Population Genomics I

Lecturer: Dr. Pardis Sabeti

Introduction
Many types of genetic differences may be present within populations of the same species, including single nucleotide polymorphisms (SNPs), variable number tandem repeats, DNA insertions/deletions, and large-scale rearrangements and copy number variants. All of these are likely to play roles in humans phenotypic variation and genetic disorders. In this lecture, we will focus on SNPs in the human genome.

Allele and genotype frequencies
Every individual has two copies of each chromosome (except sex chromosomes). An allele is the ‘value’ of a SNP found on a particular chromosome (e.g. A). The genotype of an individual is the tuple of the two alleles found in her two copies of the chromosome (e.g. AT).

Obviously, a given position in the genome could in principle have up to four alleles (A, G, C, or T) in the population. For simplicity, we will assume that a SNP in any chromosome in the population takes on one of only two alleles.

For a given SNP, the allele frequency is the fraction of chromosomes in the population with a certain allele – if \( N \) is the population size, then there are \( 2N \) chromosomes. The genotype frequency is the fraction of individuals in the population with a certain genotype. The genotype of an individual can be homozygous for one allele or the other (e.g. AA or TT) or heterozygous, where the two alleles differ (e.g. AT; we don’t distinguish the order, so AT and TA are both heterozygous).

Under various idealistic assumptions, the allele and genotype frequencies are related by the Hardy-Weinberg Law: for an allele with frequency \( p \), the frequencies of the homozygous genotypes are \( p^2 \) and \((1 - p)^2\), while the frequency of heterozygotes is \(2p(1 - p)\). The quantity \(1-p\) is often abbreviated to \(q\).

Ancestral and derived alleles
For a given SNP, we distinguish between the ancestral and derived alleles. The ancestral allele is the allele that was present in the common ancestor. The derived allele is the ‘new’ allele arisen from a relatively recent mutation.

How do we know what the ancestral allele was, since by definition the ancestor is extinct? We can look at closely related species, such as the chimpanzee, and assume that the allele we see there is the ancestral allele.
This is mostly justified if we assume that any modern-day polymorphism in the human genome is the result of a mutation that arose recently – much more recently than 6 million years ago, when the human and chimp lineages diverged. This assumption is based on thinking about genetic drift: in finite populations, allele frequency fluctuates over time, but in the long run, it’s likely to fixate at either zero or one. (This is called fixation because once the allele frequency goes to zero or one, there is no longer variation in the population that will lead to random fluctuation from generation to generation.) Older polymorphisms are more likely to have fixated by now, so polymorphisms we observe in the modern population are likely to have arisen recently.

Even with that assumption, it’s still possible that looking at chimp would give us the wrong answer for the derived allele. You’ll think about this a little on the problem set. If we want to be extra careful, we can look at other species too (e.g. other great apes) to give us even higher confidence in the ancestral state.

Differences between subpopulations
One way of looking at variation in the human population is to ask how different are its subpopulations (e.g. Europeans and Asians) in terms of allele and genotype frequencies. Say we are able to measure the allele and genotype frequencies for two given subpopulations. The fixation index $F_{ST}$ is a way to measure the difference between the two subpopulations based on how much heterozygosity (the frequency of heterozygotes) we see. Namely, knowing the allele frequencies in the population taken as a whole, we would expect a certain amount of heterozygosity if the population is at Hardy-Weinberg equilibrium. If, however, we see less heterozygosity than expected given the allele frequencies – i.e. the average heterozygosity of the subpopulations is less than the overall expected heterozygosity – it suggests that the two subpopulations have evolved distinctly, and should perhaps be treated separately. The homework will take you through some simple $F_{ST}$ calculations.

Haplotypes
As one way of finding SNPs that have been affected by natural selection, we need to introduce the concept of haplotypes, correlations between nearby SNPs along a chromosome. Let’s consider the case of two SNPs (SNP1 and SNP2) that are separated by some distance (say, 300 nucleotides) on a chromosome. By “correlations” we mean the following: if I’ve observed that SNP1=A and SNP2=G in individual 1, and I’ve observed that SNP1=A in individual 2, how confident can I be that SNP2=G in individual 2? Another, more general way is to think about it as statistical dependence between random variables SNP1 and SNP2: if I know the marginals (allele frequencies) P(SNP1) and P(SNP2), how different from P(SNP1)P(SNP2) is the actual joint P(SNP1,SNP2)?

How do haplotypes arise? Let’s say a new mutation SNP1=G arises in an individual who happens to have SNP2=C. This configuration then starts to propagate through the population. If I slightly later see someone with SNP1=G, then it’s very likely that she also has SNP2=C by descent.

Eventually, however, this correlation between SNP1 and SNP2 will be diluted. The reason is genetic recombination: chromosomes are not necessarily passed down unchanged from parent to child – rather, they undergo “crossovers” in which homologous pieces of sister chromosomes (the two copies)
are swapped. Therefore, after many generations, we will eventually see SNP1=\text{G} recombined with the other allele of SNP2, and in the long run they will be less correlated.

Recent mutations will not only be more strongly correlated with nearby SNPs, but also correlated with more distant SNPs on the chromosome. The length of a haplotype is the distance over which there are detectable correlations between SNPs. The haplotype length can be used as a proxy for the “age” of an allele: alleles arising from recent mutations tend to have longer haplotypes and older alleles tend to have shorter haplotypes, all depending on whether enough time has gone by for genetic recombination to reshuffle things.

The Extended Haplotype Homozogosity (EHH) score is a way of measuring the haplotype around a given allele. The EHH for a given allele (say SNP1=\text{G}) is defined as follows: given any two random chromosomes with SNP1=\text{G} from the population, what is the probability that all the SNPs in the region under consideration have identical alleles? You’ll deal with a variant of EHH in the homework.

**Positive selection and haplotypes**

Let’s now think about what happens to haplotypes when there has been positive selection on a particular allele – that is, the allele frequency is not just fluctuating at random, but rather it is being directionally amplified by natural selection. In a **selective sweep**, an advantageous allele goes to high frequency in the population over many fewer generations than expected under random drift. Not only that, but the rapid spread of the allele also leaves less time for recombination to shuffle it with other nearby SNPs, leading to a longer-than-expected haplotype surrounding the selected allele.

Thinking about this more systematically: young neutral alleles will have long haplotypes, but they’ll also have low frequencies, since they haven’t had time to spread through the population. Old neutral alleles may have risen to high frequency by genetic drift, but chances are there has also been a lot of recombination, leading to short haplotypes. So it’s when we see both high allele frequency and long haplotypes that it looks like a selective sweep took place. Again, you’ll look for this on the homework.

Exactly how long ago can a selective sweep happen such that we can still detect it? The answer is very complicated. Mutation and recombination rates vary in different regions of the genome, meaning we have to be very careful about what we consider to be a higher-than-expected allele frequency or longer-than-expected haplotype. There are also demographic effects to worry about: for example, Europeans and Asians have experienced population bottlenecks, which have the effect of reducing the amount of variation and adding noise to our analysis. The strength of selection on the allele (i.e. the precise fitness advantage it confers) may also be taken into account. As a rule of thumb, current methods seem to be able to detect selective sweeps from the last 10,000 years.