CSE185-LAB5-README

July 17, 2023

1 CSE185 Lab 5 Report - Code Documentation (10 pts)

- Document any commands used or additional analysis steps below!
- You should include enough detail that the instructors (or your future self) could come back to this several months from now and know exactly what you did and why you did it.
- We will not run this notebook, but will look back to see what you did especially if you end up with different answers.

For grading purposes only - Do not copy or edit this cell!

2 Question 1:

[ ]:
```
To find out the number of reads there are in each dataset I used grep command␣
 ↪with -c option for the number of lines that have '@' which indicate number␣
 ↪of reads. Commands:
grep -c '@' ~/public/lab5/Oct4.esc.fastq
grep -c '@' ~/public/lab5/Klf4.esc.fastq
grep -c '@' ~/public/lab5/Sox2.esc.fastq
grep -c '@' ~/public/lab5/H3K27ac.esc.fastq
grep -c '@' ~/public/lab5/H3K4me2.esc.fastq
grep -c '@' ~/public/lab5/input.esc.fastq

To find out the read length. I used the following command for each dataset.
awk 'NR%4==2{sum+=length($0)}END{print sum/(NR/4)}' ~/public/lab5/Oct4.esc.fastq
awk 'NR%4==2{sum+=length($0)}END{print sum/(NR/4)}' ~/public/lab5/Klf4.esc.fastq
awk 'NR%4==2{sum+=length($0)}END{print sum/(NR/4)}' ~/public/lab5/Sox2.esc.fastq
awk 'NR%4==2{sum+=length($0)}END{print sum/(NR/4)}' ~/public/lab5/H3K27ac.esc.
 ↪fastq
awk 'NR%4==2{sum+=length($0)}END{print sum/(NR/4)}' ~/public/lab5/H3K4me2.esc.
 ↪fastq
awk 'NR%4==2{sum+=length($0)}END{print sum/(NR/4)}' ~/public/lab5/input.esc.
 ↪fastq

I used this command because:
```

```
his command will output the average read length of the file Oct4.esc.fastq␣
↪located in the ~/public/lab5/ directory. The NR%4==2 "condition ensures that␣
↪only the second line of every four-line block (which contains the read␣
↪sequence) is considered in the length calculation. The sum variable␣
↪accumulates the length of all the read sequences, and NR/4 calculates the␣
↪total number of reads. The final print statement calculates and outputs the␣
↪average read length.
```

3 Question 2:

[1]:
```
Command to index reference file (This will index the reference file using␣
↪BWA-MEM):
bwa index GRCm38.fa


Commands to align reads to reference (This will use the tool BWA-MEM to align␣
↪each read to the reference and save the output to a sam file):
bwa mem GRCm38.fa Oct4.esc.fastq > Oct4.sam
bwa mem GRCm38.fa Klf4.esc.fastq > Klf4.sam
bwa mem GRCm38.fa Sox2.esc.fastq > Sox2.sam
bwa mem GRCm38.fa H3K27ac.esc.fastq > H3K27ac.sam
bwa mem GRCm38.fa H3K4me2.esc.fastq > H3K4me2.sam
bwa mem GRCm38.fa input.esc.fastq > input.sam

Command to compress, sort and index bam files (This will first compress the sam␣
↪file into a bam file, then the next command will sort the newly created bam␣
↪file and lastly the command after will index the sorted bam files)
samtools view -S -b Oct4.sam > Oct4.bam  #For the Oct4 sample
samtools sort Oct4.bam > Oct4.sorted.bam
samtools index Oct4.sorted.bam

samtools view -S -b Klf4.sam > Klf4.bam  #For the Klf4 sample
samtools sort Klf4.bam > Klf4.sorted.bam
samtools index Klf4.sorted.bam

samtools view -S -b Sox2.sam > Sox2.bam  #For the Sox2 sample
samtools sort Sox2.bam > Sox2.sorted.bam
samtools index Sox2.sorted.bam

samtools view -S -b H3K27ac.sam > H3K27ac.bam  #For the H3K27ac sample
samtools sort H3K27ac.bam > H3K27ac.sorted.bam
samtools index H3K27ac.sorted.bam

samtools view -S -b H3K4me2.sam > H3K4me2.bam  #For the H3K4me2 sample
samtools sort H3K4me2.bam > H3K4me2.sorted.bam
samtools index H3K4me2.sorted.bam
```

```
samtools view -S -b input.sam > input.bam      #For the input sample
samtools sort input.bam > input.sorted.bam
samtools index input.sorted.bam
```

```
  File "/tmp/ipykernel_174/3548961101.py", line 1
    Command to index reference file (This will index the reference file using␣
 ↪BWA-MEM):
            ^
SyntaxError: invalid syntax
```

4 Question 3:

[ ]:
```
Commands to check flagstats of each bam sample then look at the percentage␣
 ↪returned for reads that were aligned successfully:
samtools flagstat Oct4.bam
samtools flagstat Klf4.bam
samtools flagstat Sox2.bam
samtools flagstat H3K27ac.bam
samtools flagstat H3K4me2.bam
samtools flagstat input.bam
```

5 Part 2

[ ]:
```
Commands to use makeTagDirectory in order to get started with HOMER for each␣
 ↪sorted.bam:
makeTagDirectory ~/lab5/tagdirs/Oct4 ~/lab5/bams/Oct4.sorted.bam
makeTagDirectory ~/lab5/tagdirs/Klf4 ~/lab5/bams/Klf4.sorted.bam
makeTagDirectory ~/lab5/tagdirs/Sox2 ~/lab5/bams/Sox2.sorted.bam
makeTagDirectory ~/lab5/tagdirs/H3K27ac ~/lab5/bams/H3K27ac.sorted.bam
makeTagDirectory ~/lab5/tagdirs/H3K4me2 ~/lab5/bams/H3K4me2.sorted.bam
makeTagDirectory ~/lab5/tagdirs/input ~/lab5/bams/input.sorted.bam
```

6 Part 3

[ ]:
```
Commands to make UCSC files for HOMER:
makeUCSCfile ~/lab5/tagdirs/Oct4 -o auto
makeUCSCfile ~/lab5/tagdirs/Klf4 -o auto
makeUCSCfile ~/lab5/tagdirs/Sox2 -o auto
makeUCSCfile ~/lab5/tagdirs/H3K27ac -o auto
makeUCSCfile ~/lab5/tagdirs/H3K4me2 -o auto
makeUCSCfile ~/lab5/tagdirs/input -o auto
```

7 Question 4:

I downloaded IGV desktop version then uploaded the ucsc.bedGraph.gz files to created using the previous makeUCSCfile commands. Then, I selected mm10 mouse gene and went to chr17, then zoomed in on the pou5f1 gene region and looked at the peaks there and compared to each dataset.

8 Part 4

9 Question 5:

Commands to use findPeaks from Homer, this uses two different styles factor and histone depending on which dataset is called:

[ ]:
```
findPeaks ~/lab5/tagdirs/Oct4/ -i ~/lab5/tagdirs/input -style factor -o auto
findPeaks ~/lab5/tagdirs/Klf4/ -i ~/lab5/tagdirs/input -style factor -o auto
findPeaks ~/lab5/tagdirs/Sox2/ -i ~/lab5/tagdirs/input -style factor -o auto
findPeaks ~/lab5/tagdirs/H3K27ac/ -i ~/lab5/tagdirs/input -style histone -o auto
findPeaks ~/lab5/tagdirs/H3K4me2/ -i ~/lab5/tagdirs/input -style histone -o auto
```

10 Part 5

11 Question 6:

Command: annotatePeaks.pl tss ~/public/genomes/GRCm38.fa -size 8000 -hist 10 -d ~/lab5/tagdirs/Oct4 ~/lab5/tagdirs/Sox2 ~/lab5/tagdirs/Klf4 ~/lab5/tagdirs/H3K4me2 ~/lab5/tagdirs/H3K27ac -gtf ~/public/genomes/GRCm38.75.gtf > ~/lab5/annotations/tss_histogram.txt

[ ]:
```
# R code used to create plot
library(dplyr)
# library(ggplot2)
data <- read.table("tss_histogram.txt", header = TRUE, sep = "\t")

#edit column names
colnames(data) <- c("Distance", "Oct4Coverage", "Oct4PositiveTags",␣
 ↪"Oct4NegativeTags", "Sox2Coverage", "Sox2PositiveTags", "Sox2NegativeTags",␣
 ↪"Klf4Coverage", "Klf4PositiveTags", "Klf4NegativeTags", "H3K4me2coverage",␣
 ↪"H3K4me2PositiveTags", "H3K4me2NegativeTags", "H327acCoverage",␣
 ↪"H327acPositiveTags", "H327acNegativeTags")

# Create plot with Oct 4 as first line for Transcription Factors
plot(data$Distance, data$Oct4Coverage, type = "b", frame = FALSE, pch = 20, col␣
 ↪= "red", xlab = "Distance", ylab = "Coverage", ylim=c(0.5,3))
# Add Sox2 Coverage
lines(data$Distance, data$Sox2Coverage, type = "b", pch = 18, col = "blue", lty␣
 ↪= 2)
#Add Klf4 Coverage
```

```
lines(data$Distance, data$Klf4Coverage, type = "b", pch = 17, col = "green",␣
 ↪lty = 3)
#Add legend
legend("topright", legend = c("Oct4 Coverage", "Sox2 Coverage", "Klf4␣
 ↪Coverage"), col = c("red", "blue", "green"), lty = 1:2, cex = 0.8)

# Create plot with H3K4me2 as first line for Histone Modifications
plot(data$Distance, data$H3K4me2coverage, type = "b", frame = FALSE, pch = 20,␣
 ↪col = "red", xlab = "Distance", ylab = "Coverage", ylim=c(0.5,7))
# Add H327acCoverage Coverage
lines(data$Distance, data$H327acCoverage, type = "b", pch = 18, col = "blue",␣
 ↪lty = 2)
#Add legend
legend("topright", legend = c("H3K4me2coverage", "H327acCoverage"), col =␣
 ↪c("red", "blue"), lty = 1:2, cex = 0.8)
```

12 Part 6

Commands to find Motifs:

[ ]:
```
findMotifsGenome.pl ~/lab5/tagdirs/Oct4/peaks.txt ~/public/genomes/GRCm38.fa ~/
 ↪lab/motifs/Oct4 -mask -size 100
findMotifsGenome.pl ~/lab5/tagdirs/Klf4/peaks.txt ~/public/genomes/GRCm38.fa ~/
 ↪lab/motifs/Klf4 -mask -size 100
findMotifsGenome.pl ~/lab5/tagdirs/Sox2/peaks.txt ~/public/genomes/GRCm38.fa ~/
 ↪lab/motifs/Sox2 -mask -size 100
```

13 Part 7

Commands to merge Sox2 and Oct4 peaks then annotate peaks for scatter plot:

[ ]:
```
mergePeaks ~/lab5/tagdirs/Oct4/peaks.txt ~/lab5/tagdirs/Sox2/peaks.txt > ~/lab5/
 ↪overlap/oct4_sox2_peaks_merged.txt

annotatePeaks.pl ~/lab5/overlap/oct4_sox2_peaks_merged.txt ~/public/genomes/
 ↪GRCm38.fa -d ~/lab5/tagdirs/Oct4 ~/lab5/tagdirs/Sox2 > ~/lab5/overlap/
 ↪oct4_sox2_scatter.txt
```

13.1 Question 8

[2]:
```
# R Code to create scatter plot comparing Oct4 and Sox2
# Read in txt file and create table
counts <- read.table("oct4_sox2_scatter.txt", header = TRUE, sep = "\t")

# edit last two column names for Sox2 and Oct4 read counts
```

```
names(counts)[20] = "Oct4ReadCounts"
names(counts)[21] = "Sox2ReadCounts"

# fold change
fold_change <- counts$Sox2ReadCounts / counts$Oct4ReadCounts

# Create a logical vector for highlighting peaks with two-fold higher binding␣
 ↪in Sox2 vs. Oct4
highlight <- fold_change >= 2

# Plot scatter plot
plot(counts$Oct4ReadCounts, counts$Sox2ReadCounts,
     col = ifelse(highlight, "red", "blue"),
     xlab = "Oct4 Read Counts", ylab = "Sox2 Read Counts", main = "Oct 4 vs.␣
 ↪Sox 2 Normalized Read Counts")

# Add a legend
legend("topright", legend = c("Not Highlighted", "Highlighted"),
       col = c("blue", "red"), pch = 1)
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
/tmp/ipykernel_116/743276557.py in <module>
      1 # R Code to create scatter plot comparing Oct4 and Sox2
      2 # Read in txt file and create table
----> 3 counts <- read.table("oct4_sox2_scatter.txt", header = TRUE, sep = "\t"
      4
      5 # edit last two column names for Sox2 and Oct4 read counts

NameError: name 'counts' is not defined
```

13.2 Question 9

Command to run findMotifsGenome again:

[ ]:
```
findMotifsGenome.pl ~/lab5/overlap/oct4_sox2_peaks_merged.txt ~/public/genomes/
 ↪GRCm38.fa ~/lab5/Sox2 -mask -size 100
```