EDEN Exchange Participant Name: Layla Freeborn
Host Lab: The Kronforst Lab, The University of Chicago
Dates of visit: February 15, 2013 - April 15, 2013
Title of Protocol: The Analysis of RAD-tag Data for Association Studies
Rationale and Background: to analyze next-generation sequence data from a polymorphic frog species to identify loci that are associated with differences in coloration
Protocol:

The Analysis of RAD-tag Data for Association Studies

Layla Freeborn[1], Wei Zhang[2], Marcus Kronforst[2], Corinne Richards-Zawacki[1]
1 - Tulane University, Ecology & Evolutionary Biology 2- University of Chicago, Ecology & Evolution

The following protocol starts by describing the steps that are taken upon receiving next-generation sequence data that are in the .fasta file format.

Quality Control of Raw Reads

1. Unzip files using gzip (gunzip) command.
2. Consider the type of quality scores by looking at the first few lines of each file. There are numerous quality scores, but Illumina sequences are most likely to return Sanger Scores.
3. Check overall quality using FastQC. This can be run to produce either zipped .html files or to open a graphical interface. U of Chicago server's do not support the FastQC graphical interface.
4. Use wget or WinSCP to transfer downloaded zipfile for the program.
5. Use scp to transfer the zipped folder to the server on which the analysis will be done. This step will create a new directory called FastQC.
6. Within the FastQC directory, alter the mode to allow execute permissions. <chmod750 fastqc>
7. In the folder containing the fastq files, FastQC will create a new set of files for each fastqc run. One file will be a .zip file. Transfer this file to windows (using the FTP of your choice).
8. Unzip the files and open the html file with a browser of your choice. Check the quality of raw reads.

Pre-processing Raw Reads with FASTX-Toolkit

Pre-processing of raw reads will be done separately for read 1 (R1) files and read 2 (R2) files, as these are usually returned from the sequencer in separate folders. There are several tools in this toolkit that can be used to pre-process files for RAD-tag analysis. The following tools were used for this analysis:

(1) FastQ/A Barcode Splitter- splits fastq files containing multiple samples
(2) fastx trimmer- trims the barcodes and/or adapters from each sequence
(3) fastq quality trimmer- filters out those sequences with reads below a specified quality

The FASTX website has a "command-line usage" section for all the available tools.

1. Download FASTX. It is available as a pre-compiled binary or can be compiled from the source.
2. *Splitting Barcodes*. Given that the RAD-tag library submitted for sequencing contained inline barcodes all the data is returned by the sequencing center in an 'undetermined' folder. The result is that sample files are not organized or readily identified. The barcode splitter command in FASTX essentially organizes files based on barcodes, allowing identification and analysis of particular samples. This tool first requires that you create a barcode file. This file can be created in notepad (Windows computers) but the file format must be converted to a Linux/Unix file type before the command can be successfully executed.
3. *Trimming Barcodes*. Prior to analysis or RAD tags it is necessary to remove from each sequence read the 6bp barcodes and 6bp Illumina adapter. An easy way to do this is to create a script with vi editor that can be used to trim the barcode/Illumina adapter from each read at the same time.
4. *Quality Filtering*. This command filters reads based on a user-specified quality. There are two main parameters to this command: (1) [-qN] = miminum quality score to keep and (2) [pN] = minimum % of passes that must have a [-q] quality. This analysis used q = 10 and p = 90 to allow for as many SNPs as possible. As with trimming barcodes, a short script can be created that allows quality filtering to be done simultaneously for all R1 or R2 files.

Calling SNPs with STACKS

STACKS is a software pipeline designed specifically for building loci from short-read sequences such as RAD-tags. Loci identified with STACKS can then be used for the purpose of genetic mapping, population genomics, or phylogeography. The STACKS website, http://creskolab.uoregon.edu/stacks/ , contains useful step-by-step guides on using components of the pipeline for any of these purposes. The output from a STACKS analysis can be loaded into a MySQL database, which can then be accessed from a web browser of your choice. This allows the data to be viewed and sorted according to user preferences.

The STACKS pipeline builds a 'stack' that consists of a series of raw reads stacked on top of each other. Next, the SNP model examines each stack one column at a time and reports any polymorphisms. Finally, each stack is read, row by row, and each different haplotype is

recorded. The pipeline illustrated below was followed for the analysis of our wild-caught polymorphic frog data (particular STACKS components are bolded and capitalized). Modifications of this pipeline are used for purposes such as genetic mapping (below) and phylogeography.

1) Identify loci and call SNPs within them. → **USTACKS**

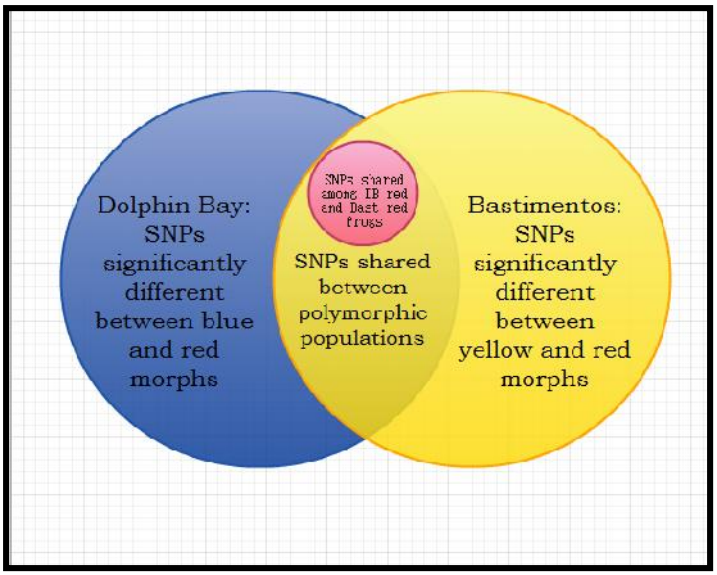
2) Create a catalog for morph type included in the RAD-library (4 total). → **CSTACKS**

3) Match all individuals against its constituent catalog. → **SSTACKS**

4) Collect within and between population statistics (e.g. F_{st}) and within and between morph statistics. → **POPULATIONS**

5) Export data from STACKS. Manually (in Linux) filter for SNPs that are significantly different for each comparison (where $F_{st} > 0.4$ and $P < 0.05$), and then identify SNPs that are shared among these.

The comparisons made in steps 4 and 5 above are summarized in the Venn diagram below.



For this particular analysis we utilized a Fisher's exact test to determine significance between morphs. This was done manually (utilizing command line) in Linux. We then identified SNPs that were shared between polymorphic populations. Of those shared SNPs, later analyses will focus on SNPs that are shared between Dolphin Bay red frogs and Bastimentos red frogs. Depending on the experimental design and the study system, meaningful comparisons between populations or groups of individuals should be considered.

The pipeline illustrated below was followed for the analysis of our lab-raised frogs with the purpose of generating mapable data (particular STACKS components are bolded and capitalized).

1) Identify loci and call SNPs within them. → **USTACKS**

2) Create a catalog for all parental sets included in the RAD-library (4 total). → **CSTACKS**

3) Match progeny against their constituent catalogs. → **SSTACKS**

4) Call genotypes. → **GENOTYPES**

5) Make manual corrections to genotypes in the STACKS web interface.

6) Export sets of genotypes for use by JoinMap.

Generating Linkage Maps with JoinMap 3.0[1]

1. Open JoinMap 3.0, making sure you have the necessary administrator access (right click on the JoinMap icon and select *run as administrator*).
2. Sets of genotypes are exported from STACKS in the form of Microsoft Excel datasheets. Attempting to import this into JoinMap will result in error messages. Data must first be converted in Excel into a tab-delimited file.

[1] Van Ooijen, J.W. & R.E. Voorrips, 2001. JoinMap® 3, Software for the calculation of genetic linkage maps in experimental populations. Kyazma B.V., Wageningen, Netherlands

3. Create a translation file following the specific instructions in the JoinMap 3.0 manual. This file will define for JoinMap the name of your population, the population cross type (e.g. F2, CP, DH, etc.), the number of loci, and the number of individuals.
4. Data can then be translated into the correct JoinMap file format, a locus genotype file (loc-file), by File → Prepare Data. Use the *Browse* box to import the excel sheet into the *File to prepare:* field. Name your .loc file in the *Output .loc file* field. Import your translation file into the *Using map file:* field.
5. Create a new project (File → New Project…). Give your project a descriptive name and specify the directory to which it will be saved.
6. Load your data into JoinMap. File →Load Data. You will see a summary of the data that you imported. Check to make sure that the information is correct.
7. You can alter the linkage map calculation options according to your experimental design. Options → Calculation Options…
8. Moving through the tabs in JoinMap will take you through the construction of the linkage maps. Press the calculate button when a blank screen is encountered. Based on these results, individuals and/or loci can be excluded from the calculation by returning to the *Loci* or *Individuals* tabs. For example, when high values of similarity are listed in the Similarity column of *Similarity of loci* or *Similarity of individuals* tabs, you may choose to exclude those individuals or loci from the analysis by checking the box next to them in the list.
9. In the *LOD groupings (tree)* tab, pressing the calculate icon will result in a list of linkage groups in 'tree' form. Right click on each linkage group that you want to be graphically mapped. Select the mapping icon to generate a linkage map.
10. Save or print the linkage map.