Evolution- the unifying theory

NOTE: In a later module we will address whether religion and evolution can co-exist. If you are uncomfortable with the Theory of Evolution due to cultural, religious, or other reasons, I invite you to remove yourself from possible biases and view this module strictly through the lens of the scientific method.

"Nothing in biology makes sense except in the light of evolution." –Theodosius Dobzhansky

Evolution has been called the unifying theory of biology because it provides the mechanism responsible for all of the variation we can observe in earth's diverse life forms. Biologists study variation; it could be argued that life without variation is impossible. From molecular biology to ecology and everything in between (microbiology, cell biology, physiology, developmental biology, behavioral ecology, biomedical science, parasitology, paleontology, etc.), patterns in the natural world can be explained using the framework of evolution. However, evolution itself requires some components in order to explain patterns in life's variation!

"Nothing in evolution makes sense except in the light of genetics." –Michael Lynch

You may remember from the history of genetics unit- Darwin struggled to come up with a mechanism for his theory of evolution by natural selection (even though the "Father of Genetics" was alive at the same time and even sent him a copy of his work!). In the early 20^{th} century evolutionary biologists and geneticists finally recognized that genetics was the missing piece of the evolutionary puzzle. The integration of genetics and evolution from the 1920s-1950s is known as the "Modern Synthesis".

In this unit we will focus on defining evolution, and use principles of genetics to examine how populations and species evolve.

What is evolution?

In this unit, we will work with two definitions of evolution:

1. Evolution is the change in the genetic characteristics of a population through time
2. Evolution is descent of all living beings from a common ancestor

These definitions are not mutually exclusive- the same processes create the patterns of variation we see in populations and species. Sometimes population-level evolution is referred to as "micro-evolution" and species-level (and deeper) evolution is referred to as "macro-evolution". What is the difference? Time! There is no other difference between macro- and micro-evolution; they occur via the same mechanism. In this unit we will examine two ways to study evolution that occurs in/between populations and evolution that occurs in/between species: (1) population genetics and (2) phylogenetics. In some ways you can consider population genetics "micro-evolution" and phylogenetics "macro-evolution", but principles from both of these fields can be used for recent and ancient evolution.

1 Population Genetics

Population genetics originated in the 1920s when a few evolutionary biologists with skills in math/stats realized that genetics (as described by Mendel) provided the mechanism of heredity for evolutionary biology (as described by Darwin). These scientists used math to explore how forces of evolution influenced the genetics of biological groups (e.g., populations). **"Forces of evolution" refers to the various means by which a population's genetic characteristics can change over time**. The components of a population's genetic characteristics that we will examine are (1) **allele frequency**, (2) **genotype frequency**, (3) **types of mutations in protein-coding genes**, and (4) population heterozygosity.

NOTE: Thinking about a population's genetic structure is a fundamental shift from what you've focused on in previous units. Up to this point, you have only considered the genetic composition of a single individual (e.g., an individual's genotype). In this unit, we will consider the collective genetic characteristics across individuals. We will use rules of probability to understand whether the genetic characteristics of a population matches what we'd expect if specific forces of evolution are not present. **To keep things simple, in this unit we will only consider diploid organisms**.

1.1 Allele frequency

Allele frequencies, referred to as p and q for genes with two alleles, are calculated using the total number of a specific allele in a population divided by the total number of all alleles.

Note: Don't confuse this p with a *pvalue* that you calculate from a statistical test (e.g., a Chi-squared test). They are completely unrelated. The p-value for the statistical test is the probability of getting your observed data given the assumption that the hypothesis you are testing is true. The p in reference to allele frequencies is the frequency of a certain allele in a population.

To calculate the allele frequency for an allele of Gene 1 in a population that has two alleles ('A' and 'a'), we take the total number of copies of a specific allele (e.g., n_A for allele 'A') and divide it by the total number of allele copies (n_A for allele 'A' and n_a for allele 'a') in the population:

$$p = \frac{n_A}{n_A + n_a} \tag{1}$$

Practice problem: A population has two alleles for Gene 1: 'A' and 'a'. This population has 15 individuals with genotype 'AA', 50 individuals with genotype 'Aa', and 35 individuals with genotype 'aa'. What is the allele frequency of 'A'?

Solution: Each individual with genotype 'AA' has two copies of 'A'. So in the 15 'AA' individuals there are 30 copies of 'A'. Each individual of genotype 'Aa' has one copy of 'A'. So in the 50 'Aa' individuals there are 50 copies of 'A'. the 'aa' individuals have 0 copies of 'A'. Therefore, there are 30+50 = 80 copies of 'A'. The total number of alleles in this population is the total number of individuals (100) x 2 = 200. So, using equation 1, the allele frequency for 'A' is $p = \frac{80}{200} = 0.4$

A handy tip for calculating the allele frequency of the second allele (for a gene that only has two alleles): you simply take 1 - p to find the second allele frequency (q). In other words, the sum of the allele frequencies should be equal to 1:

$$p + q = 1 \tag{2}$$

Practice problem: From the same practice problem above, what is the allele frequency of 'a'?

Solution: Each individual with genotype 'aa' has two copies of 'a'. So in the 35 'aa' individuals there are 70 copies of 'a'. Each individual of genotype 'Aa' has one copy of 'a'. So in the 50 'Aa' individuals there are 50 copies of 'a'. the 'AA' individuals have 0 copies of 'a'. Therefore, there are 70+50 = 120 copies of 'a'. The total number of alleles in this population is the total number of individuals (100) x 2 = 200. So, using equation 1, the allele frequency for 'a' is $q = \frac{120}{200} = 0.6$. Alternatively, we could have used equation 2 above and simply done $q = 1–p = 0.6$.

It is important to note that ideally you would have the information for every individual in a population, however this is unfeasible most of the time. To accurately estimate the allele frequency (and genotype frequencies as shown in the next section), scientists sample a number of individuals from the population to represent the population as a whole. It is important to collect a sufficient sample size that can provide an accurate view of the population as a whole. The equations shown here still work, but rather than using the population as a whole you can simply use the number of individuals collected in the sampling.

1.2 Genotype frequency

Genotype frequencies are calculated using the number of individuals with a specific genotype divided by the total number of individuals in a population. To calculate the genotype frequency (f) for a genotype (e.g., 'AA') of Gene 1 in a population that has two alleles ('A' and 'a'), we take the total number of individuals with that genotype (e.g., n_{AA} for individuals with genotype 'AA') and divide it by the total number of individuals (n_{AA} for individuals with genotype 'AA', n_{Aa} for individuals with genotype 'Aa', n_{aa} for individuals with genotype 'aa'):

$$f_{AA} = \frac{n_{AA}}{n_{AA} + n_{Aa} + n_{aa}} \tag{3}$$

Practice problem: A population has two alleles for Gene 1: 'A' and 'a'. This population has 15 individuals with genotype 'AA', 50 individuals with genotype 'Aa', and 35 individuals with genotype 'aa'. What is the genotype frequency of 'AA'?

Solution: There are 100 individuals in our population, and 15 individuals with genotype 'AA'. So, using equation 3 above, $f_{AA} = \frac{15}{15+50+35} = 0.15$

1.3 Forces of Evolution

The forces of evolution include:

1. **Mutation:** A mutation is a new version of a locus. This alters the genetic characteristics of a population by introducing a new allele and thus changing the allele frequencies of the population.
2. **Gene flow:** Gene flow is when the population of interest has individuals that reproduce with an outside population, causing genes to "flow" into the population of interest.
3. **Genetic drift:** Allele frequencies in a population change between generations due to random chance of some individuals mating more/less than others.
4. **Nonrandom mating:** If individuals of a group don't have equal probability of mating with any other individual of the opposite sex, then this can result in changes in genotype frequencies.
5. **Selection:** When one allele is favored by the environment over another, then that allele will increase in frequency over time.

1.3.1 Mutation

Imagine a population of diploid organisms where all individuals have the same allele for Gene 2 (allele 'B') for a gene (i.e., all individuals are homozygous 'BB'). The allele frequency of 'B' in this population is 1 (100% 'B'). If one individual has a mutation in one copy of Gene 2 (a new allele 'b'), this alters the allele frequencies in the population. For example, if there are 100 individuals in the population (including the individual with the new 'b' allele), then the allele frequency of 'B' is 0.995 and the allele frequency of 'b' is 0.005. From this example you should notice that mutation by itself is a weak force of evolution– it only changes the allele frequency by a very small measure.

Note: You may be thinking "Why would it be 0.995 for 'B' and 0.005 for 'b'? Shouldn't the allele frequencies be 0.99 for 'B' and 0.01 for 'b'?" Remember, this is a diploid population- and if the mutation only occurs in one copy of Gene 2 in an individual, then the other copy will be unaffected. In other words, the mutated individual will have genotype 'Bb'. This means that the population has 99 individuals with 'BB' and a single individual with 'Bb'. For the 99 individuals with genotype 'BB', there are 99x2 = 198 copies of the 'B' allele. For the single 'Bb' individual, there is one copy of the 'B' allele and one copy of the 'b' allele. To calculate the allele frequencies for each allele, you simply divide the number of one allele by the total number of alleles: $p = \frac{199}{200} = 0.995$; $p = \frac{1}{200} = 0.005$

1.3.2 Gene flow

Imagine a population (P1) of diploid organisms with two alleles for Gene 1 (allele 'A' and allele 'a') at frequencies p = 0.8 (for 'A') and q = 0.2 (for 'a'). A nearby population (P2) has frequencies p = 0.2 (for 'A') and q = 0.8 (for 'a'). P1 and P2 are separated by a large river. However, after years of drought, the river disappears and individuals from P2 begin to mate with individuals in P1. If the individuals mate freely with one another, and the populations are of approximately the same size, after sufficient time this would result in a shift in allele frequencies in P1 to 0.5 for both p and q.

Note: You may notice that gene flow, by itself, can only change allele frequencies for alleles that already exist. for genes with no variation, it is impossible for gene flow to change the allele frequencies unless there is first a mutation in one of the populations that creates a new allele.

1.3.3 Genetic drift

Imagine a population of ten asexually-reproducing diploid organisms with two alleles for Gene 1 (allele 'A' and allele 'a'). 7 individuals have the genotype 'AA' and 3 individuals have the genotype 'Aa'. This means that the allele frequency in the population is 0.85 for 'A' and 0.15 for 'a'. If random chance leads to 4/7 'AA' individuals producing a single offspring each and 3/3 'Aa' individuals producing two offspring each, this would result in the next generation having 4 'AA' individuals and 6 'Aa' individuals. The allele frequency in this new generation would now be 0.7 for 'A' and 0.3 for 'a'. This change in allele frequency (from one generation to the next) was due genetic drift- a random shift in allele frequency due to chance of who mates more/less.

Note: Genetic drift has a larger impact on small populations vs large populations. This is because a drastic change in allele frequency, as shown in the previous example (where allele frequencies changed by ± 0.15 in just one generation) require a significant proportion of the population to be affected. If the population is large enough, it is highly unlikely that random chance will cause drift to drastically affect allele frequencies.

1.3.4 Non-random mating

Imagine a population (P1) of diploid organisms with two alleles for Gene 1 (allele 'A' and allele 'a'). In a scenario where individuals mate non-randomly and genotypes only mate with other individuals of the same genotype (i.e., 'AA' with 'AA', 'aa' with 'aa', and 'Aa' with 'Aa'), this would result in the change of population genotype distribution over time until only genotypes 'AA' and 'aa' existed in the population.

1.3.5 Selection

Imagine a population of ten asexually-reproducing diploid organisms with two alleles for Gene 1 (allele 'A' and allele 'a'). 7 individuals have the genotype 'AA' and 3 individuals have the genotype 'Aa'. This means that the allele frequency in the population is 0.85 for 'A' and 0.15

for 'a'. A change in the environment causes individuals with the 'A' allele to be more likely to reproduce. Because of this, the environment 'selects' an increased number of individuals with 'A' to have offspring. This selection leads to 6/7 'AA' individuals producing three offspring each and 2/3 'Aa' individuals producing a single offspring, so the next generation has 18 'AA' individuals and 2 'Aa' individuals. The allele frequency in this new generation would now be 0.95 for 'A' and 0.05 for 'a'. This change in allele frequency (from one generation to the next) was due natural selection; which changes allele frequencies based on which allele is associated with higher reproduction rates.

Note: Selection is different than genetic drift because it is not random. In scenarios with strong selection, genotype (rather than random chance) is the primary determiner of whether an individual will produce offspring for the next generation. Reproduction capability is called fitness in evolutionary biology (more "fit" individuals produce more offspring). When alleles have **differential fitness**, this means that one allele has greater fitness than the other (i.e., it is more likely to increase in frequency).

1.4 Hardy-Weinberg Equilibrium

A population in **Hardy-Weinberg Equilibrium** (HWE) experiences no change in allele or genotype frequency. In other words, a population in HWE does not evolve (spoiler alert- this is pretty much never the case). The assumptions of a population in HWE are:

1. No mutations
2. No gene flow
3. Infinite population size (i.e., no genetic drift)
4. No non-random mating
5. No differential fitness (i.e., no selection)

1.4.1 Predicting genotype frequencies

A population that is in HWE has predictable genotype frequencies. These frequencies can be calculated by using the population allele frequencies and thinking about the population from the perspective of its **gene pool**. A gene pool is just what it sounds like- the complete mixture of all gene copies within a population. By counting all of these copies (or at least a large sample of these copies), you can extract the allele frequencies and genotype frequencies for the population. A population in HWE has no change in its gene pool across generations, and the characteristic random mating among individuals means you can use the rules of probability to estimate genotype frequencies.

Remember, a gene in a population that has two alleles has a frequency for each allele: p and q. Let's say that p represents the frequency for the dominant allele and q represents the frequency for the recessive allele. An allele "frequency" is the same as the probability you would draw that allele from the gene pool if you randomly selected a single allele. So, considering the gene pool as a mix of all the gene copies in the population, the probability of getting a homozygous dominant individual is:

$$p * p = p^2 \tag{4}$$

You can think of this in terms of probability. An individual with the homozygous dominant genotype has two of the dominant alleles. If we consider the gene pool as a pool of alleles, to "pull an individual" out of the pool we have to draw two alleles (because each individual is made up of two alleles). Therefore, the probability of getting a homozygous dominant individual is the probability of the **first draw** being the dominant allele *AND* the **second draw** being the dominant allele. The word "*AND*" should help you understand why we used multiplication for equation 4– it's the product (a.k.a. "AND") rule! **Importantly, this is not the actual genotype frequency of the homozygous dominant genotype. It is a predicted estimate of the homozygous dominant genotype *if* the population is in HWE.**

We can use the same logic for the homozygous recessive allele frequency (q). Again, considering the gene pool as a mix of all the gene copies in the population, the probability of getting a homozygous recessive individual is:

$$q * q = q^2 \tag{5}$$

If we consider the gene pool as a pool of alleles, to "pull an individual" out of the pool we have to draw two alleles (because each individual is made up of two alleles). Therefore, the probability of getting a homozygous recessive individual is the probability of the **first draw** being the recessive allele *AND* the **second draw** being the recessive allele. **Once again, this is not the actual genotype frequency of the homozygous recessive genotype. It is a predicted estimate of the homozygous recessive genotype *if* the population is in HWE.**

To estimate the heterozygous genotype frequency if a population is in HWE, we once again use the rules of probability. Considering the gene pool as a mix of all the gene copies in the population, the probability of getting a heterozygous individual is:

$$(p * q) + (q * p) = 2pq \tag{6}$$

If we consider the gene pool as a pool of alleles, to "pull an individual" out of the pool we have to draw two alleles (because each individual is made up of two alleles). Therefore, to get a heterozygous genotype we would have the **first draw** being the dominant allele *AND* the **second draw** being the recessive allele. However, this is not the only way to get a heterozygote! Let's call that way (dominant allele followed by recessive allele) "scenario 1". We could also have the scenario of the **first draw** being the recessive allele *AND* the **second draw** being

the dominant allele. Let's call this second way (recessive allele followed by dominant allele) "scenario 2". So we could have scenario 1 *OR* scenario 2. Using the sum (a.k.a. "OR") rule, we add the probabilities of these two scenarios together. **Once again, this is not the actual genotype frequency of the heterozygous genotype. It is a predicted estimate of the heterozygous genotype *if* the population is in HWE.**

It is useful to recognize that because these are the only three possible genotypes ([1] homozygous dominant, [2] homozygous recessive, [3] heterozygous), their frequencies will all add up to 1:

$$p^2 + q^2 + 2pq = 1 \tag{7}$$

1.4.2 Estimating genotype frequencies from allele frequencies

If you assume a population is in HWE, you can estimate genotype frequencies using the allele frequencies. This can be useful if you are only looking at phenotypic data for a Mendelian trait. **However, if your population is not in HWE, this approach will not be accurate**.

NOTE: These types of HWE questions are common on the MCAT

To perform this type of inference, you simply need to jump back and forth between equations 2 and 8. For example, if you are provided with phenotypic data for a Mendelian trait, use the following order of operations:

1. Divide n_{hr} by the total sample size to calculate the frequency of the homozygous recessive genotype (your estimate of q^2)
2. Take the square root of q^2 to estimate q
3. Subtract q from 1 to estimate p
4. Square p to estimate the frequency of the homozygous dominant genotype
5. Calculate $2 * p * q$ to estimate the frequency of the heterozygous genotype (known as the "carrier" genotype in recessive diseases)
6. Confirm that $p^2 + 2pq + q^2 = 1$

Practice problem: A recessive genetic disease is present in 1% of a diploid population of 300 million individuals. Assuming the population is in HWE, how many individuals in this population are carriers for the disease?

Solution: Using the order of operations described above:

1. $q^2 = 0.01$
2. $q = \sqrt{(0.01)} = 0.1$
3. $p = 1 - q = 0.9$
4. $p^2 = 0.81$
5. $2pq = 0.18$
6. $p^2 + 2pq + q^2 = 0.81 + 0.18 + 0.01 = 1$

To answer the question, we estimate the number of carriers by multiplying 0.18 x 300 million = 54 million individuals

1.4.3 Testing if a population is in HWE

If you know the genotypic information for each individual, you can test if a population is in HWE using a chi-squared test. To do this, collect a few items:

1. p and q = calculated allele frequencies
2. p^2, q^2, $2pq$ = estimated genotype frequencies
3. n_{hd}, n_{hr}, and n_{het} = actual number of individuals in population [or sample] by genotype
   → hd = "homozygous dominant", hr = "homozygous recessive", het = "heterozygous"
4. n (population [or sample] size)

Plug the numbers of the expected and observed genotypes into the chi-squared formula:

$$\chi^2 = \frac{(n_{hd} - e_{hd})^2}{e_{hd}} + \frac{(n_{het} - e_{het})^2}{e_{het}} + \frac{(n_{hr} - e_{hr})^2}{e_{hr}} \tag{8}$$

NOTE: In this chi-squared equation, your observed value the number of individuals you observed for each genotype. The expected values are the frequency estimates from HWE (p^2, $2pq$, and q^2) multiplied by the total sample size.

You will then use a chi-squared table to determine the p-value of your chi-squared test. For HWE chi-squared problems, you will use one degree of freedom. So go to the second row in the chi-squared table and move to the right until you reach your chi-squared value (or the end of

the table). Then follow whichever column you end at to the top to find your p-value.

Table 1: Table

	p-value										
DF	0.995	0.975	0.2	0.1	0.05	0.025	0.02	0.01	0.005	0.002	0.001
1	0.0004	0.00016	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.55	10.828
2	0.01	0.0506	3.219	4.605	5.991	7.378	7.842	9.21	10.597	12.429	13.816
3	0.0717	0.216	4.642	6.251	7.815	9.348	9.837	11.345	12.838	14.796	16.266

Abbreviated chi-squared table for three degrees of freedom

Remember, the p-value is the probability of getting your observations if the hypothesis you are testing is true. We are testing the hypothesis that our population of interest is in HWE. Therefore, a low p-value means that there is low probability of getting our observations if the population is in HWE. With a low enough p-value (in this class we usually use the cutoff of 0.05), we can reject the hypothesis of HWE. In other words, if the p-value is lower than 0.05, we are confident that the population is undergoing some sort of evolution (one of the assumptions of HWE is being violated). However, we can't determine the specific force of evolution using a chi-squared test on HWE– to understand what is causing our population to be out of HWE we need to do other types of statistical tests (some of which we'll talk about later).

Practice problem: A co-dominant trait for plumage coloration in a species of bird has two alleles: 'Y' and 'B'. Individuals with genotype 'YY' have yellow plumage. Individuals with genotype 'BB' have blue plumage. Individuals with genotype 'YB' have green plumage. You sample a population and collect 40 blue individuals, 30 yellow individuals, and 10 green individuals. Is this population in HWE?

Solution: Let's say p is the frequency of 'B' and q is the frequency of 'Y'. Collecting the items listed above:

1. $p = 0.5625$; $q = 0.4375$
2. $p^2 = 0.3164$; $q^2 = 0.1914$; $2pq = 0.4921$
3. $n_{`BB'} = 40$ ; $n_{`YY'} = 30$; and $n_{`YB'} = 10$
4. n = 80

Then we add our numbers to the chi-squared equation

$$\chi^2 = \frac{(40-25.312)^2}{25.312} + \frac{(10-39.368)^2}{39.368} + \frac{(30-15.312)^2}{15.312} = 44.52$$

We can use this chi-squared value on Table 1. Because it is much bigger than the value to the far right, we know that our p-value is < 0.001 (which is much smaller than our 0.05 cutoff). Therefore, we confidently reject the hypothesis that this population is in HWE.

Multiple Sequence Alignments

Taking a brief break, let's learn about **multiple sequence alignments** (a tool used in molecular genetics). A multiple sequence alignment is a format used to examine the molecular sequence data (e.g., DNA, RNA) of multiple individuals in a format where the sequences are 'aligned' to allow for comparisons to be made across sequences. You can think of a multiple sequence alignment as a table, where each row represents a different individual (e.g., a different person) and each column represents a different position in the alignment (i.e., **site**). The total length of the sequence (i.e., number of sites) is equal to the number of columns (minus the first column if it contains the individual names). If a site has variation across individuals, this site is known as a **single nucleotide polymorphism** (a.k.a. SNP [pronounced "snip"]). Below is an example of a multiple sequence alignment of human DNA at a locus with 54 sites for five individuals:

```
                                    ↓
          5    10   15   20   25   30   35   40   45   50
Ind01     ATGAACGTACGATATCGATCGTATCGTAAGCTACGTATACGCGTATATCGCTAG
Ind02     ATGAACGTACGATATCGATCGTATGGTAAGCTACGTATACGCGTATATCGCTAG
Ind03     ATGAACGTACGATATCGATCGTATCGTAAGCTACGTATACGCGTATATCGCTAG
Ind04     ATGAACGTACGATATCGATCGTATSGTAAGCTACGTATACGCGTATATCGCTAG
Ind05     ATGAACGTACGATATCGATCGTATCGTAAGCTACGTATACGCGTATATCGCTAG
```

In this alignment you can see that there is a single SNP at site 25 (marked with an arrow). The wild-type nucleotide at this position is 'C'. However, two individuals have nucleotides that are not 'C'. Ind02 has a 'G' and Ind04 has an 'S'. Hopefully you are thinking "What in the world is 'S'? Isn't DNA made up of only 'A', 'C', 'G', and 'T'?" You are right! However, something perhaps you haven't thought about is this: If humans are diploid, why is there only one row per individual? This is because each row is showing the **consensus** of an individual's DNA. In other words, each row is a combination of the individual's two copies of the locus (the maternally-inherited version and the paternally-inherited version). For almost every individual in this alignment, both the maternally-inherited and the paternally-inherited versions of the locus are the exact same. For example, here are what the maternal and paternal versions look like for Ind01 and Ind02 if we broke them up:

```
                                    ↓
          5    10   15   20   25   30   35   40   45   50
Ind01-mat ATGAACGTACGATATCGATCGTATCGTAAGCTACGTATACGCGTATATCGCTAG
Ind01-pat ATGAACGTACGATATCGATCGTATCGTAAGCTACGTATACGCGTATATCGCTAG
Ind02-mat ATGAACGTACGATATCGATCGTATGGTAAGCTACGTATACGCGTATATCGCTAG
Ind02-pat ATGAACGTACGATATCGATCGTATGGTAAGCTACGTATACGCGTATATCGCTAG
```

Because these individuals have the same nucleotide at site 25, we can call them homozygous at this site. Ind01 is homozygous for 'C' and Ind02 is homozygous for 'G'. Based on this information, can you guess what Ind04 is from the alignment above? Perhaps this alignment with their DNA broken up into maternal and paternal versions will help:

```
                                    ↓
          5    10   15   20   25   30   35   40   45   50
Ind04-mat ATGAACGTACGATATCGATCGTATCGTAAGCTACGTATACGCGTATATCGCTAG
Ind04-pat ATGAACGTACGATATCGATCGTATGGTAAGCTACGTATACGCGTATATCGCTAG
```

Ind04 is heterozygous at site 25! Because each line in the multiple sequence alignment above is a consensus of the maternal and paternal sequences, a new character is used to help

us see the heterozygosity while maintaining the integrity of the alignment. There are many characters used for the diverse scenarios of heterozygosity/ambiguity, check them out here: `https://www.bioinformatics.org/sms/iupac.html` (no need to write these down).

1.4.4 Using multiple sequence alignments to test HWE

We can use molecular data (e.g., from a multiple sequence alignment) to test if a population is in HWE. To do this, we will use techniques from the multiple sequence alignment section above.

Practice problem: Using the below sequence alignment, determine whether this population is in HWE for this locus. Use site 25 for your calculations.

```
                                    ↓
          5        10        15        20        25        30        35        40
Ind01  ATGAACGTACGATATCGATCGTATGGTAAGCTACGTATACGCGT
Ind02  ATGAACGTACGATATCGATCGTATGGTAAGCTACGTATACGCGT
Ind03  ATGAACGTACGATATCGATCGTATCGTAAGCTACGTATACGCGT
Ind04  ATGAACGTACGATATCGATCGTATSGTAAGCTACGTATACGCGT
Ind05  ATGAACGTACGATATCGATCGTATSGTAAGCTACGTATACGCGT
Ind06  ATGAACGTACGATATCGATCGTATSGTAAGCTACGTATACGCGT
Ind07  ATGAACGTACGATATCGATCGTATSGTAAGCTACGTATACGCGT
Ind08  ATGAACGTACGATATCGATCGTATGGTAAGCTACGTATACGCGT
Ind09  ATGAACGTACGATATCGATCGTATCGTAAGCTACGTATACGCGT
Ind10  ATGAACGTACGATATCGATCGTATGGTAAGCTACGTATACGCGT
Ind11  ATGAACGTACGATATCGATCGTATGGTAAGCTACGTATACGCGT
Ind12  ATGAACGTACGATATCGATCGTATSGTAAGCTACGTATACGCGT
Ind13  ATGAACGTACGATATCGATCGTATCGTAAGCTACGTATACGCGT
Ind14  ATGAACGTACGATATCGATCGTATSGTAAGCTACGTATACGCGT
Ind15  ATGAACGTACGATATCGATCGTATCGTAAGCTACGTATACGCGT
Ind16  ATGAACGTACGATATCGATCGTATSGTAAGCTACGTATACGCGT
Ind17  ATGAACGTACGATATCGATCGTATSGTAAGCTACGTATACGCGT
Ind18  ATGAACGTACGATATCGATCGTATCGTAAGCTACGTATACGCGT
Ind19  ATGAACGTACGATATCGATCGTATSGTAAGCTACGTATACGCGT
Ind20  ATGAACGTACGATATCGATCGTATGGTAAGCTACGTATACGCGT
```

Solution: Let's say p is the frequency of 'G' and q is the frequency of 'C'. Collecting the items listed above:

1. $p = 0.525$; $q = 0.475$
2. $p^2 = 0.2756$; $q^2 = 0.2256$; $2pq = 0.4988$
3. $n_{`GG'} = 6$ ; $n_{`CC'} = 5$; and $n_{`GC'} = 9$
4. n = 20

Then we add our numbers to the chi-squared equation

$$\chi^2 = \frac{(6-5.512)^2}{5.512} + \frac{(5-4.512)^2}{4.512} + \frac{(9-9.976)^2}{9.976} = 0.19$$

We can use this chi-squared value on Table 1. Because it is much bigger than the value to the far left, we know that our p-value is > 0.995 (which is much greater than our 0.05 cutoff). Therefore, we fail to reject that this population is in HWE. The verbiage here is important. We don't *prove* that the population isn't in HWE- we simply fail to reject that it *isn't* in HWE.

1.5 Detecting selection

Up to this point, we have only tested whether a population is evolving (by testing if a population is in HWE). However, we haven't looked at what forces of evolution might be acting on a population to cause it to evolve. As mentioned previously, one of those forces is selection. Here we will look at one approach to detecting and classifying selection.

One way to test for selection is to examine the rates of nonsynonymous and synonymous mutation accumulation in a protein-coding gene. If change in a protein structure is favored, as is the case in the MHC proteins that evolve rapidly to combat invasive pathogens, then nonsynonymous mutations will accumulate at a rate higher than synonymous mutations. This is known as **positive selection**. If change in a protein structure is disfavored, as is the case in the Hox genes that control body plan development, then synonymous mutations will accumulate at a rate higher than synonymous mutations. This is known as **negative selection**. These evolutionary rates can be quantified and compared by calculating values known as d_N (rate of nonsynonymous mutation) and d_S (rate of synonymous mutation). By taking the ratio of these values ($\frac{d_N}{d_S}$), which is known as ω, we can understand if the gene is experiencing positive selection ($\omega > 1$) or negative selection ($\omega < 1$).

To calculate ω, you need to (A) know which sequence is going to be the reference (this will be given to you), and (B) work in order to collect the following values:

1. N_P = Number of sites in the gene where a change would cause a **nonsynonymous** mutation ("P" for possible)
2. S_P = Number of sites in the gene where a change would cause a **synonymous** mutation ("P" for possible)
3. N = Number of **nonsynonymous** mutations
4. S = Number of **synonymous** mutations
5. $d_N = \frac{N}{N_P}$
6. $d_S = \frac{S}{S_P}$
7. $\omega = \frac{d_N}{d_S}$

For this class, do not include the start codon or the stop codon in your calculations (these codons have unique evolutionary pressures compared to the other codons). This calculation can be tricky; this video may help: `https://www.youtube.com/watch?v=ZlcHujBbQ2Q&t=702s`.

1.6 Detecting non-random mating due to population structure

Nonrandom mating can occur when a population actually has some division (i.e., there are distinct groups and individuals tend to mate with individuals in their same group). One way to test for population subdivision (a.k.a., population structure) is to calculate F_{ST} between groups that you think might have genetic divergence. If $F_{ST} = 0$, there is no population subdivision. If $F_{ST} = 1$, there is complete population subdivision.

To calculate F_{ST}, you use one allele and calculate its frequency in the groups you think may be divergent. First you will calculate the average heterozygosity of the subpopulations (where S stands for "sub"):

$$H_S = \frac{2p_1(1 - p_1 + (2p_2(1 - p_2)}{2} \tag{9}$$

Where p_1 and p_2 are the allele frequencies in subpopulations 1 and 2 (respectively). Calculating heterozygosity of a population may seem counter-intuitive, since up to this point we have only examined heterozygosity in the case of an individual's genotype. However, it simply refers to the frequency of both alleles within a population.

Next you will calculate the heterozygosity of the population as a whole (H_T where T stands for "total", considering the total population):

$$H_T = 2\frac{p_1 + p_2}{2}(1 - \frac{p_1 + p_2}{2}) \tag{10}$$

Lastly, you will take the normalized difference of these to values to calculate the divergence between the subpopulations:

$$F_{ST} = \frac{H_T - H_S}{H_T} \tag{11}$$

2 Phylogenetics

Phylogenetics originates from the root words "phylo-" meaning lineage and "-geny" meaning origin. In a broad sense, a phylogeny is simply a family tree on a larger scale, and phylogenetics is the field of evolutionary genetics that seeks to estimate phylogenies and implement them in diverse scenarios to include an evolutionary context in biological questions.

Phylogenetics is a tool that was developed by people studying "macro"-evolution, and it uses the evolution definition of "descent with modification from a common ancestor" to examine how lineages evolve. A key difference between between population genetics and phylogenetics is that the latter includes relationships when comparing multiple groups (just like a family tree).

Reading phylogenies may seem intuitive, but it can be tricky. I recommend watching this video to get a solid understanding: `https://www.youtube.com/watch?v=Xd6Tc7tGOec`.

Originally, morphological traits assumed to be inherited (such as body size and scale counts) were used to estimate a phylogeny. Today, however, most phylogenies are estimated using molecular data (usually DNA). The shift from morphological data to molecular data has helped us understand relationships between organisms that were previously very difficult to understand.

Many approaches have been developed for phylogenetic inference (e.g., maximum parsimony, maximum likelihood, Bayesian inference, and recently some machine learning approaches). For this class, you are simply required to (1) understand that genetic sequence data can be used to estimate evolutionary history, (2) know how to interpret estimated phylogenies, and (3) be able to create a phylogeny using the UPGMA distance method (described here: `https://www.youtube.com/watch?v=09eD4A_HxVQ`.