Task 07: Scrub Jays

I. General

Type up all of your work in a text editor. Basically, you should NEVER type things directly into the `R` terminal. Type them into a text editor, then either run them or copy/paste them into `R`.

Before you begin, make a new folder in Tasks called Task_07, and save an empty file named task07.r in that folder.

You must also make a separate "Project" folder within Tasks. Inside your Project folder, make "code", "data", and "text" subfolders. By Friday, you should have some data in the data folder, some code to read the data in to `R` in the code folder, and a document with a *single sentence* describing your hypothesis.

When you're **done** with this assignment, turn it in by (1) saving your text document, (2) opening your Terminal or GitBash, (3) navigating to the appropriate directory using `cd` and (4) typing:

`git add -A` (enter)
`git commit -m "Task 07 and my project hypothesis"` (enter)
`git push -u origin master` (enter)

II. Hypothesis

Your hypothesis is due this Friday. It is extremely difficult to, without any research, simply conjure a well-formed hypothesis. Instead, you'll want to do a number of searches and probably read several papers. You are all capable of doing this, and doing it well, but it requires effort and research.

You know how to revisit old labs where hypotheses were laid out for you, and you know to question some of the assumptions, as we did that in a previous `R` exercise.

You know how to download data directly in `R`, as we did it in the third `R` assignment with the paleobio-database.

You know how to read data into `R`, as we've done it in almost every `R` assignment.

You know how to look up papers pertinent to evolutionary biology, as we did that in the last project assignment.

You are all capable of formulating an appropriately tractable and specific hypothesis. Further, you know enough of evolution at this point that you should be able to determine an appropriately evolutionary hypothesis. If you want to discuss what you are thinking about, I will give you some insight and try to ensure you don't travel down a poorly chosen path, but ultimately this assignment is for you to demonstrate to me your understanding of both evolution specifically and biological hypotheses generally.

III. Background

Here, you'll be analyzing allele frequency data for a population of Florida scrub jays collected from 1998 to 2013. *Be careful when examining objects!* We'll be using some very large objects, so don't just type things in. Use `dim()`, `nrow()`, and related functions to understand the objects.

In the paper, they used sophisticated statistical models to quantify the forces that induced the observed changes in the frequencies of the alleles. You will (1) examine the observed data critically and understand what it implies about the evolution of this population, (2) use `simPop` in a manner similar to what we did previously with the BIOL-112 lab to try and assess the relative role of different evolutionary processes.

IV. Data & Simulations

The data are broken into many pieces. There are about 10,000 different alleles examined in this study, and the frequency of each allele in the population each year is given in the `overallFreq` object.

Additionally, however, there is the `alleleFreqs` object, which has the frequency of each allele in the overall population (`freq`), but also the difference in frequency between the overall population and those (1) birds born that year (`d_birth`), (2) birds that survived into that year from the previous (`d_surv`), (3) birds that arrived that year as immigrants (`d_imm`), and the number of newborns, survivors from the past year, and immigrants (`n_` columns). Additionally, the relative change in the frequency of the locus from the starting (1998) frequency is given in the `rfreq` column.

```
# this will take a few seconds to run
source("http://jonsmitchell.com/code/reformatData07.R")
source("http://jonsmitchell.com/code/simFxn.R")

# make a plot of each allele's frequency over time
plot(1, 1, type="n", xlim=c(1998, 2013), ylim=c(0, 1))
s <- apply(overallFreq, 2, function(x) lines(overallFreq[,1], x, col=
   rgb(0,0,0,0.01)))

# make a plot of each allele's rescaled frequency (observed - initial
   freq) over time
rescaleFreq <- apply(overallFreq[,3:ncol(overallFreq)], 2, function(x)
   x - x[1])
plot(1, 1, type="n", xlim=c(1998, 2013), ylim=c(-0.25, 0.25))
s <- apply(rescaleFreq, 2, function(x) lines(overallFreq[,1], x, col=
   rgb(0,0,0,0.01)))

# Force each computer to remake the right data
dYear <- c()
dAlleles <- c()

for (i in 3:ncol(overallFreq)) {
  dYear <- c(dYear, overallFreq[,1])
  Vec <- overallFreq[,i]
```

```
  Init <- overallFreq[1,i]
  dAlleles <- c(dAlleles, Vec - Init)
}

# Instead of plotting the individual allele frequencies, let's use all
    of this data to plot the probability of a given change in frequency
    (y) by year (x)
# This is basically the same as the previous plot, but instead of semi-
    transparent see-through lines, this gives a color-coded probability
    density.
smoothScatter(dYear, dAlleles, colramp = Pal, nbin=100)

# The smoothScatter() plot is a complete summary of how likely a given
    change in allele frequency is over a given period of time in this
    population. Changing the nbin argument makes it smoother or more ``
    pixelated''.
# Now, for the fun part. Using simPop, we can simulate populations and
    stick them on this graph.
# I've included the addFit() function to aide you in this task. This
    function runs a specified number (nruns) of simPop simulations with
    set parameters.
# Do this, and find me the combination of n, h, and s that best matches
     the empirical data.

smoothScatter(dYear dAlleles, colramp = Pal, nbin=100, xlab="year",
    ylab="change in allele freq. since 1998")
addFit(nruns = 50, n = 100, ngens = 18, startT = 1997, simCol = "gray40
    ", rescale = TRUE)

# Basically, find the combination of n, h, and s for addFit() that
    produces changes comparable to what is seen in the scrub jay
    population.

# You may also find the use of the d_ (change in frequency) and n_ (
    number of individuals) columns useful for determining the relative
    strength of the different evolutionary forces here. The plot below
    is one example of how they can be examined:
plot(alleleFreqs$d_freq, alleleFreqs$d_imm, xlim=c(-0.15, 0.15), xlab="
    overall freq. change", ylab="freq. change in subset")
points(alleleFreqs$d_freq, alleleFreqs$d_birth, col='blue')
points(alleleFreqs$d_freq, alleleFreqs$d_surv, col='red')
```

V. Extra Credit

This extra credit assignment is difficult, but all of you are capable of doing it.

These data are from the 2019 paper by Dr. Nancy Chen, which you are going to read this week for our discussion on Friday.

While you read that paper, you may notice that the data are freely available. Indeed, that is where I first downloaded and reformatted the data for you to use in this assignment. The code Dr. Chen and colleagues used to analyze those data is also posted along with the code.

Using whatever means you deem necessary, make me a plot showing the observed change in frequency for each allele against the probability selection caused that change.