CS262 Lecture 19: Human Population Genomics Continued

Scribe: Gus Liu
March 2016

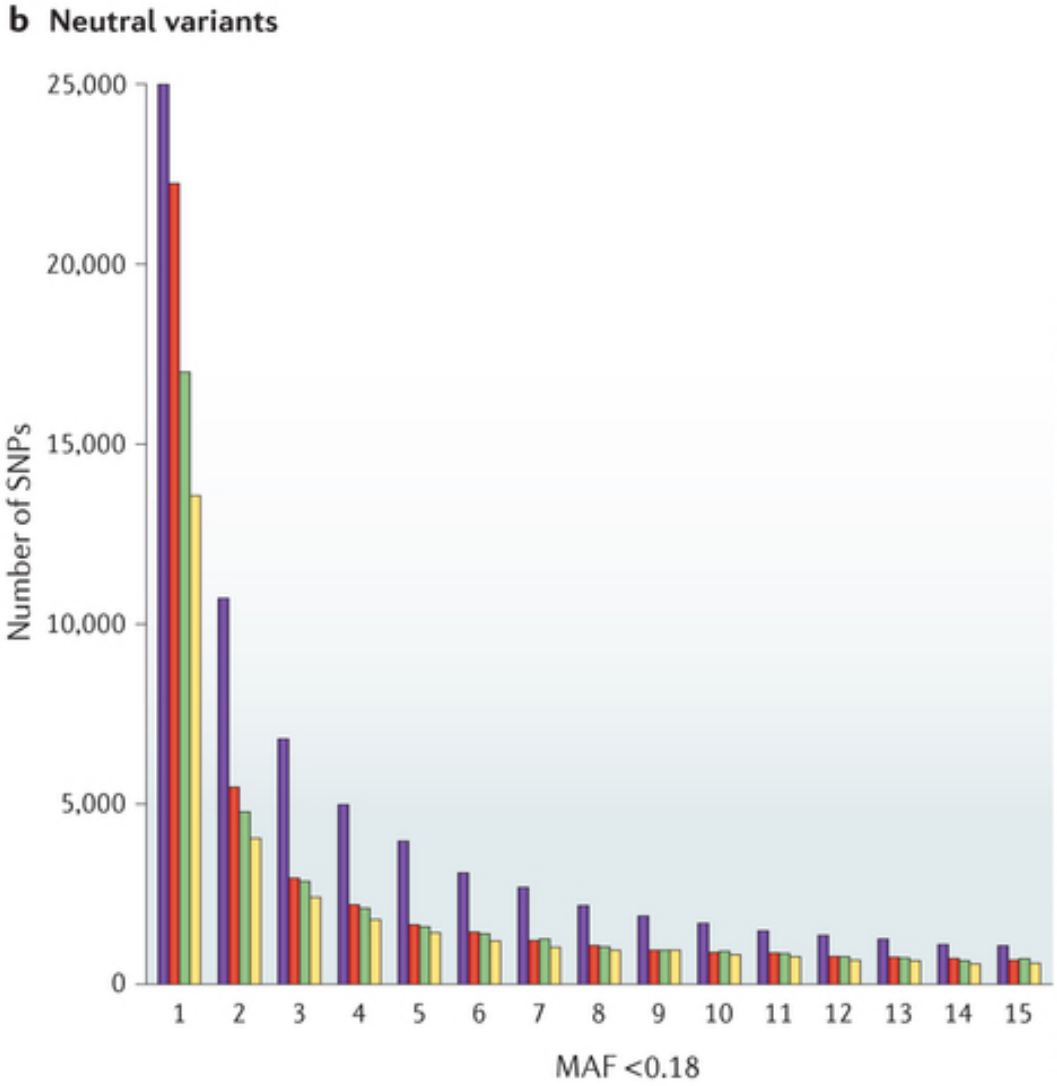
1 Lecture Overview

In this class, we continued the discussion of human population genomics, delving into the topic of detecting positive selection, negative selection, and neutral drift of alleles. We focused the discussion on ways of mathematically modeling population genetics using histograms, trees, and probability. Then we assessed the impact of population genetics analysis through application on malaria and the process of its spread. When tests for selection of long haplotypes in humans began, G6PD and TNFSF5 were found; these genes were positively selected because being heterozygous confers resistance to malaria by not allowing it to multiply efficiently in the bloodstream.

2 Positive Selection, Negative Selection, and Fixation

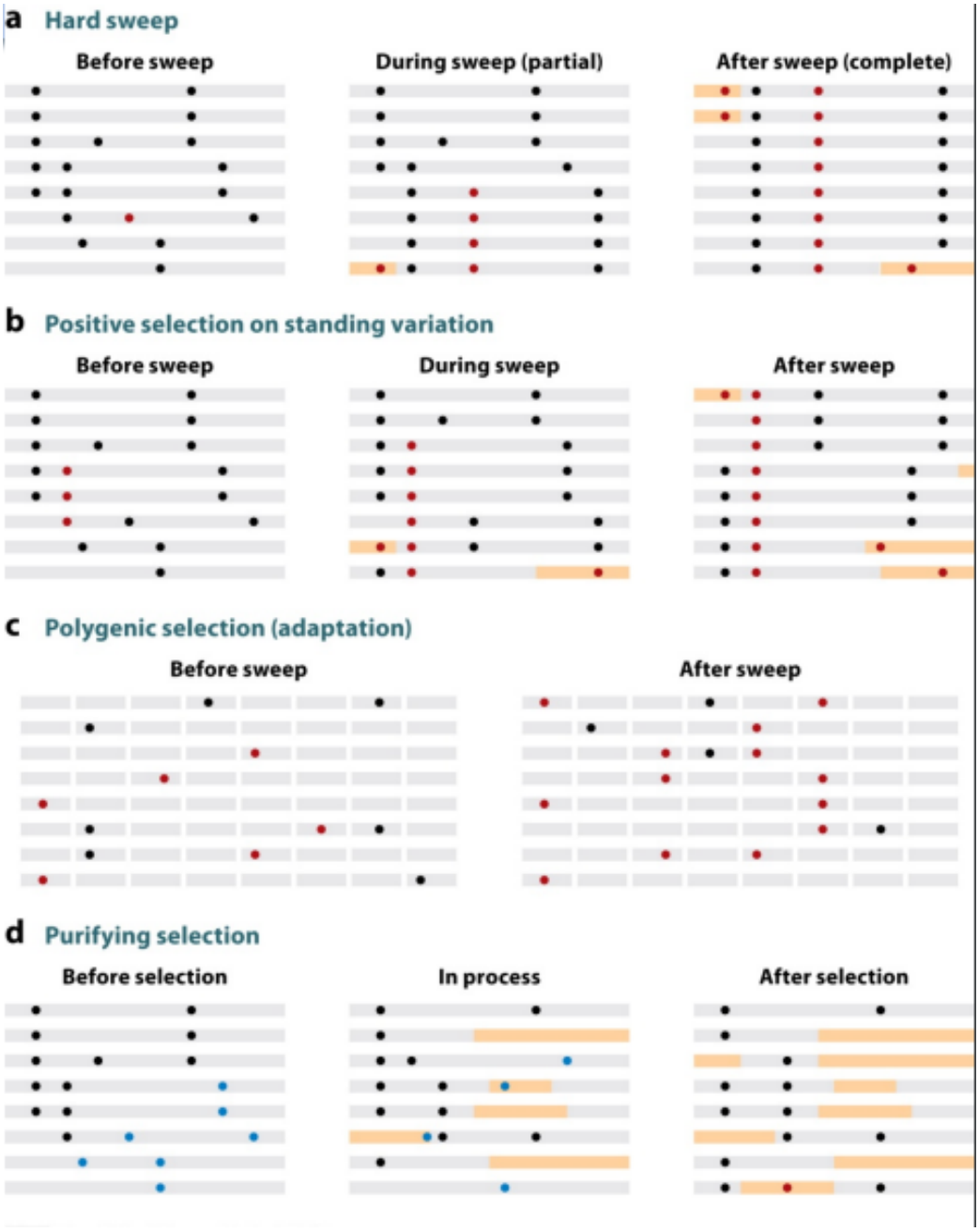
There are two types of natural selection in biological evolution: Positive (Darwinian) selection promotes the spread of beneficial alleles, and negative (or purifying) selection hinders the spread of deleterious alleles. Pseudogenization is normally detrimental and prevented by negative selection. However, changes in genetic background or environment may render a formerly useful gene worthless, leading to the relaxation of the negative selection. Consequently, mutations disrupting the gene are fixed by genetic drift, and the gene becomes a pseudogene. In population genetics, fixation is the change in a gene pool from a situation where there exists at least two variants of a particular gene (allele) to a situation where only one of the alleles remains. The effect of alleles depend on the environment, where a gene expression can be beneficial in one environment and harmful in another. Based on its effect (or lack of effect), it can undergo positive selection, negative selection, or neutral drift. If we start with a fixed number of alleles, each as a probability of becoming fixed equal to its proportion. We may ask - why are genes being preserved?

The answer is purifying selection. Suppose we look at a given class of regions in the human genome, for example that of stop codons TAG, TGA, and TAA. Over this class of regions, we will find that a region that should not be altered will have a depressed allele spectrum, evidence of purifying selection, similar to the histogram shown below.



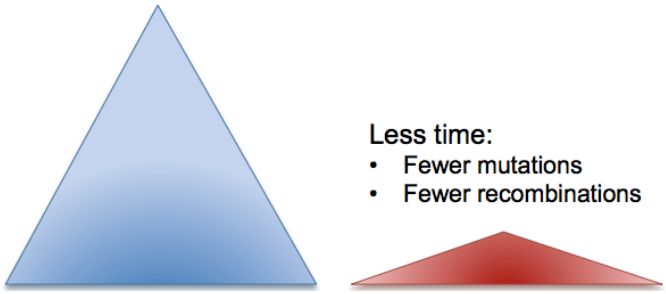
3 Detecting positive selection

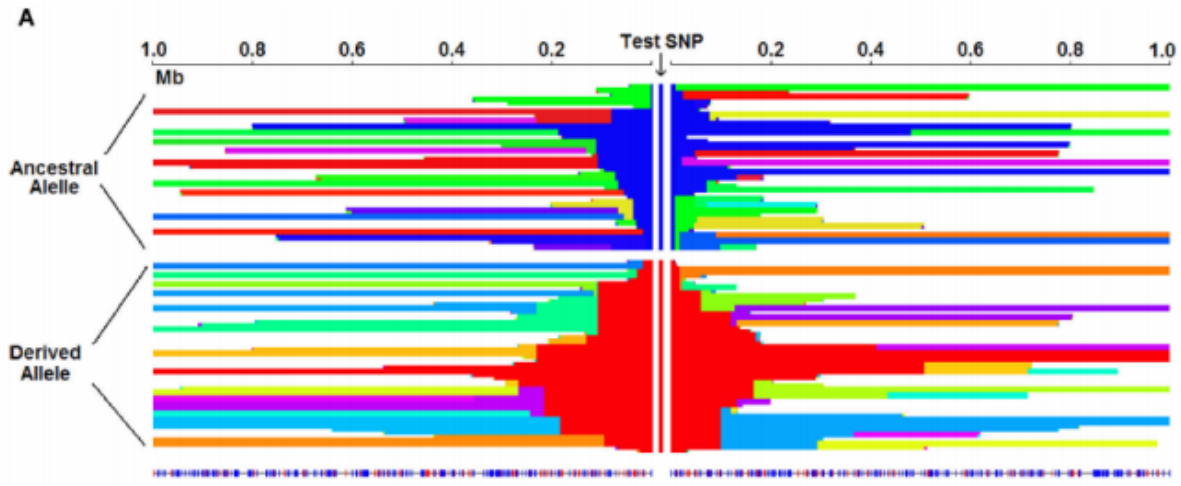
We study the detection of recent positive selection with a statistical measure of haplotype length. This method can only be used for recently positive selected alleles that have not yet reached fixation and thus are still expanding. Mutations are more likely to occur for silent alleles than those that change the amino acid. Therefore, if a particular gene tends to have more amino acid-changing mutations, we say that the Ka/Ks ratio is high and thus is under high selective pressure to be changing. Examples include an immune response or spermatogenesis. Haplotypes can carry many other alleles with the expansion, removing lots of variation around the allele in question. In other words, the only alleles that survive are those that recombine or shuffle, while others get removed.



We can see evidence of this expansion and removal of variation in haplotypes during positive selection of an allele. Observing this trend can help us identify positive selection in recent years. A derived allele is an allele that presumably appeared in humans rather than in a common ancestor like a chimpanzee. If you observe A and G at a particular position, you can ask which one is the derived allele? To answer this, we take a look at the allele frequency spectrum of a common ancestor. If it recently expanded quickly from selection, it has carried other alleles around it with it. Because derived alleles have lower frequency, we can measure the general frequency and compare it to that of the allele. If the allele has high frequency in comparison, it must have expanded recently and thus indicates positive selection.

Long haplotypes can help us identify positive selection, because recombination does not have time to shuffle alleles around if it is being positively selected. Below are different expansion trees where the different diagrams show different views of the same alleles, where the red allele is expanding fastesr than the blue allele. We can observe positions with very high linkage disenquilibrium, which is the non-random association of alleles at different loci and compare the trees. We find that the rate is proportional to the area of the triangle, or the nodes of the tree. From the second diagram, we can take the integral of lengths, find the area, and compare that way.





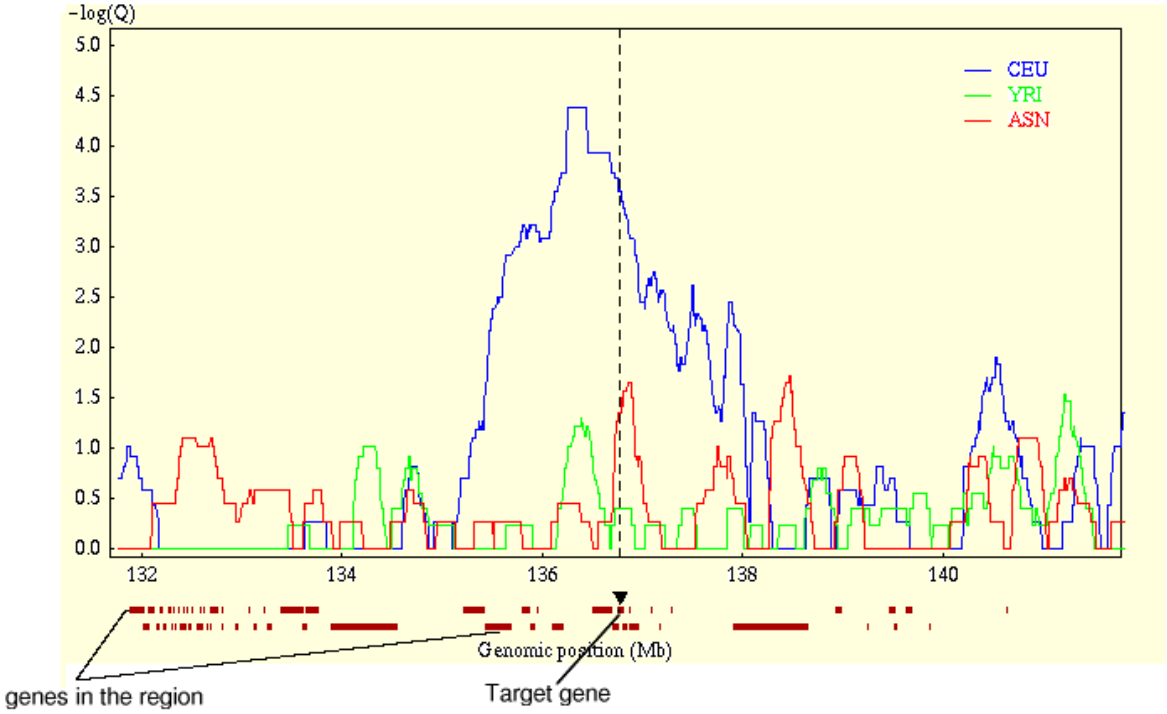
$$iHS = \ln\left(\frac{iHH_A}{iHH_D}\right)$$

4 Integrated Haplotype Score

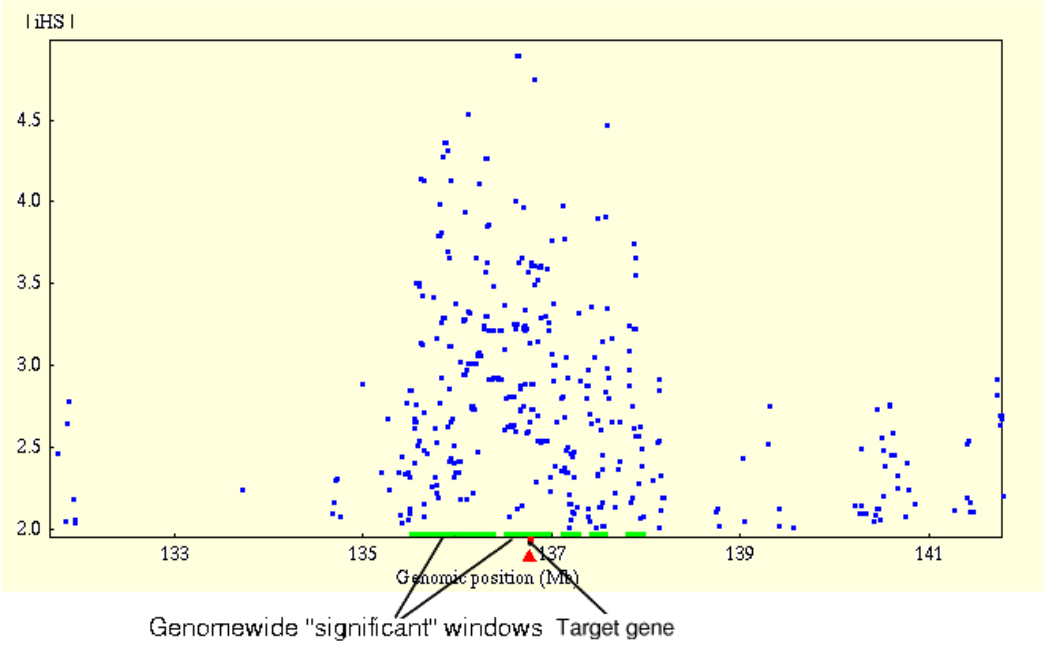
Paper: http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0040072

iHS is a statistic developed to detect evidence of recent positive selection at a particular

locus. It is based on differential levels of linkage disequilibrium, or LD, surrounding a positively selected allele compared to the background allele at the same position. A very high iHS score, i.e. iHS ¿ 2, means that haplotypes on the ancestral allele background are longer compared to the derived allele background. In contrast, a very negative score indicates that the haplotypes on the derived allele background are longer compared to those of the ancestral allele. The power of iHS also depends on many other factors such as local SNP density, SNP ascertainment, availability of haplotype phase information, and intensity of selection among others. Below is a sample iHS plot output for querying a gene.



Here is a sample iHS scatter plot from which we can gain insight given the iHS scores.

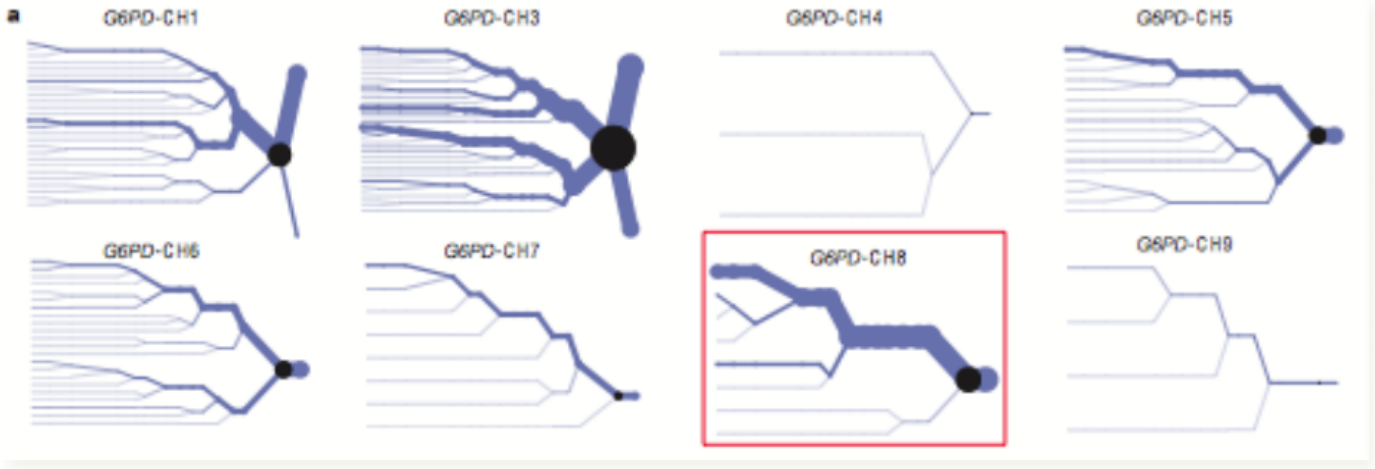


A sample iHS scatter plot

5 Application: Malaria

First, we discuss the life cycle of malaria, which involves malaria parasites infecting two types of hosts - humans and female *Anopheles* mosquitoes. In humans, the parasites grow and multiply first in the liver and then in the red cells of the blood. Once in the bloodstream, successive broods of parasites grow inside the red blood cells and destroy them, releasing daughter parasites, or merozoites, that continue the cycle by invading other red cells. After 10-18 days, the parasites, sporozoites, are found in the mosquito's salivary glands. When the mosquito bites, it start another infection when it parasitizes the liver cells.

When tests for selection for long haplotypes in humans began, genes G6PD and TNFSF5 were found. Because G6PD is well known for triggering hemolysis when eating fava beans, one might wonder why it exists from positive selection (a strong positive selection at that)? It is not purified away because being heterozygous protects individuals from malaria. In fact, in some cases affected males have shown complete immunity to the disease. One theory to explain this is that cells infected with the *Plamodium* parasite are cleared more quickly by the spleen, giving G6PD deficiency carriers an evolutionary advantage by increasing their fitness in malarial endemic environments.



These were the results observed with G6PD. The thickness of the line shows how many humans contain the nearby high frequency equilibrium for the allele.