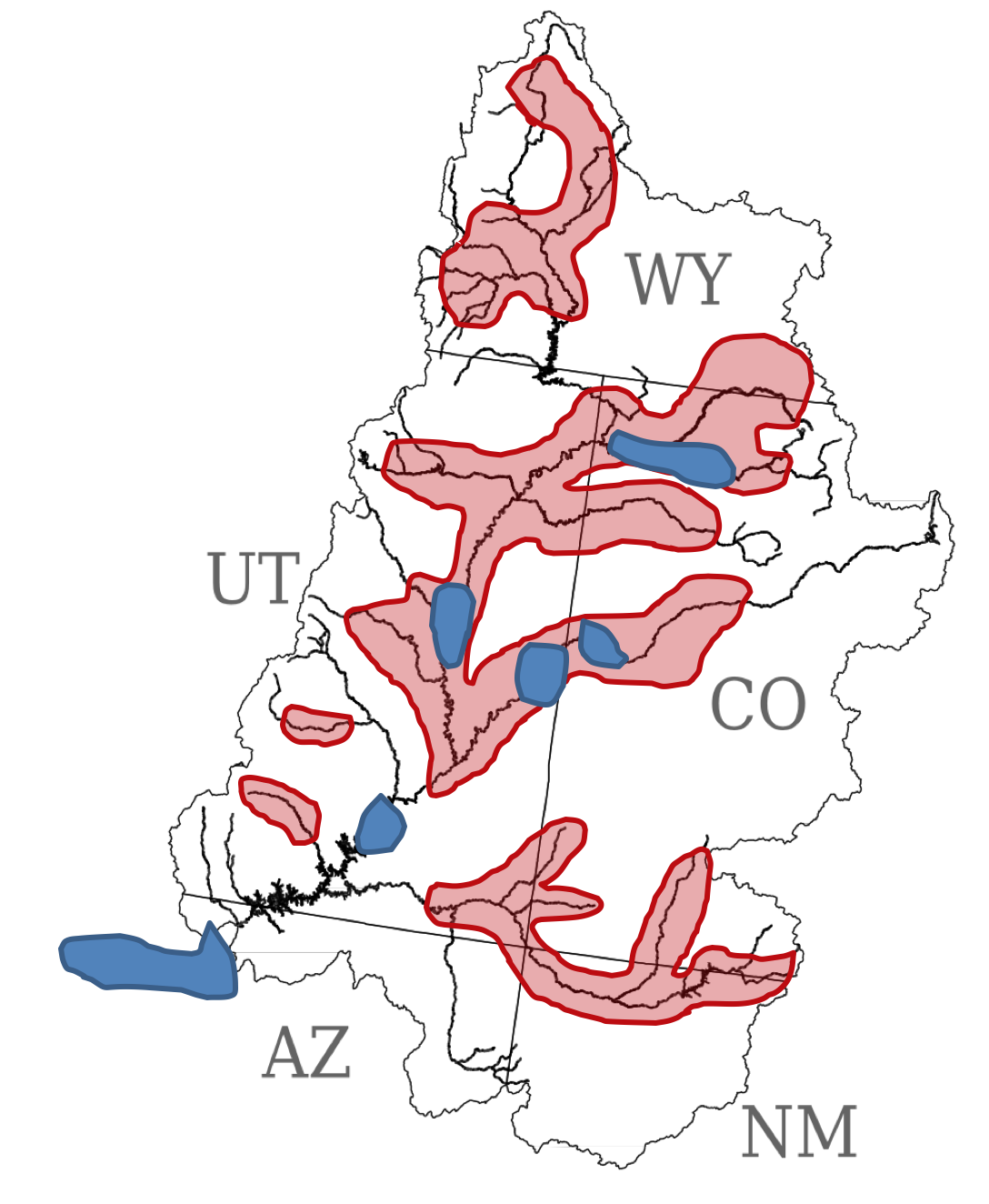
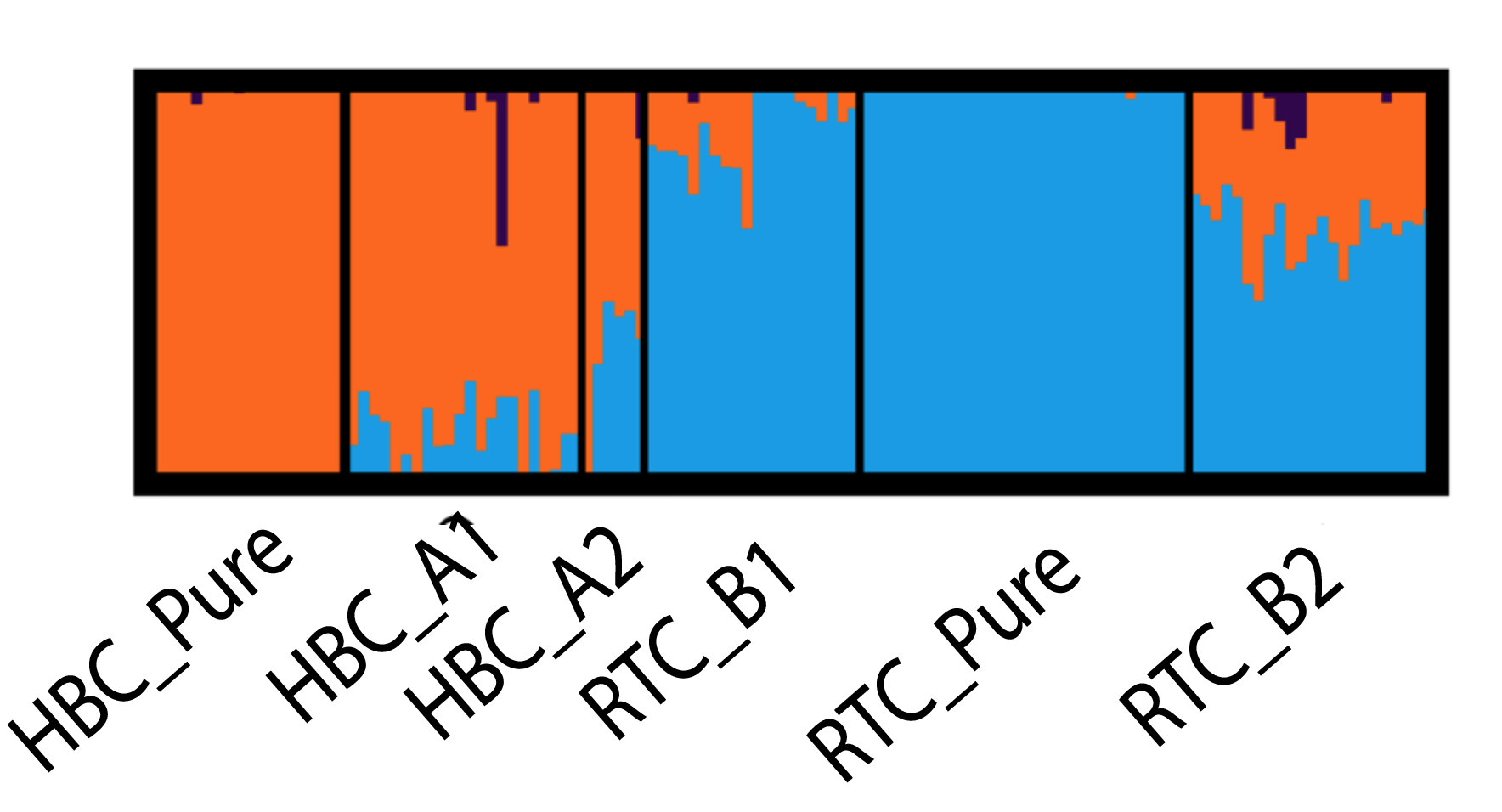
**Computer Lab 13 – Testing for introgression and genomic cline analysis**

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

Background

Today we will be looking at ddRAD data generated for species of endangered Humpback Chub (*Gila cypha*) and its sister species Roundtail Chub (*Gila robusta*) from the Colorado River. Humpback chub are only found today as 5 ‘aggregates’, and are constrained to canyon-bound reaches of the river. The Roundtail Chub are distributed more or less throughout the upper Colorado River basin. See range map below (HBC in blue, RTC in red).

I have provided a small subset of the data (subset of individuals, and randomly sampling SNPs), which we will use to explore potential hybridization between these two sister species. Take a look at the Structure barplot above, which we will use to form our hypotheses.

Part I: D-statistics

We’ll be using a parallel program written by a former lab member (Steve Mussmann) for calculating the D-statistics in parallel. This code is freely available (https://github.com/smussmann82/Comp-D) and includes options to calculate some additional variants of the D-statistic as well (partitioned-D [Eaton and Ree 2013] and Dfoil [Pease and Hahn 2015]). I’ve provided a pre-compiled binary in your Lab12/bin directory.

CompD requires an input file listing all combinations of individuals you would like to compute the D-statistic for, with one line each for the outgroup, P3, P2, and P1 lineages. In the interest of time, we’ll only do 2 of these tests. In the first, we will test for introgression from Roundtail chub into Humpback population “HBC\_A1”, and in the second we will test for introgression into population “HBC\_A2”.

To test for introgression we need to tell CompD which individuals for each of these populations we want to sample. Each combination of these will be used to compute one independent D-statistic. For example, we could choose 2-3 individuals for each population like so:

8BTC01 8BTC02 #Outgroup

8HPUR01 8HPUR02 8HPUR03 #P3

9RPUR01 9RPUR02 #P2

9B101 9B102 9B103 #P1

In this case, CompD would compute 36 separate tests, because there are 36 different combinations of the above samples (sampling one from the outgroup, one from P3 and so on). To save us some time, I’ve written a script which will create all of these input files for you. Run it like so to create your test files (make sure you have changed directories to the location of your Lab12 files first):

./bin/makeCompD.py -p gila\_popmap.txt -m 4 -i hbc\_tests.txt

The options used tell the script to sample 4 individuals per taxon (-m 4), using the populations chosen in the file ‘hbc\_tests.txt’. You could have easily done this manually by choosing samples from the gila\_popmap.txt file, but I figured I might as well save you some time. ☺

You should now have 2 separate “.compd” input files, each performing a different test:

BTC+RTC\_Pure+HBC\_Pure+HBC\_A1.compd : Test introgression from RTC -> HBC\_A1

BTC+RTC\_Pure+HBC\_Pure+HBC\_A2.compd : Test introgression from RTC -> HBC\_A2

These are all hypotheses developed based on the Structure results.

Before running compD (using the commands below), first check out the help menu so you can see what each of the options is doing:

./bin/compD -h

Next, run unzip the datafile and runcompD for each of the two tests:

./bin/compD -i gila.phy -l 7818 -H -P -b 100 --fourtax -t BTC+RTC\_Pure+HBC\_Pure+HBC\_A1.compd -o HBC\_A1.out

./bin/compD -i gila.phy -l 7818 -H -P -b 100 --fourtax -t BTC+RTC\_Pure+HBC\_Pure+HBC\_A2.compd -o HBC\_A2.out

Once this is done, you should have two output tables (HBC\_A1.out and HBC\_A2.out). Open one of them in TextWrangler to see the results. CompD should have calculate 256 separate D-statistics, and reported both chi-square and Z-scores, along with p-values for each. Are most tests significant, or non-significant? What is the sign of the D-statistic (positive or negative)? What does this mean as far as presence or absence of introgression between these populations?

To get a better view of the results, you can load each table and make some plots in R:

setwd(“/PATH/TO/LAB12”)

#Load HBC\_A1 results

hbc\_a1 <- read.table(“HBC\_A1.out”, sep=”\t”, header=T)

#Load HBC\_A2 results

hbc\_a2 <- read.table(“HBC\_A2.out”, sep=”\t”, header=T)

#Boxplot of D values

boxplot(hbc\_a1$D, hbc\_a2$D)

#Count number of significant tests at alpha = 0.05

nrow(hbc\_a1[hbc\_a1$Z.pval < 0.05,])

nrow(hbc\_a2[hbc\_a2$Z.pval < 0.05,])

#Boxplot of D values for significant tests

boxplot(hbc\_a1[hbc\_a1$Z.pval < 0.05,]$D), hbc\_a2[hbc\_a2$Z.pval < 0.05,]$D))

#You can plot some other stuff if you want

You probably noticed that HBC\_A1 gives a weird result, in that some tests are significant for introgression from P3->P2 and others are significant for introgresson from P3->P1. Lets take a look at the tests which are positive, and try and figure out what’s going on:

#Still in R

hbc\_a1[hbc\_a1$D > 0.2,]

What is common across all of these samples? What might this result mean?

Part II: H-index and Genomic Cline Analysis

We will be doing the rest of our analyses in R, but first we need to make some input files. This time we’ll use a filtered dataset, where I’ve removed singleton loci, and filtered more stringently for missing data using scripts that are freely available on my GitHub if you ever have a need to filter large SNP alignments (github.com/tkchafin/scripts).

./bin/phylip2introgress.pl -p gila\_popmap.txt -i gila\_introgress.phy -1 HBC\_Pure -2 RTC\_Pure -a RTC\_B1+RTC\_B2+HBC\_A1+HBC\_A2

Check out the help menu so you know what’s going on:

./bin/phylip2introgress.pl -h

NOTE: You may get a bunch of ‘uninitialized value’ errors from the Perl script. That’s because my script sort of sucks and is a little broken.

Now, go ahead and load the file ‘lab13\_introgress.R’ into RStudio, or into TextWrangler. You should be able to complete the remainder of the lab using the instructions in that file.