SPECIFIC DETAILS and INFORMATION

Samples and sample bam files

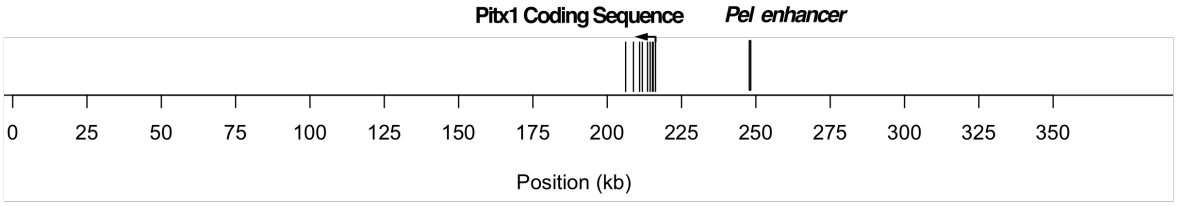
Lake	Benthic	Limnetic	Number of Individual Bam Files
Lake1	Pelvic reduced	Pelvic complete	6
Lake2	Pelvic complete	Pelvic complete	6
Lake3	Pelvic reduced	Pelvic complete	6
Lake4	Pelvic complete	Pelvic complete	6

These files already deposited on amazon server in tarball BenLim.tar. When expanded they have the file naming format **Lake1_Ben04.sorted.bam**

Reference Genome File:
(a merge of two sanger sequenced marine stickleback BAC clones covering the Pitx1 region)

~/wpsg_2016/activities/sweed/SALR.Pitx1.118G22-164F21.Combined.fa [377,804bp]

This exercise introduces you to one of the dark corners of the stickleback genome. You'll be working on the far southern telomeric end of chromosome 7. So far south you'll be deep in 68bp telomeric repeats, in a region of high recombination rate that floats on a hand-assembled-scaffold tied to the end of chromosome chrVII by genetic maps. (ie its not in the assembly).



Pel enhancer coordinates:
247768-248265

Pitx1 coding sequence coordinates:

Feature	Strand	Start	Stop
PolyA	-	206193	206188
Termination	-	209315	208803
Intron	-	211079	210877
Intron	-	211952	211812
Intron	-	213745	213561
Intron	-	214559	214433
Intron	-	215397	215219
Initiation	-	215475	215459
Promoter	-	216273	216234

Extended exercise: Other genes in the region?
This scaffold likely contains more genes than Pitx1. You can run a quick analysis to identify other possible coding regions by uploading/pasting the reference file to the genscan MIT server: *http://genes.mit.edu/GENSCAN.html*
This will predict genemodels based features like start and stop codons etc etc (some of which are nonsense). It will also give you the predicted protein sequence. To find out whether there is biological support for this protein in other organisms paste the sequence into ncbi protein blast to find out what it matches.

EXERCISE 1: Characterising the indel differences in genomic data from multiple individuals, and benthic-limnetic species pairs

A. Use samtools to estimate depth of coverage for each fish. Summarise and plot the results in R to identify the regions where pelvic reduced fish have different coverage relative to pelvic complete. Make a graph showing both depth of coverage for pelvic reduced fish and depth of coverage for pelvic complete fish across the scaffold. Repeat for the zoomed in section around the Pel enhancer.

HINTS:

- You'll be using samtools depth. At the command prompt type 'samtools depth' for detailed help.

- You can feed it a single file listing the paths of the 48 bamfiles on each line (use your unix powers for this! ls $PWD/*sorted.bam | cat >bamfiles.list;). Keep track of the order of the files in the list – you'll need this for plotting later.

  ```
  samtools depth -f ~/wpsg_2016/activities/sweed/bamfiles.list >
  ~/wpsg_2016/activities/sweed/samtools.depth.out;
  ```

- Samtools depth will create a single outfile with one column per fish. Each row is a position along the reference genome and the value in the table is the estimated depth of coverage.

  Explore what this looks like using unix:
  ```
  more samtools.depth.out
  ```

- To explore and plot this in R, you'll first want to summarise the data across individuals (eg sum the coverage for all benthics in Lake 1) or (eg average the coverage for all benthics in Lake 1). Hint: Use the 'apply' function to do this (see box below for R script examples).

QUESTIONS:

- **Based on your results, how do the coverage patterns vary across the Pitx1 scaffold?**
- **Do any see any extreme or unusual patterns of excess or undercoverage?**
- **What might be the cause of this and if bioinformatics/technical, how would you fix it?**

- **How do the coverage patterns vary across lakes?**
- **Are there any coverage differences present in lakes with pelvic reduced fish that are not present in lakes with pelvic complete fish?**
- **Do any see any extreme or unusual patterns of excess or undercoverage?**
- **What might be the cause of this and if bioinformatics/technical, how would you fix it?**

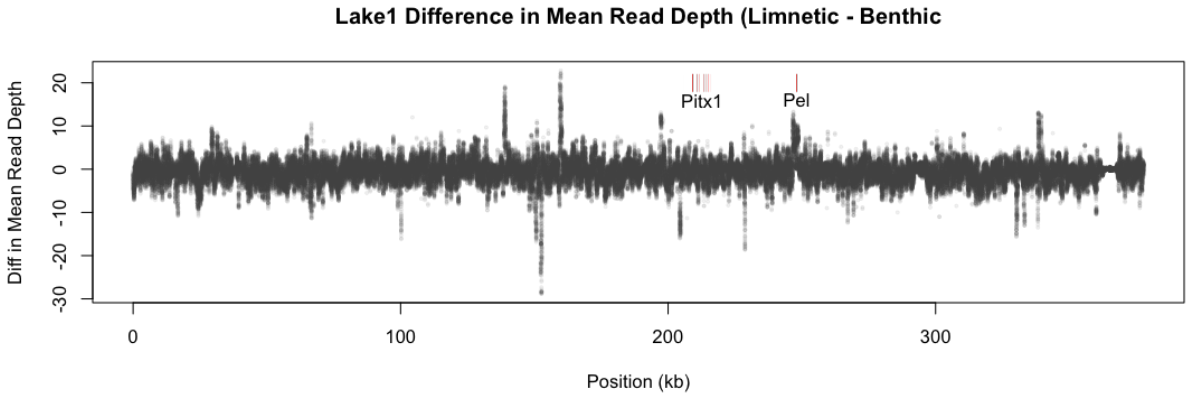
Extended exercise:

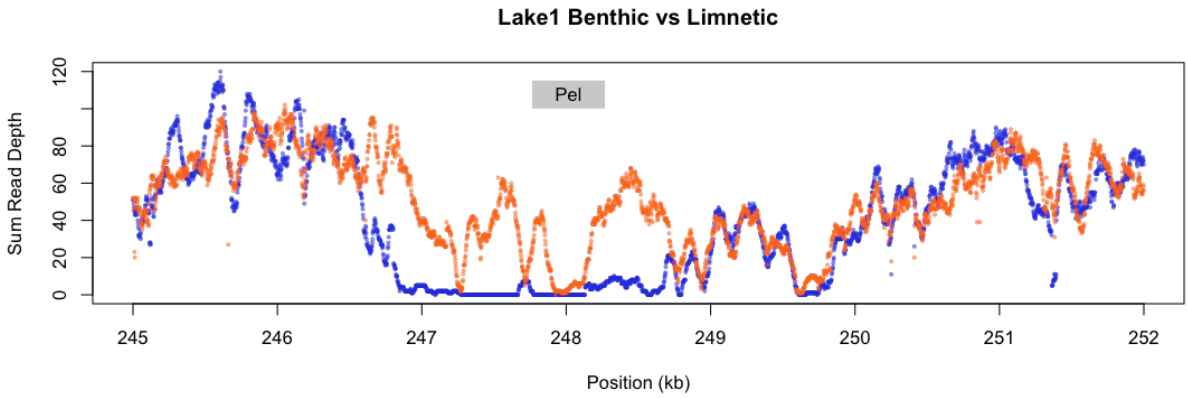
- These read placements haven't been filtered or repeat masked in any way. There are some positions with crazy high coverage likely due to repetitive elements. Repeat the depth analysis using the –q and –Q parameters in samtools depth to filter these positions.

THINGS YOU MAY ENCOUNTER

- The variance is huge and it is difficult to see any signal when viewing the entire window. Zoom in on a particular region using the 'which' function (example below), or create a single combined metric representing limnetic coverage minus benthic coverage.

```
Some example R commands that you may find useful for plotting depth in R
(modify to suit)

data<-read.table(file="samtools.depth.out",sep="\t");

key<-read.table(file="samplekey.txt",sep="\t",header=T);

#where samplekey.txt is a file specifying the order in which you fed the bam files to
#samtools: (for example)
#lake eco order
#Lake1 Lim     1
#Lake1 Lim     2
#Lake1 Ben     3
#Lake1 Ben     4
#Lake2 Lim     5
#Lake2 Lim     6
#You should create a file like this using unix and/or a text editor.

#use the apply the function to calculate mean/sum depth for different lakes and ecotypes. For
example:
Lake1Ben<-key[which((key[,1]=="Lake1")&(key[,2]=="Ben")),];
Lake1Lim<-key[which((key[,1]=="Lake1")&(key[,2]=="Lim")),];
Lake1Lim.sum<-apply(data[,(Lake1Lim[,3]+2)],MARGIN=1,FUN=sum);
Lake1Ben.sum<-apply(data[,(Lake1Ben[,3]+2)],MARGIN=1,FUN=sum);
Lake1Lim.mean<-apply(data[,(Lake1Lim[,3]+2)],MARGIN=1,FUN=mean);
Lake1Ben.mean<-apply(data[,(Lake1Ben[,3]+2)],MARGIN=1,FUN=mean);
Lake1Diff.mean<-Lake1Lim.mean - Lake1Ben.mean;

#use the following information to add Pitx1 and Pel annotations to your plots
pitx1start<-c(206193, 209315, 211079, 211952, 213745, 214559, 215397, 215475, 216273);
pitx1stop<-c(206188, 208803, 210877, 211812, 213561, 214433, 215219, 215459, 216234);
pelstart<-247768;
pelstop<-248265;


#PLOTTING LAKE 1 DIFFERENCE IN COVERAGE ACROSS ENTIRE PITX1 REGION
plot((data[,2]/1000),Lake1Diff.mean,col=rgb(80,80,80,20,maxColorValue=255),pch=20,cex=0.6,ylab
="Diff in Mean Read Depth (Lim.mean - Ben.mean)",xlab="Position (kb)",main="Lake1 Benthic vs
Limnetic");
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="coverage_of_entire.pitx1.png",type="png")



#ZOOMING IN ON PEL REGION
plot((data[which((data[,2]>245000)&(data[,2]<252000)),2]/1000),Lake1Ben.sum[which((data[,2]>
245000)&(data[,2]<252000))],col=rgb(40,20,220,120,maxColorValue=255),pch=20,cex=0.6,ylab="Sum
Read Depth",xlab="Position (kb)",main="Lake1 Benthic vs Limnetic");

points((data[which((data[,2]>245000)&(data[,2]<252000)),2]/1000),Lake1Lim.sum[which((data[,2]>
245000)&(data[,2]<252000))],,col=rgb(250,100,30,120,maxColorValue=255),pch=20,cex=0.6,
ylab="Sum Read Depth",xlab="Position (kb)",main="Lake1 Benthic vs Limnetic");

rect(pelstart/1000,100,pelstop/1000,115,col=rgb(80,80,80,80,maxColorValue=255),border=NA);text
(((((pelstop-pelstart)/2)+pelstart)/1000),108,"Pel");

quartz.save(file="coverage_of_pel.png",type="png")
```


Lake1 Difference in Mean Read Depth (Limnetic - Benthic
Diff in Mean Read Depth
-30
-20
-10
0
10
20
Pitx1
Pel
0
100
200
300
Position (kb)


Lake1 Benthic vs Limnetic
Sum Read Depth
0
20
40
60
80
120
Pel
245
246
247
248
249
250
251
252
Position (kb)

B. Does the *Pel* deletion identified in Chan et al (2010) exist in pelvically reduced benthic-limnetic species pairs that have been whole genome sequence? What about the Pelvic-complete fish?

48 bam files are ready to use:
48 fish have been whole genome sequenced (2x76bp PE Illumina)
Reads have already been placed against the unassembled Pitx1 scaffold (reference = "SALR.Pitx1.118G22-164F21.Combined.fa" using bwa.
Readgroups were labelled using Picard.
PCR duplicates have been marked using Picard
Realignment around indels has already been performed using GATK.

Use **samtools tview** to determine the precise deletion boundary coordinates in the 6 pelvic reduced benthic fish from each of Lakes 1 and 3. To do this you'll need to know the location of the minimum pel enhancer. You can find the sequence in the sup info of Chan et al 2010 Science. (also see detailed information on final page).

Sample ID	Deleted?	Start Coordinate	Stop Coordinate
Lake1_Ben04			
Lake1_Ben05			
Lake1_Ben06			
Lake1_Ben07			
Lake1_Ben08			
Lake1_Ben09			
Lake1 Limnetic Fish:			
Lake3_Ben01			
Lake3_Ben06			
Lake3_Ben08			
Lake3_Ben11			
Lake3_Ben13			
Lake3_Ben25			
Lake3 Limnetic Fish:			

QUESTION: Based on your results, has pelvic reduction evolved in benthic-limnetic species pairs via reuse of the same allele/haplotype each time?

HINTS/TIPS:
- samtools tview Lake1_Ben04.sorted.bam SALR.Pitx1.118G22-164F21.Combined.fa (the reference for this scaffold)
-'?' for help
-hold down the space bar to zip along.
-Pel enhancer coordinates where expected deletion is 247768-248265.

THINGS YOU MAY ENCOUNTER:
A trivial annoyance is the reference file has a ridiculously long sequence name! (sorry).
CH213-118G22.fulllength+CH213-164F21.pitx1.5'truncated
This prevents us from moving along the scaffold very quickly. (the typical g function to do this doesn't accept long names). For the purpose of this workshop, the name can't be shortened since all of the other files have been placed against this indexed reference.