DISCOVAR paper plan outline

Evolutionary dynamics across *Heliconius*

Part 1: the data

DISCOVAR de novo uses a single library of 450bp genomic fragments, and sequences them with 250bp illumine paired end reads. There is no true scaffolding (i.e. no mate-pairs, long-jumps, pacbio, etc), but instead employs contigging of very high copy number, high fidelity reads.

Evaluating assemblies:

* of contigs to Hmel2 mapping stats
* general genome stats: length, N50, contig size distribution, etc
* BUSCO/CEGMA output

Part 2: multiple alignment and broad scale phylogeny

With the progressiveCactus tool, we can use both the high-quality Hmel2 and Herato genomes, as well as phylogenetic information, to align DISCOVAR genomes together.

* Resolve the base of the phylogeny – the only family-wide genomic phylogeny was done using only 21 genes, and we would have much greater power with whole genome sequencing
* Create “pseudo-reference” genomes – if we can demonstrate that the DISCOVAR genomes are largely complete, and can assign them to chromosomes, they would be a good resource for population genetics. Should reduce reference mapping biases if we have more appropriate genomes to which we can map resequencing data from non-melpomene/erato individuals
  + Demonstrate this somehow…resequence data for pardalinus?

Part 3: Selection

With so many genomes, we’ll be able to get a sense of what is conserved throughout heliconiini in general and specifically heliconius. We can also see in a broad sense how each lineage has diverged, and search for regions that have been under recent strong selection.

1. conservation
   1. what percent of genome is conserved across entire group?
   2. Across all heliconius?
   3. How do patterns of conservation vary?
      1. Coding vs non-coding
2. selection modeling
   1. within coding regions dN/dS
      1. which areas are consistently under positive selection across the phylogeny? What proportion of elevated dN/dS regions are species-specific?
   2. background selection
      1. polymorphism around conserved sites
      2. Use recombination map from John Davy
   3. sweeps
      1. polymorphism, especially around sites with elevated dN/dS or changes in otherwise conserved regions
      2. interspecific vs intraspecific sweeps?

Part 4: fine-scale phylogeny and introgression

It’s possible that at a finer scale, the genomic consensus won’t be concordant with the “historical” evolutionary record of speciation. We know there has been a great deal of hybridization and introgression in many of the species, and that might drown out the “true species tree”. By looking at phylogenetic relationships across the genome, seeing how they change, and seeing how old each of the trees are, we might be able to gain insight into the history of divergence and introgression through time.

* use SOM/HMM models?
* Unbiased way to find most common trees
* How old are the trees?
* How much of the genome does each tree represent?
* How large are the genomic blocks that represent each tree type?