

Project outline for EE0B 563
Audrey McCombs

Overview

Grand Teton National Park provides habitat for a high-elevation butterfly species *Parnassius clodius*. This species lives in meadows throughout the park, where each meadow is more-or-less isolated from other meadows by distance, landscape resistance (e.g., forested areas that the butterflies don't like to fly through), and phenology (large elevation gradients mean their nectar plants flower at different times through the season). During the 2016 field season we took genetic samples of ~220 animals from 30 meadows, and created a SNPs dataset of the genetic information. Post-processing produced 3 datasets: 1) unfiltered, not imputed (218 samples, 42,215 SNPs); 2) filtered, not imputed (217 samples, 18,991 SNPs); and 3) unfiltered, imputed (146 samples, 1001 SNPs).

Our question of interest is: How are the different meadow populations related to each other? One hypothesis is that the entire population in the Tetons is panmictic, with gene flow unimpeded across all meadow populations. A second hypothesis is that gene flow occurs among some groups of populations and not others, creating genetic clusters in the park. A third hypothesis is that most of the population interbreeds, but some remote sites might be genetically isolated from the larger population.

Several years ago I conducted a genetic admixture analysis of the data using LEA and adegenet in R. That analysis suggested 3 ancestral populations, and produced admixture estimates for each of the sampled meadows - see the map on the next page.

The goal of my project in this class will be to build a phylogenetic tree from the SNPs data, to examine phylogenetically the relatedness of the park's meadow populations.

Specifically, I'll be asking the following questions:

- 1) What does the phylogenetic tree tell us about how the meadow populations are related?
- 2) How does the phylogenetic analysis differ from the genetic admixture analysis, and how are they similar? What can the similarities and differences tell us about the relatedness of the populations?
- 3) Do we get different results using the three different datasets mentioned above?

Proposed methods

I propose to conduct the phylogenetic analysis using the BioConductor suite of packages in R, especially the SNPRelate package.

GitHub repository

I've set up a Github repository for this project at:

<https://github.com/almccombs/MolyPhyloProject>

