Info Update: Detecting Local Adaptation from Population Genomic Outlier Analysis

Outline:

1. Local adaptation

2. Different approaches

- QTL Mapping, GWAS, sweeps, GxE association, differentiation outliers

3. Common obstacles

- Population structure/demography

- Missing genomic information

- Missing environmental information

- Statistical challenges

4. Solutions

5. Sampling design considerations

6. Final notes

Draw reduced representation figure too

**Local adaptation**: Organisms in a given environment have higher average fitness in their environment than organisms from other environments. This is due to spatial variation in selection.

* Selection for traits that are beneficial locally will lead to divergence of trait means and allele frequencies among populations
  + Creates the potential for identifying genomic regions involved in local adaptation
* Identifying the genomic regions involved in local adaptation is useful in understanding primary evolutionary questions
  + Natural selection act on standing or new variation
  + identifying patterns of parallel evolution
  + Reveal process of speciation
  + Improve management decisions

**Methods**: QTL mapping, GWAS, Population specific selective sweeps, Genetic environment association analyses, and Differentiation outlier method

Will focus on the last two

Genetic-environment association analyses

* look to ID alleles whose frequencies have unusually strong correlations with environmental variables (sea stars difference in disease prevalence among locations, may lead to differences in allele frequencies)

Differentiation outlier method

* Screen for alleles that show high differentiation among populations (**Fst**, XtX, DD, **pi**, G, SPA)
* Maybe the case for researchers who do not have a priori knowledge of important environmental variables

**Challenges**:

* Population structure and demography can lead to incorrect inference
* Missing genomic information arising from the genotyping approach or quality control assumptions
* Inappropriate or missing environmental information
* Statistical challenges

Demographic History & Population Structure: (range expansions/allele surfing, hybridization/introgression, Northward expansion, distance limited dispersal)

* can lead to spatial genetic differentiation without selection taking place
  + falsely identify loci as being under selection

Locus Specific Effects:

* Neutral sites near loci experiencing purifying selection will have greater variance in their contributions to future generations
  + stronger differentiation among populations for loci linked to sites experiencing purifying selection
  + coding regions may have higher Fst than intergenic areas due to the purifying selection of deleterious alleles
* ~~Uniform positive selection~~
  + ~~different beneficial alleles arising to solve the same problem, with limited migration~~
  + ~~drift brings about the new mutation and without migration you have differentiation among populations~~

Missing Genomic Data

* Genotyping approaches: chip-seq and RAD-seq are inadequate at identifying loci under selection
* Structural variants not usually included in reference (inversions, CNV, transposons)
* Repetitive regions and paralogs

Missing Landscape

* Low-resolution environmental data may reduce the accuracy of results even if the selective environment is known
* Resolution needs to be fine scale enough to characterize each local population well
* Measurements on a fine time scale are also important
* Multicolinearity

**Solutions**:

Controlling for Demographic History

* + Null Models: based on the inferred demographic history of the species
    - ~~High false positive rate if assumed demography is different than actual demography~~
    - ~~explicit demographic model, using simulations to generate a null distribution against which the observed data can be compared~~
      * Must be true to species’ actual demographic history
  + Relatedness: among sampled populations to correct for neutral population structure in the data
    - more flexible and less biased, so long as you have sufficient coverage of the genome
    - ~~Covariance in allele frequencies among populations, treelike population history, mixed models~~
    - ~~Covariance in allele frequencies among populations~~
      * ~~Can use simulations based on covariance matrix to generate a null distribution~~
    - ~~Population tree~~
      * ~~uses SNP data to estimate a tree of genetic divergence among populations~~
      * ~~simulates data based on these trees to generate null distribution~~
    - ~~Linear model~~
      * ~~captures patterns of neutral population structure by incorporating relatedness, principal components, and other factors in a model.~~

Missing genomic data

* + exome capture and RNA-seq are more informative because they focus on coding regions (more likely to be functional), but they still don’t capture the variation involved in gene regulation
  + Whole genome resequencing is best
  + sequence samples to reasonable read depth
  + Develop reference genome

Missing Landscape Data

* Knowledge of natural history

**Additional Important Notes**:

* Sampling strategy
  + the geographic scale at which local adaptation occurs and the geographic coverage of the study must be taken into consideration
  + This information can be used to optimize sampling
  + collecting at replicated paired locations where there is strong or suspected differentiation
  + For methods that require local population allele frequencies, sampling at least 10 individuals/location is important for accurate estimates
* Multiple Comparisons
  + same test applied to thousands of loci within each data set
  + False discovery rate correction based on your observed distribution of p values
  + FDR selection is important
* Sliding Window Scans
  + this may alleviate multiple comparisons issues
  + calculate summary statistics across a set of adjacent markers for defined windows across the genome
  + Sliding window size is also important (what is it based on? size? LD)
* Genetic Architecture
  + genome scans work best with simple genetic architectures
    - few genes of large affect, same adaptive alleles across populations
  + accuracy may deteriorate with instances of redundant architecture, epistasis, and pleiotropy

**Closing Notes**:

* genome scans are a good starting point, but other analyses and experiments are also useful
  + local reductions in heterozygosity
  + functional validation