**Activity 3: Assessing Genetic Variation across California**

**Learning Outcomes:**

1. Explore the data and functionality of the DNA sequence database GenBank
2. Perform a principal component analysis

**Software:** BCEEnet Shiny App, R 4.2.2+

**Background:** Introduction to Genetic Analysis

**Outline:** 1. Introduction to PCA in Genetics …………………….. 1

2. Search and download DNA sequences………………. 2

3. Align DNA sequences………………………………... 2

4. Perform a PCA ……………………….……………… 3

**1. Introduction to PCA in Genetics:**

Recall in our landscape genetic study, we have three essential questions:

1. Where is **genetic diversity** highest?
2. Is there genetic evidence of more than one **Evolutionary Significant Unit (ESU)** in the state?
3. Is there evidence of a potential **barrier**?

To answer these questions, and test your revised hypothesis for your focal species, we will conduct a principal component analysis (PCA) on single locus mitochondrial DNA sampled across a region of California (see map from Activity 2). As you read in the background, a PCA is a way to take many variables and reduce the variation across them into fewer variables called principal components. In this case, we will treat each position in a gene as a variable; your sequence data may include as many as 1500 nucleotide positions (*aka* base pairs). We will use a PCA to reduce the variation found in the samples across these many gene positions to two variables, PC1 and PC2. Individuals with similar DNA sequences will have similar values of PC1 and PC2. When we plot the values of PC1 and PC2 for each individual on a graph, samples that are close to each other on the plot share similar DNA sequences. **We interpret two individuals that are close on a PCA graph as closely-related to one another. Individuals that are located far apart on the principal component graph are more distantly related to one another.**

A principal component analysis is useful for assessing variation in lots of different kinds of data. In genetics, PCA be applied to single locus datasets like ours up to entire genomes with hundreds of thousands of base pairs! To help connect our genetic data to the landscape of California today, we will use a special PCA function in the software R designed for these voucher specimens. Once you have a PCA graph of your California samples, we will walk through how to interpret this result to answer our essential questions.

**2. Search and download DNA sequences:**

1. To download the sequence data, go to <https://www.ncbi.nlm.nih.gov/genbank/> or search in your browser for GenBank.
2. Find the search bar at the top of the page, below the NIH header. Change the dropdown menu at the left, set to the default “**Nucleotide**,” to “**PopSet**”.
3. Search the PopSet ID from Activity 1, Table 2 corresponding to your species or species complex. This will take you to the PopSet page which includes a citation for the original study that published the sequences as well as links to the sequence for each individual.
4. On the upper right side of this page, close to the sidebar, click the dropdown menu “**Send to**”. For “**Choose destination**”, select “**File**”. Then, make sure the download format is FASTA before hitting “**Create File**”. This will download the DNA sequences in a fasta format to your computer.
5. Give the DNA sequences a logical name like GENUS\_######.fasta so you can find it again.

**3. Align DNA sequences:**

To compare data from a set of individuals, it is critical that each sequence has the same frame of reference. The only way to see natural variation among individuals is to have the nucleotide position in a gene, for example, be perfectly matched between all individuals. Thus, positions 1, 2 and 3 in a gene must correspond in all individuals, even if some individuals did not get sequenced at positions 1, 2 and 3. In that case, the uncertainty of a specific position in an individual will be represented by an N (aka any nucleotide) or a dash indicating a gap in the reported sequence (Table 1).

**Table 1.** Example of an aligned sequence with three individuals and three positions where A, T, G and C refer to nucleotides, N indicates an uncertain nucleotide, and a dash indicates missing data.

|  |  |  |  |
| --- | --- | --- | --- |
| **Position** | **1** | **2** | **3** |
| Individual 1 | A | T | G |
| Individual 2 | N | T | G |
| Individual 3 | - | - | C |

To ensure all individuals in your popset have the same frame of reference, you need to align your sequence data.

**To align the PopSet DNA sequences:**

1. Search embl-ebi muscle or go to <https://www.ebi.ac.uk/Tools/msa/muscle/>
2. For “**STEP 1**”, use the “**Browse**” button to navigate to your saved fasta file and select it to upload it.
3. In “**STEP 2**”, use the dropdown menu to select “**Pearson/FASTA**” as the output format.
4. Press the green “**Submit**” button in “**STEP 3**” to submit your DNA sequences to be aligned. This may take a little while.
5. The aligned fasta will be displayed on the resulting page. Select “**Download Alignment File**” to download the aligned file. Again, save it with a logical name like GENUS\_PopSetID#\_aligned.fasta to indicate this file is aligned.
6. If the aligned file prints to your web browser, you can select all and copy the text into a text editor like BBEdit, or you can select “**Save page as**” to save it as a text file.
7. The aligned file is ready for evaluation using a principal component analysis!

**4. Perform a PCA:**

Before making a PCA with your aligned DNA sequences, you will need to download and install R and Rstudio. Here is a good resource for this ([Installing R and Rstudio](https://www.dataquest.io/blog/installing-r-on-your-computer/#:~:text=Go%20to%20the%20CRAN%20website,make%20changes%20to%20your%20device.)). You will also need to know how to find and write the path location to 1) your .fasta files and 2) save out the PCA results.

Write the full path to your .fasta file here:

Write the full path to where you’d like to save your PCA results:

1. Open Rstudio.
2. **Go to File à Open File and open the file called Example\_PCA\_Script.r.** You should see a tab open in your Rstudio window that contains an R script. This script contains information on how to install libraries in R (Step 0) and information on how to make the PCA from your data (Steps 1-4).
3. **Install the librarian, devtools, and BceenetPCAPackage by highlighting lines 4-8 and hitting the run button in the top righthand corner of the tab.** Before performing your analysis, you will need to install and load various R packages. R packages contain functions that perform certain tasks. If you wish to use a function, you must first install the package on your computer. Step 0 contains three lines (Lines 4, 5, and 8) which install the librarian, devtools, and BceenetPCAPackage respectively. To install these packages, you should ensure that you are connected to the internet. Then you can highlight lines 4-8 and hit the run button in the righthand corner of the tab. You may notice that text appears in the “console” section of your Rstudio window. This is expected! Once you have installed a package you shouldn’t have to do so again (i.e., you should only need to run step 0 once). Think of it like downloading a game, once it is on your system you don’t need to download it every time you want to play—you just need to open it up!
4. **Load your libraries by selecting lines 12-34 and hitting the run button in the top righthand corner of the tab.** While you do not need to install libraries every time you use Rstudio, you will need to load in the libraries that you would like to use at the beginning of each R session. We will use the librarian package to aid us in loading in several other packages. Select lines 12-34 and hit the run button in the righthand corner of the tab. In the console you may see many lines of text appear (and a stop sign in the righthand corner), this is normal! Wait for the stop sign in the corner of the console to disappear. Once it is gone it means that your libraries have successfully been loaded.
5. **Save the full pathnames to your .fasta file and to the location you’d like to save your PCA output into the provided variables.** Lines 39 and 41 will take the user entered path and save it as a variable to be used in later sections of this script (Steps 3 & 4). Both lines have a similar format: First you will come across the variable name (e.g., Find.AlignedDNA.Here and Save.files.here). Next you will see an arrow ( ß ) which is the symbol for “assign to” or “save as”. Finally you will see the function file.path(). This function takes text and tells the computer to recognize it as a file path. The computer will interpret these lines of code in the following way: 1) take the text inside the file.path() function and recognize it as a file path, then, save that file path the variable that is front of the arrow. You will need to take the file paths that you recorded above and copy them inside the quotation marks inside the file.path() function.

 Here is an example of what a file path may look like:

/Users/MyUserName/Desktop/Folder\_With\_FASTA\_Files

And an example of how this should look in the code:

Find.AlignedDNA.Hereßfile.path("/Users/MyUserName/Desktop/Folder\_With\_FASTA\_Files")

After you have copied the two paths into the code on lines 39 and 41. You will then highlight lines 39-41 and hit the run button in the upper righthand corner of the tab.

1. **Run line 45 to set your working directory.** Line 45 tell the computer where you’d like to save your PCA graph and output. In Step 2 you save the file path as a variable. Here in step 3, we use the setwd() function to take that variable and tell the computer to use that as your save location. Highlight line 45 and hit the run button in the upper righthand corner of the tab. To check that you’ve successfully told the computer where to put your output files you can put your cursor into the console section of Rstudio, type getwd(), and hit enter. You should the same file name that you initially entered in step 3 pop up in the console.
2. **Make your PCA using the fasta.to.pca() function from the BceenetPCApackage.** Lines 51 and 52 contain the function fasta.to.pca() from the BceenetPCApackage. This function takes two input items: 1) the path to your fasta files and 2) a title for your graph. The first input item (path to the fasta) is on line 51; you have already saved this path as the variable Find.AlignedDNA.Here and therefore do not need to change anything about this line. Line 52 has an input field called save.filename followed by an equals sign and some text inside of quotation marks. The text that lives inside of the quotation marks will be used as the graph title for your PCA and will also be the name of the saved PCAgraph and excel output table. Change the text inside of the quotation marks to be an informative title for your PCA graph. Highlight line 51-52 and hit the run button in the upper righthand corner of the tab. When you run these lines you may see some text appear in the console tab of Rstudio—This is normal! When this line of code is done running you should see a PCA graph appear in the Plots tab in the lower righthand corner of Rstudio.
3. **Find your PCA graph and PCA output table on your computer.** The code you just ran not only produced a graph inside of Rstudio, but also saved out that graph (.pdf) and additional data about the PCA (.csv). Navigate to the location you told R to save your files and open up your files.