STAT115 Lab 5: ChIP-Seq

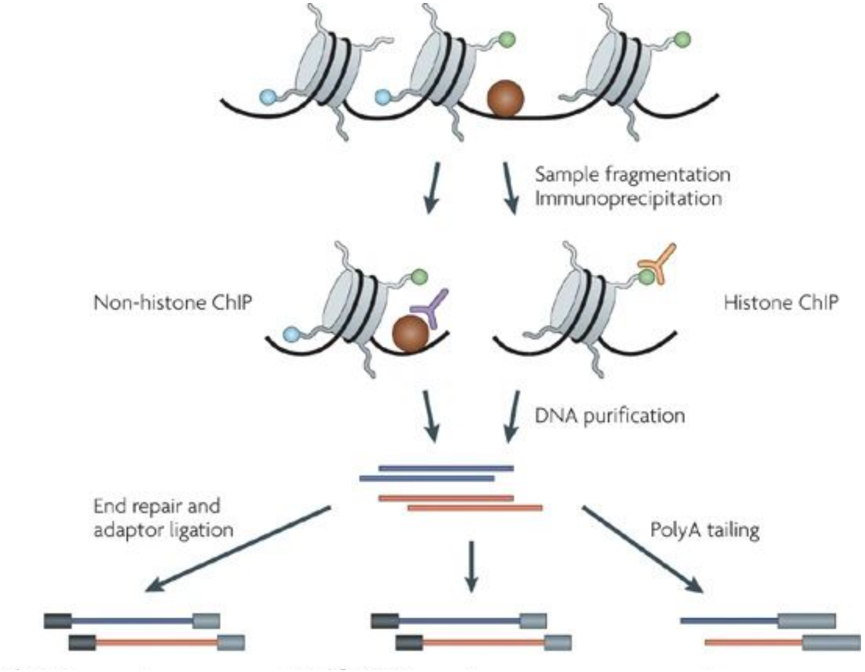
TA: Jiazhen Rong
2020.02.23

*Some Slide Content Adapted from Previous Year's TA Qian Xiao

Several Announcements

- Please fill out HW2 Feedback Survey
- HW3 is due on Mon 3/8, 2021 @ 11:59pm due to wellness day
- Please download the newest HW3 (updated today), do not work on Part V and VI yet, until the datasets are finalized and we will make an announcement.
- #hw3 or #homework-questions channel in Slack for HW3
- 2 Bonus Questions in HW3, each 5 point if your answer is selected for future HW
- We have OH on Friday 4-5pm and Sat 10:30-11:30AM

ChIP-Seq Introduction



Send library for sequencing (Illumina, etc...)

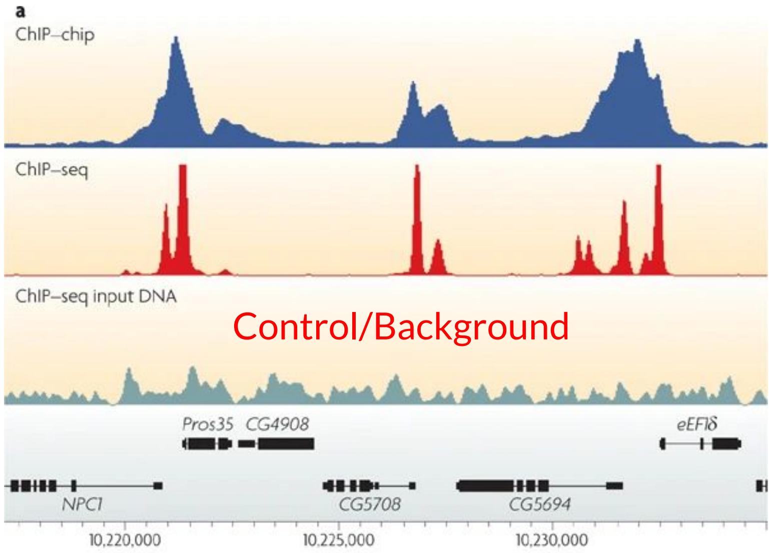
(Park, Nature Reviews Genetics, 2009)

Goal:
Detect DNA-Protein Interactions

Non-histone ChIP:
Usually for finding DNA sites interacting with Transcriptional Factors
(e.g. In HW3, Androgen Receptor ChIP Seq)

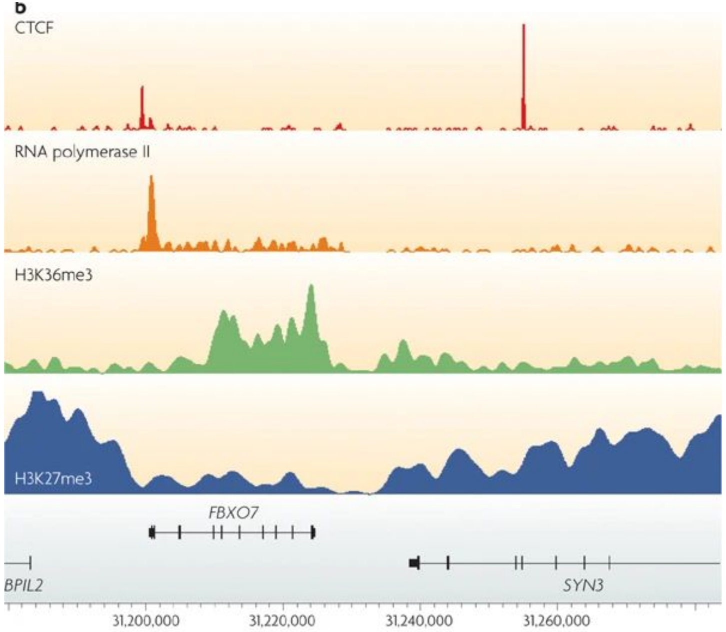
Histone ChIP:
Usually for finding nucleosome modifications.

ChIP-Seq Introduction



Control/Background

Example of ChIP-Seq Signals



Nature Reviews | Genetics

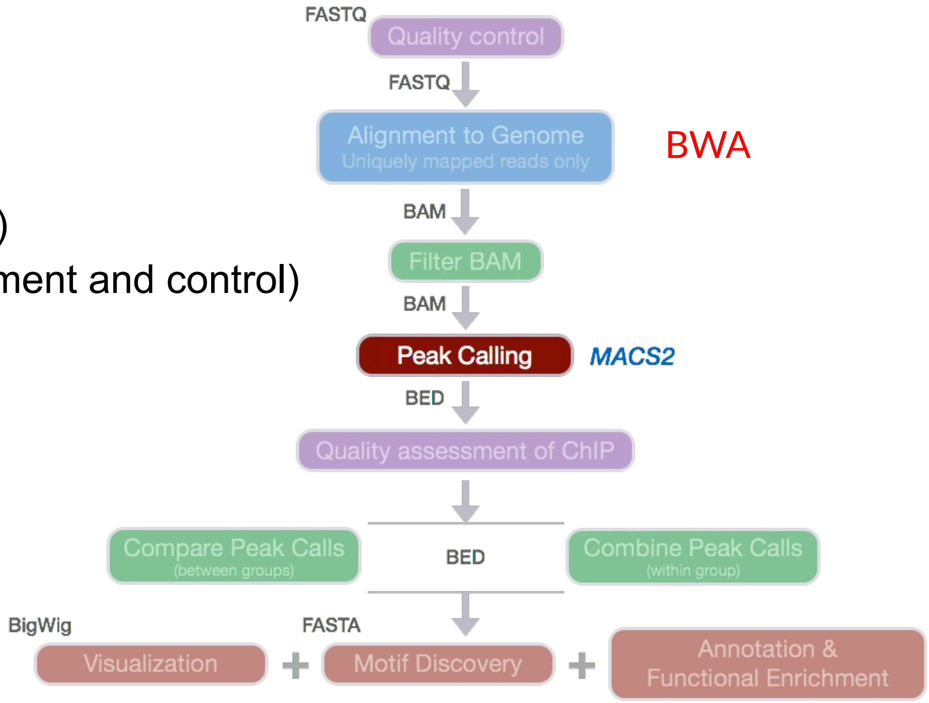
Different ChIP-Seq Profile in Human T-cells

(Park, Nature Reviews Genetics, 2009)

Peak Calling & MACS
(Part I of HW3)

Workflow of ChIP-Seq Analysis:

1. Align reads (Input Fastq output BAM)
2. Remove duplicates (Input BAM output BED)
3. Downsample (balance reads between treatment and control)
4. Call peaks (Input BAM output BED)
5. Visualize peaks (UCSC)
6. Integrate gene expression data (next lab)



FASTQ

Quality control

FASTQ

Alignment to Genome

Uniquely mapped reads only

BWA

BAM

Filter BAM

BAM

Peak Calling

MACS2

BED

Quality assessment of ChIP

Compare Peak Calls

(between groups)

BED

Combine Peak Calls

(within group)

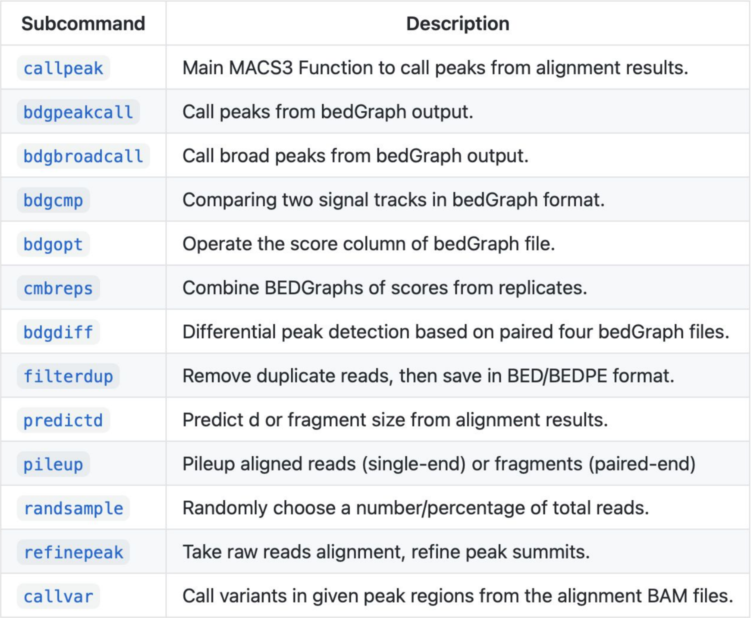
BigWig

FASTA

Visualization + Motif Discovery + Annotation & Functional Enrichment

Picture Credit: https://hbctraining.github.io/Intro-to-ChIPseq/lessons/05_peak_calling_macs.html

MACS (Model-based Analysis of ChIP-Seq)
(Part I of HW3)



Subcommand	Description
`callpeak`	Main MACS3 Function to call peaks from alignment results.
`bdgpeakcall`	Call peaks from bedGraph output.
`bdgbroadcall`	Call broad peaks from bedGraph output.
`bdgcmp`	Comparing two signal tracks in bedGraph format.
`bdgopt`	Operate the score column of bedGraph file.
`cmbreps`	Combine BEDGraphs of scores from replicates.
`bdgdiff`	Differential peak detection based on paired four bedGraph files.
`filterdup`	Remove duplicate reads, then save in BED/BEDPE format.
`predictd`	Predict d or fragment size from alignment results.
`pileup`	Pileup aligned reads (single-end) or fragments (paired-end)
`randsample`	Randomly choose a number/percentage of total reads.
`refinepeak`	Take raw reads alignment, refine peak summits.
`callvar`	Call variants in given peak regions from the alignment BAM files.

MACS GitHub/Documentation Page:
https://github.com/macs3-project/MACS

Key functions relevant with HW3:

1. callpeak -- call peaks from alignment results
2. filterdup -- remove duplicate reads

You can check the detail of each command and what will be the output files by clicking each command in github page.

Bash Script Review

```
#!/bin/bash
#SBATCH -n 1 # Number of cores requested
#SBATCH -N 1 # Ensure that all cores are on one machine
#SBATCH -t 30 # Runtime in minutes
#SBATCH -p serial_requeue # Partition to submit to
#SBATCH --mem=32G # Memory in GB (see also --mem-per-cpu)
#SBATCH -o p1.out # Standard out goes to this file
#SBATCH -e p1.err # Standard err goes to this file
```

You can change time, memory
and your log files' names
in the header section of bash script

```
# LOAD_MODULES
module load <module_name>
```

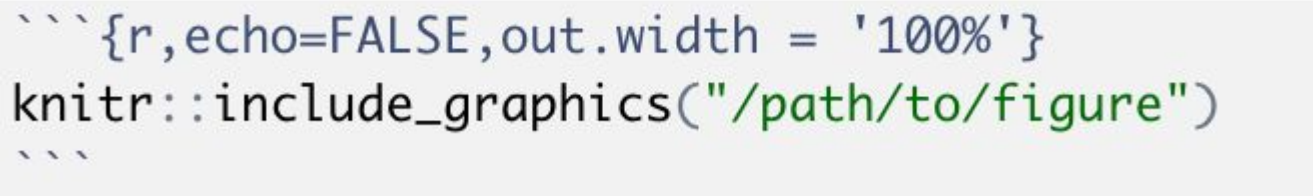
To check available modules, you can use 'module avail'; Or you can use "module spider module_name' to see available versions.

Your code….

To submit a job on slurm server, use 'sbatch your_script'
To check the progress of your job, use 'sacct'
To cancel your job, use 'scancel job_id'

More Documentation

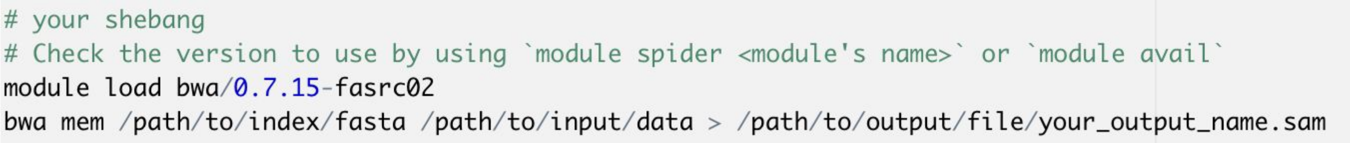
Inserting Figures in .Rmd



~~~
```{r,echo=FALSE,out.width = '100%'}
knitr::include_graphics("/path/to/figure")
```
~~~

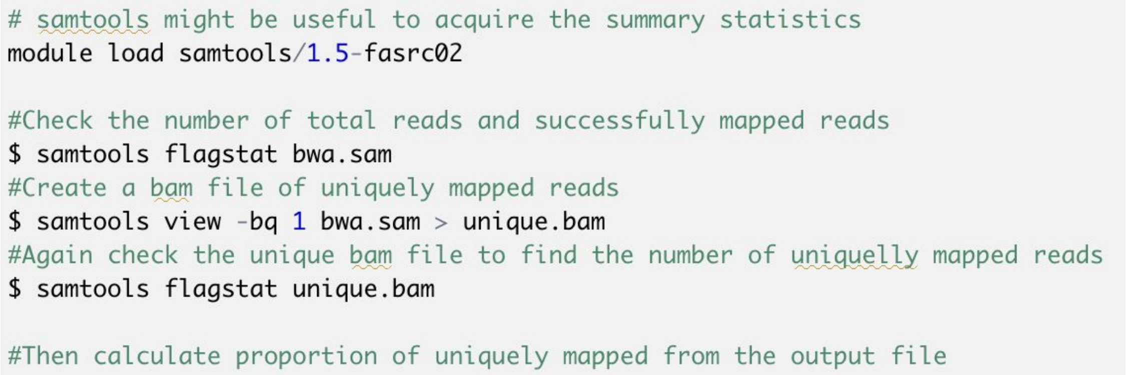
Q1:

Usually we use **BWA** to map the reads to the genome for ChIP-seq experiment. We will give you one example ChIP-seq single-end sequenced .fastq file with only 1M reads. Run BWA on this file to Hg38 of the human genome assembly. Report the **commands, logs files, and a snapshot / screenshot** of the output to demonstrate your alignment procedure. What proportion of the reads are successfully mapped (to find at least one location) and what proportions are uniquely mapped (to find a single location) in the human genome in this test sample? We will save you some time and directly give you the BWA mapped BAM files for the full samples.



```
# your shebang
# Check the version to use by using `module spider <module's name>` or `module avail`
module load bwa/0.7.15-fasrc02
bwa mem /path/to/index/fasta /path/to/input/data > /path/to/output/file/your_output_name.sam
```

Use bwa mem for alignment

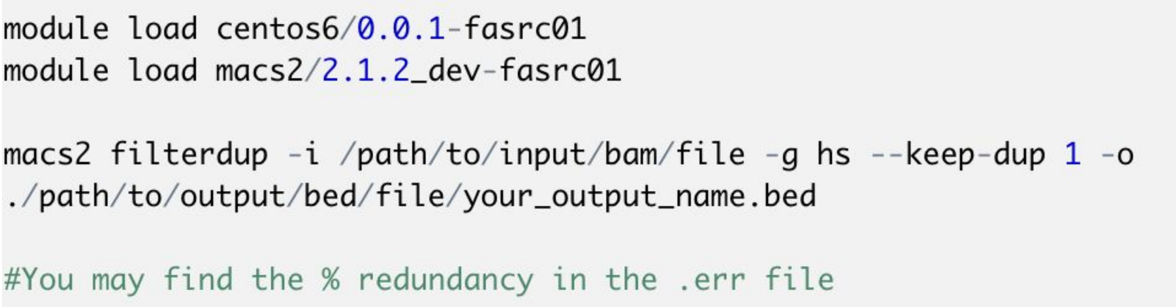


```
# samtools might be useful to acquire the summary statistics
module load samtools/1.5-fasrc02

#Check the number of total reads and successfully mapped reads
$ samtools flagstat bwa.sam
#Create a bam file of uniquely mapped reads
$ samtools view -bq 1 bwa.sam > unique.bam
#Again check the unique bam file to find the number of uniquelly mapped reads
$ samtools flagstat unique.bam

#Then calculate proportion of uniquely mapped from the output file
```

Use samtools for summary statistics

Q2:

In ChIP-Seq experiments, when sequencing library preparation involves a PCR amplification step, it is common to observe multiple reads where identical nucleotide sequences are disproportionally represented in the final results. This is especially a problem in tissue ChIP-seq experiments (as compared to cell lines) when input cell numbers are low. Removing these duplicated reads can improve the peak calling accuracy. Thus, it may be necessary to perform a duplicate read removal step, which flags identical reads and subsequently removes them from the dataset. Run this on your test sample (1M reads) (**macs2 filterdup**). What % of reads are redundant? When doing peak calling, MACS filters duplicated reads by default.

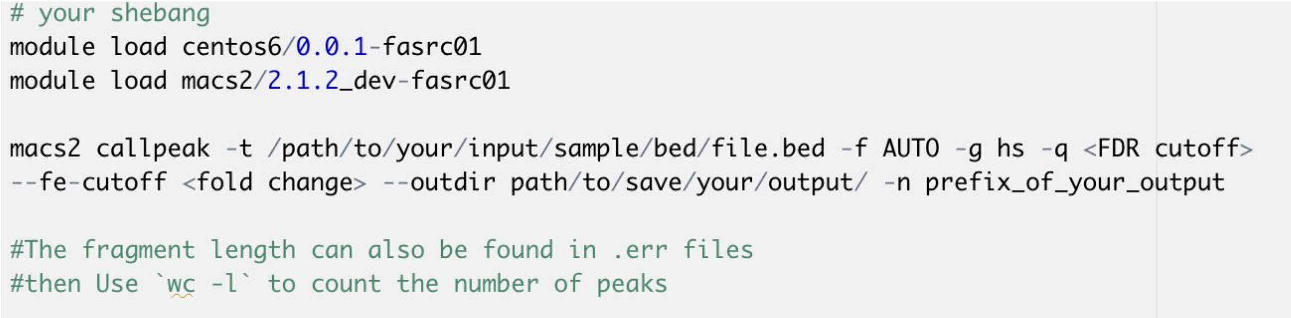


```
module load centos6/0.0.1-fasrc01
module load macs2/2.1.2_dev-fasrc01

macs2 filterdup -i /path/to/input/bam/file -g hs --keep-dup 1 -o
./path/to/output/bed/file/your_output_name.bed

#You may find the % redundancy in the .err file
```

Remove duplicated reads resulted from PCR amplification

Q3:

For many ChIP-seq experiments, usually chromatin input without enriching for the factor of interest is generated as control. However, in this experiment, we only have ChIP (of both tumor and normal) and no control samples. Without control, MACS2 will use the signals around the peaks to infer the chromatin background and estimate the ChIP enrichment over background. In ChIP-seq, + strand reads and – strand reads are distributed to the left and right of the binding site, and the distance between the + strand reads and – strand reads can be used to estimate the fragment length from sonication (note: with PE seq, insert size could be directly estimated). What is the estimated fragment size in each? Use MACS2 to call peaks from tumor1 and normal1 separately. How many peaks do you get from each condition with **FDR < 0.05 and fold change > 5**?



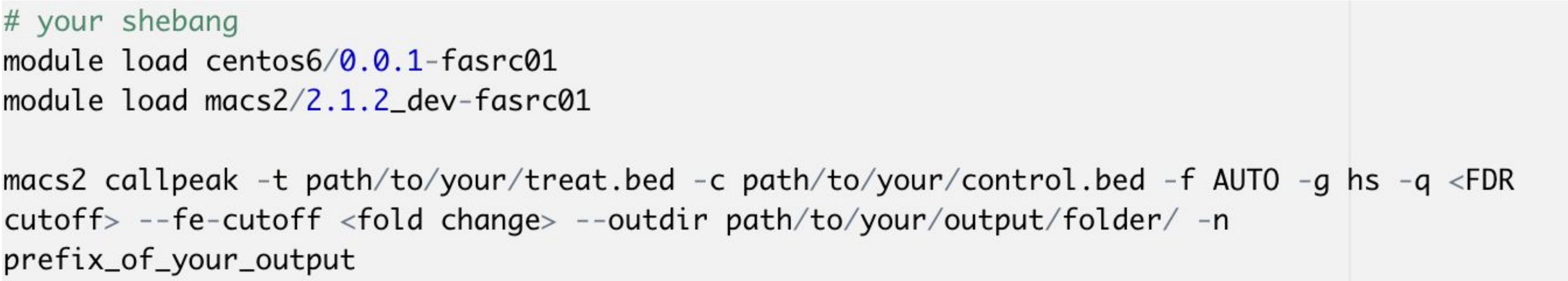
```
# your shebang
module load centos6/0.0.1-fasrc01
module load macs2/2.1.2_dev-fasrc01

macs2 callpeak -t /path/to/your/input/sample/bed/file.bed -f AUTO -g hs -q <FDR cutoff>
--fe-cutoff <fold change> --outdir path/to/save/your/output/ -n prefix_of_your_output

#The fragment length can also be found in .err files
#then Use `wc -l` to count the number of peaks
```

Perform this step for both normal and the tumor sample

-t/--treatment filename, -c/--control, -n/--output name, -f/--format of tag files
--outdir/--the folder where all the output files saved into, -n/--name of the output as NAME_peaks.bed
-g/--gsize The default hs -- 2.7e9 is recommended as for UCSC human hg18 assembly
-q/--qvalue (minimum FDR) cutoff to call significant regions. Default is 0.05.

Q4:

Now we want to see whether AR has differential binding sites between prostate tumors and normal prostates. MACS2 does have a function to call differential peaks between conditions, but requires both conditions to have input control. Since we **don’t have input controls** for these AR ChIP-seq, we will just run the AR tumor ChIP-seq over the AR normal ChIP-seq (pretend the latter to be input control) to find differential peaks. How many peaks do you get with **FDR < 0.01 and fold change > 6**?



```
# your shebang
module load centos6/0.0.1-fasrc01
module load macs2/2.1.2_dev-fasrc01

macs2 callpeak -t path/to/your/treat.bed -c path/to/your/control.bed -f AUTO -g hs -q <FDR
cutoff> --fe-cutoff <fold change> --outdir path/to/your/output/folder/ -n
prefix_of_your_output
```

Setting the values as required in Q4.

QC & Cistrome Data Browser (Part II of HW3)

Q6:

Cistrome Data Browser (http://cistrome.org/db/) has collected and pre-processed a large compendium of the published ChIP-seq data in the public. Play with Cistrome DB. Biological sources indicate whether the ChIP-seq is generated from a cell line (e.g. VCaP, LNCaP, PC3, C4-2) or a tissue (Prostate). Are there over 100 AR ChIP-seq samples which passed all QC meatures in human prostate tissues?

Live Demo in Lab

Q7:

Doing transcription factor ChIP-seq in tissues could be a tricky experiment, so sometimes even published data in high profile journals have bad quality. Look at a few AR ChIP-seq samples in the prostate tissue on Cistrome and inspect their QC reports. Can you comment on what QC measures tell you whether a ChIP-seq is of good or bad quality? Include a screen shot of a good AR ChIP-seq vs a bad AR ChIP-seq.

Live Demo in Lab

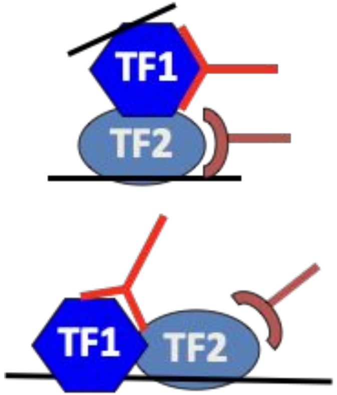
Q8:

Antibody is one important factor influencing the quality of a ChIP-seq experiment. Click on the GEO (GSM) ID of some good quality vs bad quality ChIP-seq data, and see where they got their AR antibodies. If you plan to do an AR ChIP-seq experiment, which company and catalog # would you use to order the AR antibody?

Live Demo in Lab

Motif Finding
(Part III of HW3)

Motif Finding & TF Co-interactions



- Co-interacting TFs and its binding DNA is also pulled down by ChIP-Seq
- By finding Enriched motifs, we can find which TFs might be Co-interacting with the protein pulled down
- Motif Finding methods:
  - Gibbs Sampling, MEME, Consensus ...

Q9:

We want to see in prostate tumors, which other transcription factors (TF) might be collaborating with AR. You can try any of the following motif finding tools to find TF motifs enriched in the differential AR peaks you identified above. Did you find the known AR motif, and motifs of other factors that might interact with AR in prostate cancer in gene regulation? Describe the tool you used, what you did, and what you found. Note that finding the correct AR motif is usually an important criterion for AR ChIP-seq QC as well.

HOMER: http://homer.ucsd.edu/homer/motif/
MEME: http://meme-suite.org/tools/meme-chip
Weeder: http://159.149.160.88/pscan_chip_dev/ -- Weeder is Recommended
Cistrome: http://cistrome.org/ap/root (Register a free account).

Live Demo of Weeder and Cistrome in Lab

Q10:

Look at the AR binding distribution in Cistrome DB from a few good AR ChIP-seq data in prostate. **Does AR bind mostly in the gene promoters, exons, introns, or intergenic regions?** Also, look at the **QC motifs** to see what motifs are enriched in the ChIP-seq peaks. Do you see similar motifs here as those you found in your motif analyses?

Live Demo in Lab

Identify Co-interacting TFs (Part IV of HW3)

Q11:

Sometimes members of the same transcription factor family (e.g. E2F1, 2, 3, 4, 5, etc) have similar binding motifs, significant overlapping binding sites (but they might be expressed in very different tissues), and related functions (they could also have different functions if they interact with different partners or compete for binding to the same sites in the same cell). Therefore, to confirm that we have found the correct TFs interacting with AR in prostate tumors, in addition to looking for motifs enriched in the AR ChIP-seq, we also want to see **whether the TFs are highly expressed in prostate tumor**. For this, we will use the Exploration Component on TIMER (http://timer.cistrome.org/) or GEPIA (http://gepia2.cancer-pku.cn/#general). First, look at differential expression of genes in tumors. Based on the top non-AR motifs you found before, see which member of the TF family that recognizes the motif is highly expressed in prostate tissues or tumors. Another way is to see whether the TF family member and AR have correlated expression pattern in prostate tumors. Enter AR as your interested gene and another gene which is the potential AR collaborator based on the motif, and see whether the candidate TF is correlated with AR in prostate tumors. Based on the motif and expression evidences, which factor in each motif family is the most likely collaborator of AR in prostate cancer?

Live Demo of TIMER in Lab

Note: When we conduct RNA-seq on prostate tumors, each tumor might contain cancer cells, normal prostate epithelia cells, stromal fibroblasts, and other immune cells. Therefore, genes that are highly expressed in cancer cells (including AR) could be correlated in different tumors simply due to the tumor purity bias. Therefore, when looking for genes correlated with AR just in the prostate cancer cells, we should correct this tumor purity bias.

Q12:

Besides looking for motif enrichment, another way to find TFs that might interact with AR is to see whether there are other TF ChIP-seq data which have **significant overlap** with AR ChIP-seq. Take the differential AR ChIP-seq peaks (in .bed format) between tumor / normal, and run this on the Cistrome Toolkit (http://dbtoolkit.cistrome.org/). The third function in Cistrome Toolkit looks through all the ChIP-seq data in CistromeDB to find ones with significant overlap with your peak list. You should see AR enriched in the results (since your input is a list of AR ChIP-seq peaks after all). What other factors did you see enriched? Do they agree with your motif analyses before?

Live Demo of Cistrome Toolkit in Lab

2 Bonus Questions

1. Find a dataset with batch-effect and re-run HW2 Part I analysis
2. Rewrite all HW2 Part II’s questions in sklearn.

Q&A Time