# BMEG 400E Project

Sebastian Lee (48724587) and Danny Wu (20662540)

14/04/2022

## Contents

**Introduction** **2**

**Methods** **3**

Overview of the ChIP-seq and ATAC-seq Analysis Pipeline . . . 3

ChIP-seq and ATAC-seq Analysis Specifications . . . 3

0. Installed Packages . . . 3

1. Burrows-Wheeler Alignment tool (BWA) . . . 4

2. samtools converting SAM to BAM . . . 4

3. MACS2 . . . 5

4. UCSCtool kit BedSort . . . 5

5. UCSCtool kit BedClip . . . 6

6. bedGraphToBigWig . . . 6

7. Comparing peaks . . . 6

8. deepTools computeMatrix . . . 7

9. plotHeatmap & plotProfile . . . 7

10. Repeat trials . . . 7

11. Simplified Trials . . . 8

Overview of RNA-seq Pipeline . . . 9

RNA-seq Analysis Specifications . . . 9

0. Installed Packages . . . 9

1. Spliced Transcripts Alignment to a Reference (STAR) . . . 9

2. RSeQC and Calculation of Normalized Counts per Million . . . 11

**Results** **12**

FOXA1 Analysis . . . 12

ATAC-seq Analysis . . . 15

RNA-seq Analysis . . . 16

**Conclusions** **16**

**References** **17**

# Introduction

Prostate cancer is one of the most common cancers in the world, making up nearly 8 percent of all cancers (1). In 2020, there were nearly 1.5 million cases of prostate cancer around the world (1). In Canada alone, the Canadian Cancer Society estimates that 1 in 7 men will be diagnosed with prostate cancer and that 1 in 29 will die from the disease (2). It is also cancer in which the exact causes of the occurrence are not well understood.

The cancer is commonly treated with surgery and radiation therapy. However, that may change as the researchers may have discovered a potential therapeutic method to decrease the growth of cancer. In a 2020 study "Chromatin binding of FOXA1 is promoted by LSD1-mediated demethylation in prostate cancer" by Gao et al, new insights into the behaviour of FOXA1 and LSD1 may suggest new therapeutic strategies in steroid-driven cancers (3). Previously, little was known about methods to regulate the chromatin binding of FOXA1. However, LSD1 is able to act through FOXA1 to regulate this process to decrease prostate cancer growth rates.

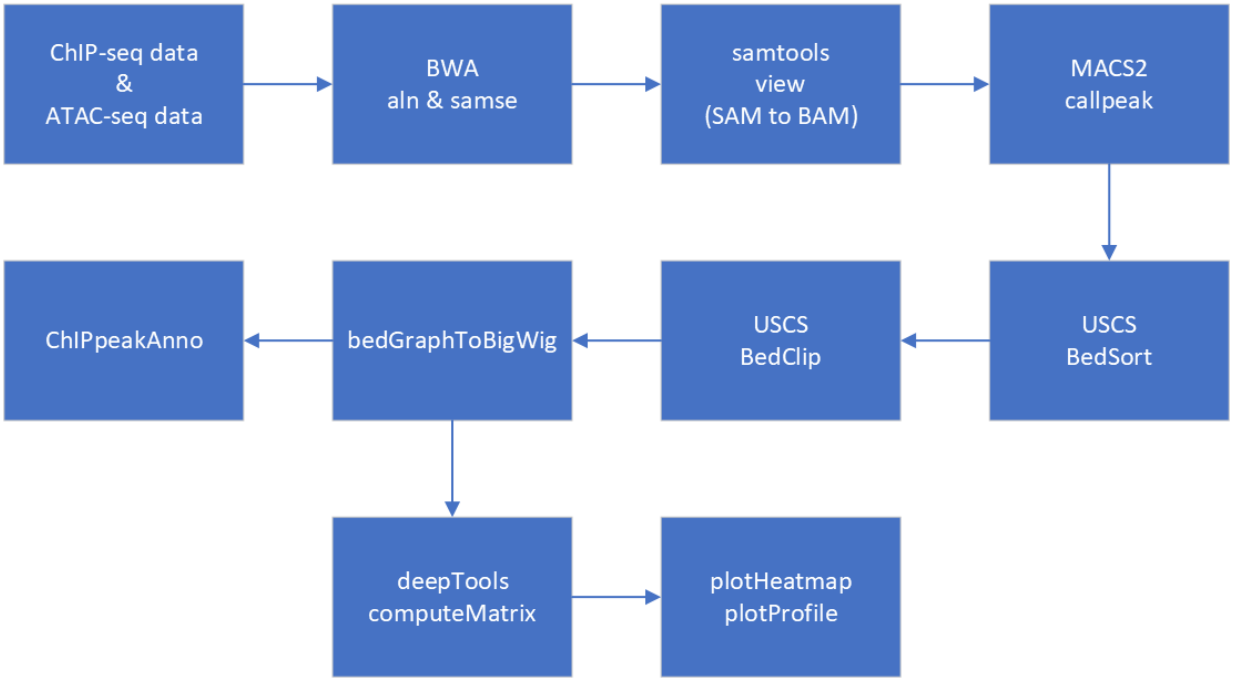
Forkhead Box A1 (FOXA1) is a protein-coding gene that acts as a transcriptional activator for a variety of transcripts, focusing on the liver. These include transthyretin, albumin, and also chromatin. Lysine-specific demethylase 1A (LSD1) is a demethylase that targets mono- or di-methylated histone H3K4 and H3K9 and regulates gene expression as a transcriptional activator or repressor. LSD1 also "binds the androgen receptor and promotes androgen-dependent transcription of hormone-responsive genes, enhancing tumour-cell growth." In the research of Gao et al., it was discovered that specifically in prostate cancer cells, LSD1 can associate with FOXA1 and that LSD1 inhibition can disrupt binding between chromatin and FOXA1. This disruption of FOXA1 chromatin binding also affects the access of chromatin to steroid hormone receptors, with one of these receptors being androgen receptors. Androgen receptors are nuclear receptors that regulate the growth of the prostate. LSD1 inhibition lowers the transcriptional output of androgen receptors, thus simultaneously lowering prostate cancer growth (3).

The goals of this project are to re-analyze and verify the claims made by Gao et al. by reproducing the genomic analysis. Our intention was to verify whether the claims of interest hold. We believe that this is an important study as if these claims are to be true, this allows a therapeutic potential that targets LSD1 and could potentially revolutionize the treatment for prostate cancer.

The re-analysis was limited to analyzing the ChIP-seq data, ATAC-seq data, and RNA-seq analysis. For the ChIP-seq and ATAC-seq data, the procedure is found under "Analysis for ChIP–seq and ATAC–seq" of the Methods section. All the data used in this paper was referenced in the data availability section of the paper. The data can be found on the Gene Expression Omnibus (GEO). The data for this paper was series GSE149007 and was published by Gao et al. This superseries of data was composed of multiple subseries, with GSE148925 being the ATAC-seq data and other series being the ChIP-seq data. fastq files were downloaded from the European Nucleotide Archive (https://www.ebi.ac.uk/ena/). We were also required to utilize hg19.fa, which was obtained from http://hgdownload.cse.ucsc.edu/goldenpath/hg19/bigZips/hg19.fa.gz. For RNA-seq analysis, GSE114267 was the GEO series used for RNA-seq analysis and hg19 gtf file was obtained from https://www.gencodegenes.org/human/release_19.html.

# Methods

## Overview of the ChIP-seq and ATAC-seq Analysis Pipeline




## ChIP-seq and ATAC-seq Analysis Specifications

Under the Method section of the study, "Analysis for ChIP–seq and ATAC–seq" procedures were followed. Some parts of the procedures were modified as some instructions were unclear and were not accomplished due to technical difficulties/issues. One series of data was used for ChIP-seq and one series of data was used for ATAC-seq. For the ChIP-seq, we used ChIP-Seq for FOXA1 in the LNCaP cell line with GSK treatment, which promotes the inhibition of LSD1, and VEH control. For the ATAC-seq, we used ATAC-Seq in the LNCaP cell line with GSK treatment and VEH control.

The ChIP-seq runs used in this section are: SRR11579386: ChIP-seq for FOXA1 in LNCaP cell line with GSK treatment SRR11579387: ChIP-seq for FOXA1 in LNCaP cell line with GSK treatment SRR11579388: ChIP-seq for FOXA1 in LNCaP cell line with VEH control

The ATAC-seq runs used in this section are: SRR11579382: ATAC-Seq in LNCaP cell line with GSK treatment SRR11579384: ATAC-Seq in LNCaP cell line with VEH control treatment

### 0. Installed Packages

There were some setup steps involving installing packages to our environments:

```
conda install -c anaconda bwa
conda install -c bioconda samtools
conda install -c bioconda macs2
conda install -c bioconda ucsc-bedsort
conda install -c bioconda ucsc-bedclip
```

```
conda install -c bioconda ucsc-bedgraphtobigwig
conda install -c bioconda deeptools
```

### 1. Burrows-Wheeler Alignment tool (BWA)

The first step was to index hg19.fa with BWA. This step indexes database sequences in the FASTA format. It was also required to obtain the data from the following link.

```
wget -O hg19.fa 'http://hgdownload.cse.ucsc.edu/goldenpath/hg19/bigZips/hg19.fa.gz'

# Commands used for BWA index
bwa index -a bwtsw -p hg19.fa hg19.fa

The next step was to align the raw reads with the hg19.fa file we indexed. This step finds the SA coordi

#Commands used for BWA align
# ChIP-seq BWA align
bwa aln -l 32 -q 5 hg19.fa SRR11579388.fastq > SRR11579388.fastq.sai

bwa aln -l 32 -q 5 hg19.fa SRR11579386.fastq > SRR11579386.fastq.sai

bwa aln -l 32 -q 5 hg19.fa SRR11579387.fastq > SRR11579387.fastq.sai

# ATAC-seq BWA align
bwa aln -l 32 -q 5 hg19.fa SRR11579382.fastq > SRR11579382.fastq.sai

bwa aln -l 32 -q 5 hg19.fa SRR11579384.fastq > SRR11579384.fastq.sai


Afterwards, BWA samse is used to generate alignments in the SAM format.

#Commands used for BWA samse
# ChIP-seq BWA samse
bwa samse hg19.fa SRR11579388.fastq.sai SRR11579388.fastq > SRR11579388.fastq.sam

bwa samse hg19.fa SRR11579386.fastq.sai SRR11579386.fastq > SRR11579386.fastq.sam

bwa samse hg19.fa SRR11579387.fastq.sai SRR11579387.fastq > SRR11579387.fastq.sam

# ATAC-seq BWA samse
bwa samse hg19.fa SRR11579382.fastq.sai SRR11579382.fastq > SRR11579382.fastq.sam

bwa samse hg19.fa SRR11579384.fastq.sai SRR11579384.fastq > SRR11579384.fastq.sam


Following the samse step, the SAM files were converted to BAM files using samtools.
```

### 2. samtools converting SAM to BAM

To make data more manageable and save storage in the server, bigger files SAMs were converted into smaller files SAMs. The input SAM files were deleted.

```
#Commands used to convert SAM files to BAM files.
# ChIP-seq
samtools view -b -h SRR11579388.fastq.sam > SRR11579388.fastq.bam

samtools view -b -h SRR11579386.fastq.sam > SRR11579386.fastq.bam

samtools view -b -h SRR11579387.fastq.sam > SRR11579387.fastq.bam

# ATAC-seq
samtools view -b -h SRR11579382.fastq.sam > SRR11579382.fastq.bam

samtools view -b -h SRR11579384.fastq.sam > SRR11579384.fastq.bam
```

### 3. MACS2

The next step was to utilize MACS2 to call peaks on the BAM files. Certain settings/parameters had to be followed, such as having fix-bimodal on and extend size at 100. The q-value cut-off for peak significance was 0.05. BDG was also set to return the output in bedGraph format.

```
#Commands used for MACS2 callpeak
# ChIP-seq
macs2 callpeak -t SRR11579386.fastq.bam -f BAM -c SRR11579388.fastq.bam -n SRR11579386_callPeak --bw 25

macs2 callpeak -t SRR11579387.fastq.bam -f BAM -c SRR11579388.fastq.bam -n SRR11579387_callPeak --bw 25

# ATAC-seq
macs2 callpeak -t SRR11579382.fastq.bam -f BAM -c SRR11579384.fastq.bam -n SRR11579382.peaks --bw 250 -
```

After the ATAC-seq callpeak, the output files were analysed as part of quality assurance process. Unfortunately, narrowPeak files for ATAC-seq were null. There are a couple explanations on why this may have occurred. One of them is a procedural error occurred during any of the first 3 stages, and the other explanation is that there is no significant difference between the GSK and the vehicle control files with the significance level of 0.05, thereby resulting in MACS2 callpeak producing empty reports.

When analyzed for errors, SAM files appear to be normal. Using samtools cat, the converted BAM files were analyzed and appeared corrupted. We believe that BAM files were not correctly converted and were the reason why MACS2 callpeak function failed to produce narrowPeaks.

### 4. UCSCtool kit BedSort

After callpeak was performed, the UCSCtool kit was utilized to sort and clip the bedGraph files before converting them to bigwig files.

The first step was to sort the files using BedSort. Bedsort sorts bed files, chromosomes and positions.

```
# Commands used to bedsort.
# ChIP-seq
sort -k1,1 -k2,2n SRR11579386_callPeak_control_lambda.bdg > SRR11579386_callPeak_control_lambda_sorted.

sort -k1,1 -k2,2n SRR11579387_callPeak_control_lambda.bdg > SRR11579387_callPeak_control_lambda_sorted.

# Assuming that the prior step worked out for ATAC-seq, the following ATAC command was used to sort.
```

```
# ATAC-seq
sort -k1,1 -k2,2n SRR11579382.peaks_control_lambda.bdg > SRR11579382.peaks.control.lambda.sorted.bdg
```

### 5. UCSCtool kit BedClip

The second step was to clip the data using Bedclip. Bedclip removes lines from the files that refer to off-chromosome places.

```
# Commands used for bedClip.
# ChIP-seq
bedClip SRR11579386_callPeak_control_lambda_sorted.bdg http://hgdownload.cse.ucsc.edu/goldenpath/hg19/bi

bedClip SRR11579387_callPeak_control_lambda_sorted.bdg http://hgdownload.cse.ucsc.edu/goldenpath/hg19/bi

# Assuming that the prior step worked out for ATAC-seq, the following ATAC command was used to clip.
# ATAC-seq
bedClip SRR11579382.peaks.control.lambda.sorted.bdg http://hgdownload.cse.ucsc.edu/goldenpath/hg19/bigZi
```

### 6. bedGraphToBigWig

The third step was to convert the bedGraph files to BigWig files. bedGraphToBigWig converts bedGraph files to BigWig files.

```
# Commands used for bedGraphToBigWig
# ChIP-seq
bedGraphToBigWig SRR11579386_callPeak_control_lambda_sorted_clipped.bdg http://hgdownload.cse.ucsc.edu/g

bedGraphToBigWig SRR11579387_callPeak_control_lambda_sorted_clipped.bdg http://hgdownload.cse.ucsc.edu/g

# Assuming that the prior step worked out for ATAC-seq, the following ATAC command was used to convert.
# ATAC-seq
bedGraphToBigWig SRR11579382.peaks.control.lambda.sorted.clipped.bdg http://hgdownload.cse.ucsc.edu/gol
```

### 7. Comparing peaks

The R package ChIPpeakAnno can be utilized to analyze peak intervals. The function findOverlapsOfPeaks can find the overlapping peaks for two or more sets of peak ranges. Only ChIP-seq files were analyzed due to

```
if (!require("BiocManager", quietly = TRUE))
  install.packages("BiocManager")

BiocManager::install("ChIPpeakAnno")

library(ChIPpeakAnno)

# ChIP-seq
g1 <- toGRanges('~/Downloads/SRR11579386_callPeak_peaks.narrowPeak', format = "narrowPeak")
g2 <- toGRanges('~/Downloads/SRR11579387_callPeak_peaks.narrowPeak', format = "narrowPeak")
overlaps <- findOverlapsOfPeaks(g1, g2)
```

```
g3 <- toGRanges('~/Downloads/SRR11579386_callPeak_peaks.narrowPeak', format = "narrowPeak")
g4 <- toGRanges('~/Downloads/GSE148926_LN_FOXA1_G_S7_peaks.narrowPeak', format = "narrowPeak")
overlaps <- findOverlapsOfPeaks(g3, g4)
```

### 8. deepTools computeMatrix

The deepTools computeMatrix tool was utilized to calculate scores per genome region.

Files downloaded for ATAC-seq: GSE148925_LN_ATAC_VEH_S1.bw: ATAC-Seq bigwig of LNCaP cells treated with GSK2879552 GSE148925_LN_ATAC_VEH_S1_peaks.narrowPeak: ATAC-Seq narrowPeak of LNCaP cells treated with GSK2879552 GSE148925_LN_ATAC_GSK_S2.bw: ATAC-Seq bigwig of LNCaP cells treated with vehicle GSE148925_LN_ATAC_GSK_S2_peaks.narrowPeak: ATAC-Seq narrowPeak of LNCaP cells treated with vehicle

```
# Commands used for computeMatrix
# ChIP-seq
computeMatrix reference-point -S SRR11579386_callPeak_control_lambda_sorted_clipped.bw SRR11579387_calll

computeMatrix reference-point -S SRR11579386_callPeak_control_lambda_sorted_clipped.bw SRR11579387_calll

# The files listed above were downloaded to generate final results, which is plotHeatmap, for this proj
# ATAC-seq
computeMatrix reference-point -S GSE148925_LN_ATAC_VEH_S1.bw -R GSE148925_LN_ATAC_VEH_S1_peaks.narrowPea

computeMatrix reference-point -S GSE148925_LN_ATAC_GSK_S2.bw -R GSE148925_LN_ATAC_GSK_S2_peaks.narrowPea
```

### 9. plotHeatmap & plotProfile

deepTools plotHeatmap was then utilized on the matrices created to create heatmaps for the scores associated with the genomic regions.

```
# Commands used for plotHeatmap
# ChIP-seq
plotHeatmap -m GSK_4h_matrix.gz -o GSK_4h_heatmap.png --hclust 2 --colorMap Blues

plotHeatmap -m GSK_48h_matrix.gz -o GSK_48h_heatmap.png --hclust 2 --colorMap Blues

# ATAC-seq
plotHeatmap -m vehmatrix.gz -o vehicle_heatmap.png --hclust 2 --colorMap Blues

plotHeatmap -m gskmatrix.gz -o gsk_heatmap.png --hclust 2 --colorMap Blues
```

### 10. Repeat trials

Some issues were encountered with the data that we prepared. However, due to the time frame, we were unable to discover how to fix the errors. The computeMatrix and plotHeatmap steps were repeated with the data provided by the authors.

```
# ChIP-seq
computeMatrix reference-point -S GSE148926_LN_FOXA1_G_GSK_S8.bw GSE148926_LN_FOXA1_G_S7.bw -R GSE148926
plotHeatmap -m ChIP_GSK_matrix.gz -o ChIP_GSK_heatmap.png --hclust 2 --colorMap Blues
```

```
computeMatrix reference-point -S GSE148926_LN_FOXA1_G_GSK_S8.bw GSE148926_LN_FOXA1_G_S7.bw -R GSE148926
plotHeatmap -m ChIP_VEH_matrix.gz -o ChIP_VEH_heatmap.png --hclust 2 --colorMap Blues

Unfortunately, the second heatmap was not available due to the following error:
numpy.core._exceptions.MemoryError: Unable to allocate 83.1 GiB for an array with shape (11151143130,)

# ATAC-seq
Vehicle (control) heatmap was not available due to the following error:
numpy.core._exceptions.MemoryError: Unable to allocate 141. GiB for an array with shape (18924948525,) a

The treated (GSK) heatmap was also not available due to the following error:
numpy.core._exceptions.MemoryError: Unable to allocate 99.9 GiB for an array with shape (13404493245,) a
```

### 11. Simplified Trials

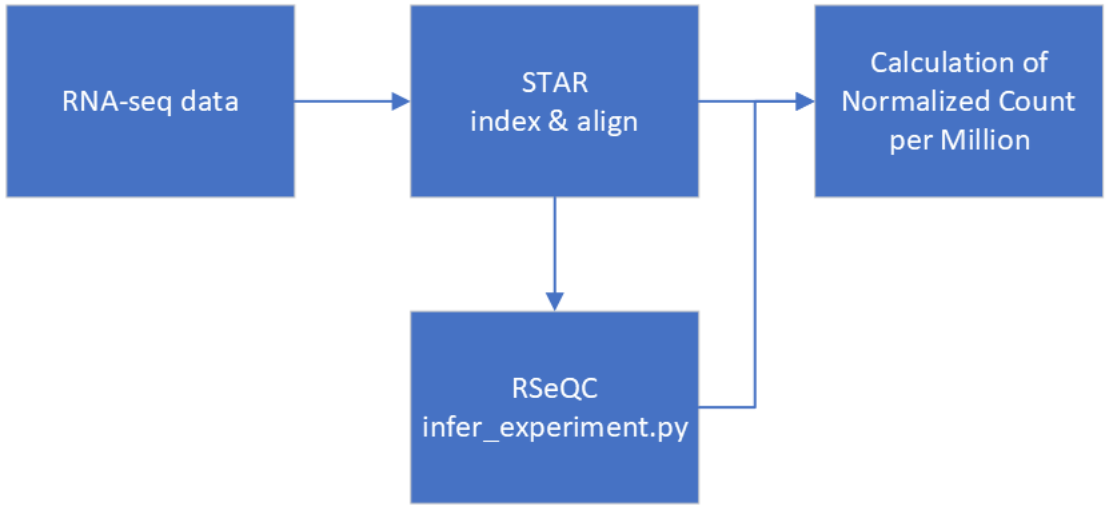
Another attempt was made deviating from the specific instructions of the researcher. However, some errors still occurred.

Sorting and indexing the BAM files created from step 2.

Due to the big ATAC-seq file sizes, simplified trials were not able to be conducted.

```
#ChIP-seq
samtools sort SRR11579386.fastq.bam -o SRR11579386.fastq.sorted.bam
samtools sort SRR11579387.fastq.bam -o SRR11579387.fastq.sorted.bam

samtools index SRR11579386.fastq.sorted.bam
samtools index SRR11579387.fastq.sorted.bam


The BAM files can now be converted to bigwig files after being sorted and indexed.

#ChIP-seq
bamCoverage -b SRR11579386.fastq.sorted.bam -o SRR11579386.fastq.sorted.bw
bamCoverage -b SRR11579387.fastq.sorted.bam -o  SRR11579387.fastq.sorted.bw

computeMatrix reference-point -S SRR11579386.fastq.sorted.bw  SRR11579387.fastq.sorted.bw -R SRR1157938

plotHeatmap -m simple1matrix.gz -o simple1heatmap.png --hclust 2 --colorMap Blues
```

## Overview of RNA-seq Pipeline




## RNA-seq Analysis Specifications

### 0. Installed Packages

A couple packages were installed in preparation for RNA-seq analysis:

```
conda install -c bioconda star
conda install -c bioconda rseqc
```

### 1. Spliced Transcripts Alignment to a Reference (STAR)

Similar to ChIP-seq and ATAC-seq analysis, hg19 was indexed prior to the RNA alignment. Although we did have pre-existing index files with bwa, we still performed indexing with STAR and created the index files.

```
# hg19 reference genome index was created prior to an alignment
STAR --runThreadN 32 --runMode genomeGenerate --genomeDir /home/slee_bmeg22/project/data/rna/index --ge

# Then, two RNA-seq files downloaded from NCBI GEO were aligned with --quantMode GeneCounts
STAR --genomeDir /home/slee_bmeg22/project/data/rna/index --runThreadN 15 --readFilesIn /home/slee_bmeg

STAR --genomeDir /home/slee_bmeg22/project/data/rna/index --runThreadN 15 --readFilesIn /home/slee_bmeg
```

Here are the results of STAR alignment:

```
                         Number of input reads |       16255066
                   Average input read length |       51
                                UNIQUE READS:
                 Uniquely mapped reads number |       12343890
                      Uniquely mapped reads % |       75.94%
                        Average mapped length |       50.89
                     Number of splices: Total |       1128528
          Number of splices: Annotated (sjdb) |       1117447
                     Number of splices: GT/AG |       1114832
                     Number of splices: GC/AG |       9621
                     Number of splices: AT/AC |       981
                Number of splices: Non-canonical |       3094
                    Mismatch rate per base, % |       0.13%
                       Deletion rate per base |       0.01%
                      Deletion average length |       1.55
                      Insertion rate per base |       0.00%
                     Insertion average length |       1.26
                           MULTI-MAPPING READS:
      Number of reads mapped to multiple loci |       3720276
           % of reads mapped to multiple loci |       22.89%
      Number of reads mapped to too many loci |       60963
           % of reads mapped to too many loci |       0.38%
                               UNMAPPED READS:
Number of reads unmapped: too many mismatches |       0
     % of reads unmapped: too many mismatches |       0.00%
          Number of reads unmapped: too short |       84779
               % of reads unmapped: too short |       0.52%
            Number of reads unmapped: other |       45158
                 % of reads unmapped: other |       0.28%
                               CHIMERIC READS:
                     Number of chimeric reads |       0
                          % of chimeric reads |       0.00%
```

```
                         Number of input reads |       13468909
                   Average input read length |       51
                                UNIQUE READS:
               Uniquely mapped reads number |       9958669
                    Uniquely mapped reads % |       73.94%
                      Average mapped length |       50.88
                  Number of splices: Total |       944167
       Number of splices: Annotated (sjdb) |       935252
                  Number of splices: GT/AG |       932964
                  Number of splices: GC/AG |       7639
                  Number of splices: AT/AC |       866
          Number of splices: Non-canonical |       2698
                 Mismatch rate per base, % |       0.13%
                    Deletion rate per base |       0.01%
                   Deletion average length |       1.55
                   Insertion rate per base |       0.00%
                  Insertion average length |       1.29
                       MULTI-MAPPING READS:
    Number of reads mapped to multiple loci |       3356136
         % of reads mapped to multiple loci |       24.92%
    Number of reads mapped to too many loci |       43226
         % of reads mapped to too many loci |       0.32%
                            UNMAPPED READS:
Number of reads unmapped: too many mismatches |       0
     % of reads unmapped: too many mismatches |       0.00%
          Number of reads unmapped: too short |       80318
               % of reads unmapped: too short |       0.60%
            Number of reads unmapped: other |       30560
                 % of reads unmapped: other |       0.23%
                              CHIMERIC READS:
                   Number of chimeric reads |       0
                        % of chimeric reads |       0.00%
```

**2. RSeQC and Calculation of Normalized Counts per Million**

Using infer_experiment.py, which is available within RSeQC package, was used to determine how the reads were stranded. This information will be useful in the next steps in which gene-level count data are pulled from one of the STAR output files. Then, normzliaed counts per million and fold-change are calculated. Pseudo count of 1 is added when calculating the fold-change. However, due to a downloading issue, RSeQC could not be downloaded to Anaconda3 environment. Therefore, this step and the subsequent calculation step could not be accomplished. Below is the troubling shooting process gone through.

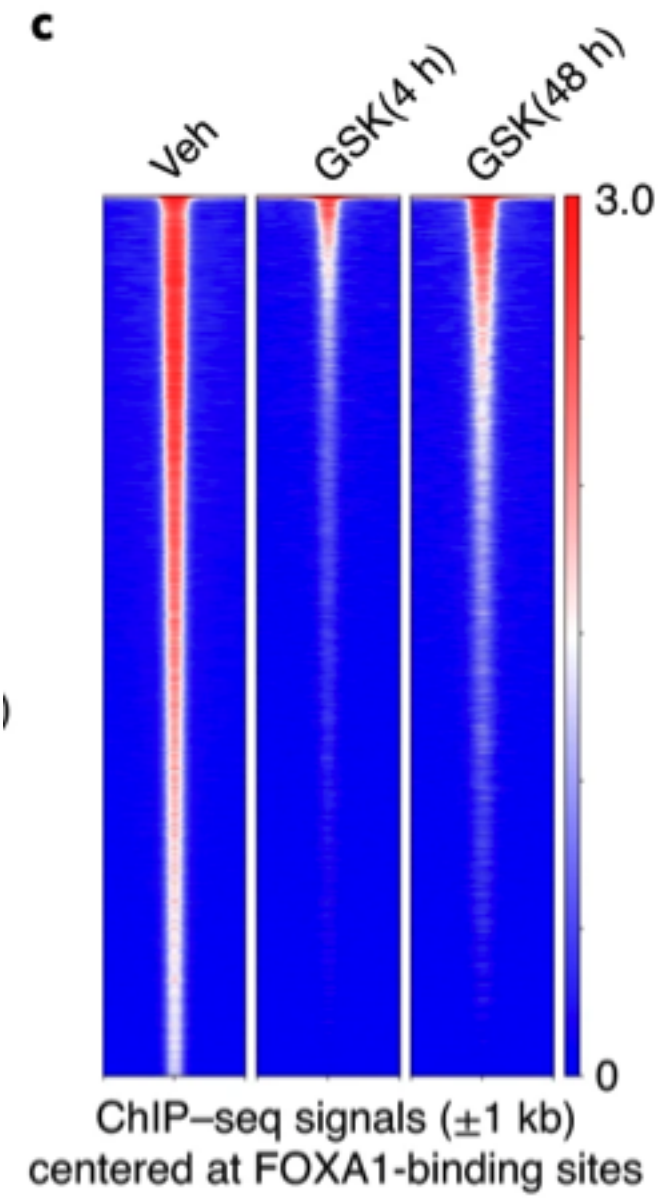
```
# Installation Sources
conda install -c bioconda rseqc
pip install -i https://pypi.anaconda.org/liguow/simple rseqc
pip3 install RSeQC
```

```
# Additional package downloads were initiated to troubleshoot
conda install -c anaconda zlib
conda install python=3.10.4

# Error message for bioconda
UnsatisfiableError: The following specifications were found to be incompatible with each other:

Output in format: Requested package -> Available versionsThe following specifications were found to be i

  - feature:/linux-64::__glibc==2.23=0
  - feature:|@/linux-64::__glibc==2.23=0
  - rseqc -> libgcc-ng[version='>=9.3.0'] -> __glibc[version='>=2.17']

Your installed version is: 2.23
```

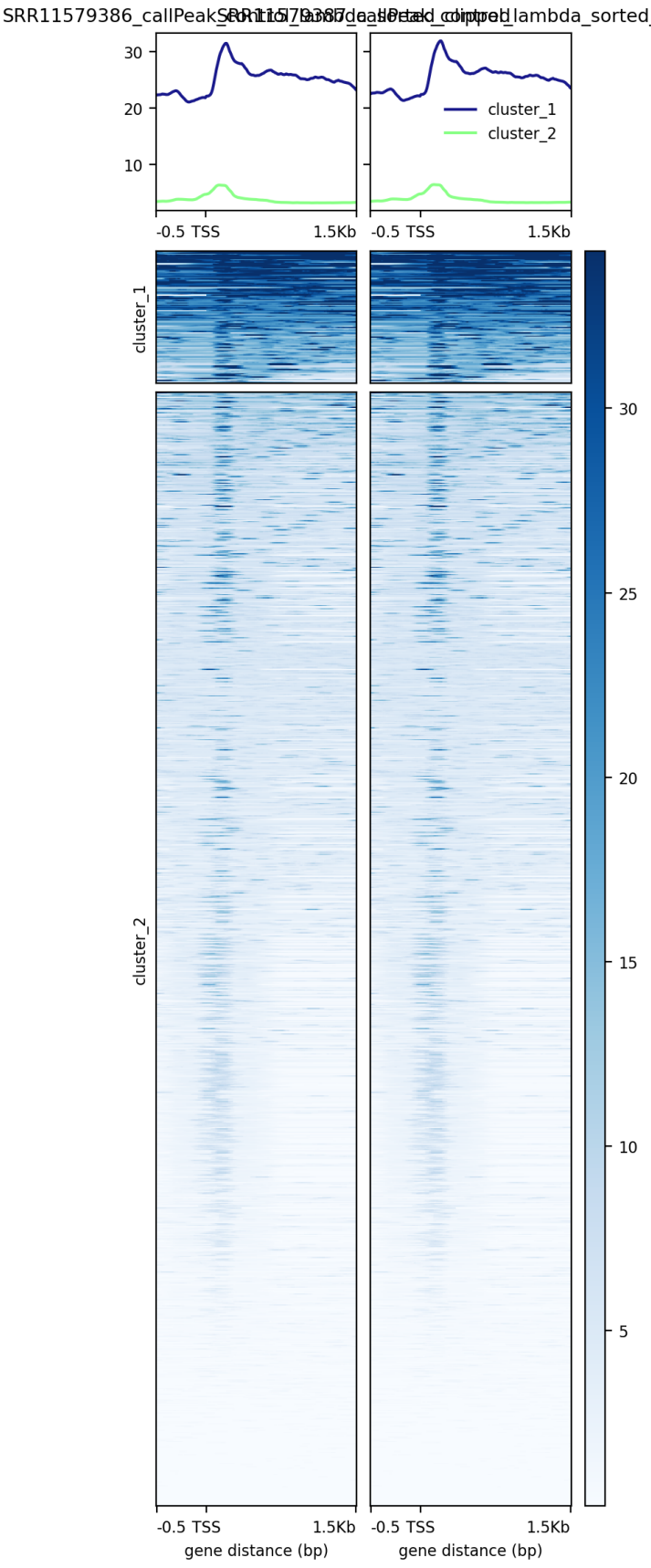
# Results

## FOXA1 Analysis

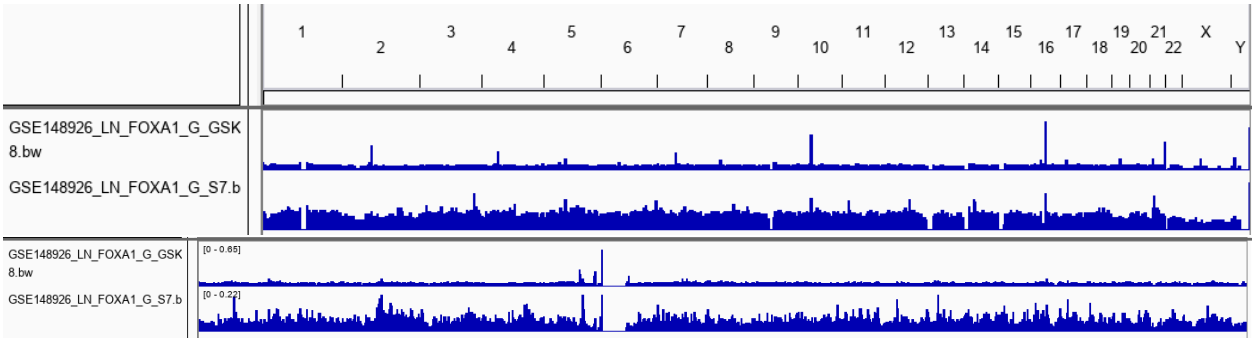
Original:


c
Veh
GSK(4 h)
GSK(48 h)
3.0
0
ChIP–seq signals (±1 kb)
centered at FOXA1-binding sites

Our version:


SRR11579386_callPeaks_SRR11579387_callPeaks_control_lambda_sorted_clipped
30
20
10
cluster_1
cluster_2
-0.5 TSS
1.5Kb
cluster_1
cluster_2
30
25
20
15
10
5
gene distance (bp)

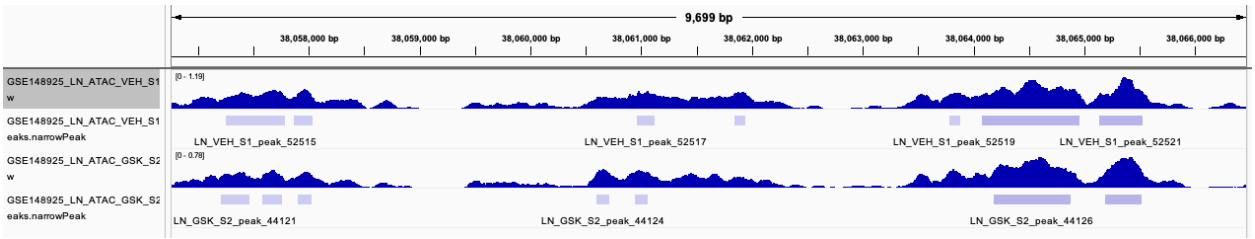
Looking at the heatmap compared to the one from the article, we can see that there are brief similarities. The blue is concentrated at the top and stays relatively close to the center. However, the values are off and there seems to be an issue with the data. Using the correct bigwig files and peak files, we can see below that the generated heatmaps match up more closely with the actual expected images. The heatmap we generated has Veh on the right and the GSK (48 h) on the left. Unfortunately, the GSK (4h) was too large to run on the server. Both the heatmaps for the Veh follow the same pattern, with the concentration relatively close to the middle and starting off thicker and darker before narrowing as you move down. The GSK(48 h) also shares some similarities, as it starts off thicker and darker before narrowing as you move down. Similar to the image provided in the article, the GSK(48 h) is shorter and less dense than the Veh.



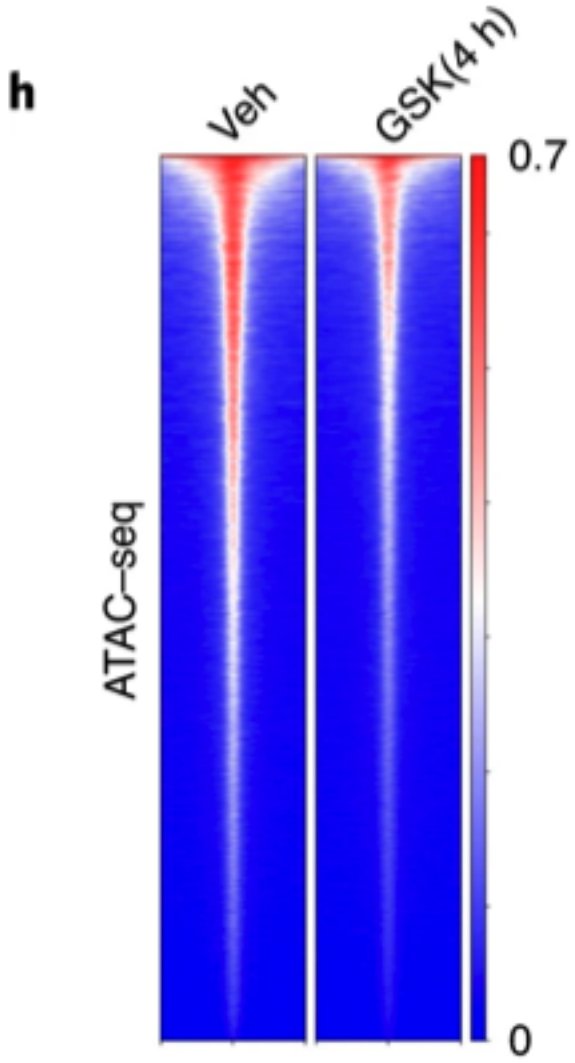

Viewing the bigwig files in IGV, we can see some locations on the Veh and the GSK(48 h) that share some peaks. However, there are also some peaks in the Veh that do not seem to stand out as much in the GSK.

## ATAC-seq Analysis

Due to the error mentioned above, the heatmap could not be generated for ATAC-seq.




Because the bigwig and callPeak files were not produced with the original fastq files, the bigwig and callPeak files downloaded from NCBI GEO were run through IGV. The top two are the vehicle (control) tracks and the bottom two are GSK treated tracks. The area of interest was focused on FOXA1, where the treatment of GSK should affect the expression of FOXA1. Comparing these two pairs of tracks, GSK tracks have narrower identified peaks than the control. The GSK narrowPeak track catches an additional peak (peak 44124) that occurs within FOXA1 but overall peaks are reduced, especially peak 44126 at the beginning of FOXA1 when compared to the vehicle peak 52520. Based on this analysis, we expect the GSK to have an overall decreased expression of FOXA1 compared to the vehicle sample, which is similar to the original heatmap shown below.




### RNA-seq Analysis

Unfortunately, the analysis could not be accomplished due to technical error.

## Conclusions

We reproduced the results from the 2020 study "Chromatin binding of FOXA1 is promoted by LSD1-mediated demethylation in prostate cancer" by Gao et al. Regardless of challenges, we were able to produce somewhat similar ChIP-seq FOXA1 data that prove that inhibition of LSD1 affects the expression of FOXA1 and similar with the ATAC-seq data analysis, proving the same. It would have been However, the full analysis could not be accomplished and we were not able to confirm all the claims made the authors. The project was somewhat difficult to perform due to limitations and poor documentation provided by the authors. Also, due to the high usage of the server and high volume of data, it was difficult to generate results that we desired to gather. For example, none of the heatmaps and profiles were generated from the bigwig files and narrowPeak files downloaded from NCBI GEO. If we were to have greater storage and memory, this would have been possible to produce. Another limitation of this study was that the analysis section was poorly

written and the code availability was not available. Granted, code availability is not mandatory, but without it, repeatability of the study becomes extremely difficult. If we were to do this project again, we would choose a study has at least a well-written method section.

## References


(1) https://www.wcrf.org/cancer-trends/worldwide-cancer-data/#:~:text=Breast%20and%20lung%20cancers%20were,con

(2) https://www150.statcan.gc.ca/n1/pub/82-003-x/2019004/article/00002-eng.htm

(3) https://www.nature.com/articles/s41588-020-0681-7?fbclid=IwAR0xgx8q4bxVD_PbEMLk2yaMA6tsMfxdZy3USVEn
K1WY#Sec2