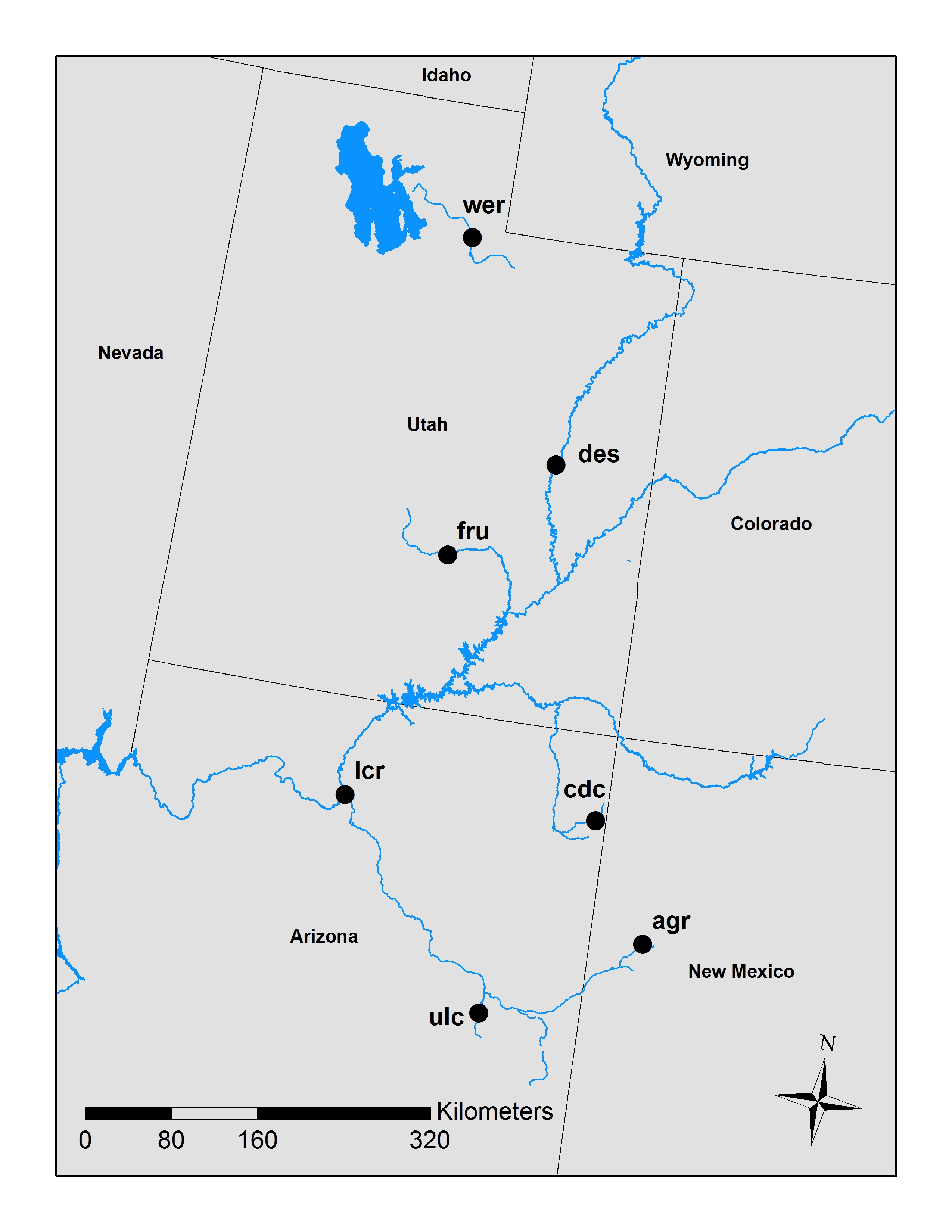
**Computer Lab 9 – Genetic Distance Using Microsatellites**

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

**Population Information and Descriptive Statistics**

Populations of a native fish (Bluehead Sucker) are distributed throughout the Colorado River (populations 1-6) and Bonneville (population 7) basins in the Southwestern US (see map below). The Bonneville basin has no contemporary connections to other basins or oceans.



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.
7. **WER** – different ESU. Bonneville Basin.

Part I – Genetic Diversity Indices

All calculations today will be performed in GenAlEx, a plugin for Microsoft Excel freely available from <http://www.anu.edu.au/BoZo/GenAlEx/>. A valid Microsoft Excel license is required to use the program. The plugin is compatible with all PC versions of Excel from 1997 onward, and Excel 98 to 2004 on the Mac, as well as Excel 2011 for Mac. It will not work in the Mac version of Excel 2008. This program can calculate a wealth of summary statistics and basic analyses, as well as convert your data into different formats.

**Using GenAlEx**

Download the Lab 11 homework file from Blackboard and unzip it to your personal directory on your desktop. Open the “msat-data.xlsx” file in Microsoft Excel. Now start GenAlEx by opening the “GenAlEx 6.501.xla” file. When prompted, you must enable macros or else the program will not work. Click the “Add-Ins” tab at the top of the Excel window to find the GenAlEx button on this menu.

Click “GenAlEx” and go to “Frequency…” Click “OK” in the resulting window. In the next window you will see several options for basic summary statistics, as well as options for other things such as calculating distances and pairwise *F*ST. Begin by calculating the same basic descriptive statistics. Make sure the following boxes are checked:

* Frequency by Pop
* Het, Fstat & Poly by Pop
* Het, Fstat & Poly by Locus
* Private Alleles List

Click “Ok” when these are selected and give the program a few seconds to do the calculations. After it is finished, you should have five new sheets within your Excel workbook.

* **AFP** – this sheet lists the frequency of each allele for each locus within each population.
* **HFP** – This sheet displays a variety of summary statistics for each locus in each population, as well as mean values for each. Some of the important summary statistics are as follows:
  + **N** = number of samples.
  + **Na** = Number of alleles.
  + **Ne** = effective number of alleles (if all alleles were equally frequent in a population, this would be the number of alleles needed for the same level of heterozygosity).
  + **Ho** = the observed heterozygosity.
  + **He** = expected heterozygosity.
  + **UHe** = unbiased heterozygosity (this calculation is useful for low sample sizes).
* **HFL** – This sheet contains the same summary statistics as the HFP sheet, except they are calculated for each individual locus at each population.
* **PAS** – This sheet shows the private alleles list by population. These are alleles that only appear within one population and nowhere else.
* **PAL** – This sheet contains all of the individuals in the dataset that contain one or more private alleles.

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

Part II – Distance Measures

**Genetic Distance Measures**

The following are common measures of genetic distance, most of which can be calculated in GenAlEx.

* Nei’s Standard Genetic Distance – this measure requires the calculation of **Nei’s Genetic Identity (*I*)**, after which distance is easily calculated using the formula *D*=-ln(*I*). This assumes the markers used are selectively neutral. It is based on the **Infinite Alleles Model** (= IAM): all mutations result in new alleles of any size; thus mutations can consist of addition or deletion of any numbers of repeats. Genetic Identity ranges from 0 to 1, while Genetic Distance ranges from 0 to infinity. Negative values are interpreted as 0.
* *G*ST – this measure is the proportion of genetic diversity found within populations. When only two alleles exist at a locus *G*ST = *F*ST. When there are multiple at a locus, *G*ST is the weighted average of *F*ST for all alleles. *G*ST will not be calculated directly using GenAlEx, however a method will be used to calculate *F*ST that is “more or less equivalent” to *G*ST according to the GenAlEx manual.
* *F*ST – this is a measure of the variance among subpopulations relative to the total population variance. *F*ST assumes an infinite alleles model (IAM).
* *R*ST – analog of *F*ST adapted specifically for microsatellite loci, except it assumes a **stepwise mutation model** (= SMM); mutations consist of addition or deletion of a single repeat. Consequently, alleles of similar size are genetically more closely “related” than alleles of very different size.

Make sure you have your “Data” spreadsheet active. Click “GenAlEx,” go to “Frequency…” and click “OK” in the first dialog box. Click the “Uncheck All” button to clear selections, and then check only the following:

* Nei Distance
* Pairwise *F*st
* Output Labeled Pairwise Matrix

Click “OK” to have the program produce four new spreadsheets within your file – **NeiP**, **NeiL**, **FstP**, and **FstL**.

On the NeiP sheet, there are two matrices. The top matrix is Nei’s Genetic Distance. Increasing values represent greater dissimilarity among populations. The bottom matrix is Nei’s Genetic Identity. Populations with a Genetic Identity value close to 1 are very similar to one another, while values close to 0 represent greater dissimilarity.

The FstP sheet contains a pairwise matrix of *F*ST values. The method used to calculate these values is the one described by the GenAlEx manual as being “more or less equivalent” to *G*ST values.

**Do questions 2 through 5 in the homework document.**

**Calculating *F*ST and *R*ST from an AMOVA**

Now, *F*ST will be calculated from an AMOVA. This method uses a permutation procedure to assess significance, in addition to using a different equation. Make sure your “Data” spreadsheet is active. Go to “GenAlEx 🡪 AMOVA…” and click “OK” on the first window to pop up. Make sure the following are selected:

* Codom-Allelic
* As Tri Matrix
* Label Matrix
* Sample
* Analysis for Total Only

Then click OK. Another window will now pop up. Make sure the following settings are selected:

* Permutation number should be set to 999 (there are two boxes where this may need to be changed)
* Make sure “Output Pairwise FST Matrix” and “Output Pairwise Linearized Fst Matrix” are checked.
* Leave all other settings as default.

It will take a few minutes for GenAlEx to perform the calculations. The program may appear to freeze during this time, but it is still working.

This will produce three new spreadsheets – one labeled **Fst**, another labeled **FstP (2)**, and a third named **LinFst**. Ignore the LinFst sheet for now – it will be used in the Mantel test.

**Do questions 6 and 7 in the homework document.**

Now, *R*ST will be calculated from an AMOVA. Make sure your “Data” sheet is active. Go to “GenAlEx 🡪 AMOVA…” and click “OK” on the first window to pop up. Make sure the following are selected:

* Codom-Microsat
* As Tri Matrix
* Label Matrix
* Sample
* Analysis for Total Only

Then click OK. Another window will now pop up. Make sure the following settings are selected:

* Permutation number should be set to 999 (there are two boxes where this may need to be changed)
* Make sure “Output Pairwise RST Matrix” is checked.
* Leave all other settings as default.

It will again take a few minutes for GenAlEx to perform the calculations. The program may once again appear to freeze, but it is still working.

This will produce two new spreadsheets – one labeled **Rst** and another labeled **RstP**.

**Do questions 8 through 11 in the homework document.**

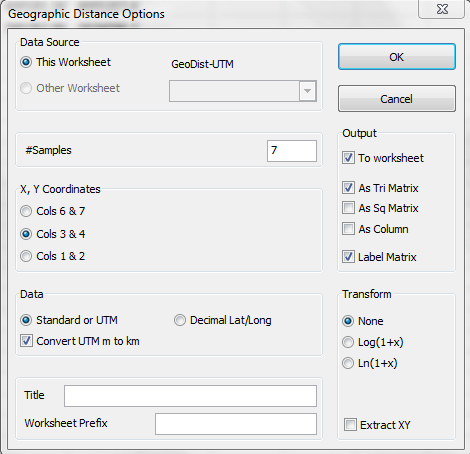
Part III – Mantel Test

A Mantel test allows for the testing of a statistical relationship between two distance matrices. This method calculates a correlation coefficient for the two matrices (this ranges from -1 to 1, and uses random permutation to test for significance. In this test, the null hypothesis is that no relationship exists between the X and Y matrices. In population genetics, this method is used to test for **Isolation By Distance (IBD)**, which assumes that the genetic distance between two populations will increase linearly as geographic distance increases.

GenAlEx tests for significance using a permutation procedure. For a Mantel test, this means the data in the Y matrix is shuffled around while the X matrix is held constant. The Mantel test is performed again following each iteration. If the correlation coefficient for the permuted matrix is found to be more extreme (closer to +1 or -1) than the original data less than 5% of the time (ɑ=0.05), then the relationship between the two matrices is significant.

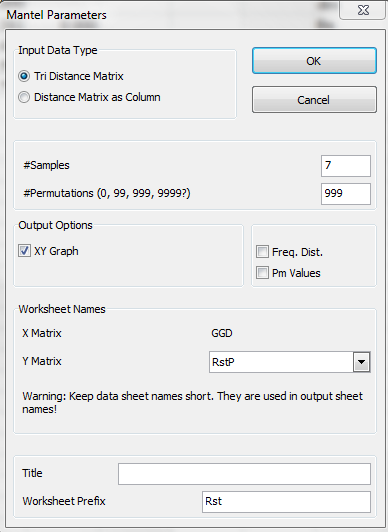
**Performing a Mantel Test in GenAlEx**

First, a matrix of pairwise geographic distances must be created. Click the “GeoDist-UTM” sheet to make it active. Then go to “GenAlEx 🡪 Distance 🡪 Geographic…” and select the following options:



Once everything has been set, click “OK.” This generates a new sheet labeled **GGD** with a matrix containing the geographic distance between each pair of sites in kilometers (i.e., the distance along a straight line between two populations). This will be the X matrix.

With the GGD sheet active, go to “GenAlEx 🡪 Mantel 🡪 Paired…” and select the following options:



Click “OK” once everything has been set. This should produce a new sheet named **Rst GDDvRstP MT**.

**Do questions 12 through 19 in the homework document.**

Part IV – Principal Component Analysis

Matrices such as those generated in part I of this lab can be difficult to interpret by sight alone, and this may be an impossible feat for larger matrices. As a result, visual representation of the data can aid in identifying trends. Tree-building using distance methods and microsatellite data is possible, however the quickly-evolving nature of microsatellite data combined with the large number of taxa typically used in population genetic studies makes it impractical. Instead, a method known as Principal Component Analysis (PCA) can be used.

PCA works by transforming data into a new coordinate system with multiple coordinates. Each coordinate (or axis) explains progressively less of the variation. Usually, the first few axes explain most of the variance.

To perform a PCA in GenAlEx, first select a matrix. For the first PCA, select the FstP (2) matrix. Then go to “GenAlEx 🡪 PCoA 🡪 Analysis…” and leave all of the settings at their default values. Click “OK” to produce the sheet named **PCoA**.

To view plots of different axes, go to GenAlEx 🡪 PCoA 🡪and choose one of the other PCA options (Axes 1 vs. 3, or Axes 2 vs. 3). Remember: the axes are perpendicular to one another, so as you cycle through the different plots, it’s similar to viewing a 3-dimensional plot from different angles.

**Do questions 20 through 28 in the homework document.**