Population Genomics I (Pardis Sabeti)
Scribed by Arjun K. Manrai

1 Introduction

It is readily observable that a single species, while phenotypically discernible as a distinct group different from other species, often displays a great amount of phenotypic diversity of its own. In this lecture, we study the genetic basis of this diversity by considering polymorphisms and their evolution, basic population genetics, the fixation index, and haplotype analysis.

2 Polymorphisms

Polymorphisms are DNA variations between different individuals of the same species. There are several types of polymorphisms, including single nucleotide polymorphisms (SNPs), variable number tandem repeats, insertions and deletions, large-scale polymorphisms, and copy number variants. In addition to helping explain the genetic basis of our diversity, understanding polymorphisms is also medically relevant. These changes, when they affect critical genes, can result in severe clinical outcomes like sickle cell anemia (SNP), Huntington’s disease (triplet repeat), and cystic fibrosis (deletion).

3 Population genetics

Consider a population of $N$ individuals. At a given gene locus, there may be many possible alleles (polymorphisms). Each individual in the population has two alleles for every autosomal gene locus, and the genotype of an individual refers to this collection of two alleles, taken over all loci of interest. The allele frequency of a particular allele is the proportion of all $2N$ alleles of the population that are of that type, and the genotype frequency of a particular genotype is the proportion of all $N$ individuals that have that genotype. A genotype is said to be homozygous if the genotype corresponds to two identical alleles, and heterozygous if the two alleles are different.

Hardy-Weinberg equilibrium refers to the state of a population when several (idealized) assumptions specifying mating patterns and evolutionary parameters are true. To make this more concrete, say we have two alleles $A$ and $T$ with frequencies $p$ and $q = 1 - p$, respectively, and therefore three genotypes: $AT$, $AA$, and $TT$. We further assume the following HW conditions hold true: (1) random mating, (2) no mutation, (3) no migration, (4) no selection, and (5) a large population. Then the allele and genotype frequencies will both remain constant, with the allele frequencies given by $p$ and $q$ for $A$ and $T$ respectively. The corresponding genotype frequencies for $AT$, $AA$, and $TT$ are $2pq$, $p^2$, and $q^2$, respectively.

The chi-square test can be used to determine how closely a particular population’s genotype frequencies resemble those expected from a HW model.
In real biological systems, HW assumptions are frequently violated. Small populations (violating assumption 5), for example, mean that genetic drift (essentially, stochasticity) will either fix alleles in the entire population, or eliminate them all together, given enough time.

4 Natural history of polymorphism

We can trace the evolution of current SNPs by finding the ancestral state of a given polymorphism. To do this, we compare the species in which we observe the SNP to species that are closely related, termed outgroups. We assume that the allele we observe in the outgroup represents the ancestral state. This assumption is rationalized by reasoning that observable SNPs likely arose recently, since old SNPs are likely to have reached fixation given enough time. If the split between a species and its outgroup occurred long before the introduction of the SNP into the species, then the outgroup likely contains the ancestral state.

5 Population differences

We can compare subpopulations within a species by measuring differences in allele frequencies. The metric used to do this is called the fixation index of a subpopulation compared to total, or $F_{ST}$, for short. $F_{ST}$ is given by

$$F_{ST} = \frac{H_{tot} - H_{sub}}{H_{tot}},$$

(1)

where $H_{tot}$ is the total heterozygosity of all subpopulations, and $H_{sub}$ is the heterozygosity of a particular subpopulation. For two alleles in proportions $p$ and $q$, heterozygosity is given by $2pq$. More generally, for a population with $n$ alleles in frequencies $p_i$ ($1 \leq i \leq n$), we can define heterozygosity as $1 - G$ where $G$ is the homozygosity of the population, given by

$$G = \sum_{i=1}^{N} p_i^2,$$

(2)

As we can see from this formula, $F_{ST}$ simply measures the proportion of the total heterozygosity of the entire population that is explained by the particular subpopulation of interest. For example, suppose we have two alleles $A$ and $T$, and two subpopulations 1 and 2. 1 is entirely $A$ and 2 is entirely $T$. In this extreme case, both subpopulations have $H_{sub} = 0$ and the $H_{tot}$ is $2pq = 0.5$. Hence $F_{ST} = 1$.

6 Haplotype analysis

Under natural selection, individuals with certain phenotypes (those that are selectively favorable) are more likely to produce fecund offspring, and consequently their underlying
genotypes are more likely to propagate through the population than both selectively neutral and selectively unfavorable phenotypes. A haplotype is a collection of alleles that lie together on the same chromosome, passed down as a unit. Because of recombination—the normal exchange of DNA between homologous chromosomes—we would expect selectively neutral portions of the genome to have long haplotypes, corresponding to young alleles, in low frequency. If, however, there was positive selection, the corresponding haplotype would be both long and high frequency. Sabeti et al developed a metric that uses long-range markers to measure haplotype lengths, the Extended Haplotype Homozygosity (EHH) [1]. After determining haplotype lengths with such a metric, we could identify positive selection by looking for young alleles that exhibit both long haplotypes and high frequency.

Through large-scale collaborative efforts like the International HapMap Project, haplotype data has been gathered that has allowed us to understand population differences, and search for selection sweeps within populations. A selection sweep occurs when an allele quickly ascends in frequency due to positive selection. Sweep software developed by Sabeti et al [2] and applied to the HapMap Project Data has identified population-specific positively selected regions of the genome. One noteworthy example is the identification of LARGE and DMD, two genes which have been connected to the Lassa virus, in West Africa. For a thorough review of this recent work, please consult [2]. Such results, when coupled with existing annotation data and functional analysis, could further both basic science as well as modern medicine.

7 References
