WCF, CM, DRT, JPS
July 12, 2021

# 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)

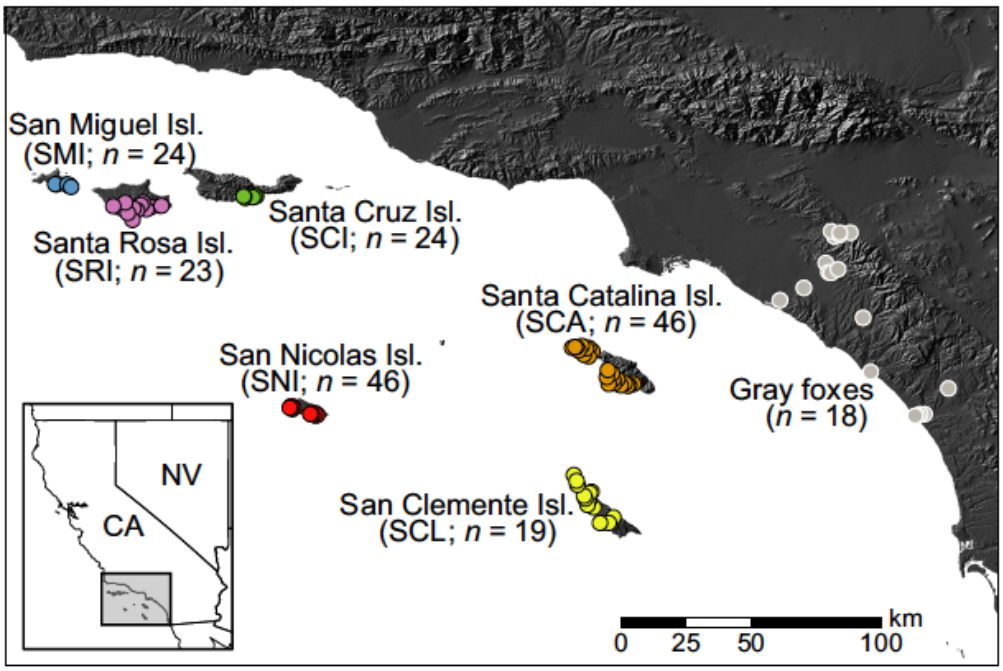


**Fig. 1** Map of island fox and grey fox individuals included in genomic analyses. Abbreviations and sample sizes are shown in parentheses. Inset shows location of study area in southern California, USA.

## Population structure: Divergence among populations

### *Adegenet: PCA*

- **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. Note that adegenet uses basic methods of imputation which are not always appropriate for SNP data. For example, we will be imputing missing values in the fox SNP matrix using the mean, which can artificially inflate heterozygosity. There are better options for imputation (the simplest being the median, or most common genotype), which we won't have time to implement today.

- **Primary assumptions:** No assumptions for PCA. However, PCA is a linear model and will summarize patterns of association among alleles as a linear relationship. When allele frequencies are spatially autocorrelated (e.g., under isolation-by-distance), distortions of the principal components space can occur, known as the "horseshoe effect". In these cases, alternative methods, such as spatial factor analysis (Frichot et al. 2012, Frontiers in Genetics) should be investigated.

- **Step by step instructions:**
  1. In RStudio, open the R script "Island_Fox_adegenet.R" (in the "/GDW_IF_Adegenet" folder).
  2. Work through the script; we'll discuss 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. 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 implements 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:** Linkage equilibrium among markers.

- **Step by step instructions:**
  1. Open terminal
  2. Navigate to "admixture_macosx_1.3.0" directory with data input files (i.e., Plink .bed files)
  3. Run Admixture for K=1-10

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

  4. The above runs Admixture sequentially as a loop using K values from 2 to 10; the command pipes the standard output of each run of K to a log file ("logK.out"). Each run of K will produce a .Q, .P, and .out file.
  5. Paste the following command into the terminal window to find the best supported value of K using cross validation error (we want to minimize the cross validation error):

```
grep -h CV log*.out
```

  6. Copy .Q files into "GDW_IF_Admixture" directory
  7. Open R script "Island_Fox_PlotAdmixture.R" (can drag R script onto R icon)
  8. Work through the R code to make barplots of the different values of K using the .Q files.

- **Questions re: results:**
  1. What is the best supported number of populations (i.e., K)? Is it the same or different from the K estimated using PCA?
  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 based on the Admixture results?

## Population structure: Genetic variation within populations

### *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 ("GDW_IF_NeEstimator")
  3. Choose file → Choose input file ("GP_by_island_top5_nonoutliers.txt")
  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?