EXERCISE 2: Scanning for selective sweeps across the Pel locus in benthic-limnetic species pairs.

A. Use SweeD (Sweep Detector, Pavlidis & Alachiotis) to perform Composite Likelihood Ratio Tests to detect selective sweeps based on the Site Frequency Spectrum of SNPs (following the methods of Nielsen et al 2005).

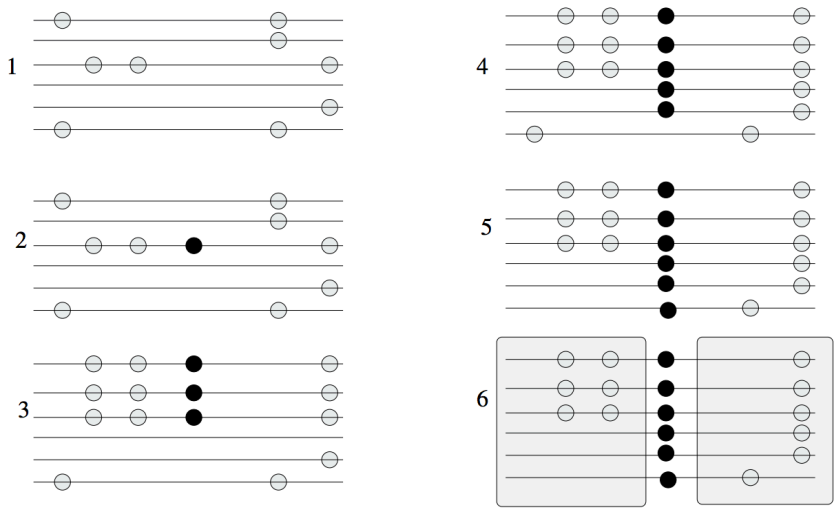


Figure 2.1: SFS patterns generated by a selective sweep. 1. Neutral mutations (light circles) are present in the population. 2. A beneficial mutation (black circle) appears in the population. 3. The frequency of the chromosome that carries the beneficial mutation increases. 4. Due to recombination (between chromosomes 5 and 6) neutral mutations that were previously on a neutral chromosome are located now on a beneficial chromosome. 5. Recombination occurs between chromosomes 5 and 6 and brings other neutral mutations on the beneficial chromosome. 6. The gray square denotes the region where the SFS has been shifted to low- and high-frequency dervied variants.

from Pavlidis & Alachiotis Sweed manual.

Here: http://sco.h-its.org/exelixis/resource/download/software/sweed3.0_manual.pdf

Sweep Detector (SweeD) calculates the site frequency spectrum in a grid (ie windows) across your sequence genomic region of interest. We'll be using input files in the "SweepFinder" format.

```
Example Input:

position x       n       folded
67      12       12      0
186     4        10      0
374     8        12      0
495     4        12      0
582     4        12      0
…

where
position=chromosome position
x=number of sequences carrying the derived allele
n=number of valid sequences
folded=0 the spectrum has been polarised using the ancestral allele.

So at the first snp (position 67) all 6 individuals are homozygous for the derived allele (12 derived
alleles out of 12 total), and the next position has only 4 derived alleles.
```

RE: ANCESTRAL ALLELE

The ancestral allele can be estimated by alignment to other outgroups or alternatively using a known ancestor. For today's exercise, I have used marine sticklebacks to calculate the ancestral allele: the most frequent allele in 6 whole

genome sequenced marine fish. This will obviously be wrong for parts of the genome subject to recent sweeps from de novo mutations in marine fish. However, since the marine ecotype is morphologically unchanged from its fossilized ancestors, an argument can be made that it is unlikely that recent sweeps are pervasive in the genome.

```
8 sweed input files are ready to use:

        1)  ancestral allele was calculated
                        - only biallelic snps were considered
                        - the most frequent allele in marine was identified and called 'ancestral'.
        2) The ancestral allele was coded into the INFO/AA tag of the vcf file using samtools perl
        script 'fill-aa';
        3) For each Lake ecotype, SNPs were quality filtered using bcftools filter '%QUAL>=30 &&
        MQ0F<0.5 && AN>=10'
        4) A perl script was used to turn this vcf file into sweep finder format.
        5) The resulting files are ready for sweed analysis:
        ~/wpsg_2016/activities/sweed/Lake1_Ben.sweed.input
        ~/wpsg_2016/activities/sweed/Lake1_Lim.sweed.input
        ~/wpsg_2016/activities/sweed/Lake2_Ben.sweed.input
        ~/wpsg_2016/activities/sweed/Lake2_Lim.sweed.input
        ~/wpsg_2016/activities/sweed/Lake3_Ben.sweed.input
        ~/wpsg_2016/activities/sweed/Lake3_Lim.sweed.input
        ~/wpsg_2016/activities/sweed/Lake4_Ben.sweed.input
        ~/wpsg_2016/activities/sweed/Lake4_Lim.sweed.input
```

A. SWEED ANALYSIS (perform for each Lake and ecotype):

```
SweeD -name Lake1Ben.SweeDRun1 -input Lake1_Ben.sweed.input -grid 250
```

where:
-name is the run name you would like to assign to this particular analysis run.
-input is your inputfile
-grid is the number of blocks/windows you would like to break up your sequence into. 250 windows across at 377kb region is ~ 1.5kb window size. You may choose to make the windows larger (grid number smaller).

SWEED OUTPUT:

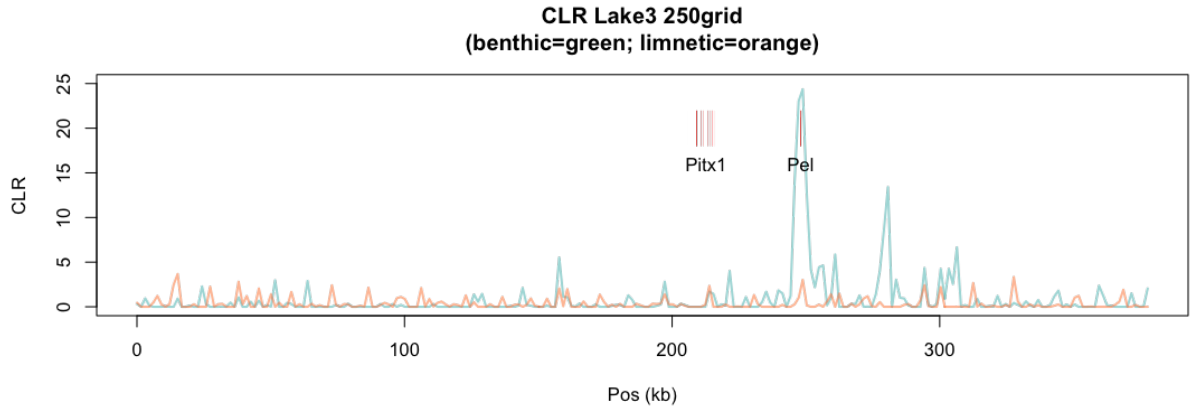
1. SweeD_Info.run1 <-summary file
2. SweeD_Report.run1 <-main results file

```
        SweeD_Report.run1

        //1
        Position Likelihood      Alpha
        186.0000 1.711778e-01    1.473081e+02
        1701.9237       1.833198e-01    1.113694e+00
        3217.8474       1.409242e-01    3.072038e-02
        4733.7711       0.000000e+00    4.137932e-01
        6249.6948       0.000000e+00    9.677419e-02
        …

        Likelihood = Composite Likelihood Ratio
                Numerator = likelihood of a sweep at a certain part of the genome
                Denominator = neutral model (empirical SFS based on all snps)

        Alpha = r/s.ln(2N) = Probability of escaping a selective sweep.
```

B. PLOTTING SWEED RESULTS IN R.

For each Lake, generate a plot showing the CLR (Likelihood) statistic across the 377kb Pitx1 scaffold for each ecotype.

```
Example R code

# before importing into R, use a text editor to remove/delete the first 2 lines from your
SweeD_Report.run1

Lake3ben.250<-read.table(file="SweeD_Report.Lake4.Ben.250",sep="\t",header=T);
Lake3lim.250<-read.table(file="SweeD_Report.Lake4.Ben.250",sep="\t",header=T);

#start and stop coordinates of the Pitx1 coding sequence
pitx1start<-c(206193, 209315, 211079, 211952, 213745, 214559, 215397, 215475, 216273);
pitx1stop<-c(206188, 208803, 210877, 211812, 213561, 214433, 215219, 215459, 216234);

#start and stop coordinates of the Pel enhancer
pelstart<-247768;
pelstop<-248265;

plot(Lake3ben.250[,1]/1000,Lake3ben.250[,2],main="CLR Lake3 250grid\n(benthic=green,
limnetic=orange)",xLab="Pos
(kb)",ylab="CLR",pch=20,col=rgb(80,140,140,240,maxColorValue=255),type="n",ylim=c(0,25));
lines(Lake3ben.250[,1]/1000,Lake3ben.250[,2],lwd=2,col=rgb(80,180,180,140,maxColorValue=255));
lines(Lake3lim.250[,1]/1000,Lake3lim.250[,2],lwd=2,col=rgb(250,100,30,120,maxColorValue=255));

rect(pelstart/1000,18,pelstop/1000,22,col=rgb(250,80,80,250,maxColorValue=255),border=NA);
text(((((pelstop-pelstart)/2)+pelstart)/1000),16,"Pel");
rect(pitx1start/1000,18,pitx1stop/1000,22,col=rgb(250,80,80,250,maxColorValue=255),border=NA);
text(mean(pitx1start)/1000,16,"Pitx1");

quartz.save(file="Sweed.CLR.Lake3.benlim.250grid.png",type="png");
```



QUESTIONS:

1. **Based on your results, is there any evidence that *Pel* has been selected in particular ecotypes or lakes?**
2. **What is the size / extent / width of the selective sweep?**
3. **How does the CLR and alpha value change with grid size?**