Julie Garcia
March 5, 2018

Lab 6
Advanced Genomics and Genetics Analyses

In this lab, we will be working with a ChIP-Seq dataset to identify TF binding sites and chromatin modifications from the authors of PeakSeq. However, we will be using MACS, rather than PeakSeq to analyze it. As bioinformaticians, we often find that when working with multiple software tools, we have to format and reformat data to accommodate the requirements of the tool being used. You will be using some simple shell commands and R (or another language of your choice) to format this data into a format that MACS will accept. We will then run MACS to identify the peaks in the dataset, and plot the results in R.

1.) Get the file called FC305JN_s_1_eland_result.txt.gz, unzip it and extract only lines with "chr6.fa" since we will only be analyzing the sequence information from chromosome 6. How many lines are in the new file that you created?

In the command line:

```
grep -n "chr6.fa" FC305JN_s_1_eland_result.txt> FC305JN_output.txt
```

```
wc -l FC305JN_s_1_eland_result.txt
```

```
wc -l FC305JN_output.txt
```

There were 5,964,656 lines to start, and after filtering there were **166,056** lines.

2.) Now since we don't need all of the columns, extract columns 1,2,7,8,9 from this chromosome 6 file that you created in #1. These columns correspond to the read name, read sequence, chr6.fa, locus, and strand. One way to do this is with the awk command, although there are other ways to do this as well.

```
awk '{print $1, $2, $7, $8, $9}' FC305JN_output.txt > FC305JN_cols.txt
```

3.) In R, or a language of your choice, read in the 5 column file you created in #2 and count up the number of unique locus values for each strand separately. Then output a MACS formatted file (see lecture slides) that contains the appropriate fields for each strand. You will need to change F/R to +/- in addition to providing the frequencies for each locus (read). You may also need to cast a character field or two to numeric with something like as.numeric(as.matrix(x)) to appropriately

order the locus values in ascending order. Be sure to check your tag counts and coordinates in R to make sure that they are not accidentally cast to character.

```
data = read.table("~//Datasets//FC305JN_cols.txt")
dim(data)

colnames(data) <- c("Name", "Sequence", "chr6.fa", "Locus",
"Strand")
colnames(data)
data.F <- data[data$Strand=="F",]
numUnique.F <- length(unique(data.F$Locus))
data.R <- data[data$Strand=="R",]
numUnique.R <- length(unique(data.R$Locus))
```

There are 68183 unique loci on the forward strand. There are 68328 unique loci on the reverse strand.

```
#create macs input file
macs <- data.frame("Chr"=as.character(), "start"=as.numeric(),
"end"=as.numeric(),
                   "??"=as.numeric(), "tags"=as.numeric(),
"sense"=as.character())

for (row in 1:nrow(data)) {
  print(row)
  start <- data[row,4]
  if(nrow(macs[macs$start==start,]) > 0) {
    macs[macs$start==start,5] <- macs[macs$start==start,5] + 1
  } else {
    end <- as.numeric(data[row, 4]) +
nchar(as.character(data[row, 2]))
    strand <- if(data[row, 5]=="F") "+" else "-"
    new <- data.frame("chr6", start, end, 0, 1, strand)
    names(new) <- c("Chr", "start", "end", "??", "tags", "sense")
    macs <- rbind(macs, new)
  }
}

#sort macs input by start locus and write to file
dim(macs)
macs.sorted <- macs[order(macs$start),]
write.table(macs.sorted, file="~//Datasets//FC305JN_macs.txt",
       row.names=FALSE, quote=FALSE, sep='\t', col.names=FALSE)
```

4.) You are now ready to run MACS on the file that you created in #3. Use the –t argument and the –name arguments and run MACS to identify significant regions. We will use the default p-value threshold of 1e-5. Print out the table in the *.xls file of the significant regions (this table should have <400 rows). You can also run the R script that is output with MACS to see the peak distributions.

In Macs:

```
macs14 -f BED -t FC305JN_macs.txt -name FC305JN_macs_out
```

In R:

```
macs.data <- read.table("~//Datasets//ame_peaks.xls", header=T)
macs.data
dim(macs.data)
```

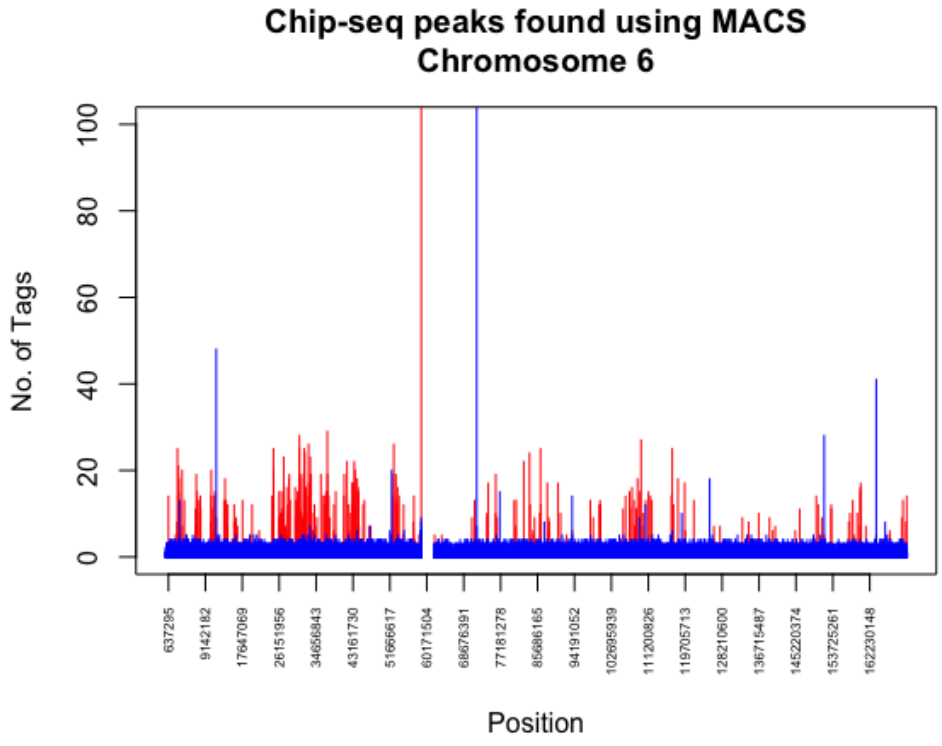
There were 348 rows.

5.) In R, make a plot similar to that below where the blue peaks are the read counts based on the unique locus calculations you conducted in #4 and the red peaks are those found using MACS. Use the polygon function for those regions identified by MACS since these span multiple physical positions. Also only make your y-axis up to 100 tags, so we can see the patterns better.

```
plot(macs.data$tags, macs.data$start, type="n",
     xlab="", ylab="No. of Tags", xaxt="n",
     ylim=c(0,100),
     xlim=c(0,max(macs.data$start)),
     main="Chip-seq peaks found using MACS\nChromosome 6",
     col = "black")

lines(macs.data$start, macs.data$tags, type="h", col="red")
lines(macs.sorted$start, macs.sorted$tags, type="h", col="blue")

minStart = min(macs.data$start)
maxStart = max(macs.data$start)
range = maxStart - minStart
xticks<-seq(minStart, maxStart, by=round(range/20))
axis(side=1, at=xticks, labels=xticks,las=2, cex.axis=0.5)
mtext("Position", side=1, line=4)
```

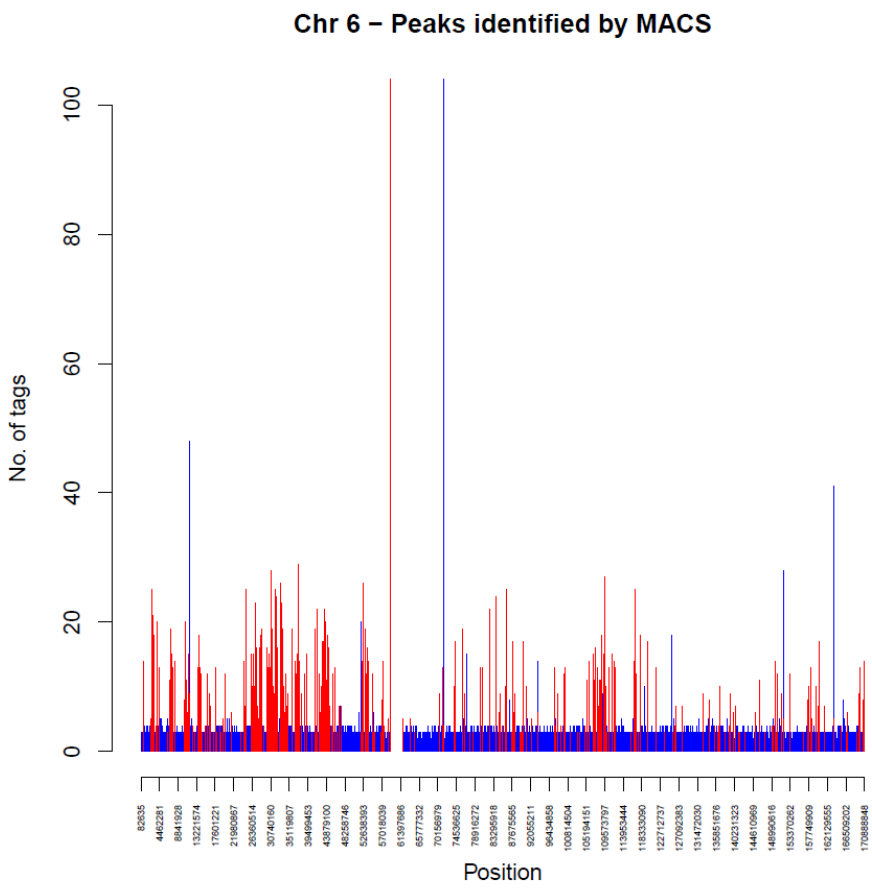


6.) Why is it likely that the 4 largest blue peaks were not identified as significant regions?

These are likely spurious peaks caused by noise and were filtered out by MACs by removing excess duplicate tags.

7.) Why do you think there is a gap at the physical position ~60,000,000 for this chromosome?

This is the centromere.


Chr 6 – Peaks identified by MACS
No. of tags
0
20
40
60
80
100
82835
4462281
8941928
13221574
17601221
21980867
26360514
30740160
35119807
39499453
43879100
48258746
52638393
57018039
61397686
65777332
70156979
74536625
78916272
83295918
87675565
92055211
96434858
100814504
105194151
109573797
113953444
118333090
122712737
127092383
131472030
135851676
140231323
144610969
148990616
153370262
157749909
162129555
166509202
170888848
Position