Biology 3250, Ecology and Evolution

Spring 2021

**Microevolutionary processes**

*Working with population genetic data*

**Week 2**

R**eal genotype data: nine-banded armadillo (Cont.)**

Last week we learned how to do some basic population genetic calculations in MS Excel. For today’s lab, we will focus on how to use computer software to test for HW equilibrium and linkage disequilibrium (LD).

To start today’s lab you will need to reopen the Genofun.xlsx file (from BlazeView) and go back to the sheet for “Armadillos”. If you are starting from scratch, you will again need to sort the “Age class” column and then remove the juveniles, yearlings, and question marks (?)…then delete the column. You will also need the genotype counts for the first locus. Follow the instructions on the following pages and answer questions *f-i*.

*Your homework is to repeat these analyses for the Grasshopper Sparrow data set.*

1. **Testing for HW equilibrium**

As explained in the Lecture 7 (*Population genetics and microevolutionary processes, part II*), a common way to test for HW equilibrium is via a chi-square goodness-of-fit test. Recall that the categories are the genotypes and you are comparing the observed count to the expected count under HW.

1. ***Conduct a chi-square goodness-of-fit test (manually….using MS Excel) for the first locus. Report the test score, the number of degrees of freedom, and the critical value for a significance level of 0.05 (find the critical value using the table from lecture, or just find one via a search engine). Report whether or not you can reject the null hypothesis.***
2. **The beauty of software**

In reality, most data sets are too large and too cumbersome to do calculations manually, so many people have developed software to facilitate the calculations.

A simple example is a software extension that was written for MS Excel, called the Excel Microsatellite Toolkit.

* Make sure you are on the sheet with the raw genotype data and that you have deleted the “age” column.
* If you look in Excel, I have added a tab in the menu bar for Add-Ins\*; choose this tab. Next click on the drop down menu for “Microsatellites” and choose the option entitled “Microsatellite Toolkit”.
  + A menu should appear. Choose the appropriate input format (diploid two-column format) and unclick the option to “Check data for errors”. Then press “OK!”. The next menu allows you to choose the populations and loci you want to work with. Since we are only working with one population and are interested in all the loci, just click “OK!”.
  + A final menu will appear which gives you various options for summarizing or formatting the data. Click the box labeled “Calculate Allele Frequencies and Diversity Statistics” and then click “Go!”.

The Toolkit will have added five new sheets to the Excel file. One of the sheets contains the counts for each allele (by locus) and another sheet contains the allele frequencies. This is much faster than “doing it by hand”.

* Have a look at the different summary statistics that are generated by the Excel Microsatellite Toolkit. Note that one of the sheets is named “Hz and PIC”; these sheet contains estimates of observed heterozygosity () and Nei’s Unbiased Gene Diversity for each locus. You will need this information to calculate the system of mating inbreeding coefficient (see part III, next page).

1. **The system of mating inbreeding coefficient, *f***

Recall from lecture that the system of mating inbreeding coefficient, *f*, is a correlation coefficient (i.e., its value ranges between -1 & 1). It measures the deviation from HW random mating expectations at the level of the deme. Positive *f* indicates system of mating inbreeding, negative *f* indicates avoidance of inbreeding, and an *f* of zero indicates random mating.

There are several ways to estimate the system of mating inbreeding coefficient; the simplest estimator is based on heterozygosity and is calculated by subtracting from one the ratio of observed heterozygosity to expected heterozygosity:

Recall that the toolkit generated a sheet labeled “Hz and PIC”, created by the MS Excel Toolkit .

1. ***Using the results for observed heterozygosity and Nei’s Unbiased Gene Diversity, calculate (manually using MS Excel) the system of mating inbreeding coefficient (f) for each locus. Then average f over all loci. Based on these loci, what is the system of mating of the deme? Again, show your work (in MS Excel).***
2. **Non-parametric test of HW equilibrium via randomization of the system of mating inbreeding coefficient (*f*)**

For the purpose of real scientific investigation, we have to test whether we could have arrived at our observed value of *f* due to chance alone. The simplest way to do this is with a randomization test. To conduct a randomization test for *f*, you randomize the alleles among the individuals, and then recalculate *f*. This procedure is done over and over (e.g., 999 times) to create a distribution of *f*-values based on the randomized data. If there were only a very small proportion of randomizations (say less than 2.5%) that yielded a value of *f* equal to or larger than the value you observed, then you reject the null hypothesis of no difference between your observed value and the randomized values.

You are going to conduct a randomization test for *f* (for each locus as well as for the average *f*-value over all loci) using custom scripts in R. Follow the directions of Dr. Anderson to learn how to run the scripts.

1. ***Include a printout of the results from R. For (α = 0.05) for which loci can we reject the null hypothesis of no difference between our observed value and the random expectation? For the average over all loci, can we reject the null hypothesis?***
2. **Testing for linkage disequilibrium**

Multi-locus measures of population structure assume that loci are independent (that there is no linkage disequilibrium). Because we cannot look at the gametes directly, we instead test for *genotypic* linkage disequilibrium. The idea is that we should not see certain combinations of genotypes more than would be expected if the genotypes were independent. If we cannot reject the null hypothesis of independence of genotypes, we might choose to discard problematic loci before conducting subsequent multi-locus analyses that assume independence.

* To test for linkage disequilibrium, we are going to use the program “GenePop on the Web”. However, to use the program, we have to get the microsatellite data into a format that the program can read.
* Go back to the raw data sheet, again select the Microsatellite toolkit from the “Add-In” menu and repeat the same steps to get to the last menu (again allowing it to delete the sheet 1 Col Data)…but in the last step, under “Format Options” choose “GenePop 3-digit format” and then click “Go!” The program will ask you to name the GenePop file and it will then create a new sheet called “GenePop3D”. Copy the text from the GenePop3D sheet and paste it into a new text file on the desktop. Be sure to delete the extra line at the end of the file before saving. Name the file and save it; this time, you do not need to add a file extension.
* To find “GenePop on the Web” simply type “genepop” into a search engine and then open the link to the website. Select “Option 2. Linkage Disequilibrium”.
* To understand better the logic behind testing for genotypic linkage disequilibrium, start with sub-option number 2: “Only create genotypic contingency tables”. This creates a contingency table for each pair of loci, where the cells in the table represent the counts of individuals in the population with a particular combination of genotypes. We want our results returned as HTML plain text (so check that box). Now browse for the text file that you saved to the desktop and then submit your data. It should give you the results in a matter of seconds. These are our “observed” counts; we want to compare these to the “expected” counts under the null hypothesis that genotypes are independent at each locus.
* Now we want the results for sub-option number 1: “Test for each pair of loci in the population” using the “log-likelihood ratio statistic” (G-test). The G-test is an analog of the chi-square goodness-of-fit test and, like the chi-square test, is used when you want to test whether an observed count fits an expected count for different nominal categories (here counts for different combinations of genotypes for a pair of loci).
* Again we want the results returned as HTML plain text (so check that box). Now browse for the text file that you saved to the desktop and then submit your data.

1. ***Include a printout of the results summary for the LD tests that clearly shows the P-value for each locus. State which pairs of loci exhibit significant LD (assume α = 0.05)***
2. **Homework**

Your **HOMEWORK** is to repeat the analysis (and answer the questions *f-i*) for **Grasshopper Sparrows**. For part III (exercise “f”) do the first locus only (Asu09); for all other exercises include all loci.