Population Genetics Project #5

González-Martinez et al. (*Genetics* 172:1915-1926; 2006) investigated patterns of nucleotide sequence variation in 18 genes that were candidates for drought-stress responses in loblolly pine (*Pinus taeda*). On the course website you will find two data sets for each of four loci included in their analysis:

- *ccoaomt-1*: caffeoyl-CoA-O-methyltransferase 1
- *cpk3*: calcium-dependent protein kinase
- *erd3*: early response to drought 3
- *pp2c*: protein phosphatase 2C-like protein

One data set for each locus, the `Pinus-taeda-<gene name>-coding.fasta` data set, includes only the coding portion of the nucleotide sequence as downloaded from Genbank. The other data set, the `Pinus-taeda-<gene name>.fasta` data set, includes the complete nucleotide sequence as downloaded from Genbank. Each data set contains 32 sequences that were aligned using `Muscle` at `http://www.ebi.ac.uk/Tools/msa/`. The following table shows the number of nucleotides included in each data set.

	Nucleotide sequence length	
Locus	Coding	Complete
ccoaomt-1	258	517
cpk3	378	630
erd3	625	882
pp2c	461	638

Using these data, answer the following questions:

1. Is there evidence for selection, a recent population expansion, or a recent population bottleneck at any locus when the complete sequence is considered?
2. Is there evidence for selection, a recent population expansion, or a recent population bottleneck at any locus when only the coding sequence is considered?

3. Assume that there has not been a recent population expansion, a recent bottleneck, or a recent change in range size. What kind of selection might account for the patterns revealed in your answers? Are the patterns of selection you detect consistent with these loci being adaptively important in drought responses?

4. González-Martinez et al. present evidence from 21 highly polymorphic microsatellite markers indicating that there is not significant population structure within the sample and that is no evidence for demographic processes like range expansion. How strong is their evidence for these conclusions?

Hints

- Use `strataG` to analyze these data. To read data from a FASTA file in `strataG` and to estimate Tajima's D and determine if it is significantly different from 0, just do the following:

```
library(strataG)

seqs <- read.fasta("filename.fasta")
tajimasD(seqs)
```

  If you are also inerested in getting an estimate of nucleotide diversity, simply

```
nucleotideDiversity(seqs)
```

- If a locus is adaptively important in drought response, we might expect it to reflect the effects of directiional selection rather than balancing selection. Presumably there would be one allele that provides the greatest protection from drought and is adaptively favored.

- Answering Question #4 will require you to stretch a bit. Think about what `Structure` can and can't tell you. Think about what, if anything, departures from Hardy-Weinberg within populations tell you anything about recent changes in population or range size.

  Don't worry about the details of the Ewens-Watterson test. All you need to know is that it compares the observed number of alleles in a sample with the number expected under neutrality given the observed level of heterozygosity. Like Tajima's D it depends on the assumption that the population size is stable. The Ewens-Watterson test may

reject neutrality either because the allele frequency distribution is "too even" or because one allele is "too common." Think about what, if anything, it tells us about recent changes in population or range size.