**Long test (and other downstream analyses) for detecting admixture outliers**

STEP 1:

Open R, the .R script provided (LongTest\_Rcode.R), and load the data (ConGen\_LongTest\_data.txt).

Note the file path I provide in the code won’t work on your machine – you can edit directly, use the file.choose() command, or you can use the data import function in R Studio (hint – use R Studio!).

One other note – the calculations for the Long test require intermediate allele frequencies (i.e., no values of 0 or 1). The allele frequencies in your file have been adjusted to avoid those values, by using Bayesian posterior estimates of allele frequencies with a flat prior.

STEP 2:

Let’s run through the simple .r script for our first population (Finley Creek – a tributary to the Jocko River just south of here). The Long test calculates the observed variance in allele frequencies across all loci to build a neutral expectation for the distribution of allele frequencies (i.e. introgression) across loci. For each locus, the test then provides a residual chi-square and a p-value that indicate whether the locus differs significantly from the neutral expectation. Note that loci can differ significantly by having either a higher or lower introduced allele frequency than expected.

STEP 3:

Then populations 2 and 3 (Cyclone Creek and Hay Creek – tributaries to the North Fork Flathead River (about 50 miles north of here). Check out your output file to make sure the results have all been recorded.

STEP 4:

Now we have three-population specific p – values for each locus. Let’s combine information across loci with Fisher’s combined test. Fisher’s combined test combines results from multiple independent tests that are addressing the same null hypothesis (in this case – that locus specific allele frequencies are a function of overall admixture, sampling variation, and genetic drift). Simple to execute!

STEP 5:

We now have 9380 p-values, which is to say, a whole truckload. Multiple tests no longer satisfy our statistical criterion for type I error (false positives) so a correction is needed. With genomic data sets, simple corrections such as the Bonferroni are overly restrictive (conservative) potentially leading to type II error (false negative). Instead, let’s calculate a false discovery rate first in excel to show you the basic procedure (yes, antiquated but effective) and then in one simple r command.

Here’s a simple description taken directly from Narum 2006 (Conservation Genetics)

Macintosh HD:Users:ryankovach1:Desktop:Narum 2006 ConGen.pdf

EXERCISE (if time now, or in the evening):

Look through the admixture outliers that we identified and describe any general patterns? Are there any consistent patterns of selection, loci of interest, etc.? Report your results back and discuss the interpretation of them in the evening session (approximately 9 PM).