Regulatory Genomics Lab

Part 1: ChIP-seq peak calling

Part 2. Analyzing ChIP-seq peaks

Saurabh Sinha

PowerPoint by Saba Ghaffari
Edited by Brianna Bucknor & Giovanni Madrigal

Part 1: ChIP–Seq Peak Calling

Data set:

- 2 cell lines: ”G1E” and “G1E_ER4”
- 2 data sets for each cell line: ChIP-seq for CTCF and control

Steps:

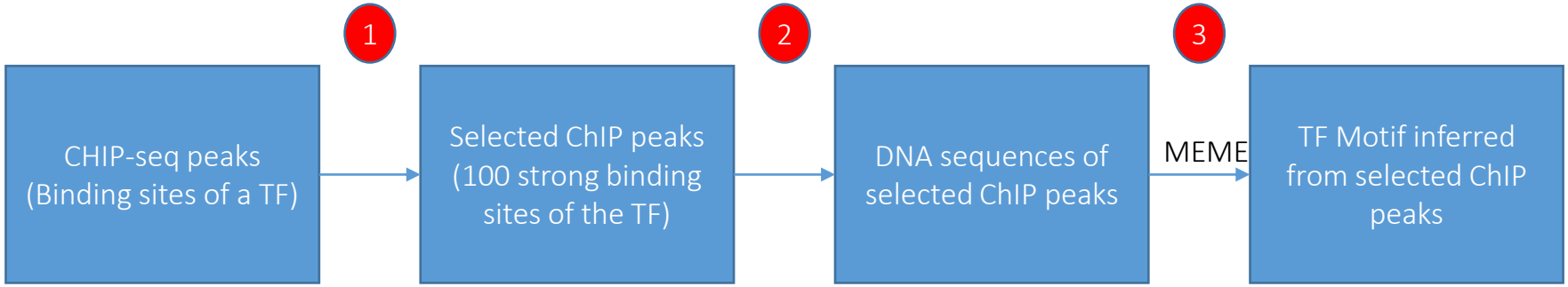
1. Start with a data set with ChIP-seq reads, do quality control
2. Map (align) reads to a reference genome using **Bowtie2.**
3. Call peaks from aligned reads using **MACS2**.
4. Identify ChIP peaks that differ between G1E_ER4 (stimulated) and G1E (un-stimulated) cell lines

Variations on the theme:

- Peak calling on ChIP-seq data with and without using corresponding control data

Part 2: Analyzing ChIP–Seq Peaks

Use command line tools to manipulate a ChIP-seq peak set for TF called “BIN” in D. melanogaster



Subject peak sets to MEME suite for computational motif discovery

In the end, compare MEME-reported motifs with Fly Factor Survey motifs for BIN TF

Part 1: ChIP–Seq Peak Calling

Slides by Shayan Tabe Bordbar

Introduction

This goals of the lab are as follows:

1. Learn how to map Next Generation Sequencing (NGS) reads to a reference genome using **Bowtie2.**

2. Demonstrate how to call peaks from aligned reads (in SAM format) using **MACS2**.

Start the VM

- Follow instructions for starting VM (This is the Remote Desktop software).

- The instructions are different for UIUC and Mayo participants.

- Find the instructions for this on the course website under Lab Set-up: https://publish.illinois.edu/compgenomicscourse/2022-schedule/

Step 0A: Accessing the IGB Biocluster

Open **MobaXterm**

1. Select **Session**

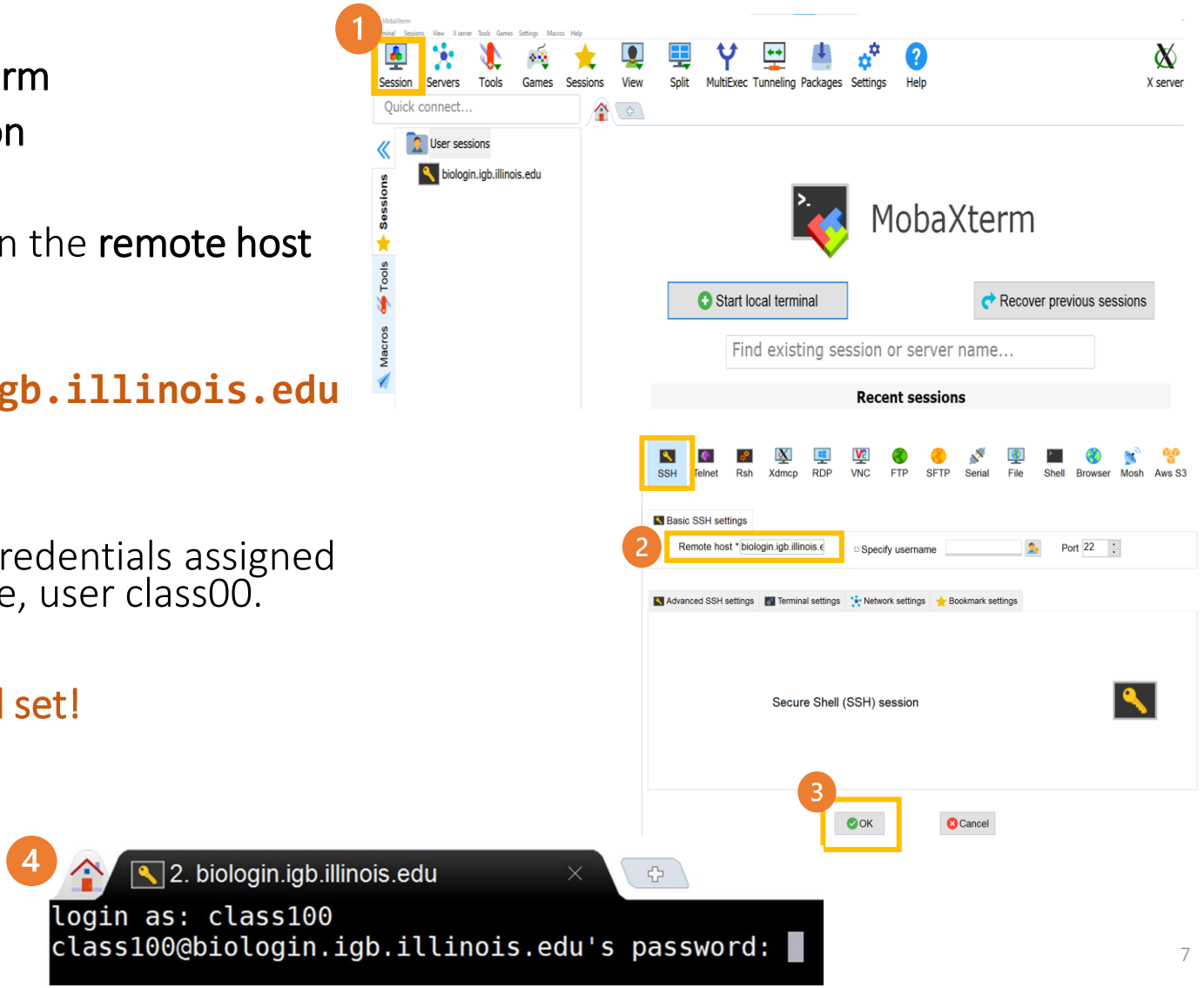
2. Under **SSH**, in the **remote host** textbox type:

`biologin.igb.illinois.edu`

3. Click **OK**

4. Enter login credentials assigned to you; example, user class00.

Now you are all set!



Step 0B: Lab Setup

The lab is located in the following directory:

/home/classroom/mayo/2020/05_Epigenomics/

Following commands will copy a shell script -designed to prepare the working directory- to your home directory. Follow these steps to copy and then submit the script as a job to biocluster:

```
$ cd ~/

# Note ~ is a symbol in Unix paths referring to your home directory

$ cp /home/classroom/mayo/2020/05_Epigenomics/src/prep-directory.sh ./

# Copies prep-directory.sh script to your working directory.

$ sbatch prep-directory.sh

# submits a job to biocluster to populate your home directory with necessary
files

$ squeue –u <userID> # to check the status of the submitted job
```

This is the same as your login. Do not include < >
ex: `$ squeue –u classxx`

Step 0C: Working directory: data

Navigate to the created directory for this exercise and look what data folder contains.

```
$ cd 05_Epigenomics

$ ls

# output should be:

# data results src

$ ls data/

$ ls data/index
# output should be:
# mm9.1.bt2 mm9.2.bt2 mm9.3.bt2  mm9.4.bt2
# mm9.rev.1.bt2  mm9.rev.2.bt2  mm9.zip
```

Note: G1E cell lines are erythroid, red blood cell, cell lines missing the GATA-1 gene.

GATA-1 is crucial for the maturation of erythroid cells.

G1E_E4R cell lines conditionally express GATA-1 in the presence of estradiol, enabling erythroid maturation.

Filename	Description
G1E_ER4_CTCF_chr19.fastqsanger	A sample ChIP-seq dataset on CTCF in G1E_ER4 cells, reads have been reduced to those mapping to chr19 for demonstration use.
G1E_ER4_input_chr19.fastqsanger	Control DNA taken from chr19.
G1E_CTCF.fastqsanger	CTCF Chip for G1E line.
G1E_input.fastqsanger	Control for G1E line.

Step 0D: Working directory: scripts

Navigate to the directory containing the scripts and look what's inside.

```
$ cd src
$ ls *.sh
# lists the scripts to be used in this lab:
#      fastx_summary.sh
#      run_bowtie2.sh
#      run_macs2_noControl.sh      run_macs2_withControl.sh    run_macs2_noER.sh
#      bedtools_overlap_1.sh  bedtools_overlap_2.sh
#      bedtools_subtract_1.sh  bedtools_subtract_2.sh
```

Read Mapping and Peak Calling

In this exercise, we will map ChIP Reads to a reference genome using **Bowtie2** and call peaks among the mapped reads using **MACS2**.

Step 1: FASTQ Summary Statistics

In this step, we will gather summary statistics of ChIP data for quality control.

We use FASTX-Toolkit to get statistics on the quality and content of each column of fastq files (sequencing reads).

"fastx_quality_stats" is the name of the tool used from FASTX-Toolkit.

How to Use [Do NOT run the following commands]:

```
$ fastx_quality_stats    -i <input.fastq> -o <output_summary.txt>
```

fastx_summary.sh uses fastx_quality_stats to get summary reports for all four provided fastq files.
RUN the following command:

```
$ cd ~/05_Epigenomics/src/

$ sbatch fastx_summary.sh

# OUTPUT in ~/05_Epigenomics/results/

$ squeue –u <userID> # to check the status of the submitted job
```

This is the same as your login. Do not include < >
ex: `$ squeue –u classxx`

Please do not try to Run the commands in this slide. This is just to explain what the script that we just ran (fastx_summary.sh) is supposed to do in more detail.

What's inside the **fastx_summary.sh** script?

```
#!/bin/bash
#SBATCH -c 4
#SBATCH --mem 8000
#SBATCH -A Mayo_Workshop
#SBATCH -J fastx_summ
#SBATCH -o fastx_summ.%j.out
#SBATCH -e fastx_summ.%j.err
#SBATCH -p classroom


# load the tool environment
module load FASTX-Toolkit

# this is our input (fastq)
export FASTQ_1=../data/G1E_CTCF.fastqsanger
export FASTQ_2=../data/G1E_input.fastqsanger
export FASTQ_3=../data/G1E_ER4_CTCF_chr19.fastqsanger
export FASTQ_4=../data/G1E_ER4_input_chr19.fastqsanger
# this is our output (summaries)
export OUT_1=../results/G1E_CTCF_summary.txt
export OUT_2=../results/G1E_input_summary.txt
export OUT_3=../results/G1E_ER4_CTCF_chr19_summary.txt
export OUT_4=../results/G1E_ER4_input_chr19_summary.txt

fastx_quality_stats -i $FASTQ_1  -o $OUT_1
fastx_quality_stats -i $FASTQ_2  -o $OUT_2
fastx_quality_stats -i $FASTQ_3  -o $OUT_3
fastx_quality_stats -i $FASTQ_4  -o $OUT_4
```

Tells the cluster 'job manager' what resources you want (4 CPUs, 8GB memory, run on the 'classroom' nodes, and name the job 'fastx_summ'

Load the software. We are using a tool called 'FASTX-Toolkit' to generate some basic stats on the fastq files.

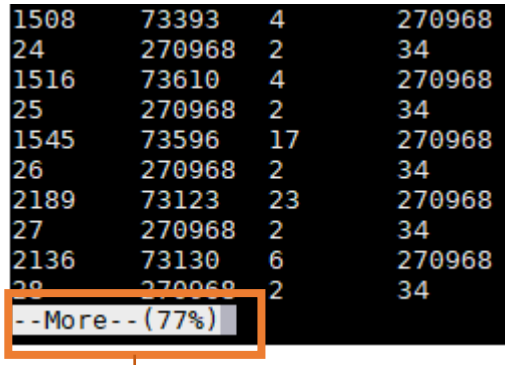
Create shortcut names for input and output files.

These commands, not to be run by you, execute the fastx_quality_score tool on all four input fastq files.

Step 1: FASTQ Summary Statistics

- To view one of the summary files use:

```
$ more ~/05_Epigenomics/results/G1E_CTCF_summary.txt

# shows beginning of G1E_CTCF_summary.txt file

# To get the length of the reads use:

$ cat ~/05_Epigenomics/results/G1E_CTCF_summary.txt | wc -l

# you should get 37
```



Note: This is a lowercase L, not the #1

Tip:

Pressing **enter** will make more lines appear. Pressing **"d"** will allow you to exit out of seeing the summary stats and enter the next line of code.

Discussion

- How long are these reads?
- What is the median quality at the last position?
- View and explore other summary files:
  - G1E_CTCF_summary.txt
  - G1E_input_summary.txt
  - G1E_ER4_CTCF_chr19_summary.txt
  - G1E_ER4_input_chr19_summary.txt

Step 2: Map ChIP-Seq Reads to MM9 Genome

Next, we will map the reads in **G1E_E4R_CTCF_chr9.fastqsanger** to the mouse genome using **Bowtie2.**

Please do not try to Run the commands in the first box. This is just to explain the arguments to bowtie2

usage:

```
bowtie2 [options]    -x <base name of index files> \
                     -U <in_file_name> \
                     -S <out_file_name.sam>
                     --sensitive
```

The index files are available on Biocluster and you do not need to download them now. However it can be downloaded from [Please **do not** download now]:
ftp://ftp.ccb.jhu.edu/pub/data/bowtie2_indexes/mm9.zip
Script run_bowtie2.sh uses bowtie2 on all four input fastq files and maps them to mm9 genome.

```
$ cd ~/05_Epigenomics/src/

$ sbatch run_bowtie2.sh

# OUTPUT in ~/05_Epigenomics/results/Bowtie_output

$ squeue –u <userID> # to check the status of the submitted job
```

There are other parameters that can be specified for a more controlled use of Bowtie2. In particular, following are some preset options that can be used to modify the speed and sensitivity of the tool:

--very-fast

--fast

--sensitive (default)

--very-sensitive

More information on the Bowtie2 can be found in its well-written manual:

http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml

Please do not try to Run the commands in this slide. This is just to explain what the script that we just ran (run_bowtie2.sh) is supposed to do in more detail.

What's inside the **run_bowtie2.sh** script?

```
#!/bin/bash
#SBATCH -c 4
#SBATCH --mem 8000
#SBATCH -A Mayo_Workshop
#SBATCH -J bowtie2_test
#SBATCH -o bowtie2_test.%j.out
#SBATCH -e bowtie2_test.%j.err
#SBATCH -p classroom
# load the tool environment
module load Bowtie2/
# this is our input (fastq)
export BOWTIE_INP_1=../data/G1E_ER4_CTCF_chr19.fastqsanger
export BOWTIE_INP_2=../data/G1E_ER4_input_chr19.fastqsanger
export BOWTIE_INP_3=../data/G1E_CTCF.fastqsanger
export BOWTIE_INP_4=../data/G1E_input.fastqsanger
mkdir -p ../results/Bowtie_output
# this is our output (SAM)
export BOWTIE_OUT_1=../results/Bowtie_output/G1E_ER4_CTCF_chr19.sam
export BOWTIE_OUT_2=../results/Bowtie_output/G1E_ER4_input_chr19.sam
export BOWTIE_OUT_3=../results/Bowtie_output/G1E_CTCF_chr19.sam
export BOWTIE_OUT_4=../results/Bowtie_output/G1E_input_chr19.sam
# PATH TO the mm9 dm3 index file along with their common prefix (mm9)
export GENOME_MM9_BOWTIE2=../data/index/mm9

bowtie2 -q -x $GENOME_MM9_BOWTIE2 -U $BOWTIE_INP_1 -S $BOWTIE_OUT_1
bowtie2 -q -x $GENOME_MM9_BOWTIE2 -U $BOWTIE_INP_2 -S $BOWTIE_OUT_2
bowtie2 -q -x $GENOME_MM9_BOWTIE2 -U $BOWTIE_INP_3 -S $BOWTIE_OUT_3
bowtie2 -q -x $GENOME_MM9_BOWTIE2 -U $BOWTIE_INP_4 -S $BOWTIE_OUT_4
```

Tells the cluster 'job manager' what resources you want (4 CPUs, 8GB memory, run on the 'classroom' nodes, and name the job 'bowtie2_test'

Load the software. We are using a tool called 'Bowtie2' to Map fastq files to mm9 mouse genome.

Create shortcut names for input and output files, as well as mm9 genome index files' prefix.

Run bowtie2 tool on all four input fastq files

Step 2: Map ChIP-Seq Reads to MM9 Genome

```
$ head -40 ~/05_Epigenomics/results/Bowtie_output/G1E_ER4_CTCF_chr19.sam

# view the first 40 lines of an output SAM file
```

- View and explore other SAM files:
  - G1E_CTCF_chr19.sam
  - G1E_input_chr19.sam
  - G1E_ER4_CTCF_chr19.sam
  - G1E_ER4_input_chr19.sam

You can find more information on the structure of SAM files in the following links:

https://www.samformat.info/sam-format-flag

https://en.wikipedia.org/wiki/SAM_%28file_format%29

Step 3A: Calling Peaks with MACS2

With our mapped ChiP-Seq reads, we now want to call peaks.

We use **MACS2** for this purpose.

usage:

Please do not try to Run the commands in the following box. This is just to explain the arguments to macs2

```
macs2 callpeak     -t <path_to_treatment_input_alignment> \
                   -c < path_to_control_input_alignment > \ # optional
                   -g <effective_genome_size> \ # use mm for mouse
                   -n <prefix_for_naming_output_files>
```

A useful tutorial for MACS2 can be found here:
https://hbctraining.github.io/Intro-to-ChIPseq/lessons/05_peak_calling_macs.html

Step 3A: Calling Peaks with MACS2

Script run_macs2_noControl.sh runs MACS2 to call peaks for G1E_ER4_CTCF_chr19.sam with the default parameters.

Note that this macs2 run is performed without using input from control experiment.

```
$ cd ~/05_Epigenomics/src/

$ sbatch run_macs2_noControl.sh

# OUTPUT in ~/05_Epigenomics/results/MACS2_output/CTCF_ER4_noControl*

$ squeue –u <userID> # to check the status of the submitted job
```

Please do not try to Run the commands in this slide. This is just to explain what the script that we just ran (run_macs2_noControl.sh) is supposed to do in more detail.

What's inside the **run_macs2_noControl.sh** script?

```
#!/bin/bash
#SBATCH -c 1
#SBATCH --mem 8000
#SBATCH -A Mayo_Workshop
#SBATCH -J macs2_noC
#SBATCH -o macs2_noC.%j.out
#SBATCH -e macs2_noC.%j.err
#SBATCH -p classroom
```

Tells the cluster 'job manager' what resources you want (1 CPU, 8GB memory, run on the 'classroom' nodes, and name the job 'macs2_noC'

```
# load the tool environment
module load MACS2/2.1.2-IGB-gcc-4.9.4-Python-2.7.13
```

Load the software. We use a tool called 'MACS2' to call ChIP peaks.

```
# this is our input (SAM)
export MACS_TREAT=../results/Bowtie_output/G1E_ER4_CTCF_chr19.sam

# this is our output_directory
export MACS_OUT_DIR=../results/MACS2_output
# this is our output prefix
export MACS_OUT_1=CTCF_ER4_noControl
```

Create shortcut names for input alignment file, output directory, and output prefix.

```
macs2 callpeak -t $MACS_TREAT  -g mm -f SAM --outdir $MACS_OUT_DIR -n $MACS_OUT_1
```

This command, not to be run by you directly, executes the MACS2 tool on G1E_ER4_CTCF_chr19.sam without using the control experiment

Step 3A: Calling Peaks with MACS2

Number of peaks called without using a control experiment input can be obtained using:

```
$ cat ../results/MACS2_output/CTCF_ER4_noControl_peaks.narrowPeak | wc -l

# You should get 626

$ head ../results/MACS2_output/CTCF_ER4_noControl_peaks.narrowPeak
```

Here are the fields (columns) of a .narrowPeak file:

1. chromosome
2. start coordinate
3. end coordinate
4. name
5. score
6. strand

Standard BED file fields

7. **signalValue - Measurement of overall enrichment for the region**
8. **pValue - Statistical significance (-log10)**
9. **qValue - Statistical significance using false discovery rate (-log10)**
10. **peak - Point-source called for this peak; 0-based offset from chromStart**

narrowPeak specific fields

screenshot from:
https://hbctraining.github.io/Intro-to-ChIPseq/lessons/05_peak_calling_macs.html

Call ChIP-Seq Peaks with a Control Sample

We will perform the same procedure we did in the previous exercise. This time though, we will work with a control sample in addition to the treated one.

Step 3B: Calling Peaks with MACS2 using Control Chip-Seq Reads

Script run_macs2_noControl.sh runs MACS2 to call peaks for G1E_ER4_CTCF_chr19.sam with the default parameters.

Note that this macs2 run is performed using additional input from control experiment (G1E_ER4_input_chr19.sam).

```
$ cd ~/05_Epigenomics/src/

$ sbatch run_macs2_withControl.sh

# OUTPUT in ~/05_Epigenomics/results/MACS2_output/CTCF_ER4_withControl*

$ squeue –u <userID> # to check the status of the submitted job
```

Number of peaks called using the additional control experiment input can be obtained using:

```
$ cat ../results/MACS2_output/CTCF_ER4_withControl_peaks.narrowPeak | wc -l

# You should get 528
```

Please do not try to Run the commands in this slide. This is just to explain what the script that we just ran (run_macs2_withControl.sh) is supposed to do in more detail.

What's inside the **run_macs2_withControl.sh** script?

```
#!/bin/bash
#SBATCH -c 1
#SBATCH --mem 8000
#SBATCH -A Mayo_Workshop
#SBATCH -J macs2_wC
#SBATCH -o macs2_wC.%j.out
#SBATCH -e macs2_wC.%j.err
#SBATCH -p classroom
```

Tells the cluster 'job manager' what resources you want (1 CPU, 8GB memory, run on the 'classroom' nodes, and name the job 'macs2_wC'

```
# load the tool environment
module load MACS2/2.1.2-IGB-gcc-4.9.4-Python-2.7.13
```

Load the software. We use a tool called 'MACS2' to call ChIP peaks.

```
# this is our input (SAM)
export MACS_TREAT=../results/Bowtie_output/G1E_ER4_CTCF_chr19.sam
export MACS_CONTROL=../results/Bowtie_output/G1E_ER4_input_chr19.sam
# this is our output directory
export MACS_OUT_DIR=../results/MACS2_output
# this is our output prefix
export MACS_OUT_1=CTCF_ER4_withControl
```

Create shortcut names for treatment and control input alignment files, output directory, and output prefix.

```
macs2 callpeak -t $MACS_TREAT -c $MACS_CONTROL -g mm -f SAM –outdir \
$MACS_OUT_DIR  -n $MACS_OUT_1
```

This command, not to be run by you directly, executes the MACS2 tool on G1E_ER4_CTCF_chr19.sam while using G1E_ER4_input_chr19.sam as the control experiment.

Step 3C: Calling Peaks with MACS2 on Chip-Seq Reads for un-stimulated cells

Script run_macs2_noER.sh runs MACS2 to call peaks for G1E_CTCF_chr19.sam with the default parameters.

Note that this macs2 run is performed using additional input from control experiment (G1E_input_chr19.sam).

```
$ cd ~/05_Epigenomics/src/

$ sbatch run_macs2_noER.sh

# OUTPUT in ~/05_Epigenomics/results/MACS2_output/CTCF_noE2_*

$ squeue –u <userID> # to check the status of the submitted job
```

Exercise:
Find out the number of peaks called for this ChIP-Seq experiment

Please do not try to Run the commands in this slide. This is just to explain what the script that we just ran (run_macs2_noER.sh) is supposed to do in more detail.

What's inside the **run_macs2_noER.sh** script?

```
#!/bin/bash
#SBATCH -c 4
#SBATCH --mem 8000
#SBATCH -A Mayo_Workshop
#SBATCH -J macs2_noER
#SBATCH -o macs2_noER.%j.out
#SBATCH -e macs2_noER.%j.err
#SBATCH -p classroom

# load the tool environment
module load MACS2/2.1.2-IGB-gcc-4.9.4-Python-2.7.13

# this is our input (SAM)
export MACS_TREAT_NOER=../results/Bowtie_output/G1E_CTCF_chr19.sam
export MACS_CONTROL_NOER=../results/Bowtie_output/G1E_input_chr19.sam
# this is the output directory
export MACS_OUT_DIR=../results/MACS2_output
# this is our output prefix
export MACS_OUT_1=CTCF_noE2


macs2 callpeak -t $MACS_TREAT_NOER -c $MACS_CONTROL_NOER -g mm -f SAM –outdir \
$MACS_OUT_DIR  -n $MACS_OUT_1
```

Tells the cluster 'job manager' what resources you want (1 CPU, 8GB memory, run on the 'classroom' nodes, and name the job 'macs2_noER'

Load the software. We use a tool called 'MACS2' to call ChIP peaks.

Create shortcut names for treatment and control input alignment files, output directory, and output prefix.

This command, not to be run by you directly, executes the MACS2 tool on G1E_CTCF_chr19.sam while using G1E_input_chr19.sam as the control experiment.

MACS2 summary

MACS2 creates two output files:
_peaks.narrowPeak: BED6+4 format file which contains the peak locations together with peak summit, pvalue and qvalue.
_peaks.xls: a tabular file which contains information about called peaks. Additional information includes pileup and fold enrichment

Discussion

1. Examine the **BED** tracks.
2. How many peaks are called when using a control sample?
3. How does this compare to the previous situation where we only had experimental Chip-Seq reads?

Identifying Differential Binding Sites

In this exercise, we will identify binding sites exclusive to undifferentiated and differentiated cell lines as well as those common to both, using “bedtools” toolkit.

Step 4A: Subtract Peaks Between Cell Lines

we will use “bedtools intersect” tool from bedtools toolkit to identify CTCF peaks that are unique to the differentiated cell line:

Usage:

Please do not try to Run the commands in the following box. This is just to explain the arguments to bedtools

```
$ bedtools intersect[options]  -a <first_interval.bed> \

                               -b <second_interval.bed>
```

Specific options determine the behaviour of bedtools intersect, e.g.

-wa Write the original entry in A for each overlap.

-wb Write the original entry in B for each overlap.

-v Only report those entries in A that have no overlap in B.

As an example, following command finds entries in A.bed that are absent in B.bed:

Please do not try to Run the commands in the following box. This is just to show an example.

```
$ bedtools intersect -v -a A.bed -b B.bed
```

Here is a link to the manual for bedtools intersect:
https://bedtools.readthedocs.io/en/latest/content/tools/intersect.html

Step 4A: Subtract Peaks Between Cell Lines.

Use the following command to find peaks in E2 treated cells that are absent in untreated cells:

```
$ cd ~/05_Epigenomics/src/

$ sbatch bedtools_subtract_1.sh

# OUTPUT in ~/05_Epigenomics/results/peak_inspection/CTCF_subtract_1.bed

$ squeue –u <userID> # to check the status of the submitted job
```

The resulting **BED** file (CTCF_subtract_1.bed) contains peaks exclusive to the **differentiated** cell line (G1E-ER4).

Discussion

1. How many peaks are exclusive to G1E-ER4?

```
$ cat ~/05_Epigenomics/results/peak_inspection/CTCF_subtract_1.bed | wc -l

# You should get 136
```

Please do not try to Run the commands in this slide. This is just to explain what the script that we just ran (bedtools_subtract_1.sh) is supposed to do in more detail.

What's inside the **bedtools_subtract_1.sh** script?

```
#!/bin/bash
#SBATCH -c 1
#SBATCH --mem 8000
#SBATCH -A Mayo_Workshop
#SBATCH -J bedtools_subt1
#SBATCH -o bedtools_subt1.%j.out
#SBATCH -e bedtools_subt1.%j.err
#SBATCH -p classroom

# load the tool environment
module load BEDTools

# this is our input (bed like)
export PEAK_1=../results/MACS2_output/CTCF_ER4_withControl_peaks.narrowPeak
export PEAK_2=../results/MACS2_output/CTCF_noE2_peaks.narrowPeak

mkdir -p ../results/peak_inspection
# this is our output (bed like)
export OUT_1=../results/peak_inspection/CTCF_subtract_1.bed


bedtools intersect -v -a $PEAK_1 -b $PEAK_2 > $OUT_1
```

Tells the cluster 'job manager' what resources you want (1 CPU, 8GB memory, run on the 'classroom' nodes, and name the job 'bedtools_subt1'

Load the software. We use a tool called 'BEDTools' to work with generated peak files.

Create shortcut names for the two input bed files, and the output bed file.

run 'bedtools intersect' using the –v flag to get the difference between the two bed files, and store the results in CTCF_subtract_1.bed

Step 4A: Subtract Peaks Between Cell Lines

Redo Step1 only **SWITCH** the input order to get the peaks unique to the untreated cells.

use the following command to do just that:

```
$ cd ~/05_Epigenomics/src/

$ sbatch bedtools_subtract_2.sh

# OUTPUT in ~/05_Epigenomics/results/peak_inspection/CTCF_subtract_2.bed
```

The resulting **BED** file (CTCF_subtract_2.bed) contains peaks exclusive to the **undifferentiated** cell line (G1E).

Exercise:
How many peaks are exclusive to the undifferentiated cell line?

```
$ cat ~/05_Epigenomics/results/peak_inspection/CTCF_subtract_2.bed | wc -l

# You should get 23
```

Please do not try to Run the commands in this slide. This is just to explain what the script that we just ran (bedtools_subtract_2.sh) is supposed to do in more detail.

What's inside the **bedtools_subtract_2.sh** script?

```
#!/bin/bash
#SBATCH -c 1
#SBATCH --mem 8000
#SBATCH -A Mayo_Workshop
#SBATCH -J bedtools_subt2
#SBATCH -o bedtools_subt2.%j.out
#SBATCH -e bedtools_subt2.%j.err
#SBATCH -p classroom

# load the tool environment
module load BEDTools

# this is our input (bed like)
export PEAK_1=../results/MACS2_output/CTCF_ER4_withControl_peaks.narrowPeak
export PEAK_2=../results/MACS2_output/CTCF_noE2_peaks.narrowPeak

mkdir -p ../results/peak_inspection
# this is our output (bed like)
export OUT_1=../results/peak_inspection/CTCF_subtract_2.bed


bedtools intersect -v -a $PEAK_2 -b $PEAK_1 > $OUT_1
```

Tells the cluster 'job manager' what resources you want (1 CPU, 8GB memory, run on the 'classroom' nodes, and name the job 'bedtools_subt2'

Load the software. We use a tool called 'BEDTools' to work with generated peak files.

Create shortcut names for the two input bed files, and the output bed file.

This command, not to be run by you directly, executes the 'bedtools intersect' using the –v flag to get the difference between the two bed files, and store the results in CTCF_subtract_2.bed

Step 4B: Intersect Peaks Between Cell Lines

Following command finds entries in A.bed that overlap with at least one entry in B.bed:

Please do not try to Run the command in the following box. This is just to show an example.

```
$ bedtools intersect –wa –a A.bed –b B.bed
```

Use the following command to find peaks in E2 treated cells that overlap with peaks in the untreated cells:

```
$ cd ~/05_Epigenomics/src/

$ sbatch bedtools_overlap_1.sh

# OUTPUT in ~/05_Epigenomics/results/peak_inspection/CTCF_overlap_1.bed
```

The resulting **BED** file (CTCF_overlap_1.bed) contains peaks from the differentiated cell line (G1E_ER4) that overlap with peaks in the undifferentiated cell line (G1E).

Please do not try to Run the commands in this slide. This is just to explain what the script that we just ran (bedtools_overlap_1.sh) is supposed to do in more detail.

What's inside the **bedtools_overlap_1.sh** script?

```
#!/bin/bash
#SBATCH -c 4
#SBATCH --mem 8000
#SBATCH -A Mayo_Workshop
#SBATCH -J bedtools_ovl1
#SBATCH -o bedtools_ovl1.%j.out
#SBATCH -e bedtools_ovl1.%j.err
#SBATCH -p classroom

# load the tool environment
module load BEDTools

# this is our input (bed like)
export PEAK_1=../results/MACS2_output/CTCF_ER4_withControl_peaks.narrowPeak
export PEAK_2=../results/MACS2_output/CTCF_noE2_peaks.narrowPeak
mkdir -p ../results/peak_inspection
# this is our output (bed like)
export OUT_1=../results/peak_inspection/CTCF_overlap_1.bed

bedtools intersect -wa -a $PEAK_1 -b $PEAK_2 > $OUT_1
```

Tells the cluster 'job manager' what resources you want (1 CPU, 8GB memory, run on the 'classroom' nodes, and name the job 'bedtools_ovl1'

Load the software. We use a tool called 'BEDTools' to work with generated peak files.

Create shortcut names for the two input bed files, and the output bed file.

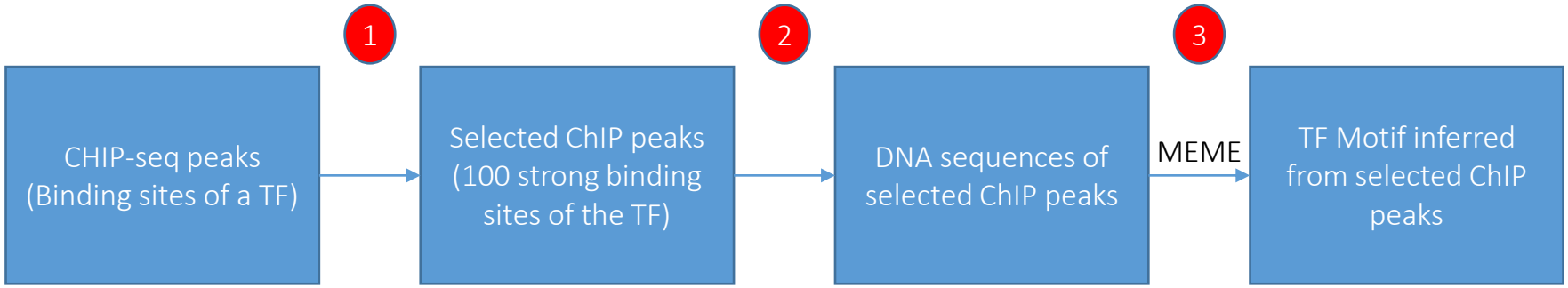
This command, not to be run by you directly, executes the 'bedtools intersect' using the –wa flag to get the entries in the first file (-a) that have an overlap with an entry in second input file (-b) and store the results in CTCF_overlap_1.bed

Part 2: Analyzing ChIP–Seq Peaks

Slides by Shayan Tabe Bordbar

In this lab, we will do the following:.

- Use command line tools to manipulate a ChIP track for BIN TF in *D. melanogaster*
- Subject peak sets to MEME suite.
- Compare MEME motifs with Fly Factor Survey motifs for BIN TF.


1
2
3
CHIP-seq peaks (Binding sites of a TF)
Selected ChIP peaks (100 strong binding sites of the TF)
DNA sequences of selected ChIP peaks
MEME
TF Motif inferred from selected ChIP peaks

In MobaXterm

Step 0A: Lab Setup

The lab is located in the following directory:

`/home/classroom/mayo/2021/06_Regulatory_Genomics/`

Following commands will copy a shell script -designed to prepare the working directory- to your home directory. Follow these steps to copy and then submit the script as a job to biocluster:

```
$ cd ~/

# Note ~ is a symbol in Unix paths referring to your home directory

$ rm ./prep-directory.sh

# Remove any existing copy of this file from your working directory

$ cp /home/classroom/mayo/2020/06_Regulatory_Genomics/src/prep-directory.sh ./

# Copies prep-directory.sh script to your working directory.

$ sbatch prep-directory.sh

# submits a job to biocluster to populate your home directory with necessary files

$ squeue –u <userID> # to check the status of the submitted job
```

This is the same as your login. Do not include < >
ex: `$ squeue –u classxx`

Step 0B: Working directory: data

Navigate to the created directory for this exercise and look what data folder contains.

```
$ cd 06_Regulatory_Genomics
$ ls
# output should be:
# data results src
$ ls data/
# BIN_Fchip_s11_1000.gff
# dm3.fasta
# flygenes_vm.bed
```

Name	Description
BIN_Fchip_s11_1000.gff	ChIP peaks for BIN transcription factor in GFF format
dm3.fasta	*Drosophila Melanogaster* genome
flygenes_vm.bed	Coordinates of all *Drosophila* genes in BED format

Step 0C: Working directory: scripts

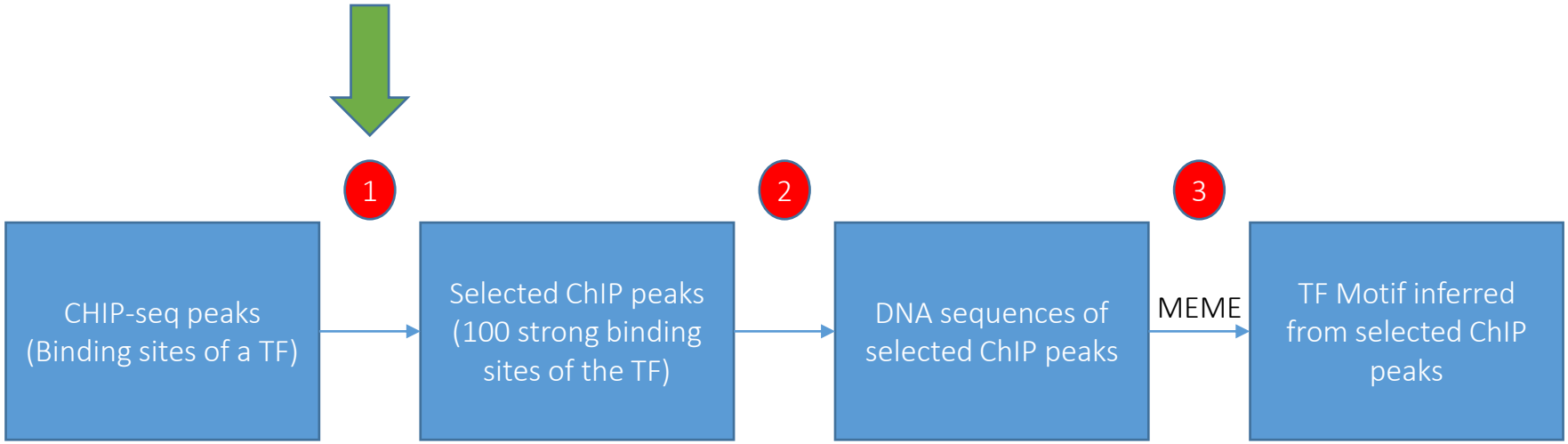
Navigate to the directory containing the scripts and look what's inside.

```
$ cd src

$ ls *.sh

# lists the scripts to be used in this lab:

# get_closest_genes.sh  get_sequence.sh  get_top100.sh
```

Computational Prediction of Motifs

In this exercise, after performing various file manipulations, we will use the MEME suite to identify a motif from the top 100 ChIP regions.

Subsequently, we will compare our predicted motif with the experimentally validated motif for BIN at Fly Factor Survey.


1
2
3
CHIP-seq peaks (Binding sites of a TF)
Selected ChIP peaks (100 strong binding sites of the TF)
DNA sequences of selected ChIP peaks
MEME
TF Motif inferred from selected ChIP peaks

Step 1: Obtain the top 100 strongest ChIP peaks

- We will use “sort” command, to sort the peaks based on their score and then take the top 100 peaks.
- Use the following line to get the top 100 chip peaks from the original ChIP gff file.

```
$ cd ~/06_Regulatory_Genomics/src/

$ head ~/06_Regulatory_Genomics/data/BIN_Fchip_s11_1000.gff
# The above command shows the first few lines of the file

$ sbatch  get_top100.sh

# OUTPUT in ~/06_Regulatory_Genomics/results/Top_100_peaks.gff
```

Please do not try to Run the commands in this slide. This is just to explain what the script that we just ran (get_top100.sh) is supposed to do in more detail.

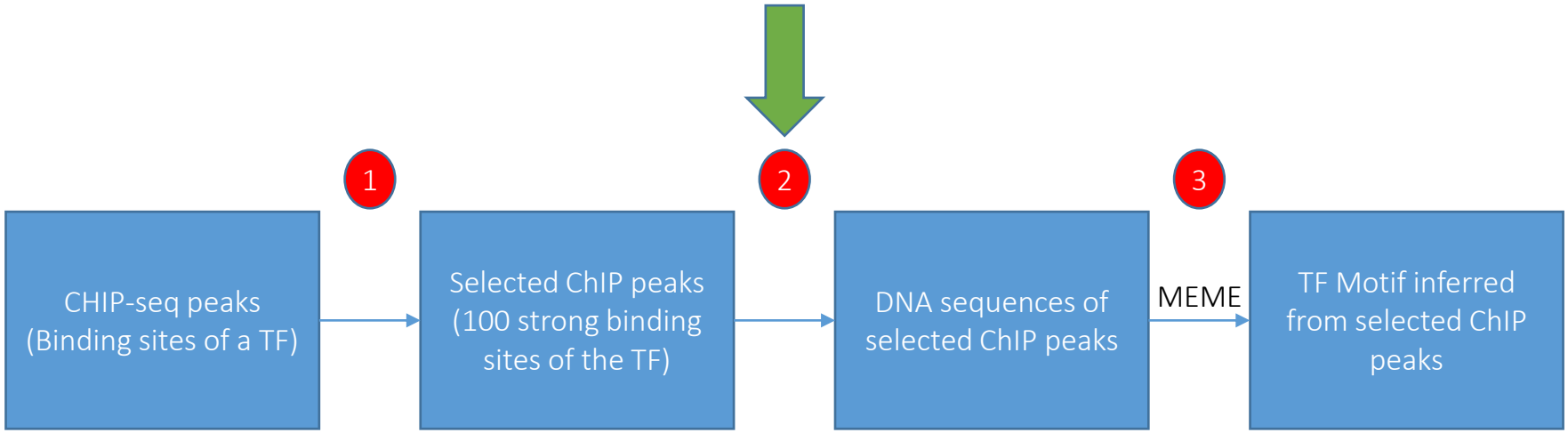
What's inside the **get_top100.sh** script?

```
#!/bin/bash
#SBATCH -c 1
#SBATCH --mem 8000
#SBATCH -A Mayo_Workshop
#SBATCH -J getTop100
#SBATCH -o getTop100.%j.out
#SBATCH -e getTop100.%j.err
#SBATCH -p classroom


# this is our input (gff)
export TOBESORTED=../data/BIN_Fchip_s11_1000.gff


sort -k 6,6nr  $TOBESORTED | head -100 > ../results/Top_100_peaks.gff
```

Tells the cluster 'job manager' what resources you want (1 CPU, 8GB memory), run on the 'classroom' nodes, and name the job 'getTop100'

Create shortcut name for input ChIP peak file in GFF format.

This command, not to be run by you directly, executes the Linux sort command to sort the file based on the numeric score stored in the 6th column of the gff file (ChIP score). [–k flag introduces the column to be sorted by. '**nr**' notes that we desire a **n**umeric sort in **r**everse order.] Output is directed to (>) Top_100_peaks.gff file.


1
2
3
CHIP-seq peaks (Binding sites of a TF)
Selected ChIP peaks (100 strong binding sites of the TF)
DNA sequences of selected ChIP peaks
MEME
TF Motif inferred from selected ChIP peaks

Step 2: Extract DNA sequence of Top 100 ChIP Regions

We will use a "getfasta" tool from "bedtools" toolkit to get the DNA sequence for the top 100 ChIP peaks.
Usage:

Please do not try to Run the commands in the first box.
This is just to explain the arguments to bedtools getfasta

```
$ bedtools getfasta [options]      -fi <genome_file_name> > \

# specifies the path to the genome sequence in FASTA format

                                   -bed <file_name.bed>

# specifies the path to coordinates of input regions in (BED/GFF/VCF) # formats
```

Script get_sequence.sh uses Bedtools getfasta to get the sequence corresponding to peaks stored in Top_100_peaks.gff. Run the following command:

```
$ cd ~/06_Regulatory_Genomics/src/

$ sbatch get_sequence.sh

# OUTPUT in ~/06_Regulatory_Genomics/results/BIN_top_100.fasta

$ squeue -u <userID>
```

Please do not try to Run the commands in this slide. This is just to explain what the script that we just ran (get_sequence.sh) is supposed to do in more detail.

What's inside the **get_sequence.sh** script?

```
#!/bin/bash
#SBATCH -c 1
#SBATCH --mem 8000
#SBATCH -A Mayo_Workshop
#SBATCH -J get_sequence
#SBATCH -o get_sequence.%j.out
#SBATCH -e get_sequence.%j.err
#SBATCH -p classroom

# load the tool environment
module load BEDTools

# this is our input (dm genome in fasta format)
export GENOME_DM3_FASTA=../data/dm3.fasta
export INPUT_CHIP=../results/Top_100_peaks.gff
export OUTPUT_NAME=../results/BIN_top_100.fasta

# use bedtools
bedtools getfasta -fi $GENOME_DM3_FASTA -bed $INPUT_CHIP | fold -w 60 > $OUTPUT_NAME
```

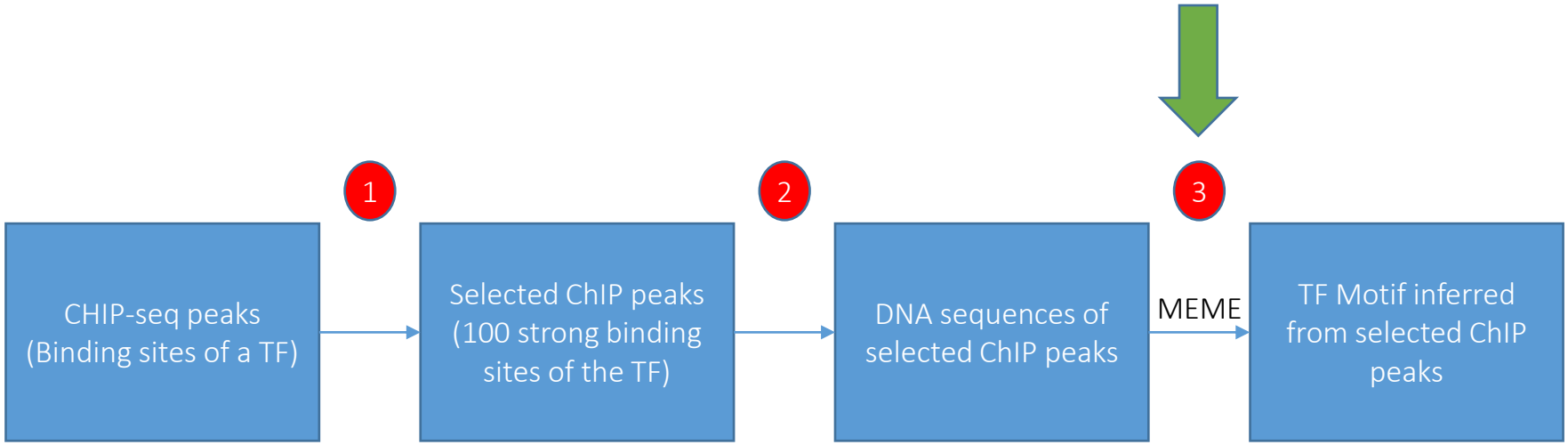
Tells the cluster 'job manager' what resources you want (1 CPU, 8GB memory, run on the 'classroom' nodes, and name the job 'get_sequence'

Load the software. We use a tool called 'BEDTools' to work with peak files.

Create shortcut names for input genome, input ChIP peak file and output FASTA file.

This command, not to be run by you directly, executes the 'bedtools getfasta' to get the DNA sequence in dm3.fasta genome corresponding to coordinates contained in Top_100_peaks.gff
fold –w 60 ensures that the width of lines in the output file does not exceed 60 characters.
results are directed to (>) BIN_top_100.fasta

Note that output of get_sequence.sh (BIN_top_100.fasta) has already been copied to the VM to be used in the next step.


1
2
3
CHIP-seq peaks (Binding sites of a TF)
Selected ChIP peaks (100 strong binding sites of the TF)
DNA sequences of selected ChIP peaks
MEME
TF Motif inferred from selected ChIP peaks

Step 0: Local Files

For viewing and manipulating the files needed for this laboratory exercise, the path on the VM will be denoted as the following:

[course_directory]

We will use the files found in:

[course_directory]\06_Regulatory_Genomics

[course_directory]= Desktop\Labs **UIUC**

[course_directory]= Desktop\VM **Mayo**

On Desktop

Step 3: Submit to MEME

In this step, we will submit the sequences to MEME

Go to the following address on your VM internet browser:

http://meme-suite.org/tools/meme

You can find BIN_top_100.fasta in the following directory on the VM:

[course_directory]\06_Regulatory_Genomics\BIN_top_100.fasta

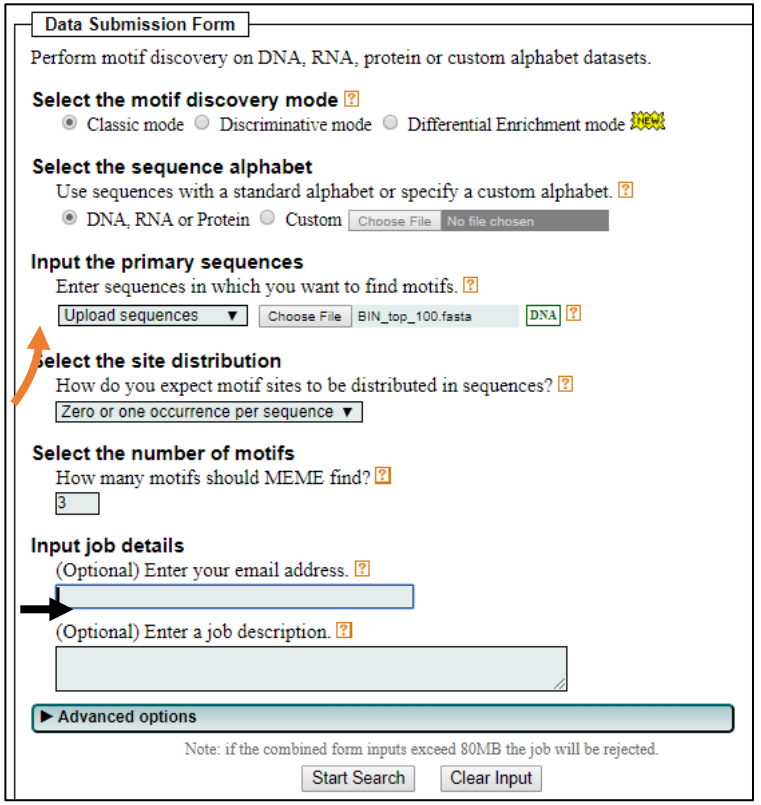
Upload your **sequences file** here

Enter **your email address** here.

Leave other parameters as default.

Click "Start Search".

DON'T WAIT FOR IT TO COMPLETE. MEME TAKES A VERY LONG TIME.



Step 3A: Analyzing MEME Results

Upon completion, MEME server will send you a notification email with web address. The webpage contains a summary of MEME's findings.

NO NEED TO WAIT FOR THE EMAIL.

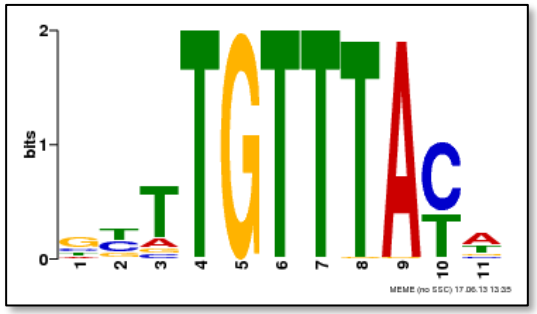
The webpage contents are also available in the following directory:

```
[course_directory]\06_Regulatory_Genomics\MEME.html
```

Let's investigate the top hit.

Step 3B: Analyzing MEME Results

To the right is a LOGO of our predicted motif, showing the per position relative abundance of each nucleotide



At the bottom are the aligned regions in each of our sequences that helped produce this motif. As the p-value increases (becomes less significant) matches show greater divergence from our LOGO.

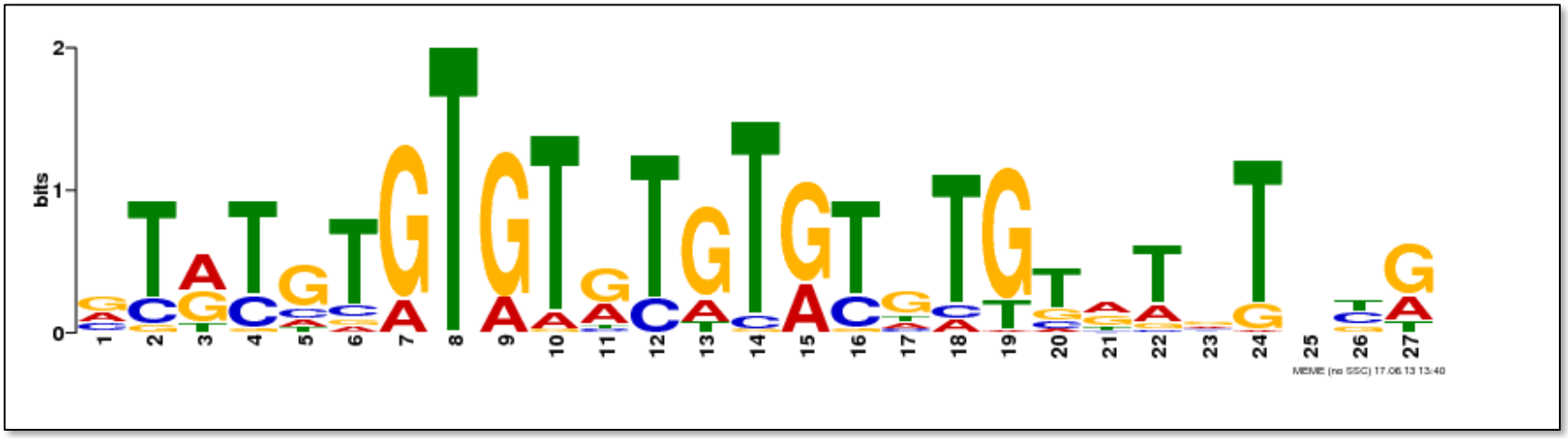
MOTIF LOCATIONS

Only Motif Sites | Motif Sites+Scanned Sites | All Sequences | Download PDF | Download SVG

	Name	p-value
1.	chr3R:14813999-14814500	4.63e-5
2.	chr2L:2445699-2446200	3.60e-2
3.	chr3R:26607599-26608100	6.68e-8
4.	chr2R:7896299-7896800	4.36e-5
5.	chr2R:2676399-2676900	7.14e-2
6.	chr2R:2871699-2872200	1.96e-4
7.	chr2R:9451549-9452050	5.55e-4
8.	chr3R:17214099-17214600	1.17e-1
9.	chr2L:18336599-18337100	1.16e-4
10.	chrX:7057299-7057800	3.58e-3
11.	chr3L:14359199-14359700	1.50e-1
12.	chr2R:9112349-9112850	1.04e-3
13.	chr3R:4854499-4855000	1.03e-6
14.	chr3R:17222349-17222850	1.88e-5

Step 3C: Analyzing MEME Results

Other predicted motifs do not seem as plausible.



Step 4A: Comparison with Experimentally Validated Motif for BIN

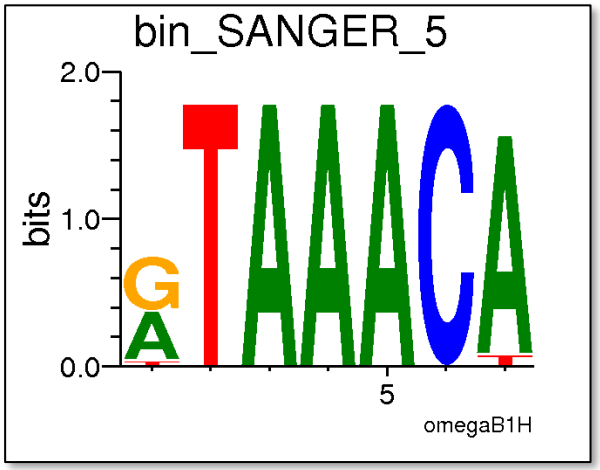
FlyFactorSurvey is a database of TF motifs in *Drosophila melanogaster*.

Use the internet browser on your VM to go to the following link to view the motif for BIN:

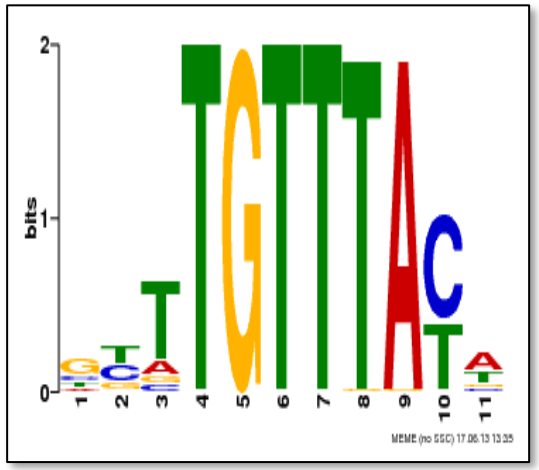
http://pgfe.umassmed.edu/ffs/TFdetails.php?FlybaseID=FBgn0045759

Step 4B: Comparison with Experimentally Validated Motif for BIN

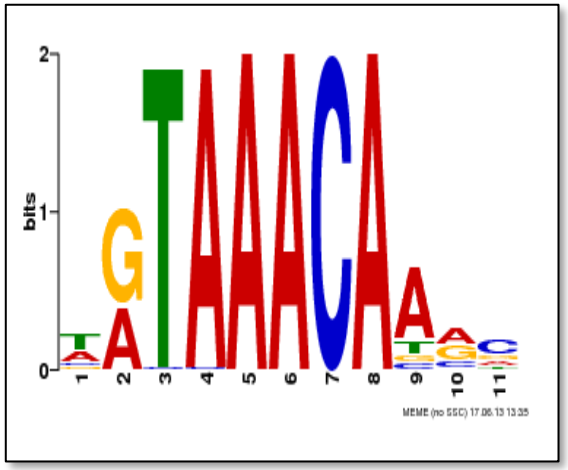
Actual BIN Motif



Best MEME Motif



Best MEME Motif Reverse Complemented



There is strong match between the actual motif and MEME's best motif. This indicates that MEME is working correctly.