations in the genome that have a high degree of linkage disequilibrium with it.



Now consider the following algorithm:

Reveel:

1. Identify candidate polymorphic sites
2. Calculate k nearest neighbors
   - Jaccard indices Sim_1, Sim_2, Sim_3
3. Initialize $G^{(0)}$
4. Summarization/Maximization
   $p^{(n+1)}_{ijg} = Prob(g_{ij} = g \mid G^{(n)}, data)$
   $g^{(n+1)}_{ijg} = argmax\ p^{(n+1)}_{ijg}$
5. Recalculate k nearest neighbors
   - Approximate Correlation Coefficient (Schaid 2004)
6. Summarization/Maximization
7. Recalculate k nearest neighbors
   - Approximate CC, Entropy
8. Summarization/Maximization

A candidate polymorphic site is a position j where some individuals have at least 2 reads with the same minor allele.
This algorithm maximizes the probability of having true reads by modeling linkage disequilibrium using the nearest neighbors procedure described above.

Molecular Evolution and Phylogenetic Tree reconstruction

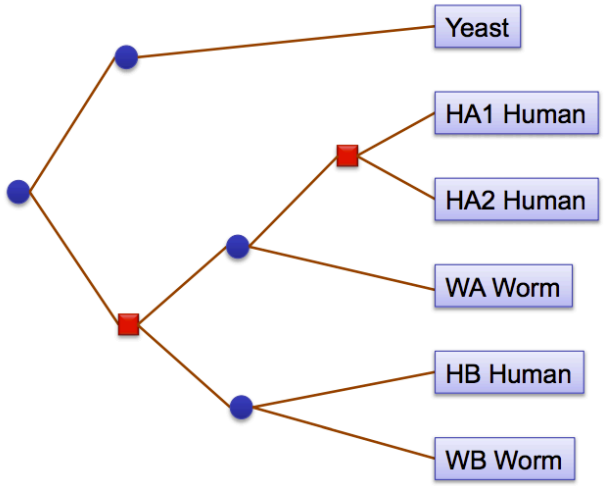
The second portion of the lecture focused on mathematically modeling molecular evolution. Molecular evolution occurs due to changes in DNA copies from one generation to the next. These changes are usually due to small insertions and deletions, often of a single base. Even though larger mutation events occur, we will focus on modeling these single base change events.

Proteins (genes) evolve by both duplication and species divergence. A gene that creates a protein is highly specialized. A few changes in the regulatory sequence, improper splicing, or a few amino acid errors may result in a dysfunctional protein.

How do we get amazing gene diversity? The answer centers on the process of duplication. Suppose we have 1 functional gene, but after some event it is duplicated, so now the genome contains two functional copies of the gene. Now, one of the copies can evolve while the other maintains its original function. This allows the organism to keep making the protein encoded by the original gene while simultaneously allowing new proteins to be formed as the other copy is mutated.

Orthology and Paralogy

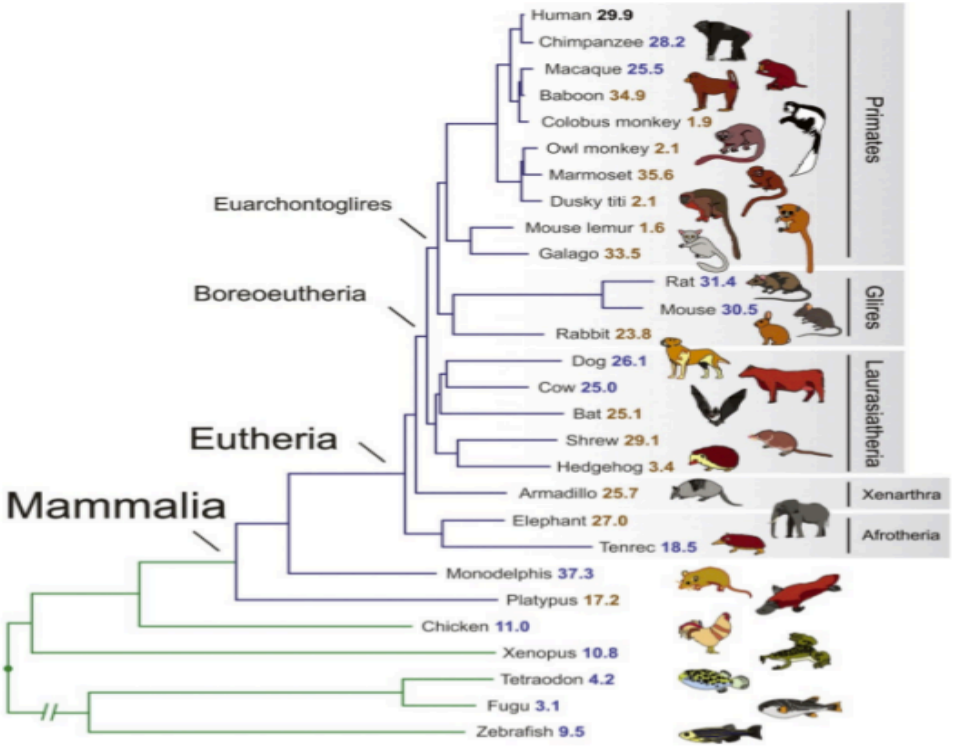
Homologs can be classified as either orthologs or paralogs. Orthologs are derived from speciation, paralogs are everything else.



In this tree, duplication events are represented as red squares, and speciation events are represented as blue circles. To determine if two genes are orthologs or paralogs, follow the edges of the tree back to where they converge. If they converge at a red square, then they are paralogs, if they converge at a blue circle, they are orthologs.

Phylogenies

Phylogenies are a compact way of representing the evolution of species. In a phylogenetic tree, nodes are species, edges are time of independent evolution, edge length represents evolutionary time (genetic distance, not necessarily chronological time).



Above is a phylogeny tree of species sequenced so far.