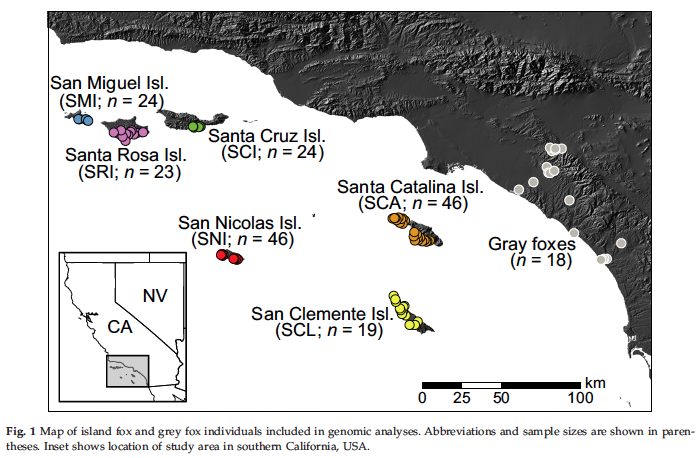
WCF, DRT, JPS

June 7, 2019

**Host population genomics lab: Island Fox**

**Description of dataset and input files**

* Island fox (*Urocyon littoralis*) from the California Channel Islands
* Sample size = 200 total before filters; 188 total after filters (**Fig. 1**)
* Data type = SNPs generated from RAD-seq
* Number of loci = 4858 after all filters
* Input files: Started with STACKS input file, then converted to other input file formats used below using combination of PGDSpider, TextWrangler or BBEdit (GREP; Regular Expressions), and R scripts (for transposing some tables)



**Population structure: Divergence among populations**

***Adegenet: PCA and DAPC***

* **Citations:**
  + Jombart T. (2008) Adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, **24**, 1403-1405.
  + Jombart T. and Ahmed I. (2011) Adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics*, **27**, 3070-3071.
* **URL:** <http://adegenet.r-forge.r-project.org/>
* **Description of method:** Adegenet is an R package dedicated to the exploratory analysis of genetic data. It implements a set of tools ranging from multivariate methods to spatial genetics and genome-wise SNP data analysis.
* **Primary assumptions:** No assumptions. Note that imputing missing values in SNP matrices uses the mean, which may artificially increase heterozygosity.
* **Step by step instructions:**

1. Open R script “adegenet.R” in R (can drag R script onto R icon)
2. Enter “command-return” to execute commands (we will discuss each command as we go)

* **Questions re: results:**

1. How many populations (i.e., K) of island foxes are there?
2. Is there any evidence of gene flow among islands?
3. Can you tell which islands are most similar to each other and most divergent from each other?

***Admixture***

* **Citations:**
  + Alexander D.H., Novembre J., and Lange K. (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, **19**, 1655-1664.
  + Alexander D.H. and Lange K. (2011) Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*, **12**, 1-6.
* **URL:** <http://software.genetics.ucla.edu/admixture/>
* **Description of method:** ADMIXTURE is a software tool for maximum likelihood estimation of individual ancestries from multilocus SNP genotype datasets. It uses the same statistical model as STRUCTURE but calculates estimates much more rapidly using a fast numerical optimization algorithm.
* **Primary assumptions:** No assumptions. Faster than MCMC algorithm in Structure program with large SNP datasets.
* **Step by step instructions:**

1. Open terminal
2. Navigate to Admixture directory with data input files (i.e., Plink .bed files)
3. Run Admixture for K=1-10

for K in 1 2 3 4 5 6 7 8 9 10; do admixture --cv fox.bed $K | tee log${K}.out; done

1. Find most supported value of K using cross validation error

grep -h CV log\*.out

1. Open R script “PlotAdmixture.R” (can drag R script onto R icon)
2. Enter “command-return” to execute commands in terminal and in R, as directed (we will discuss each command as we go)

* **Questions re: results:**

1. What is the most supported number of populations (i.e., K)? Is it the same or different from the K estimated using PCA and DAPC? Why?
2. Is there any evidence of gene flow among islands?
3. Which islands are most similar to each other and most divergent from each other?

***Genepop (F*ST*)***

* **Citation:** Rousset F (2008) GENEPOP ' 007: a complete re-implementation of the GENEPOP software for Windows and Linux. *Molecular Ecology Resources* **8**, 103-106.
* **URL:** <http://genepop.curtin.edu.au/>
* **Description of method:** Calculation of *F*ST, a measure of population differentiation between 0 and 1 (as described in Pop Gen 101 lecture)
* **Primary assumptions:** No assumptions of *F*ST. It is simply an index of population differentiation. However, using *F*ST to estimate gene flow based on Wright’s island model requires MANY assumptions.

* **Step by step instructions:**

1. Open Genepop
2. Click on Fst & other correlations
3. Allele identity (F-statistics) 🡪 For all pop pairs
4. Output format and delivery 🡪 HTML
5. Open input file in BBEdit 🡪 Command-A 🡪 Command-C
6. Input data 🡪 Paste Genepop input file in window (Command-V)
7. Submit data
8. (Wait)
9. Go to very end of output file to see pairwise *F*ST values averaged across all loci

* **Questions re: results:**

1. Are these *F*ST values low or high? Why do you think they are low or high?
2. Which island appears to be the most divergent from other islands?
3. Which island is most similar to mainland gray foxes?

**Population structure: Genetic variation within populations**

***Genepop (Heterozygosity)***

* **Citation:** Rousset F (2008) GENEPOP ' 007: a complete re-implementation of the GENEPOP software for Windows and Linux. *Molecular Ecology Resources* **8**, 103-106.
* **URL:** <http://genepop.curtin.edu.au/>
* **Description of method:** Calculation (Fis), observed heterozygosity (Ho), and expected heterozygosity (He); measures of genetic diversity (as described in Pop Gen 101 lecture)
* **Primary assumptions:** Hardy-Weinberg equilibrium assumptions for He.

* **Step by step instructions:**

1. Open Genepop
2. Click on Basic Information
3. Gene diversities & Fis 🡪 Using allele identity
4. Output format and delivery 🡪 HTML
5. Open input file in BBEdit 🡪 Command-A 🡪 Command-C
6. Input data 🡪 Paste Genepop input file in window (Command-V)
7. Submit data
8. (Wait until done transferring data…will take a couple of mins)
9. Go to very end of output file to see gene diversity averaged across all loci (“Statistics per sample over all loci…). Qintra is essentially observed heterozygosity; Qinter is expected heterozygosity (assuming HW proportions)

* **Questions re: results:**

1. Which populations are least and most variable?
2. Is Ho or He usually larger for most populations? Why? What is Fis?

***NeEstimator***

* **Citation:** Do C, Waples RS, Peel D*, et al.* (2014) NEESTIMATOR v2: re-implementation of software for the estimation of contemporary effective population size (N-e) from genetic data. *Molecular Ecology Resources* **14**, 209-214.
* **URL:** <http://www.molecularfisherieslaboratory.com.au/neestimator-software/>
* **Description of method:** Estimates Ne based on theory showing that the amount of LD at independent loci is purely a function of magnitude of genetic drift and can therefore be used to estimate Ne.
* **Primary assumptions:** Random mating, no gene flow into population, loci independent, no selection, no overlapping generations (assumptions of HW)
* **Step by step instructions:**

1. Open “NeEstimator 2x1.jar”
2. Directory 🡪 Find folder that input file is in
3. Choose file 🡪 Choose input file
4. Click on “Genepop” format
5. Methods 🡪 Linkage disequilibrium 🡪 Random mating (unclick all other Methods)
6. Run Ne Estimator (wait several minutes)
7. Find Ne estimates and CIs with minor allele frequency of zero

* **Questions re: results:**

1. Which populations have the largest and smallest Ne?
2. Would you say these estimates are precise or imprecise? Why?
3. What assumptions might not be met for island foxes? How will these affect Ne estimates?

**Testing for signature of divergent selection: High *F*ST outlier test**

***PCAdapt***

* **Citation:** Luu K, Bazin E, Blum MGB (2017) pcadapt: an R package to perform genome scans for selection based on principal component analysis. *Molecular Ecology Resources* **17**, 67-77.
* **URL:** <http://membres-timc.imag.fr/Michael.Blum/PCAdapt.html>
* **Description of method:** A PCA-based method for implementing a genome scan for detecting genes involved in local adaptation.
* **Primary assumptions:** No assumptions. In contrast to population-based approaches, PCAdapt can handle admixed individuals and does not require grouping individuals into populations.
* **Step by step instructions:**

1. Open R script “pcadapt.R” in R (can drag R script onto R icon)
2. Enter “command-return” to execute commands (we will discuss each command as we go)

* **Questions re: results:**

1. What proportion of loci show a signature of divergent selection?
2. What would you do next to confirm that theses loci are actually adaptive?