Mol/Alb/Mic Project Outline

Kevin Murray

2016-07-20

This project entails examining the population structure of the *E. moluccana*, *E. albens*, *E. microcarpa* (Mol/Alb/Mic) complex, and possible further landscape genomics analysis of this complex.

Population genomics of the Mol/Alb/Mic complex

Using a subset of the data generated for Project 1 of the Discovery grant (i.e. Jasmine's data) I aim to determine:

1. The broad, genome-wide divergence between these species.
2. The distribution of windowed gene trees between the three species.
3. Outliers of genetic differentiation between these species.
4. (If no-one else has done or is about to do so) the approximate chloroplastic tree of the 10 species

Genome-wide divergence

Here I aim to determine the overall genetic relationships between these species. I will aim to use several analyses to answer, with increasingly improved certainty, the patterns of inter- and intra-specific relatedness of the Mol/Alb/Mic complex.

A rough initial estimate of whole-genome divergence can be calculated using `kWIP`. The relative relatedness of individuals within and between species can be determined, as can the relative relatedness of species. These analyses can be performed directly on sequencing data, hence will be the first analyses performed and will guide further analyses. The usual quality control checks of replicate detection and species ID can be addressed using `kWIP` too.

I will then use alignment to the *E. grandis* reference to compute a set of ordered SNP markers across the whole genome. I propose doing this using a standard pipeline I have developed in my work at the GDU/ABC which uses `bwa mem`, `samtools mpileup`, and `freebayes` for read mapping and variant calling. This pipeline could (and should) be refined for use in *Eucalyptus*, and is directly applicable to the broader project.

The divergence of all three species from *E. grandis* may pose a problem when aligning to this reference, especially with reference bias. Rob Lanfear and Reed Cartwright have used and iterative approach of mapping, updating the reference with SNPs fixed in all samples, and remapping. They have shown that this approach increases the read mapping rate and quality, which should reduce mapping bias. I'm in the process of doing exactly this with the *E. melliodora* mosaic tree data to create a reference for *melliodora* GBS. Again this could be used across all 10 species.

Once we have trustworthy SNP data, several analyses would follow. Species trees could be estimated using SNAPP or similar (though this is likely going to be done as part of the larger analysis, right?). I will calculate genome-wide F_{ST} (or whatever the most appropriate derivative is) along with appropriate other measures of genome-wide differentiation. Various analyses based on eigendecomposition (i.e PCA and friends) will be performed (particularly I'd like to examine Gil McVean's work on genealogical interpretations of PCA[1]). As an aside, Alistair Miles has a phenomenal blog post on PCA for SNP data here.

Divergence outlier analysis

In this section I am to find specific loci that do not follow genome-wide patterns of genetic divergence.

References

1. McVean, G. A Genealogical Interpretation of Principal Components Analysis. *PLOS Genet* **5,** e1000686 (2009).