

Whichloci tutorial

Generating Simulated Datasets (AKA Resampled Files)

- 1) Open Whichloci
- 2) Click the **Select Baseline File** button
- 3) Open your microsatellite dataset that's in Genepop format (generated during the Convert tutorial).
- 4) An **Allele Frequencies File** is automatically generated. You can save this file if you wish (using the File menu) and later, when you use the program again, just import it instead of the baseline file.
- 5) Click the **Resample** button. This will create your simulated populations based on the allele frequencies just calculated, assuming HWE and no LD. You have options for how many datasets to generate and how large you want the populations to be.
- 6) Click **Ok**. A new window will appear with your resampled data set. Save it (again using the File menu) using a name like **resampled_100a** to remember what it is, then close the window.
- 7) For today's demonstration, repeat steps five and six to generate two more resampled files with population sizes of 100 and one file with population sizes of 500.
- 8) Whichloci is now ready to perform analysis on these files. Below are four examples of analyses that could be done.

What percentage of individuals in these datasets can be correctly assigned?

- 1) Click the **Resampled File** button. Open the first resampled file that you generated above.
- 2) Uncheck the **Loci Ranked By** box. Some options should grey out. **Assignment Stringency** should be at 0 and **Use Critical Population Method** should not be checked.
- 3) Click **Execute**. A report file will appear. You can save it if you like.
- 4) Scroll down to the bottom to see what percentage of simulated individuals were assigned to the correct population.
- 5) Keep the report open but close your resampled file. Open a different resampled file and repeat steps two through four.
- 6) Notice that the results are slightly different for each simulated dataset. The dataset with the larger population gives more precise results.
- 7) If you have extra time, try clicking the **Resample** button and entering a number greater than one for the number of datasets to generate. These resampled files will not appear in windows, as the program is now running in batch mode. Click **Execute**. In the report that is generated, you will now see more statistical information about the results, such as mean and variance.

Raising the Assignment Stringency

- 1) Open the resampled file of your choice.

- 2) Leaving the rest of the options as they were in the above exercise, change the Assignment Stringency (LOD score) to 1. This sets a minimum level of confidence for the assignments generated. Individuals that cannot be assigned with this level of confidence will be counted as incorrectly assigned.
- 3) Click **Execute**. Notice in the report that fewer individuals were assigned correctly, and the difference is listed as being due to stringency.

Ranking loci and determining how many you need for a given assignment accuracy

- 1) Open the resampled file of your choice, if one isn't open.
- 2) Check the **Loci Ranked By** box.
- 3) For today's demonstration, set **Min % Assigned to Population** to 80 and **Assignment Stringency** to 0. (You may want to aim higher in your actual experiments.)
- 4) Click **Execute**.
- 5) You'll see in the report that the loci are now ordered according to usefulness. As you scroll down, it will tell you how many loci were necessary to achieve the criteria you set in step 3 (it will always use the best loci first). At the bottom of the report is the percentage of individuals correctly assigned using only those loci.
- 6) There are two options for how the loci are ranked. **Whichrun assignment** calculates the percentage of individuals that would be assigned correctly if only that locus were used. It is particularly useful if you plan to analyze these loci in the program WHICHRUN. **Allele frequency differential** calculates the sum of all differences between allele frequencies across all the populations for that locus. If time allows, try changing which of these you use (whichrun assignment is used by default) and see how the results differ.

Looking at assignments and misassignments to one population at a time

- 1) Open the resampled file of your choice, if one isn't open.
- 2) Check the **Use Critical Population Method** box.
- 3) Click the button under it. Select **CANhat_30**.
- 4) The **Max % Misassigned To Critical Pop** box is no longer grayed out. Set this value to 10. Leave the other values as they were in the above example.
- 5) Changes to note when using this method: Not only can you set the maximum percentage of individuals that can be misassigned to **CANhat_30**, but the minimum percentage assigned to population now refers only to individuals that belong to the critical population. The allele frequency differential is also calculated differently when using the critical population method. Essentially, all the populations that are not the critical population are now treated as one population.
- 6) Click **Execute**.
- 7) Look at the report file. The results may be a bit different from what you saw when considering all populations. Note that the values at the bottom of the report now only refer to individuals correctly assigned or misassigned to **CANhat_30**.

Recommended reading:

Banks, M.A., Eichert, W., and Olson, J.B (2003) Which genetic loci have greater population assignment power? *Bioinformatics* 19(11):1436-8