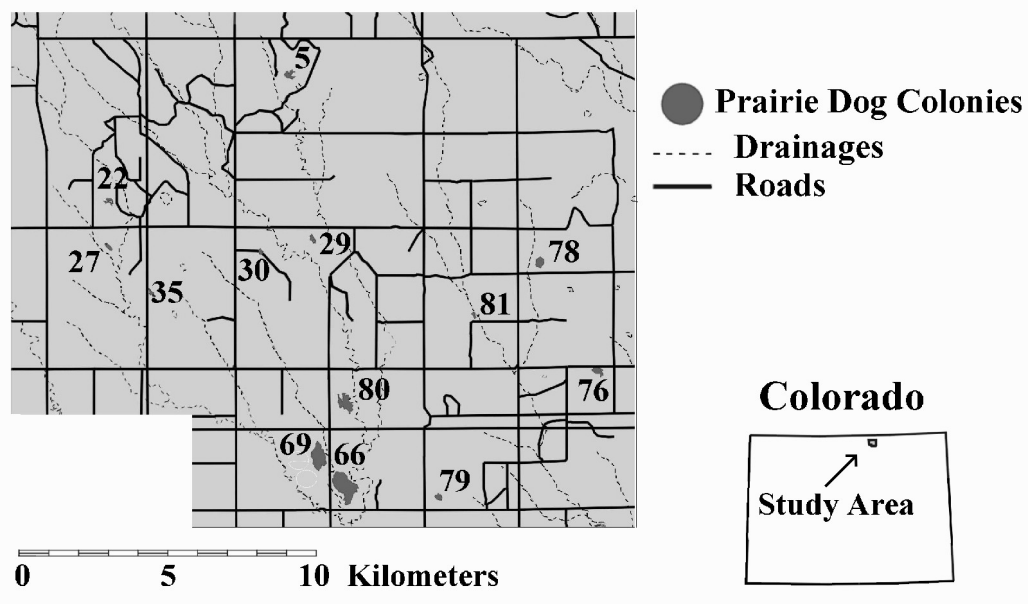
Using Program STRUCTURE to Determine Admixture (Gene Flow) Between Populations

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 of organisms, the spatial extent of populations, and temporal changes in populations.

Today, rather than analyze pathogen data, 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 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 of whether samples of organisms from different localities are separate populations. The statistics of assignment tests can help begin to answer this question.

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.



Initiate a Project in STRUCTURE

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 large number of population genetic programs in use, each with a unique data format. Run the program CONVERT to create a data file usable by STRUCTURE.

Enter the same data set 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

Once your project is begun, 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 = 10, 11, 12, and 13, and record likelihood scores for each of these runs.

- Which gives the lowest likelihood?

- 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 form he data set to compare across all of your analyses)?

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 "admixture" model (under "Ancestry model").

- What happens to your likelihoods?
- Are "populations" now unambiguously identified?
- What about assignments of individuals?

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. In this case, 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.

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, cut and paste into text files, and format them in the same manner as the files drain.txt and linear.txt.

Use the program MANTEL too examine the correlations between linear and drainage distances, and the genetic distances. Which gives a better picture of isolation by distance, and how do these compare to the correlations in Roach et al. (2001)?

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 pylori 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.

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.