ISSJ\_Analysis\_Project\_Outline

**Bioinformatics**

**001-Process\_RadTags**

Description: De-multiplex raw reads from sequencer by individual barcodes

Associated Data: Raw reads, barcode file. Output processed fastq.gz

**002-Align\_BWA**

Description: BWA is a software package for mapping low-divergent sequences against a large reference genome, BWA-MEM, which is the latest, is generally recommended for high-quality queries as it is faster and more accurate and has better performance for 70-100bp Illumina reads. Use BWA-mem to align reads to FLSJ reference genome

Associated Data: processed reads. Output bam files

**003-ref\_map**

Description: Program will execute Stacks pipeline to genotype and call SNPs using the population-wide data per locus. Makes use of the population map to specify all the samples in an analysis and determine which groupings to use for calculating summary statistics. If no population map is supplied the populations program computes these statistics with all samples as a single group.

Associated Data: aligned bam files. Output stacks files

**004-popualtions**

Description: Stacks, through the populations program, is able to export data for a wide variety of downstream analysis programs. Will use to export to VCF format for filtering

Associated Data: ref\_mapped stacks files. Output vcf

**005-VCFtools**

Description: Export data for 012 format for geoscapeRtools filtering tests

Associated data: vcf file. Output 012 file.

**006-genoscape\_R\_tools**

Description: Visualize the amount of missing data in the SNP matrix to investigate where a good cutoff might be for filtering to maximize SNPs and individuals in dataset.

Associated Data: 012 Files from vcf tools. Output: test filtering regimes

**007-filtering in plink for final SNP dataset**

Description: Plink open-source whole genome association analysis toolset, designed to perform a range of basic, large-scale analyses in a computationally efficient manner. Based on results from GenoscaprRTools, will filter for missingness and quality for final SNP matrix.

Associated Data: 012 Files from vcf tools. Output: final SNP dataset in structure, plink, and vcf format

**008-populations\_stats**

Description:Import data from Plink to calculate population wide summary statistics such as heterozygosity, π, and FIS.

Associated Data: Files from Plink.

Results: population summary statistics.

**009-WHOA**

Description:Custom tool from E. Anderson to investigate the distribution of genotypes in GBS data where they expect (by and large) Hardy-Weinberg equilibrium, in order to assess heterozygote miscall rates (rates at which true heterozygotes are incorrectly called as homozygotes) prevalent in some RAD-seq data sets.

Associated Data: vcfR object. You can make such an object yourself by reading in a VCF file using vcfR::read.vcfR()

Results: Better sense of what the miscall rate could be in the dataset and how much that may limit inference in downstream analyses. Latest version of STACKs (v. 2.53) should be better than V.1 at this miscall rate (according the E.A.) but we will see.

**Analyses**

**Section 1- Neutral Population Genomics**

**100- Filter to just the neutral SNPs by removing outliers**

**101- Mantel Test**

Description: Finds correlations between (dis)similarity matrices, in this case genetic dissimilarity and geographic dissimilarity to determine whether spatial genetic structure can be attributed to IBD.

Associated data: Filtered SNPs dataset that only includes the “neutral” variants

Results: Result indicating presence or absence of IBD

**102- MEMGENE**

Description: Use multivariate regression approach to extract the spatially relevant genetic variation using Moran Eigenvector Maps proportion of shared alleles between individuals to test for neutral population genetic structure across the island and between habitats.

Associated data: Filtered SNPs dataset that only includes the “neutral” variants

Results: Detecting weak spatial genetic patterns

**103- sPCA**

Description: sPCAs utilizes a similar multivariate approach as MEMGENE, but instead uses a reduced number of orthogonal axes to maximize variation between allele frequencies and spatial autocorrelation in allele frequencies using Moran’s I to visualize genetic variation across geographic space.

Associated data: Filtered SNPs dataset that only includes the “neutral” variants, sampling data for each individual and raster file for SCI.

Results: Map illustrating population genetic structure across the island similar to Langin et al. IBD figure

**104- Clustering algorithms**

Description: Admixture in the R package LEA which utilizes a least-squares estimate of ancestry, and the Baysian approach of STRUCTURE (separate program with GUI interface so no code)

Associated data: Neutral SNP dataset

Figures & Results: Admixture plots with tests of K from 2-6

**105- MLPE**

Description: Estimate the impact of landscape features on neutral population structure using maximum likelihood of population effects with individual based genetic distances as the response, and environment distances that we have strong *a priori* hypotheses for regarding population structure in this system as the explanatory variables

Associated data: environmental data of each sampling point including, elevation, mean temperature of warmest month, and difference of habitat composition, random effect matrix of individual comparisons to control for non-independence, individual based genetic distances (proportion of shared alleles and relatedness)

Results: Global model will be ranked using the Bayesian information criterion (BIC) to infer how gene flow is affected by landscape features, and which variables have the greatest impact on neutral genetic structure

**Section 2 – Identification of Population Structure Associated with Habitat**

**201-RDA\_habitat**

Description: Partial redundancy analysis implemented with the LEA R package to test multiple loci simultaneously for genetic associations with habitat type (pine vs. oak) while controlling for geographic distance using constrained orthogonal axes.

Associated data: Pine & oak proportion data with approximate grouping data for the jays (western pine vs. western oak ect, based on which pine stand individuals were sampled closest to). SNP matrix with all SNPs (neutral and outlier)

Figures & Results: RDA plots showing genotypes and individual SNPs, evidence of selection between habitat types.

**202- Variance\_Partitioning**

Description: Variance partitioning of the partial RDA to test how much of the explainable genetic variance may be uniquely explained by habitat independent of geographic distance and disentangle the effects of isolation by distance and isolation by environment.

Associated data: RDA Results?

Figures & Results: Ven diagram of variance results with genotypic variation explained by geographic distance, habitat and “other”

**Section 3- Identification of loci Underlying Variation in Morphology**

**301-GWAS\_RDA**

Description: Identify loci associated with bill morphology using partial RDA to correct for population structure related to geographic distance and flag outlier loci using a standard deviation cutoff of 2.0 and 2.5 to find SNPs under moderate and strong selection.

Associated data: Bill morphology data, in this case body size corrected measures for bill length. Complete SNP matrix.

Figures & Results: Multiple loci flagged across the genome as under selection

**302- GWAS\_MRMLM**

Description: Multivariate GWAS that incorporates kinship, can be implemented in older versions of R but currently not working

Associated data: Bill morphology data, in this case body size corrected measures for bill length. Complete SNP matrix.

Results: Multiple loci flagged across the genome as under selection

**303- GWAS\_GEMMA**

Description: Multivariate GWAS that incorporates kinship, implemented in the shell

Associated data: Bill morphology data, in this case body size corrected measures for bill length. Complete SNP matrix.

Figures & Results: Multiple loci flagged across the genome as under selection

**304- Merge\_GWAS**

Description: Merge the SNPs flagged by all GWAS tests in R using dplyr to identify the SNPs flagged by multiple tests.

Associated Data: SNPs lists from MRMLM, GEMMA, and RDA

Figures & Results: determine which SNPs occur in RDA and both GWAS programs that relate to environment and bill morphology. Manhattan plot illustrating where these SNPs are on the Genome using chromosome coordinates from N. Chen

**305-BLAST\_scripts**

Description: Determine whether there are any genes within same linkage group of each SNP to test for associations with genes known to affect bill morphology (See Bill Morph Genes table) by correlating loci and surrounding regions to the annotated Zebra Finch and Hooded Crow genomes using the basic alignment search tool (Blast) from NCBI to query the sequences.

Associated Data: list of shared candidate SNPs from MRMLM, GEMMA, and RDA. Annotated genomes of Zebra Finch and Hooded Crow.

Figures & Results: list of loci flagged by GWAS and aligned genomic regions (GO Table)