Biology 3250, Ecology and Evolution

Spring 2021

**Microsatellite DNA Lab**

*Working with population genetic data*

Week 1

1. Class demonstration with candy (if time permits).
2. Real genotype data: nine-banded armadillo.
3. Homework: repeat analysis (from II) for grasshopper sparrow data set: redo exercises *a*-*e* for locus Asu09 from the grasshopper sparrow data set.

Week 2

1. Real genotype data: nine-banded armadillo (continued).
2. Homework: repeat analysis for grasshopper sparrow data set; do exercises *f-i.*

Learning outcomes:

1. The primary aim of this lab is to learn how to do basic population genetic calculations (e.g., genotype frequencies, allele frequencies, observed heterozygosity and expected heterozygosity). In the second week we will learn how to test for Hardy-Weinberg equilibrium and linkage disequilibrium, as well as how to calculate the system of mating inbreeding coefficient (*f*) and test if it is statistically significant.
2. Secondary aims of today’s lab include improving your skills with spreadsheet software (i.e., MS Excel) and specialized statistical software, including practice interpreting the results of a statistical test.

**Week 1**

1. **Class demonstration with candy**

* Two alleles: butterscotches (*b*) and cinnamon discs (*c*).
  + Each student picks two.
* Calculate frequency of the *b* and *c* alleles.
* Calculate observed genotype frequencies.
* Calculate allele frequencies from genotype frequencies using the “Jedi” equation.
* Calculate HW expected frequencies based on allele frequencies.
* Demonstrate random mating (by random draw of two alleles from the gene pool).
  + Calculate observed genotype frequencies and to HW expected genotype frequencies.
* Review methods for testing the null hypothesis of HW.

*Introduction to Microsatellite DNA (uDNA) markers*

Genetic sequences that are neutral and **hypervariable** may be used to make inferences about microevolutionary processes. For example, **microsatellite DNA (uDNA)** are short segments of non-coding DNA with a nucleotide sequence (e.g., “CA”) that is repeated a variable number of times (e.g., “CACACA” or “CACACACA”, etc.). During replication, these types of repeated sequences are prone to a phenomenon known as “DNA slippage” and this results in an elevated **mutation rate** ( = 10-3 to 10-5).

The perk of uDNA is that it is variable enough to resolve differences between individuals, but not so variable that **identity by state** is a big problem. If the temporal scale of interest is too deep or if the spatial scale of inquiry is too large, identity by state becomes more of a problem and other more slowly evolving markers are preferred.

In diploid organisms, each individual has two alleles at each microsatellite locus. Alleles are typically scored into size classes using some form of **gel electrophoresis**, where each electromorph is assumed to be a distinct **allele**; similarity of alleles within and between individuals is assumed to be due to **identity by descent**.

Identifying polymorphic microsatellite loci is relatively easy for model organisms (such as humans) where the genome has been sequenced and many uDNA loci have been identified, including information about sequences flanking the microsatellite locus that can be used to develop oligonucleotide primer sets. Otherwise, you have to use molecular genetic techniques to locate and screen potential microsatellite loci. This process starts with the creation a **clone library** (using restriction enzymes, phages and *E. coli* bacteria) and then washing the clone library with repetitive sequences that are “labeled”. Once a microsatellite sequence is found, it is spliced out and then sequenced, and the sequences in the flanking regions of the locus are used to design oligonucleotide primers for the forward and reverse strands surrounding the locus.

If you are lucky, somebody may have already developed locus-specific primers for your study system. These are usually published as primer notes: a special paper in a scientific journal that provides the primer sequences and summary information for each locus based on a sample of the focal species and then usually some information as to whether or not the primer set worked for closely related taxa.

Once you have the primers, you use PCR to amplify the targeted locus and then some sort of method to visualize the fragments, such as a “gel”. PCR reactions usually include DNA, a primer set (forward and reverse primers) for a particular locus, dNTPs, some sort of polymerase (the most popular is *TAQ*), and some sort of PCR buffer. These reagents are put in a small tube and then placed in a thermocycler. After PCR is completed, you load the product of your reaction into the well of a gel (or capillary tube), along with a size standard, and then subject it to an electric current. Larger fragments will migrate through the gel (or capillary tube) more slowly than small fragments. In the final step, you “score” the size of your alleles by running them in juxtaposition to the known size standards. Nowadays, most people use automated genotyping machines, which often include some sort of software package that helps you to identify the size of the fragments.

1. **Real genotype data: nine-banded armadillo**

You will be using two real microsatellite DNA data sets: one (**Armadillos**) to learn how to do the calculations in class. For your homework, you will repeat the same analysis (and answer questions *a*-*e*) for a second data set (**Grasshopper Sparrows**).

*The genotype data is contained in MS Excel spreadsheet that has been posted on BlazeView. The genotype data for each species is contained on a different sheet in the Excel file.*

*Calculating allele frequencies and genotype frequencies*

I have posted an MS Excel table containing the genotype data on BlazeView (“genofun.xls”). Go ahead and save a copy of this table to the desktop and make an extra copy (for backup). Open the table.

* Note that the first column contains a unique identifier for each individual. The next columns contain the individual’s genotype for each locus, where each locus is represented by two columns (“diploid two column format”).
  + Each allele is specified by its size (*e.g.*, 228) and each genotype corresponds to the two alleles that an individual carries (*e.g.*, 228│240). The top row of the table contains the name of each locus (or an abbreviation). For example, note that the first two columns represent locus *Dnov1* and that the two columns associated with that locus are called *Dnov1a* and *Dnov1b*. The last column of the table contains information about whether the individual was an adult, juvenile, yearling, or whether the age-class was not specified (marked as “?”).
* To process this data, the first step is to remove all of the non-adults (including the individuals who are missing this information (= “?”). The best way to do this is to **sort** the age class column (the default setting is to “expand the selection” to include all of the columns) and then deleting the rows of individuals that were non-adults when captured.
* Note that second and third columns contain genotype information for locus *Dnov1*. For some individuals, a genotype was not obtained or could not be obtained (denoted by “0”). Individuals with no genotype should not be included in an analysis for that locus.

1. ***List the alleles that were observed at locus Dnov1 and their frequencies (calculate the allele frequencies by counting the number of alleles of a particular size and then dividing it by the total number of alleles screened at locus Dnov1).***
2. ***List the observed genotypes at locus Dnov1 and the frequency of each genotype (calculate the genotype frequencies by counting the number of individuals with a particular genotype and then dividing by the total number of individuals genotyped at locus Dnov1).***
3. ***Calculate the allele frequencies again for locus Dnov1, but this time use the “Jedi Equation” to calculate allele frequencies from the genotype frequencies.***

**Observed heterozygosity** is the number of observed heterozygotes at a locus divided by the total number of individuals genotyped at that locus. Conversely, one can calculate observed heterozygosity as one minus the relative frequency of homozygotes (since the relative frequency of homozygotes and heterozygotes must sum to one).

**Expected heterozygosity** is the number of individuals expected to be heterozygous under Hardy-Weinberg. Expected heterozygosity is usually calculated as:

where is the frequency of the of alleles. In other words, expected heterozygosity is calculated as one minus the expected relative frequency of homozygotes.

1. ***For locus Dnov1, calculate (manually…using Excel) observed and expected heterozygosity.***

*Nei’s unbiased gene diversity*

When allele frequencies are calculated based on a sample from a larger population, expected heterozygosities may vary depending on sample size. For this reason, it is preferable to use an “**unbiased estimator**” of expected heterozygosity, such as that proposed by Nei (1973):

Where is the number of individuals sampled at that locus. Note that as increases, the expression outside the parentheses approaches 1. In other words, as sample size increases, the unbiased estimate approaches the theoretical expectation.

1. ***Recalculate expected heterozygosity (again only for Dnov1) using the unbiased estimator of Nei (1973).***
2. **Homework**

Your **HOMEWORK** is to repeat the analysis (exercises *a*-*e*) for a different data set: **Grasshopper Sparrows**. The genotype data can be found on a separate sheet of the Excel file that you downloaded from BlazeView. Again, do only the first locus: Asu09. For all questions that require calculations in MS Excel, you should show your work. Clearly label your answers and show some evidence of how you set up your calculations.