**Quantifying and interpreting patterns of population structure**

**Week 9 Laboratory Tutorial**

**Distributed Graduate Seminar**

Sean Schoville

sean.schoville@wisc.edu

**Overview**

The goals of this lab are to:

1. Assess how patterns of genetic variation can be used to delimit natural populations.

2. Compare methods that assess population structure.

3. Understand how population structure can be used to interpret biogeographic history.

**Data Sets:**

* Simulated data using the two-island model and admixture model
* SNP data from Catchen et al. 2013 Catchen et al. 2013. The population structure and recent colonization history of Oregon threespine stickleback determined using restriction-site associated DNA-sequencing. Molecular Ecology 22:1365-294X. <http://dx.doi.org/10.1111/mec.12330>

**Part 1. Simulated data: 2-island model**

We simulated data under a classic two-island model, using the coalescent simulation program **ms**, developed by Richard Hudson. The program simulates a coalescent tree under various demographic models, and uses those trees to create allelic data. For those interested in using **ms**, the source code is available here: <http://home.uchicago.edu/rhudson1/source/mksamples.html>

We simulated **200** haploid individuals genotyped at **100** loci. The effective mutation rate was *μ*=0.5. We sampled 2 islands with 100 individuals in each. The effective migration rate was *Nm*=2 (*N* is the effective size in each of the two island, *m* is the bidirectional rate of gene flow). Our ms command was as follows:

ms 200 100 -t .5 -I 2 100 100 -ma x 10 10 x > dataNm10.txt

These raw data need to be converted in a format amenable to statistical analyses in R. To get started, either set your working directory to where your data files are (Windows, RStudio), or change your terminal directory to launch R from inside the data files directory.

**file <- scan(file = "dataNm2.txt", what ="character", sep="\n", skip = 2)**

**genotype <- NULL**

**for(locus in 1:100){**

**res.locus <- file[4:203]**

**file <- file[-(1:203)]**

**genotype <- cbind(genotype, as.numeric(as.factor(res.locus)))}**

Now we have a new data file, ***genotype***, loaded in R. This file contains 200 rows and 100 columns. Each row corresponds to a simulated individual. The columns code for their multi-locus genotypes. Our first objective is to use ordination (Principal components analysis, or PCA) to examine population structure for the Nm = 2 data set.

The R command for PCA is fairly simple and fast.

**pc = prcomp(genotype, scale =T)**

In order to visualize how the first two eigenvectors capture genotype variation, we will color each population.

**plot(pc$x, pch = 19, cex = 2, col = rep(c("blue2","orange"), each = 100))**

**Question 1: Is population structure (genetic differentiation) evident? How much of the genetic variance can be explained by our first two components?** To answer the second part of this question, use:

**summary(pc)**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Next we would like to see how population genetic structure relates to geographic space. To this aim, we could display PC maps. A PC map is a spatial interpolation of a particular component. Let’s map PC 1.

***# Create synthetic spatial coordinates (X,Y) and map them***

**coord <- cbind(sort(c(rnorm(100, -2, 1), rnorm(100, 2, 1))), runif(200))**

**install.packages(‘fields’, dependencies = TRUE)**

**# You will be asked to select a CRAN session. Choose any convenient location**

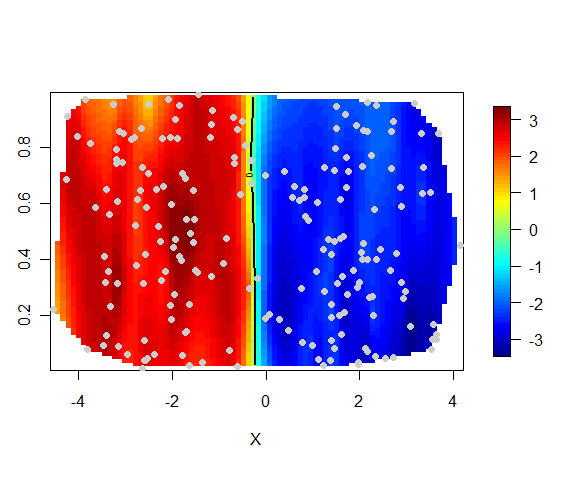
**library(fields)**

**fit <- Krig(coord, pc$x[,1], m=1)**

**surface(fit)**

**points(coord, pch=19)**

This map predicts the value of the first principal component at each location in our study area. We observe that the study area is partitioned into two zones that correspond to the 2 clusters visible from PC 1. We have superimposed individual sample sites to see our species distribution.



To check that the PC map is consistent with having 2 islands, we can examine the PC assignment vs the sampling location. Remember that, in the data sets, the 100 first individuals were sampled from island 1 and the last 100 were sampled from island 2. We compare these assignments to the PCA classification as follows.

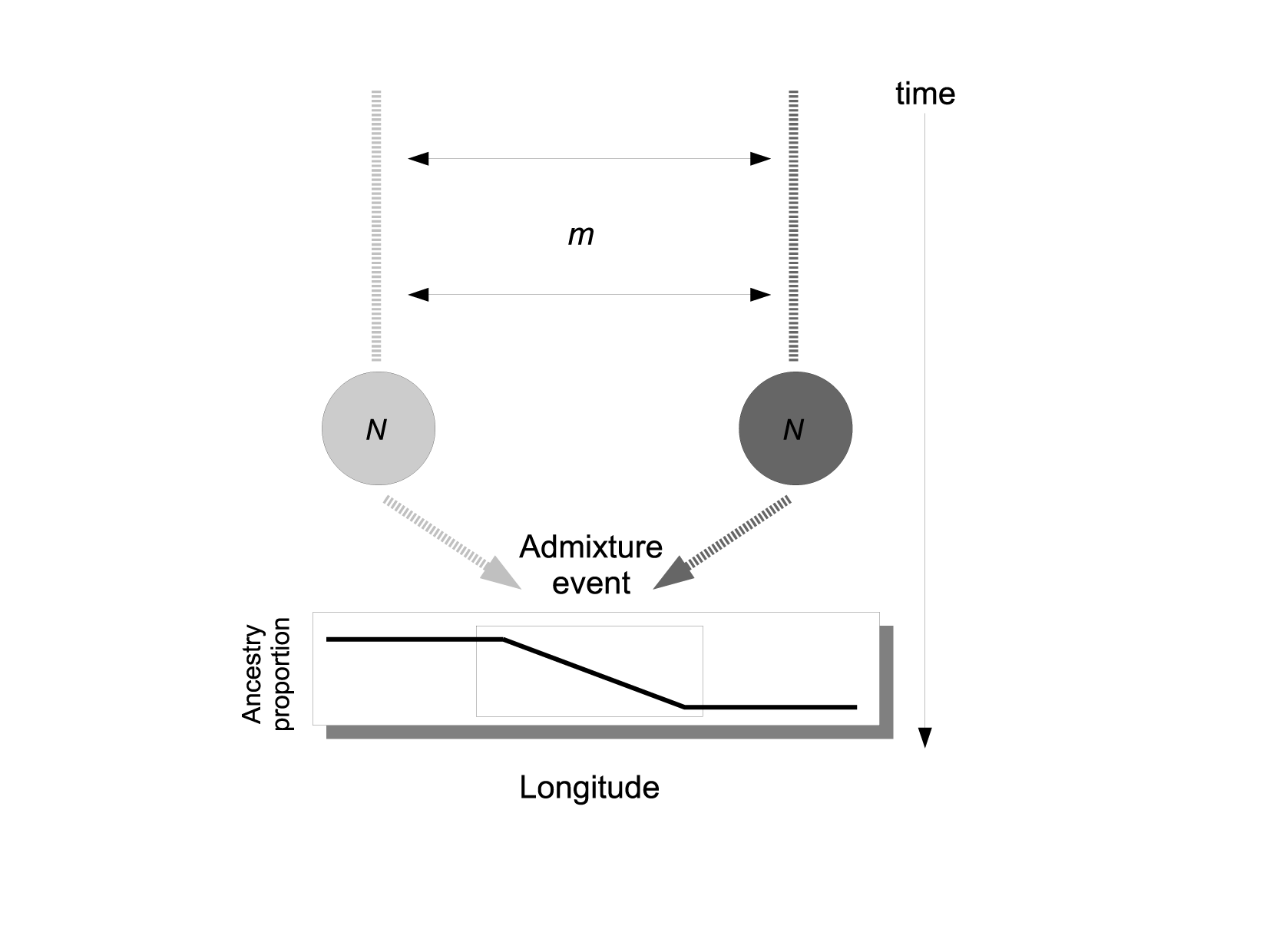
**table((pc$x[,1] > 0) == (1:200 <= 100))**

In this example, we found that only one individual was not assigned to its island of origin. Well, this individual might be a migrant from the last generation. These results indicated that a very simple method based on principal component analysis can correctly describe spatial population genetic structure.

**Question 2: Does PCA provide an accurate description of population genetic structure when the genetic differentiation between the 2 islands is less pronounced?** To answer this question, re-run the first analytical steps up to PCA for data simulated with Nm = 10 (a higher value of gene flow), which is stored in datafile **dataNm10.txt**. This was generated with the following ms command:

ms 200 100 -t .5 -I 2 100 100 -ma x 10 10 x > dataNm10.txt

**Part 2. Simulated data: 2-island model with admixture**



A 2-island model is a relatively simple scenario and unlikely to capture the variation we will see in empirical studies. Will our very basic assignment method based on PCA hold up to more complex scenarios? Let’s consider a scenario where the 2 populations had been evolving for a long time under an equilibrium island model, and then their environment suddenly changed. Our 2 populations had to move to track their shifting habitat, and after these movements they come into contact in an intermediate region. This contact event resulted in an admixed population with the density of mixed individuals greater in the center of the contact zone than at the ancestral origins at the edges of the landscape.

Using R and our previous simulation, a multi-locus cline that resumes this scenario can be simulated has follows. The source population data are: ***dataNm1.str***

**# A function for the shape of a cline**

**sigmoid <- function(x){ 1/(1 + exp(-x))}**

**p1 <- sigmoid( 0.5 \* coord[,1])**

**# Our admixed genotypes are built from a 2 island model with low gene flow (Nm=1)**

**genotype = read.table(file = "dataNm1.str")[,-(1:2)]**

**admixed.genotype <- matrix(NA, ncol = 100, nrow = 200)**

**for (i in 1:100){ for (j in 1:100)**

**admixed.genotype[i,j] = sample( c(genotype[i, j],genotype[i+100, j]), 1, prob = c(p1[i], 1 - p1[i]) )}**

**for (i in 101:200){ for (j in 1:100)**

**admixed.genotype[i,j] = sample( c(genotype[i - 100, j],genotype[i, j]), 1, prob = c(p1[i], 1 - p1[i]) )}**

**res <- data.frame(coord, admixed.genotype)**

Now our data set is the R object ***res***. The next exercise is to apply PCA to these data to evaluate how geographical genetic variation can be captured by this approach. In comparison with the previous example where we had two geographically discrete populations, we now have a “continuous” population in a landscape. Geographical genetic variation is thus expected to be more gradual than in the previous example. Generate and examine the PC 1 map.

**pcA = prcomp(admixed.genotype, scale =T)**

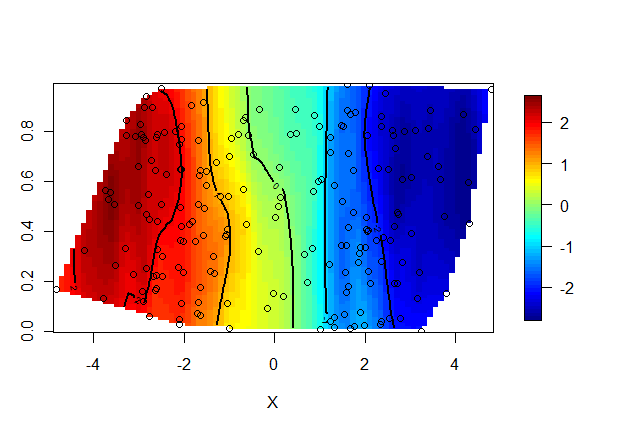
**plot(pcA$x, pch = 19, cex = 2, col = rep(c("blue2","orange"), each = 100))**

**#Look at the PC Map**

**fit <- Krig(res, pcA$x[,1], m=1)**

**surface(fit)**

**points(res)**



**Question 3: How does admixture change our prediction of population structure (PCA plot)? Is genomic ancestry correlated with geographical location?** To answer this latter part, check the R2 and significance of statistical association between PC1 component scores and geographical position (p1):

**Summary(lm(pcA$x[,1]~ p1))**

**Part 3. Empirical data: Threespine sticklebacks**

The Threespine stickleback (*Gasterosteus aculeatus*) is a fish that has emerged as a model of rapid and parallel adaptation. Catchen et al. (2013) were interested in how populations colonize freshwater habitats in the Pacific Northwest, USA. These sticklebacks have diversified into three life history forms, one exclusively dwelling in the ocean, another being adapted to freshwater habitats, and one unusual population that can utilize both habitats. It was unclear if this one particular population (Riverbend), from a stream in central Oregon, was introduced due to unintentional human transport, and if this could be an example of rapid adaptation to freshwater from oceanic populations. Single nucleotide polymorphism data were generated using genotyping-by-sequencing, for 9 populations occupying coastal areas and inland streams.



In this tutorial, we will analyze the genetic data generated by Catchen et al. (2013) using a few exploratory methods to quantify and visualize genetic differentiation among the stickleback populations sampled. By the end of this tutorial, hopefully you will be able to make a convincing argument for the regional origin of the recently-introduced inland stickleback population.

**data <- read.table("stickleback\_data.txt",sep="\t",as.is=T,check.names=F)**

**#Create a list of population IDs**

**pops <- unique(unlist(lapply(rownames(data),function(x){y<-c();y<-c(y,unlist(strsplit(x,"\_")[[1]][1]))})))**

To understand the experimental design a bit better, let’s look at the sample size at each site.

**sample\_sites <- rep(NA,nrow(data))**

**for (i in 1:nrow(data)){**

**sample\_sites[i] <- strsplit(rownames(data),"\_")[[i]][1]}**

**N <- unlist(lapply(pops,function(x){length(which(sample\_sites==x))}))**

**names(N) <- pops**

**N**

Let’s start examining population structure, first using PCA. We’ll look at the amount of variation explained in the first few components, and then we’ll plot individuals for four components, coloring them by population.

**pcaS <- prcomp(data,center=T)**

**plot(pcaS$sdev^2/sum(pcaS$sdev^2),xlab="PC",ylab="Fraction Variation Explained",main="Scree plot")**

**# Get % variance explained for first few PCs**

**perc <- round(100\*(pcaS$sdev^2/sum(pcaS$sdev^2))[1:10],2)**

**names(perc) <- apply(array(seq(1,10,1)),1,function(x){paste0("PC",x)})**

**perc**

**#Get RColorBrewer package and select a color palette**

**install.packages('RColorBrewer', dependencies = TRUE)**

**library('RColorBrewer')**

**colors <- brewer.pal(9,"Paired")**

**#Plot first three PCs with colored symbols**

**par(mfrow=c(2,2))**

**plot(pcaS$x[,1:2],col=colors[factor(sample\_sites)],pch=16,cex=1.2)**

**legend("bottomleft",legend=levels(factor(sample\_sites)),col=colors,pch=16,ncol=3,cex=0.8)**

**plot(pcaS$x[,2:3],col=colors[factor(sample\_sites)],pch=16,cex=1.2)**

**plot(pcaS$x[,3:4],col=colors[factor(sample\_sites)],pch=16,cex=1.2)**

**Question 4: Do you see evidence of population structure? Is the number of components (here 4) a good representation of the number of populations?**

Now we are going to use a clustering method to examine population structure. There are many approaches, with various assumptions, and it is important to consider the underlying biology of your research organism (and your dataset size) before choosing an appropriate method. Here, we will use sparse negative matrix factorization (**SNMF**) because it is fast to compute for large datasets and it approximates the results of the well-known **STRUCTURE** algorithm. Notably, it relaxes population genetic assumptions such as Hardy-Weinberg proportions, so it may not converge on the same results as other programs. We will use SNMF to estimate the number of genetic clusters (K) among our sampled populations.

**source("http://bioconductor.org/biocLite.R")**

**biocLite("LEA")**

**library('LEA')**

**snmf2 <- snmf("stickleback.geno",K=1:8,ploidy=2,entropy=T,alpha=100,project="new")**

We will use SNMF's cross-entropy criterion to infer the best estimate of K. The lower the cross-entropy, the better our model accounts for population structure. Sometimes cross-entropy continues to decline, so we might choose K where cross entropy first decreases the most.

**par(mfrow=c(1,1))**

**plot(snmf2,col="blue4",cex=1.4,pch=19)**

The number of clusters is hard to determine, but four seems to be important and is similar to the results revealed by PCA. I will proceed assuming K=4. We will rerun SNMF using this setting.

**K=4**

**snmf = snmf("stickleback.geno", K = K, alpha = 100, project = "new")**

**# Create matrix of ancestry proportions**

**qmatrix = Q(snmf, K = K)**

**# Plot results with a barplot similar to that used to represent STRUCTURE results**

**barplot(t(qmatrix),col=brewer.pal(9,"Paired"),border=NA,space=0,xlab="Individuals",ylab="Admixture coefficients")**

**# Add population labels to the axis**

**for (i in 1:length(pops)){**

**axis(1,at=median(which(sample\_sites==pops[i])),labels=pops[i])}**

**# Lets visualize our admixture proportions on a map**

**install.packages(c('fields','mapplots'), dependencies = TRUE)**

**library('mapplots')**

**# Import geographical coordinates for the populations**

**sites <- read.csv("stickleback\_coordinates.csv",as.is=T,check.names=F,h=T)**

**# Calculate population average ancestry proportions & create an array with population coordinates**

**qpop <- matrix(NA,nrow=length(pops),ncol=K) #initialize array for ancestry proportions**

**coord.pop <- matrix(NA,nrow=length(pops),ncol=2) #intialize array for coordinates**

**index=0**

**for (i in 1:length(pops)){**

**if (i==1){**

**qpop[i,] <- apply(qmatrix[1:N[i],], 2, mean) # input pop ancestry proportions for each K cluster**

**coord.pop[i,1] <- sites[which(sites[,1]==names(N[i])),6] #input pop coordinates**

**coord.pop[i,2] <- sites[which(sites[,1]==names(N[i])),5] #input pop coordinates**

**index = index + N[i]**

**} else {**

**qpop[i,] <- apply(qmatrix[(index+1):(index+N[i]),], 2, mean)**

**coord.pop[i,1] <- sites[which(sites[,1]==names(N[i])),6]**

**coord.pop[i,2] <- sites[which(sites[,1]==names(N[i])),5]**

**index = index + N[i]**

**}**

**}**

**# Create map with pie charts depicting ancestry proportions**

**plot(coord.pop, xlab = "Longitude", ylab = "Latitude", type = "n")**

**map(database='state',add = T, col = "grey90", fill = TRUE)**

**for (i in 1:length(pops)){**

**add.pie(z=qpop[i,],x=coord.pop[i,1],y=coord.pop[i,2],labels="",col=brewer.pal(K,"Paired"),radius=0.1)**

**}**

**Question 5: How does the pattern of clustering vary in space? Is there evidence of population admixture? How do you interpret this pattern biologically?**

**Question 6: Based on the analyses done in this tutorial, where do you hypothesize the "Riverbend" ("rb") population originated from? What evidence supports your rationale?**