Using Program STRUCTURE to Determine Admixture (Gene Flow) Between Populations

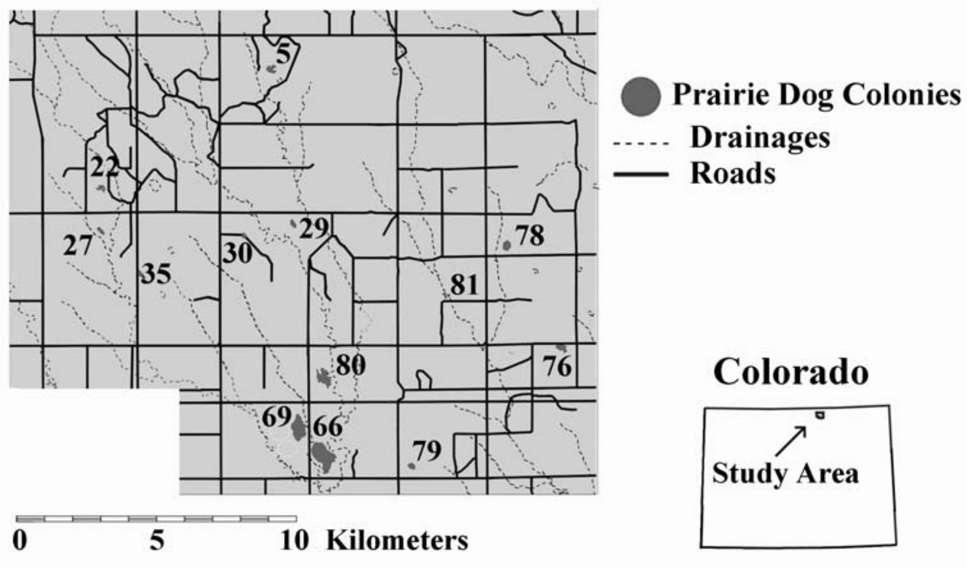
Analyses based on multi-locus genotypes are becoming increasingly common, as we develop the ability to simultaneously sample multiple regions of genomes of organisms, even for things as small as pathogens. Even if organisms (parasites) are discretely distributed among patches (hosts), do we know what makes a breeding population of those organisms (parasites)? Population geneticists obsess about this problem because it can tell us much about evolutionary potential, and the spatial extent temporal changes of populations.

Today, we will analyze data from both prairie dogs and plague, which can be quite complicated because of the lack of independence between parts of genomes (why is this particularly a problem for Eukaryotic organisms, but look at Falush et al. 2003). We will be looking at the genetic structure of a prairie dog metapopulation from the Pawnee National Grasslands, northeast of Fort Collins in north-central Colorado (see map below). We will be using the program STRUCTURE (Pritchard et al. 2000), the currently most popular program in use for genetic assignment tests. Importantly, STRUCTURE has been updated to handle the haploid genomes of pathogens, including problems associated with low levels of recombination among parts of genomes (Falush et al. 2003a, 2007).

The purpose and appeal of STRUCTURE is to infer genetic mixing within and between populations directly from the genetic data, without necessarily using prior population information that could bias our interpretations. That is to say, samples originate in space and time: organisms are captured in particular places by particular means. Yet, do we know whether samples collected form the same locality comprise a single population, or whether samples of organisms from different localities are separate populations. The statistics of assignment tests can help begin to answer this question.

1.)Population genetics in diploid organisms

To begin, you are being provided with data from Jen Roach's M.S. thesis work (Roach et al. 2001). The data are from 155 prairie dogs collected from 13 towns on the LTER/CPER/Pawnee National Grassland (see map) in 1997-1998, genotyped for seven simple-sequence repeat (SSR, also called microsatellites) markers. Overall F_{st} = 0.118 in this sample, so we should have discriminatory power (so what's this thing we call F_{st}?)



Initiate a Project in STRUCTURE

Formatting Data: Data for these analyses are in "Cynomys population study.xls". Have a look at the data in this file, as this is a format that is used by program CONVERT to format data for the various population genetic programs in everyday use, and each with a unique data format. Run the program CONVERT to create a data file usable by STRUCTURE (you'll have to export the file as a plain text file before CONVERT will like it).

Enter the same data set (now with a *.str suffix) into STRUCTURE as a project (the manual does a pretty good job of showing you how this is done). You'll need to know the following: 155 individuals, ploidy level = 2 [i.e. diploid, what would this be for the plague bacterium?], number of loci = 7, missing values = -9. Also, if you look at the data file there is a header row with locus labels, an individual label for each sample, and a column for the population of origin for each individual.

Assignment Tests and K, the Number of Inferred Populations

A.) Once you begin your project, initiate a parameter set and select the following options:

Run length: set both the burn-in period and the number of replicates after burn-in to 5,000.
Ancestry model: use population information
Allele frequency model: allele frequencies independent.

Once your parameters are set, run your analysis for K = 7, 8, 9, 10, 11, 12, and 13, and record likelihood scores for each of these runs (what hypothesis would we be testing if we simulated K = 15, or K = 20?).

- Which gives the highest likelihood (the number closest to zero)? You can compare the runs of your model by pulling down the View drop and opening Simulation Summary. Examine the columns headed LnP(D).

- Examine your diagnostic plots to see whether the data converge (how do you know the model converged?).

- How do individual assignments of individuals compare between different K-values (i.e. how does the assignment probability for each individual change with K. Pick a few individuals from the data set to compare across all of your analyses)?

- Calculate the posterior probabilities of your various K-values to see how well they fit the data (see page 14 of the STRUCTURE manual). Posterior probabilities for the likelihood for each K can be calculated from Bayes rule as follows. If you output the following data from Simulation Summary you could ask which of these has the higher probability given the data.

K	ln Pr(XjK)
1	-4356
2	-3983
3	-3982
4	-3983
5	-4006

The posterior probability for K = 2 will be:

$$\Pr(K = 2) = \frac{e^{-3983}}{e^{-4356} + e^{-3983} + e^{-3982} + e^{-3983} + e^{-4006}}$$

In reality, the probability is easier calculated if the likelihoods are normalized by subtracting the lowest value (K =3). Thus the probability becomes:

$$\Pr(K = 2) = \frac{e^{-1}}{e^{-374} + e^{-1} + e^{-0} + e^{-1} + e^{-24}} = 0.21$$

The following R code will calculate posterior probabilities for a series of runs at a number of values of K (much thanks to Jennie Lavine, Class of '09), and plot these probabilities as a function of K.. Insert your values into this code, change the range and values of K, and see what you get:

```
#read in data - this is ln(P(D)) for K = 10:13, 4 runs each
dat10<--c(2569.9, 2576.9,2576.3, 2572.0)
dat11<--c(2558.5,2575.2,2593.9,2554.3)
dat12<--c(2539.0,2533.7,2548.8 ,2548.2)
dat13<--c(2580.2,2583.9,2562.8,2569.1)

# Concatenate and give each column a name
data<-as.data.frame(cbind('10'=dat10, '11'=dat11, '12'=dat12, '13'=dat13) )

means = apply(data, 2,mean) # Mean of each K value
stand<-means-min(means) # Standardize the by subtracting off the lowest value
expstand<-exp(stand)      # Exponentiate
post_prob<-expstand/sum(expstand) # Divide by sum of all model likelihoods

#plot with a smooth line
plot(10:13,post_prob,ylab="Posterior P(K)",xlab='Number of Clusters (K)',
      pch=16,col=4)
lines(lowess(10:13,post_prob),col=4)
```

B.) Repeat the analysis, but this time select the option to "infer lambda" under allele frequency model (what *IS* lambda? Read the STRUCTURE manual!). "Configure" to infer a separate lambda for each population.

- What happens to your likelihoods for each level of K?

Now repeat the first analysis using an "admixture" model (under "Ancestry model"). What basic aspect of using Bayesian analysis are you now changing?

- What happens to your likelihoods?

- Are "populations" now unambiguously identified?

- What about assignments of individuals?

C.) Isolation by distance

Isolation by distance is a model that describes regular gene flow among a series of populations, where gene flow between close neighborhoods can be greater than between those further away. Populations may not be defined by discrete units, but could be arrayed along a continuum (along shorelines, up or down elevation gradients, along streams, etc.), and STRUCTURE may have a difficult time identifying discrete populations. Isolation by distance is measured by Mantel correlations between two pairwise distance matrices, one that measures physical distance, one that measures genetic distance between populations.

To generate Mantel correlations between genetic distance and geographic distance, first install the ecodist package into R:

```
        require(ecodist)
```

Next, you will need to enter geographic distances into R. IN this case, we have distance matrices in the files drainage.csv and linear.csv. Import these into data frames called drainage and linear (what do these look like?):

```
        linear<-read.csv("linear.csv")
        drainage<-read.csv("drainage.csv")
```

Convert the data frames into matrices, dropping the first row and column:

```
        linear.mat<-as.matrix(linear[2:13,2:13])
```

The following command will extract the lower diagonal of the matrix (you will need this for the mantel test below, what's it look like, how many values do you get?):

```
        linear.mat[lower.tri(linear.mat)]
```

STRUCTURE provides genetic distance matrices as allele-frequency divergence between populations. Find this matrix in the output of the runs with K = 13 from above, get it into Excel to make file a file genetic.csv, and import it into R and modify it to a matrix you can use.

Use the function mantel to calculate your Mantel correlations:

```
mantel(genetic.mat[lower.tri(genetic.mat)]~linear.mat[lower.tri(linear.mat)],
        nperm=100)
```

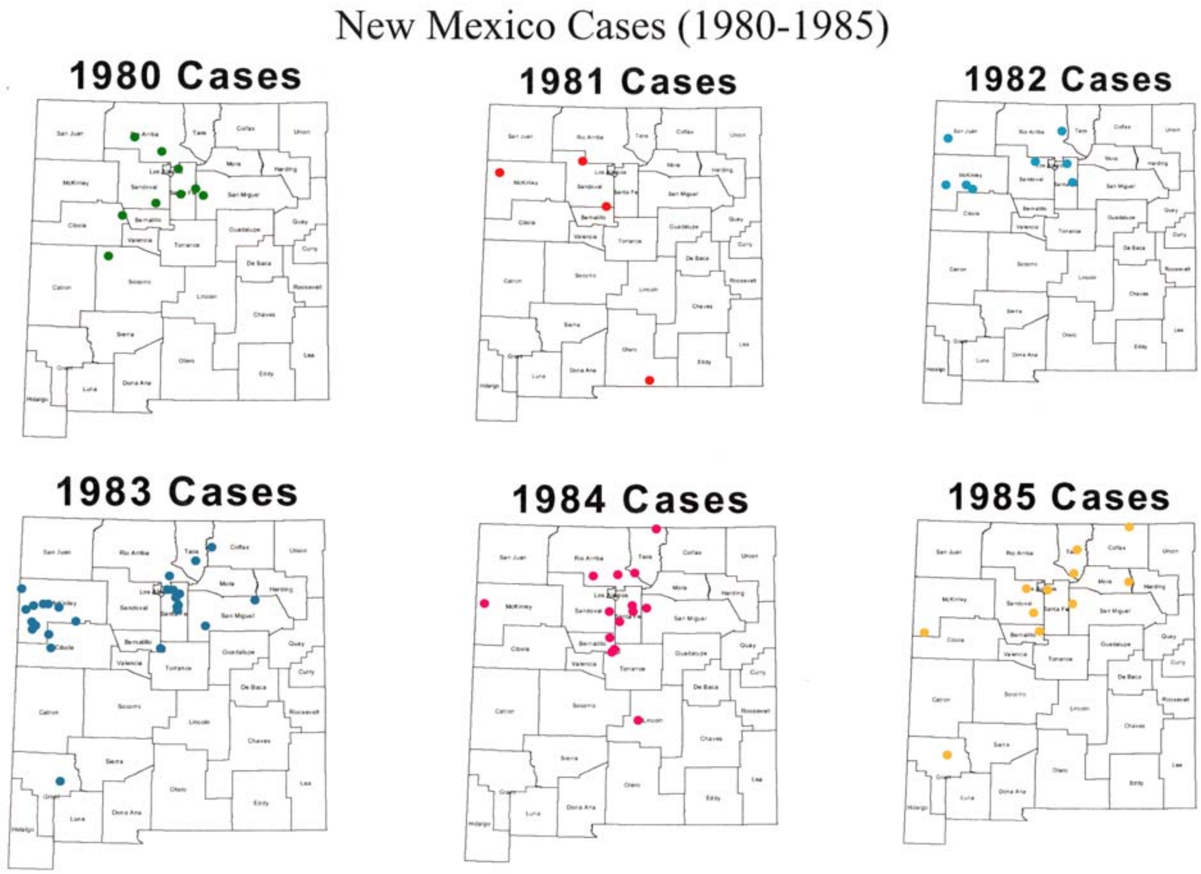
What are the correlations between linear and drainage distances, and the genetic distances (and their confidence intervals)? Which gives a better picture of isolation by distance, and how do these compare to the correlations in Roach et al. (2001)? What happens to the correlation and the confidence intervals when you increase the number of permutations?

Why did we go through all this trouble rather than just running a Pearson's correlation?

```
cor(genetic.mat[lower.tri(genetic.mat)],linear.mat[lower.tri(linear.mat)])
```

2.) **Population genetics in haploid organisms (plague)**

This data set is unpublished, based on a large number of plague isolates that were collected from infected humans, mostly around the time of the big outbreaks during El Niño years in New Mexico in the early 1980s. Additional samples come from fingerprinting study carried our more recently, where human plague isolates were matched to environmental samples: infected rodents, fleas, pets, whatever in the vicinity where exposure was thought to have occurred (Lowell et al. 2005).



In this analysis, we will be following the types of analyses carried out by Falush et al. (2003b), who examined a large number of isolates of *Helicobacter pilori*, the bacterium that causes gastric ulcers in humans. What is one of the big assumptions made in the *H. pylori* study that allowed them to carry out a STRUCTURE analysis (and was this tested)?

Enter the data set called NM.plague.str into STRUCTURE as a new project. You'll need to know the following: 104 individuals, ploidy level = 1 [i.e. haploid], number of loci = 18, missing values = -9. Also, if you look at the data file there is a header row with locus labels, an individual label for each sample, and a column for the population of origin for each individual.

In this case, we're going to explore three options:

- Vary K from 6 to 12?
- Set population ID as a prior for the analysis, or test admixtuer?
- Change correlation between loci?

We can think of other ways that we could try to test for other covariates. Which one could be really easy to implement (look at the individual labels for each isolate)?

Finally, we can have a quick look at another way to analyze these data, using parsimony in PAUP. Open PAUP and run the file NM.plague.PAU.nex. Notice the difference in how the data were formatted for this analysis. How is uncertainty determined in these analyses, and is there much evidence of structure from these analyses?

References:

Falush, D., Stephens, M., and Pritchard, J. K. (2003a). Inference of population structure: Extensions to linked loci and correlated allele frequencies. Genetics 164:1567-1587.

Falush D, Wirth T, Linz B, et al. (2003b). Traces of human migrations in Helicobacter pyl3ori populations. Science 299: 1582-1585.

Falush, D., Stephens, M., and Pritchard, J. K. (2007). Inference of population structure using multilocus genotype data: dominant markers and null alleles. Molecular Ecology Notes.

Lowell, J.L., D.M. Wagner, B. Atshabar, M. Antolin, A.J. Vogler, P. Keim, M.C. Chu, and K. L. Gage. 2005. Identifying sources of human exposure to plague. *J. Clinic. Microbiol*. 43: 650–656.

Lowell, J.L., A. Zhansarina, B. Yockey, T. Meka-Mechenko, G., B. Atshabar, L. Nekrassova, R. Tashmetov, K. Kenghebaeva, M.C. Chu, M. Kosoy, M.F. Antolin, and K.L. Gage. 2007. Phenotypic and molecular characterizations of *Yersinia pestis* isolates from Kazakhstan and adjacent regions. *Microbiology* 153: 169–177.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics 155: 945-959.

Roach, J.L., B. van Horne, P. Stapp, and M.F. Antolin. 2001. Genetic structure of a black-tailed prairie dog metapopulation. Journal of Mammalogy 82: 946-959.