Khan Inan
BI-GY 7653
Week 12 assignment

Q1.1a: How many peaks were called by MACS2 for each of the two Androgren Receptor ChIP-Seq samples in your MACS2 outputs from last week?

```
wc -l SRR7207011_filteredmp_macs2_peaks.narrowPeak
#the command that counts the number of total peaks required sample SRR7207011
#the output is named Output: 2001 SRR7207011_filteredmp_macs2_peaks.narrowPeak

wc -l SRR7207017_filteredmp_macs2_peaks.narrowPeak
#the command that counts the number of total peaks required sample SRR7207017
#The output is named Output: 20783 SRR7207017_filteredmp_macs2_peaks.narrowPeak
```

The number of peaks that were called by the MACS2 is 2001 for sample SRR7207011 and 20783 for SRR7207017

Q1.1b: What is the mean peak width for each sample? Show the R command (or other approach) you used to arrive at your answer.

The average width for the first sample is around 255.75962. For the second sample the average width is 337.264. The alternative approach I used for my answer did not involve R and instead involved basic calculations within an excel file/spreadsheet.

Q1.1c: What is meant by the “signalValue” in column 7?

The signalvalue column is the measured enrichment value for any given region. The signal value is essentially the intensity of the signal and these intensity show our peaks for the regions.

Q1.2.For your Q1.2 answer, report the FRiP score for each of the two androgen receptor ChIP-seq libraries and the command lines you used to generate your answer. Do the two androgen receptor libraries pass the 1% threshold typical of high quality ChIP-seq libraries? [2 points].

```
bedtools intersect -a SRR7207011_filteredmp.bam -b SRR7207011_filteredmp_macs2_summits.bed > SRR7207011_bedtools.out
samtools view -c SRR7207011_bedtools.out
#The numerator is 19880 which is the output

samtools view -c SRR7207011_filteredmp.bam
#The denominator is 22517898 which is the output
#The final caclculated FRip scor for the sample SRR7207011: 19880/22517898=0.0008828533=0.08828533%

bedtools intersect -a SRR7207017_filteredmp.bam -b SRR7207017_filteredmp_macs2_summits.bed > SRR7207017_bedtools.out
samtools view -c SRR7207017_bedtools.out
#The numerator is 212633 which is the output
samtools view -c SRR7207017_filteredmp.bam
#The denominator is 20702991 which is the output
#The final caclculated FRip scor for the sample SRR7207017: 212633/20702991=0.01027064=1.027064%
```

The first sample SRR7207011 has a FRIP score of around 0.088 which is less than 1% which is what is considered standard for high quality chip-seq
The first sample SRR7207011 has a FRIP score of around 1.027 which is greater than 1% which is what is considered standard for high quality chip-seq

Q2.1a.Include the cross-correlation profile plot in your answers file [1 point].
The chip list content:

```
class(crossvalues_Chip)
crossvalues_Chip
$CC_StrandShift
[1] 200

$tag.shift
[1] 100

$N1
[1] 8697346

$Nd
[1] 9035245

$CC_PBC
[1] 0.963

$CC_readLength
[1] 36

$CC_UNIQUE_TAGS_LibSizeadjusted
[1] 5987801

$CC_NSC
[1] 1.447

$CC_RSC
[1] 1.079

$CC_QualityFlag
[1] 1

$CC_shift
[1] 200

$CC_A
[1] 0.229

$CC_B
[1] 0.224

$CC_C
[1] 0.158
```

```
$CC_ALL_TAGS
[1] 9420858

$CC_UNIQUE_TAGS
[1] 9035245

$CC_UNIQUE_TAGS_nostrand
[1] 8909432

$CC_NRF
[1] 0.959

$CC_NRF_nostrand
[1] 0.946

$CC_NRF_LibSizeadjusted
[1] 0.599
```

Q2.1b. Which of the following are true statements about the cross-correlation profile. Select all that apply [1 point]:

a. <u>aligned read asymmetries between DNA strands are the basis for the cross-correlation profile approach to ChIP-seq QC</u>
b. cross-correlation profiling requires peaks to be called first using a tool like MACS2
c. <u>correlations between strand-specific depths are recalculated after performing a "strand shift"</u>
d. a single correlation is calculated between strand-specific depth
e. high quality ChIP-seq libraries will typically have a single cross-correlation peak
f. NSC values below a pre-defined threshold are of acceptable quality for downstream analysis
g. <u>NSC values above a pre-defined threshold are of acceptable quality for downstream analysis</u>

Q2.2a What is the NSC value for the ChIP sample?

It is 1.447

Q2.2b What is the RSC values for the ChIP sample?

It is 1.079

Q2.2c Do either of the metrics incorporate the "shadow peak" height in how they are calculated? Which one(s)?

It used the shadow peak values otherwise known as the phantom peak for the RSC value

Q2.2d Landt et al. 2012 (see “Cross-correlation Analysis”) provide minimum NSC and RSC values for libraries with acceptable signal-to-noise ratios. What are these minimum values and does this library pass ENCODE standards for quality control? Select one:

a. minimum NSC = 0.95, minimum RSC = 0.7, the library fails QC
b. minimum NSC = 0.95, minimum RSC = 0.7, the library passes QC
c. minimum NSC = 1.0, minimum RSC = 0.75, the library fails QC
d. minimum NSC = 1.0, minimum RSC = 0.75, the library passes QC
e. minimum NSC = 1.05, minimum RSC = 0.8, the library fails QC
f. minimum NSC = 1.05, minimum RSC = 0.8, the library passes QC

Q2.3a What steps are typically included/excluded in the preparation of the control/input sample? (see Week 11 pre-recorded video). Select all true statements.

a. chemical cross-linking of DNA and protein
b. fragmentation of DNA
c. chromatin immunoprecipation with an appropriate antibody
d. unlinking of DNA and protein
e. library preparation and sequencing of DNA

Only c is excluded

Q2.3b What are the NSC and RSC values for the input sample? Would this library pass the quality control standards for the ENCODE project if it were a ChIP sample?

The output list is show below

```
$CC_StrandShift
[1] 140

$tag.shift
[1] 70

$N1
[1] 8725518

$Nd
[1] 9119345

$CC_PBC
[1] 0.957

$CC_readLength
[1] 36

$CC_UNIQUE_TAGS_LibSizeadjusted
[1] 5992518

$CC_NSC
[1] 1.04

$CC_RSC
[1] 0.214

$CC_QualityFlag
[1] -2

$CC_shift
[1] 100

$CC_A
[1] 0.127

$CC_B
[1] 0.145

$CC_C
[1] 0.122
```

```
$CC_ALL_TAGS
[1] 9606580

$CC_UNIQUE_TAGS
[1] 9119345

$CC_UNIQUE_TAGS_nostrand
[1] 9098023

$CC_NRF
[1] 0.949

$CC_NRF_nostrand
[1] 0.947

$CC_NRF_LibSizeadjusted
[1] 0.599
```

The input NSC value is 1.04 and the output RSC value is 0.214. Based on these values the library did not pass the standards according to the encode project.

Q2.4. ChIC calculates “Global Metrics” for ChIP-seq quality control. These metrics center around the fingerprint plot and nine metrics that can be derived from such plots (These are not discussed in the pre-recorded videos or elsewhere in the course).

The best interpretation of the plot is that the chip sample is one of a higher quality based of the amount of reads that are within the higher ranks of the graph compared to the input

Q2.5. What is shown in the TSS plot? Please provide a detailed interpretation of the plot. Based on the TSS profile plot, describe where are H3K36me3 modifications typically located relative to protein coding genes. [1 point]

Basically the TSS plot shows that the density of the reads around certain sites are higher than that of the control. The chip samples showed a few peaks around the transcript sites as opposed to the control. These peaks (where the one on the right is broader) suggests that there is some form of modification in accordance to the relative location of the genes that code for protein