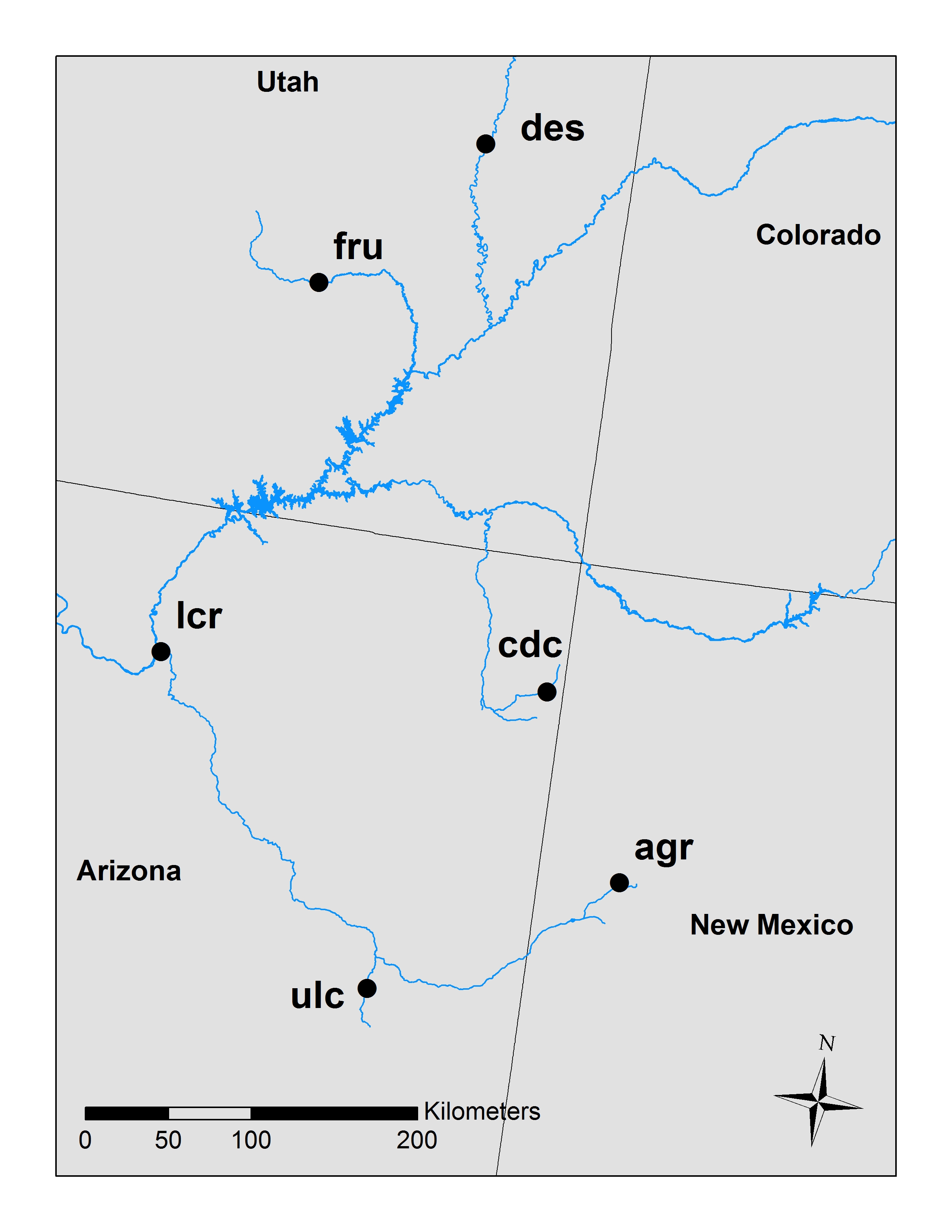
**Computer Lab 12 – Population Structure**

**Conservation Genetics (BIOL 4174 / 5174)**

**Population Information and Descriptive Statistics**

Populations of a native fish (Bluehead Sucker) are distributed throughout the Colorado River basin in the Southwestern US (see map below).



1. **AGR** –supposed to be a different subspecies.
2. **CDC** – potentially introgressed by another ESU.
3. **DES** – mainstem population, upper basin.
4. **FRU** –tributary population, upper basin.
5. **LCR** –tributary population, lower basin.
6. **ULC** – different ESU.

Part I – Structure

Structure 2.3.4 uses Bayesian clustering algorithms to study population structure. Individuals can be assigned to populations without any *a priori* locality information. It is available from <http://pritchardlab.stanford.edu/structure.html> for Windows, Mac OSX, and Linux.

**Using Structure**

Structure requires the use of a special file format that can be exported from any of a variety of programs, including GenAlEx. However, the Structure file format itself is very flexible, meaning it can be produced in several different variations to incorporate (or exclude) numerous categories of data. The procedure given below describes how to read data files that were produced by the default settings in GenAlEx.

**Beginning a new Structure project.**

Download the Lab 12 assignment file from Blackboard and unzip it to your personal folder on your desktop. Open Structure from the dock and follow the instructions below:

* Go to “File 🡪 New Project.”
  + Name the project grand\_canyon
  + Click “Browse…” to select the working directory. Find your lab\_12 folder, and then click the “select” button.
  + Click the other “Browse…” button to read your data file. Select the   
    structure-data.str file in your lab\_12 folder
  + Click the “Next” button
  + Enter the following:
    - Number of individuals: 211
    - Ploidy of data: 2
    - Number of loci: 15
    - Missing data value: -9
  + Click the “Next” button
  + Check the following boxes:
    - Row of marker names
    - Data file stores data for individuals in a single line
  + Click the “Next” button
  + Check the following boxes:
    - Individual ID for each individual
    - Putative population origin for each individual
    - USEPOPINFO selection flag
* Click “Finish” and then “Proceed.” If no error messages are produced, then the file has been read successfully and a new project has been created.

**Setting Parameters**

Structure allows for a wide variety of parameters to be set, the most important of which being the ancestry and allele frequency models. Unfortunately, Structure has no analog for the “Modeltest” programs that are used in choosing among models of DNA sequence evolution for tree building. Rather, you must rely upon knowledge of your data and study species in selecting appropriate models.

Four choices are available for the ancestry model:

* **Use No Admixture Model** – Under this model, each individual is fully assigned to one of several populations. Misapplication of this model will lead to strange results, so it should be used only when it is known that your data is comprised of non-interbreeding groups that have been separated for many generations.
* **Use Admixture Model** – This model allows individuals to have mixed ancestry. In other words, one individual may inherit a fraction of its genome from population A, and another fraction from population B. This is a more general model that can be applied in many situations to elucidate gene flow among populations or cryptic population structure. This is also the most frequently used model, and is suggested as a starting point by the manual for most analyses.
* **Use Linkage Model** – This is an extension of the Admixture model that is designed to compensate for low levels of linkage disequilibrium. However, it is still unable to deal with markers that are very tightly linked.
* **Use Population Information to test for migrants** – This model can be applied when population structure reflects sampling locations with the exception of a few individuals that appear to be misclassified (i.e., potential recent immigrants into a population).

**Do question 1 in the homework document.**

Three choices are available for the Allele Frequency Model

* **Allele Frequencies Correlated** – this model assumes allele frequencies to be similar for different populations, possibly as a result of gene flow or shared ancestry.
* **Allele Frequencies Independent** – this model assumes allele frequencies are very different among populations (i.e., no gene flow and deep historical population separation).
* **Infer Lambda** – Lambda is the model parameter that specifies the distribution from which allele frequencies are sampled. Lambda is also fixed at a set value or inferred from the correlated or independent allele frequency models (this is hidden under the “Advanced” option for both models). The default option is to fix lambda = 1 for either allele frequency model. The manual recommends leaving this option alone for most analyses.

**Do question 2 in the homework document.**

**Setting up a Structure run**

As with all Bayesian analyses, exploratory runs must be done initially to both get a feel for the data and determine how program settings must be adjusted.

* Go to “Parameter Set 🡪 New…” and enter the following options
* Run Length
  + Burnin: 25,000
  + MCMC Reps after Burnin: 50,000
* Ancestry Model
  + Select the model you chose in question 1.
* Allele Frequency Model
  + Select the model you chose in question 2.
* Click “O*K*” then enter the name “exploratory” for your parameter set name and click “O*K*” again.

**Starting a run**

* Go to “Project 🡪 Start a job…”
* Click “exploratory” in the window to highlight it, then set *K* from 2 to 5 and click “Start”
  + The ***K*** values are the numbers of populations to be tested (i.e., for *K*=2 the program will attempt to assign each individual to one of two populations). Structure itself does not have a means of directly estimating *K*. Rather, a range of *K* values suspected to encompass the actual number of populations is tested. You will evaluate only a short range of K values, partially due to time constraints of the lab, and partially because it is more difficult for runs with higher K values to converge during the short runs that are being conducted.

**Bugs in the Structure GUI**

Some rather irritating bugs exist in the GUI version of Structure. Descriptions (and solutions) are provided below:

* **Bug 1**: Structure has a bug that frequently prevents a run from starting. If a window pops up at this stage that says “Start Job…” and nothing else, click “OK” in that window. Then do the following:
  1. Save your project.
  2. Close Structure.
  3. Reopen Structure.
  4. Go to “File 🡪 Open Project” and navigate to the location where your project is saved. Look for and open a file named “grand\_canyon.spj.”
  5. Try starting the job again.
  6. Repeat steps 2 through 5 until Structure actually works.
* **Bug 2**: Structure 2.3.4 has a second bug that causes it to interpret your choice of K values as suggestions rather than literal commands. It will sometimes ignore them completely, and instead test random K values of its own choice. If this happens, conduct a run for just the K value (or values) that Structure decided to skip.

**Viewing Results**

If necessary, expand the “Parameter Sets,” “exploratory,” and “Results” folders in the upper left pane of your Structure window. If you click on the results for one of the runs, they will appear in the top right pane of the window. The exact contents of this window will vary depending on your model choices and other selected options, since different options result in the estimation of different parameters. The program will report the inferred ancestry of each pre-defined population (in this case, the six sampling locations listed on the first page of this lab) and the inferred ancestry of each individual in the data file. All details of the program settings are at the very end of the results file.

Results can also be displayed graphically. Click on a results file, and then go to “Bar plot 🡪 Show” in the results window. Now click “Group by POP Id” in the bar plot window. This groups individuals by the locations from which they were sampled (Numbers 1 through 6 correspond to the sampling locations provided on the first page of this lab). Each vertical bar represents the population ancestry of one individual as estimated by Structure.

**Do questions 3 through 5 in the homework document.**

The program provides useful time series plots for assessing MCMC convergence under the “Data Plot” menu. Plots for different parameters will be available contingent upon the model selected. For the Admixture/Independent allele frequency model, the only useful plot for this purpose will be **alpha** (the parameter for degree of admixture – small alpha indicates most individuals have ancestry from a single population, while an alpha > 1 means most individuals are of mixed ancestry). If convergence has occurred, the plot should approximate a “white noise” pattern similar to that observed in MrBayes during lab 7.

**Do question 6 in the homework document.**

Structure can also produce triangle plots, which are most useful for visualizing *K*=3. In this type of plot, each individual is represented by a circle. Each circle is colored by the location from which it was sampled. Each corner of the plot represents one of three clusters found by Structure (Cluster 1 in the lower left corner of the triangle, cluster 2 in the lower right corner, and cluster 3 at the top). Individuals that are unambiguously assigned to a certain cluster are in the corners of the plot. Individuals partially assigned to two clusters fall along the edge of the plot, somewhere between the two corners. Individuals that are equally assigned to all three populations will be shown in the center of the triangle.

**Do questions 7 through 10 in the homework document.**

Part II – Structure Harvester

Structure itself has no built-in method for determining the appropriate *K*, although its manual makes suggestions that include calculating the posterior probability of each *K*, and plotting the log likelihood of the data against each *K*. Other researchers developed the delta *K* method (Evanno et al. 2005), which has been implemented in the program Structure Harvester. This program is easily used through a web interface at <http://taylor0.biology.ucla.edu/structureHarvester/>. The original Python code (Python is pre-installed on Mac OS X) can also be downloaded from this site and run locally on your own machine. Running the program locally via command line is sometimes necessary for very large files.

**Evanno Method**

Running the Evanno method is very computationally intensive due to the number of Structure runs that must be conducted to carry out the calculations. First, you must decide on a range of *K* values that are expected to include the true value of *K*. The *K* values that immediately precede and follow this range must also be tested. For example, if *K* = 2 through 7 are to be tested using the Evanno method, then Structure must also be run for *K* = 1 and *K* = 8. As a result, it is impossible to ever directly evaluate *K* = 1 using this method since *K*=0 cannot be calculated. Multiple independent runs (usually 10 or more) must be conducted for each *K* value. Additionally, each independent simulation is typically run for several hundred thousand or perhaps millions of generations. Run time also increases exponentially as *K* increases.

For large datasets testing many *K* values, analyses can run for days or even weeks on a desktop computer. Some relatively simple scripting can run the command line version of Structure in parallel on a computer cluster, thus significantly reducing run time. This is yet another case which demonstrates the importance of programming knowledge to the biologist. However, running an analysis of such length is impossible during lab hours even when the program is run in parallel, so a long Structure run has already been completed for you.

In Structure, close your current project (“File 🡪 Close Project”) then go to “File 🡪 Open Project…” and navigate to the “lab\_12\_analysis” folder on your desktop. Open the “Consgen\_analysis.spj” file and expand the results folder. This analysis was conducted on the same data file used for the exploratory analyses earlier in this lab, but some additional settings were modified to reduce the variation in results observed at each *K* value. Primarily this involved setting informative priors for location using the LOCPRIOR option in Structure, as well as doubling the default value of ALPHAPROPSD (this influences the standard deviation of the prior on alpha) to improve convergence.

**Do question 11 in the homework document.**

**Using Structure Harvester**

Structure Harvester will perform all of the calculations described in Evanno et al. 2005, prepare graphs and summary tables, and prepare files for other programs designed to process Structure results (Clumpp and Distruct). Navigate to the results folder of the full analysis located at ~\Desktop\lab\_12\_analysis\iaf and click on the “results” folder to highlight it. Control-click the folder and choose “Compress” from the resulting shortcut menu. This will create a file named “Results.zip.”

In Firefox, go to the website <http://taylor0.biology.ucla.edu/structureHarvester/> (or Google structure harvester, and it should be the first hit), click “Browse” and upload the zipped results folder that you just created. Click “Harvest” and the website will begin analyzing the data and generating the graphs.

The first graph [L(*K*)] can be used to determine *K* via one of the methods described by the Structure manual. Examine the graph. *K* values are plotted on the X axis, while the mean ln(likelihood) of the data is on the Y axis. The point at which the ln(likelihood) reaches an asymptote is supposed to represent the number of populations present.

**Do question 12 in the homework document.**

The fourth graph (Delta*K*) is used by the Evanno method. The *K* value that corresponds to the highest value of Delta*K* is supposed to represent the true number of populations.

**Do questions 13 through 16 in the homework document.**