1. Introduction

The process of evolution consists of random mutations, together with forces of selection that determine which mutations are beneficial. We’d like to understand the relative importance of each of these forces. In particular, we’d like to know which genes are undergoing active selection. We accomplish this using neutrality tests, which test how the observed frequencies of alleles deviate from a neutral model where no selection occurs.

First, we need to understand a neutral model, focusing on inheritance alone. Historically, this area of study dates back to Darwin’s first models of evolution. Darwin’s model of blending inheritance stated that each organ was determined by a different gemmule. In this model, children inherit a blending of their parents’ corresponding gemmules, so if mom had a “yellow” gemmule and dad had a “blue” gemmule, then baby would inherit a “greenish” gemmule. As Darwin’s critics pointed out, this model predicts that everyone’s gemmules blend and blend until all gemmules are a drab shade of gray, losing all genetic diversity.

Mendel’s more accurate model, produced by his famous study of pea plants, states that each trait is determined by two corresponding alleles, which together determine a person’s phenotype. Offspring receive one allele from each parent, selected randomly from the parent’s two copies. The allele model is quite close to what actually happens when gametes pair during meiosis, and correctly predicts the observed phenomena of dominance and recessivity.

2. Hardy-Weinberg Law

Mendel’s model motivates the first neutral model, the Hardy-Weinberg Law (1908). This model assumes infinite population size, completely random mating, non-overlapping generations, and an absence of selection, mutation, or migration. Considering two versions A and a of an allele, let \( u_0, v_0, \) and \( w_0 \) be the respective frequencies of genotypes \( AA, Aa, \) and \( aa, \) respectively. Then the frequency of \( A \) is \( p = u_0 + v_0/2, \) and the frequency of \( a \) is \( q = w_0 + v_0/2. \) A Punnett square predicts that after a single generation, we observe frequencies \( u = p^2, v = 2pq, \) and \( w = q^2, \) and that these frequencies remain fixed over successive generations. Because \( q = 1 - p, \) our entire model is characterized by a single parameter \( p. \)

In practice, several of the Hardy-Weinberg Law’s assumptions are violated. For instance, in (slide 3, page 2), we see that for the recessive allele causing sickle-cell anemia, many more people have genotype \( Ss \) than is predicted by the Law. This
discrepancy is indeed due to selection, because people with one copy of \( s \) possess immunity to the dreaded tropical disease malaria.

We next study two different approaches to neutral theory: the \textit{prospective approach} of classic population genetics, and the \textit{retrospective approach} focusing on the \textit{coalescent}.

3. Prospective method

3.1. \textbf{Neutral Theory history}. In the 1960s, people thought all mutations were different in fitness, so selection would rule out bad mutations and fix good mutations, with variation kept by balancing selection. So when Neutral Theory was first proposed by Motoo Kimura at that time, it was shocking and controversial. Having mathematical background, Kimura incorporated diffusion approximations to the field of population genetics, focusing on finite population. The main idea of Neutral Theory is that most of the mutations are neutral or nearly neutral, and the change of mutations frequencies is a result of random genetic drift in finite population (Slide 6, Page 2). All the mutations will eventually go extinct or be fixed by this process, and variations seen in population are mutations which are on the way to fixation (Slide 1, Page 3).

3.2. \textbf{Ewens Sampling Formula}. In 1972, Warren Ewens proposed the famous Ewens Sampling Formula, which is based on diffusion theory and introduced the \textit{infinite alleles} model. The infinite alleles model claims that there are an infinite number of states into which an allele can mutate, so each mutation generates a unique allele. So if two sequences have the same allele, they are necessarily descendants of a common ancestor, a rule known as \textit{identity by descent}. This model depends on parameters \( N \), the population size, and \( \mu \), the mutation rate. We assume a diploid population (two alleles/ person), so there exist \( 2N \) alleles in total. For each generation, we introduce \( 2N \mu \) new alleles, each with an initial frequency of \( 1/2N \). The Ewens Sampling Formula calculates the probability that a sample of \( n \ll N \) gene copies contains \( k \) alleles and there are \( a_1, a_2, \ldots, a_k \) alleles represented \( 1, 2, \ldots, n \) times in the sample:

\[
P(a_1, a_2, \ldots, a_n) = \frac{n! \Theta^k}{\Theta_{(n)}} \prod_{j=1}^{n} \frac{1}{j^{a_j} a_j!},
\]

where

\[
\Theta_{(n)} = \Theta(\Theta + 1) \ldots (\Theta + n - 1).
\]

The main purpose of the Ewens Sampling Formula is to get an expected site frequencies spectrum (Slide 6, Page 2).

4. Retrospective approach

4.1. \textbf{Coalescent Theory}. Coalescent Theory was first introduced by Kingman in the early 1980s. It traces all alleles of a gene in a sample from a population to a single ancestor shared by all members of the population, known as the most common recent ancestor (MCRA) by inferring a tree. All polymorphisms in the sample can then be demonstrated by putting mutation events on different branches of the coalescent tree (Slide 1, Page 4). The expected MCRA of a population could be found at \( T=4N \) generations before, but in practice, the MCRA of a 20-individual sample can be very close to that of the population. The coalescent is a
probability-based tool, very useful in population genetics topics such as parameter estimation, genealogy tree reconstruction, et al.

4.2. tests of selection within species. We construct three polymorphism summary statistics:

1. \( S \), the number of segregating sites in the sample;
2. \( \pi \), the average number of pairwise differences;
3. \( \nu_i \), the number of sites that divide the sample into \( i \) and \( n - i \) sequences.

These statistics can be used to estimate the important statistic \( \Theta \) (Slide 6, Page 4). Different ways in changing the allele frequency spectrum in population could lead to difference in each estimation. Based on this inference, three selection tests are proposed by Tajima and Fu and Li respectively. (Slide 1, Page 5). In these tests, neutral expectation is used as null hypothesis (Slide 3, Page 5), an alternative hypothesis (positive selection, balancing selection, population structure/subdivision or population expansion) is tested (Slide 4,5,6, Page 5 and Slide 1, Page 6)

The HKA Test measures the polymorphism and divergence in two loci, and tests whether there is excess in one of two classes (Slide 2, 3, Page 6). The MK test measures the synonymous and nonsynonymous polymorphism and divergence in one locus, and tests whether there is an excess in one class (Slide 4, 5, Page 6).

4.3. rate-based selection metric. To determine if a certain gene is under selection, we may use the rate-based selection metric. Recall that an amino acid is determined by a codon of three adjacent nucleotides, and that different codons, such as CGC and CGA, can represent the same amino acid, arginine in this case. A synonymous mutation leaves the amino acid sequence intact, while a non-synonymous mutation changes it.

We use various methods, notably PAML, to compute \( d_N \), the rate at which synonymous mutations occur, and \( d_S \), the rate at which synonymous mutations occur. We define the rate-based selection metric to be \( d_N / d_S \). There are three cases of interest:

1. \( d_N / d_S < 1 \), purifying selection;
2. \( d_N / d_S = 1 \), neutral expectation;
3. \( d_N / d_S > 1 \), positive selection.

In general, genes performing essential roles, such as those encoding hemoglobin, undergo purifying selection, because changes tend to disrupt the gene’s function and cause harm. Genes that perform peripheral functions can undergo positive selection if some change is useful, but not necessary for survival. For instance, changes to a mammal’s hair coloring could provide useful camouflage.

However, some synonymous substitutions are more likely than others because of preferred codons, codons more likely to be used in a gene than equivalent codons.

However, some synonymous substitutions are more likely than others because of preferred codons, codons more likely to be used in a gene than equivalent codons. We don’t view unpreferred substitutions as neutral, so we form the additional ratio \( Kp/Ku \). Here,

\[
Kp = \frac{\# \text{ preferred substitutions}}{\# \text{ preferred sites}},
\]

and

\[
Ku = \frac{\# \text{ unpreferred substitutions}}{\# \text{ unpreferred sites}}.
\]

The three relevant cases are:

1. \( Kp/Ku < 1 \) Excess Unpreferred Substitution
(2) \( K_p/K_u = 1 \) Neutral Expectation
(3) \( K_p/K_u > 1 \) Excess Preferred Substitution