Lab 3 Example

1 Fst outliers

Let's take a visual approach to detecting Fst outliers (non-rigorous)

- Download the `repens_150.stru` file.
- This contains a subset of 187 individuals and 150 snp loci
- Read `repens_150.stru` into R using the `adegenet` package and `read.structure()` function, save it to object called `repens`
  - Use `str()` to look at `repens` object
- Call the `pegas` library and use the code `Fst(as.loci(repens))` to look at locus-by-locus Fst values, and save the output as a `data.frame`
- Make a histogram of the Fst values. You may want to increase the number of breaks.
- Do there appear to be any outliers? What do these mean?

2 Run Structure

- Call the `strataG` library and convert and save `repens` to a gtypes object using `genind2gtypes`
- Use the `runStructure()` function to run Structure through `strataG`
  - num.k.rep = 3
  - k = 2:19 (will take some time)
  - burnin = 2000, numreps = 2000
- Make plots using `evanno()`
- Choose a few values of `K` and make bar plots using `structurePlot()`
- How does your choice of `K` compare to those of Prunier et al.? To Project #2?

3 Calculate Tajima's D on sequence data

Now let's calculate Tajima's D on a small sample of *Protea repens* sequence data
Each sequence contains up to 274,405 bp

- Download the `repens.fasta` data file and read it into R using `read.fasta`, saving it as the object `repens.fasta`, then convert and save it as a matrix

- Esimate the average per nucleotide diversity $\hat{\theta}_\pi$ using:
  ```
  pws.diff <- dist.dna(repens.fasta, model = "N", pairwise.deletion = TRUE,
  as.matrix = TRUE)
  pi <- mean(pws.diff[lower.tri(pws.diff)])
  ```

- Estimate the number of segregating sites `k` using:
  ```
  S <- ncol(variableSites(repens.fasta)$sites)
  ```

- Now estimate $\hat{\theta}_k$
  ```
  n <- nrow(repens.fasta)
  n.vec <- 1:(n-1)
  a1 <- sum(1/n.vec)
  ```

- Finally, look at the difference between $\hat{\theta}_\pi$ and $\hat{\theta}_k$
  ```
  pi - (S/a1)
  ```

- What do you get? What a crazy number!

- There are other pieces going on here that need to be corrected for. Use the function `tajimasD()` in `strataG` to calculate Tajima's D on `repens.fasta`. Look at the `tajimasD()` source code for details!

- What value do you get? Is it significant? What does this value imply about the evolutionary history of these samples?