Chip – Seq Peak Calling in Galaxy

Lisa Stubbs

Introduction

This goals of the lab are as follows:

1. Gain experience using Galaxy.

2. Teach how to map next generation (NSG) reads to a reference genome using **Bowtie.**

3. Demonstrate how to call peaks from Chip-Seq data.

Step 1A: Logging into Galaxy

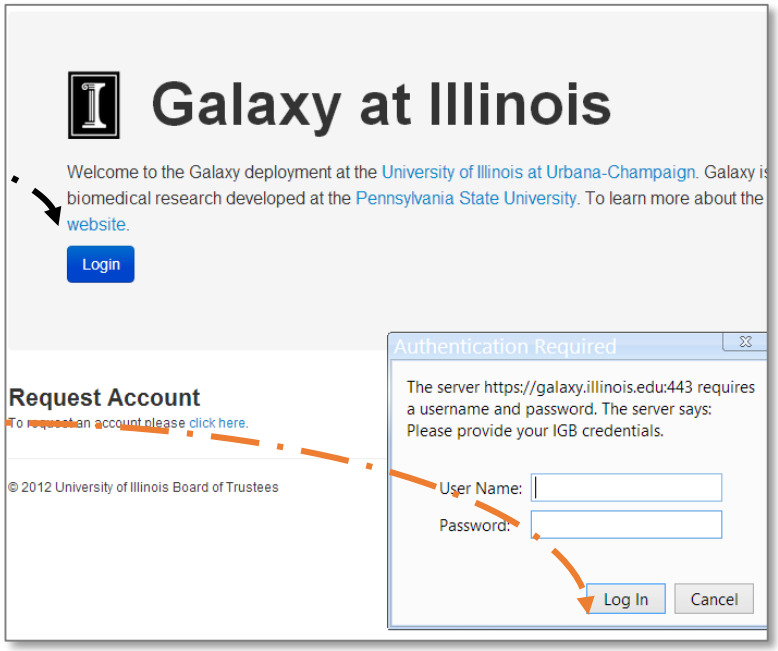
Go to: **biocluster.igb.illinois.edu**

Click **Enter**

Click **Login**

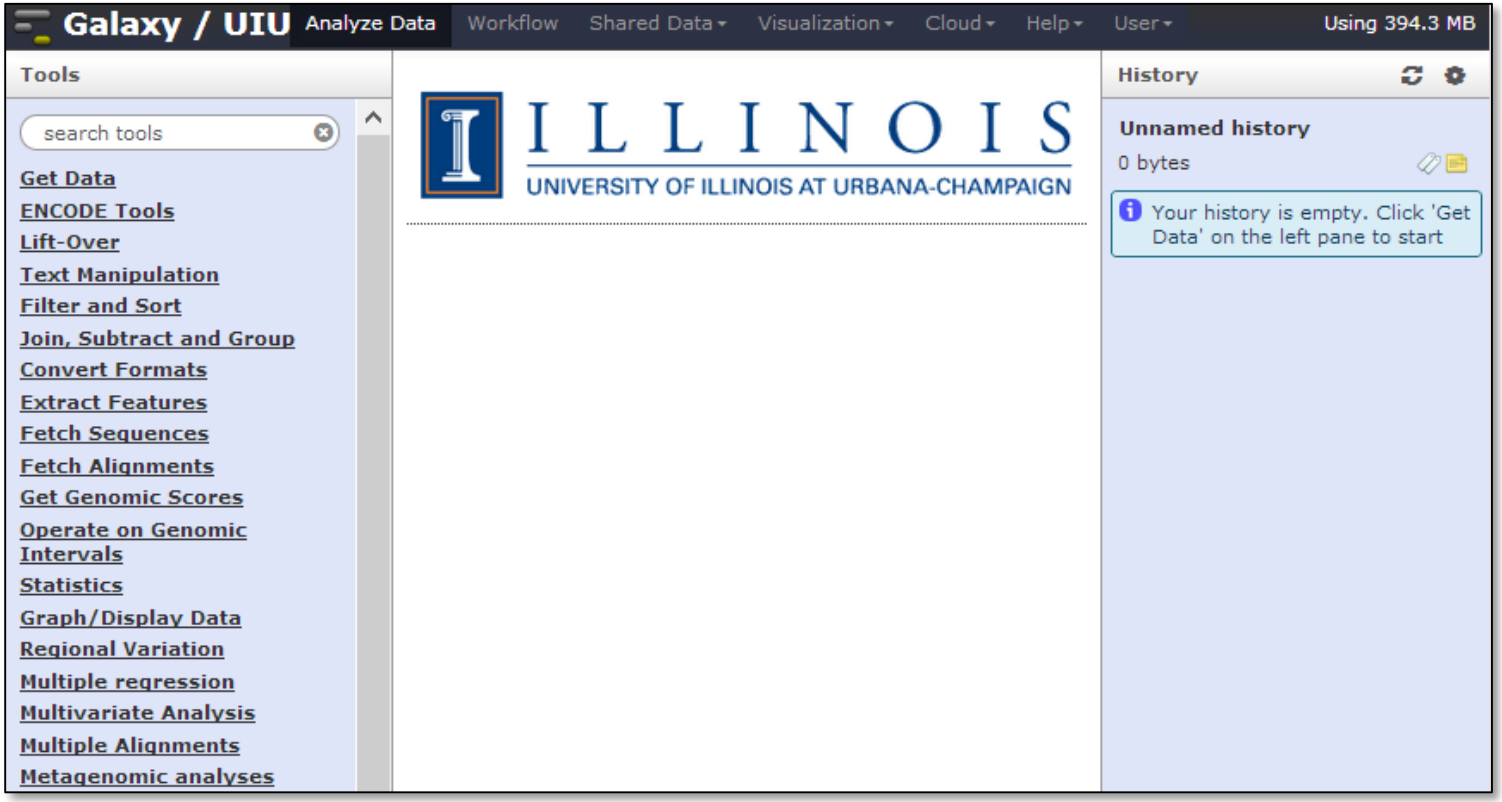
Input your login credentials.

Click **Login**.



Step 1B: Galaxy Start Screen

The resulting screen should look like the **figure** below:



Step 2A: Importing the Data

In this step, we will import the following data files:

Filename	Meaning
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.

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

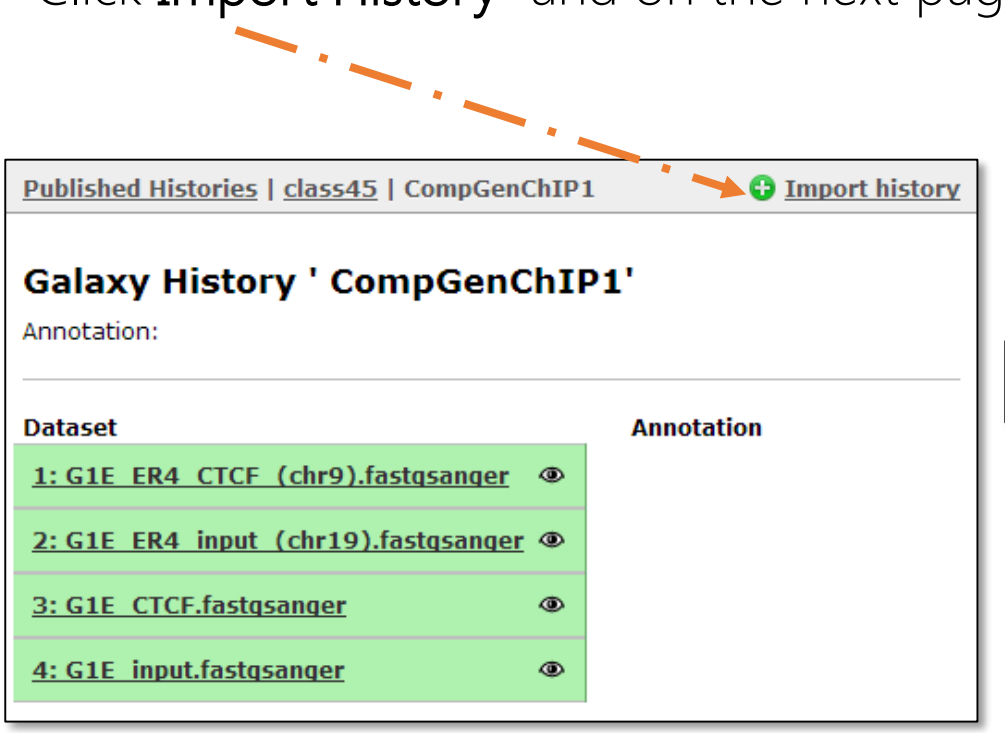
GATA-1 is crucial for the maturation of erythroid cells.

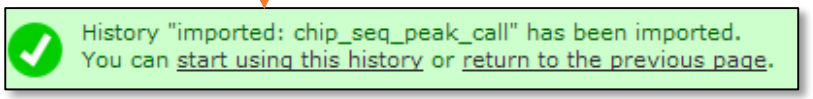
G1E_E4R cell lines conditionally express GATA-1 in the presence of estradiol, enabling erythdoi maturation.

Step 2B: Import Data in Galaxy

Go to https://galaxy.illinois.edu/galaxy/u/instr04/h/compgenchip1

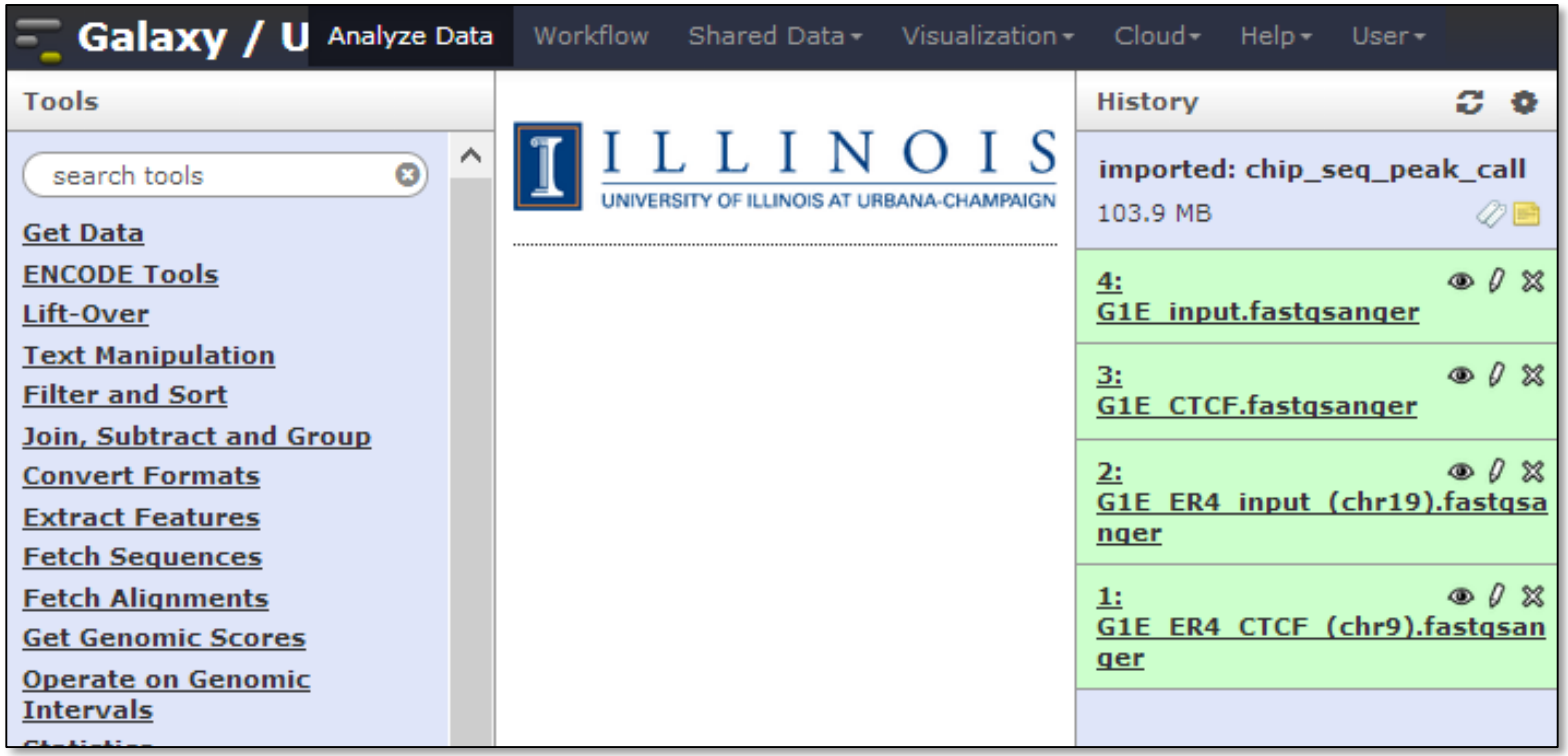
Click **Import History** and on the next page click **Start Using History.**





Step 2C: Import Data into Galaxy

Your Galaxy page should look like the following now:



Tools Pane **Main** Pane **History** Pane

Read Mapping and Peak Calling

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

Step 3A: Summary Statistics

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

Click **NGS: QC and manipulation** from the **Tools** pane.

Then click **FASTQ Summary Statistics.**

Step 3B: FASTQ Summary Statistics

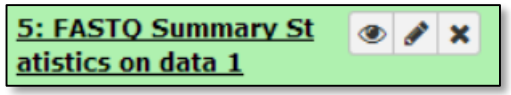
On the next page, make sure **1: G1E_ER4_CTCF_(chr..fastqsanger)** is selected.

Press **Execute.**

Step 3C: FASTQ Summary Statistics

The summary file will be the 5th file in the History pane.

Click  to display the file in the **Main** pane.



#column	count	min	max	sum	mean	Q1	med	Q3	IQR	lW	rW
1	270631	2	33	8498504	31.4025518141	32.0	33.0	33.0	1.0	31	33
2	270631	2	34	8324960	30.7612948997	30.0	33.0	33.0	3.0	26	34
3	270631	2	34	8181664	30.231806408	29.0	32.0	33.0	4.0	23	34
4	270631	2	34	8184981	30.2440629492	29.0	32.0	33.0	4.0	23	34
5	270631	2	34	8161333	30.1566819766	29.0	32.0	33.0	4.0	23	34
36	270631	2	34	7244057	26.7672846052	25.0	30.0	32.0	7.0	17	34

Discussion

How long are these reads?

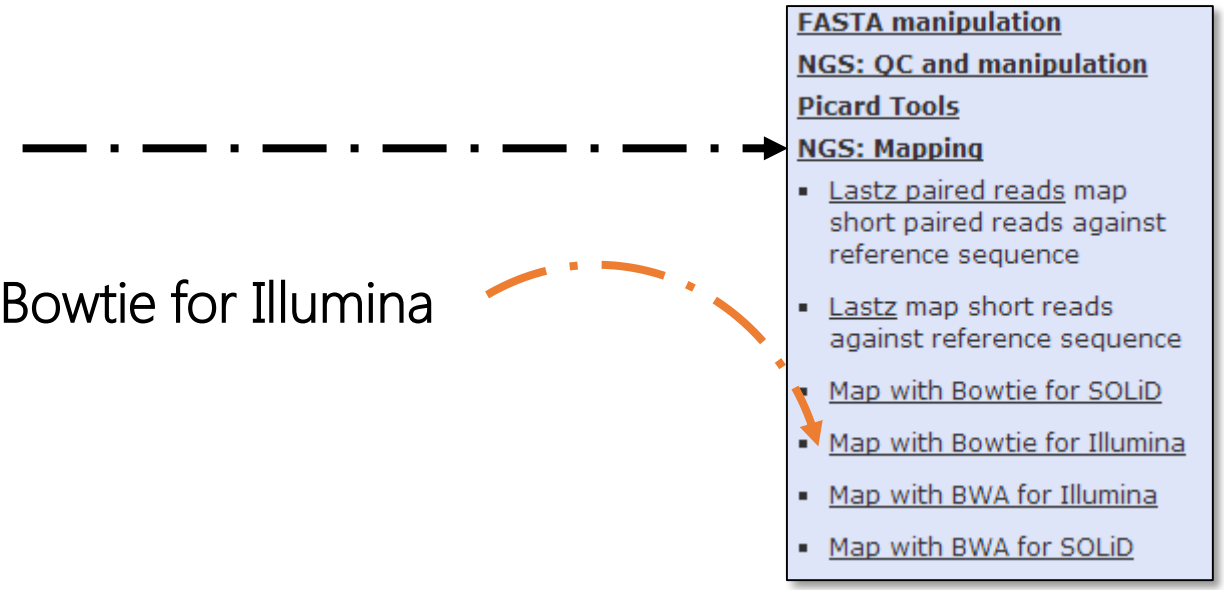
What is the median quality at the last position?

Step 4A: Map ChIP-Seq Reads to MM9 Genome

Next, we will map the reads in **G1E_E4R_CTCF_(chr9).fastqsanger** to the mouse genome.

Select **NGS: Mapping**

Then select **Map with Bowtie for Illumina**



Step 4B: Map ChIP-Seq Reads to MM9 Genome

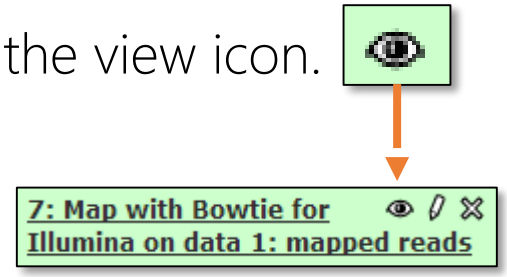
Make sure to select **mm9** as the reference genome.

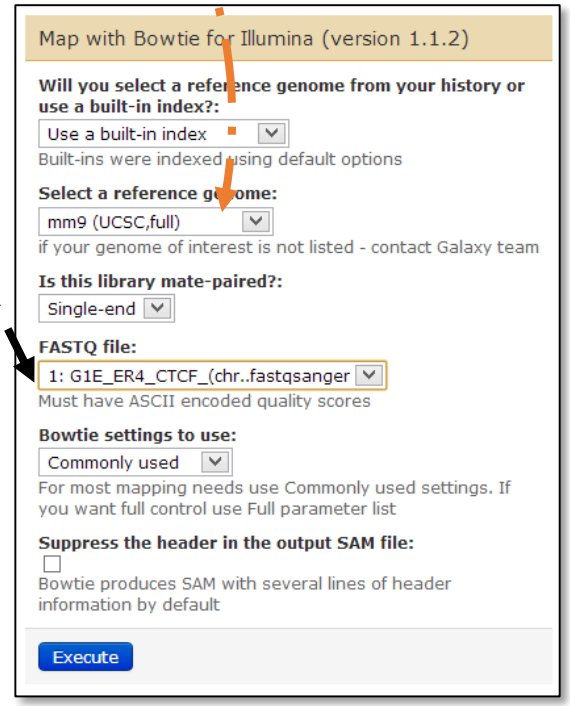
Make sure **1: G1E_ER4_CTCF_(chr..fastqsanger)** is selected.

Hit **Execute**.

It will take a few moments to complete.

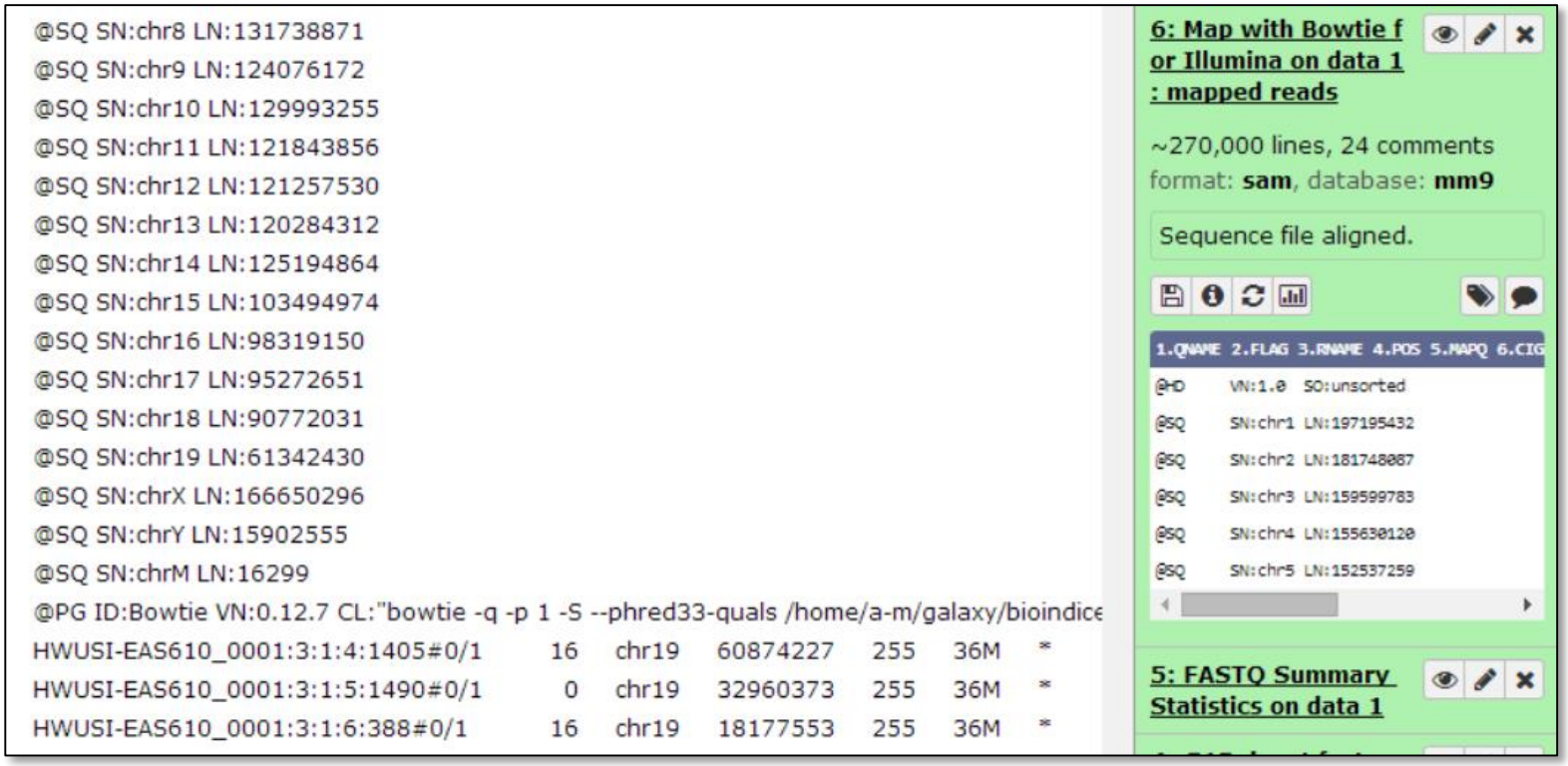
When done, click the view icon.





Step 4C: Map ChIP-Seq Reads to MM9 Genome

Your **Main** pane and **History** pane should look like the following:

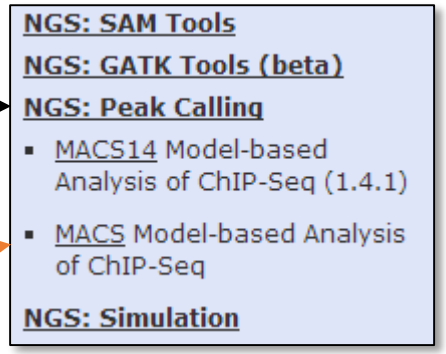


Step 5A: Calling Peaks with MACs

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

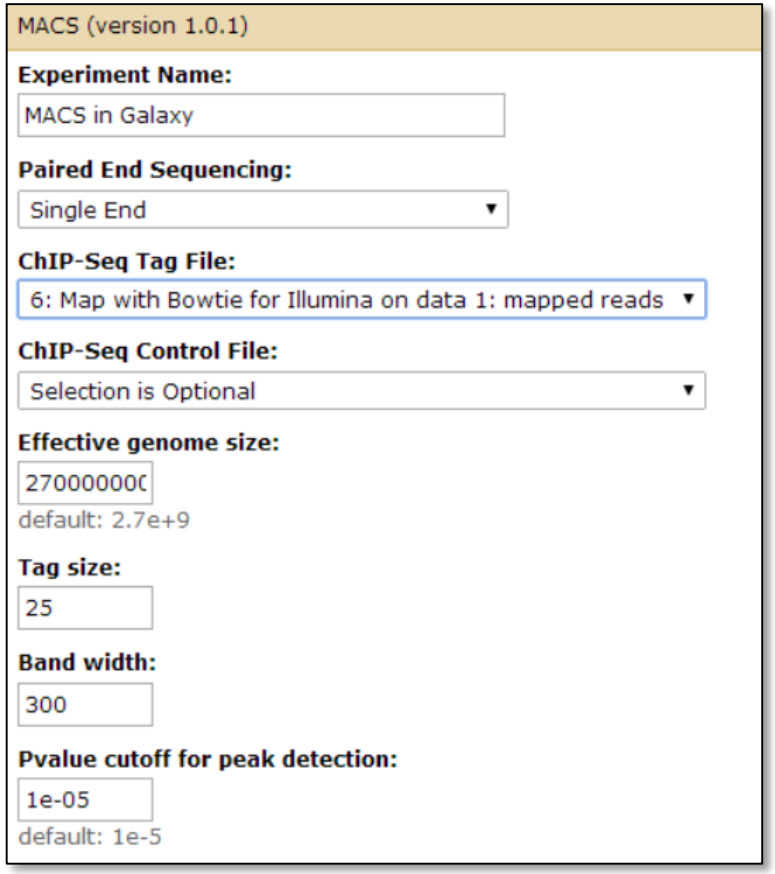
Select **NGS: Peak Calling**

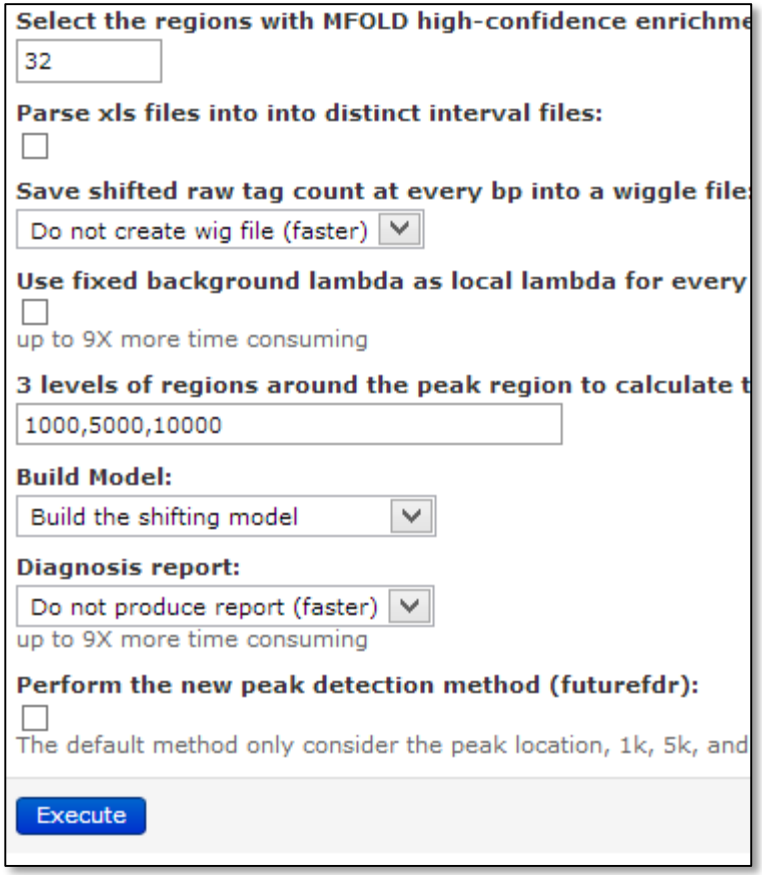
Select **MACS**



Step 5B: Calling Peaks with MACs

Run MACs with the default parameters.



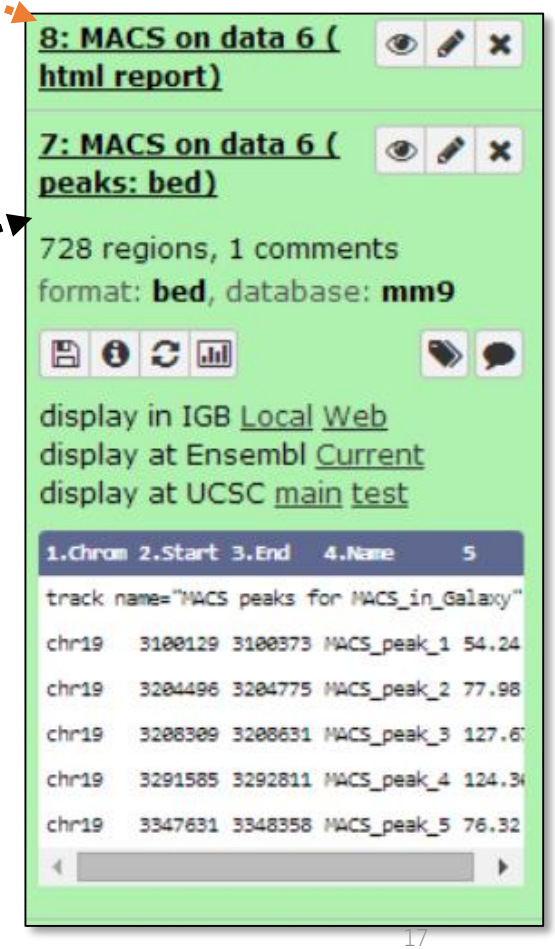


Step 5C: Calling Peaks with MACs

When done, MACs will create two files.

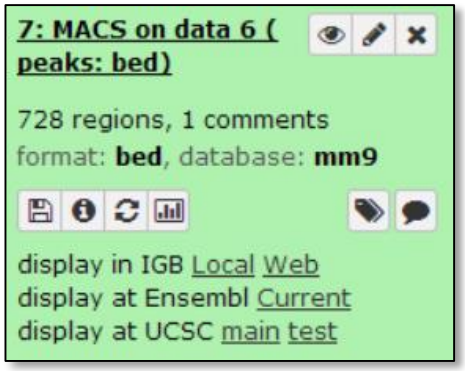
8: Macs on data 6 (html report) is an html document with information on the peak calling process.

7: MACs on data 6 (peaks:bed) is a BED file with coordinates and scores of ChiP-Seq peaks in chr19.



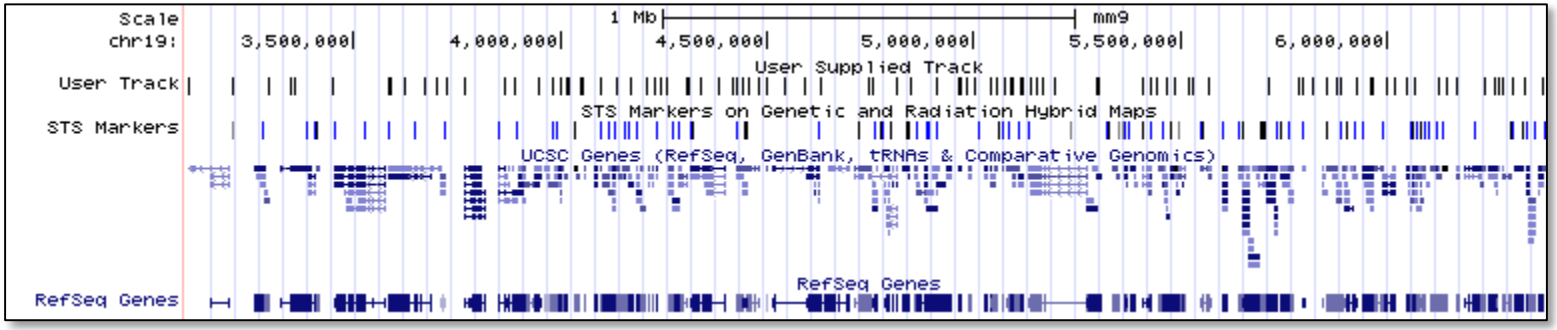
track name="MACS peaks for MACS_in_Galaxy"				
chr19	3100129	3100373	MACS_peak_1	54.24
chr19	3204496	3204775	MACS_peak_2	77.98
chr19	3208309	3208631	MACS_peak_3	127.67
chr19	3291585	3292811	MACS_peak_4	124.30
chr19	3347631	3348358	MACS_peak_5	76.32
chr19	3355095	3355575	MACS_peak_6	111.02
chr19	3450738	3451989	MACS_peak_7	177.92

Step 5D: Calling Peaks with MACs



In the **7: MACs on data 6 (peaks:bed)** section in the **History** pane click main next to display at UCSC browser.

The result should look similar to below:



Discussion

1. Look at the **BED** file. How many peaks were found?

Call Chip-Seq Peaks with a Control Sample

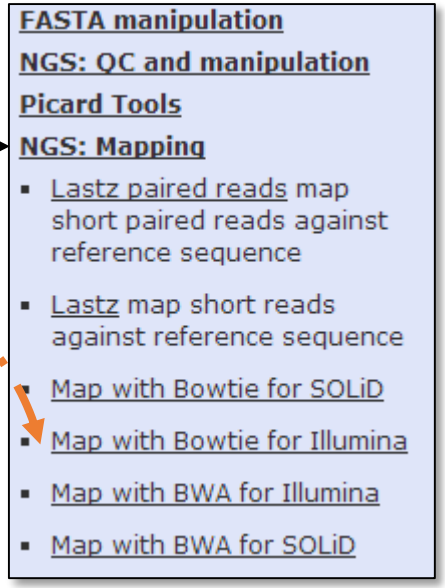
We will perform the same procedure we did in the previous exercise. This time though, we will work with a control sample instead of an experimental one.

Step 6A: Map Control ChIP-Seq Reads to MM9 Genome

Let's map the reads in **G1E_E4R_input_(chr19).fastqsanger** to the mouse genome.

Select **NGS: Mapping**

Then select **Map with Bowtie for Illumina**



Step 6B: Map Control ChIP-Seq Reads to MM9 Genome

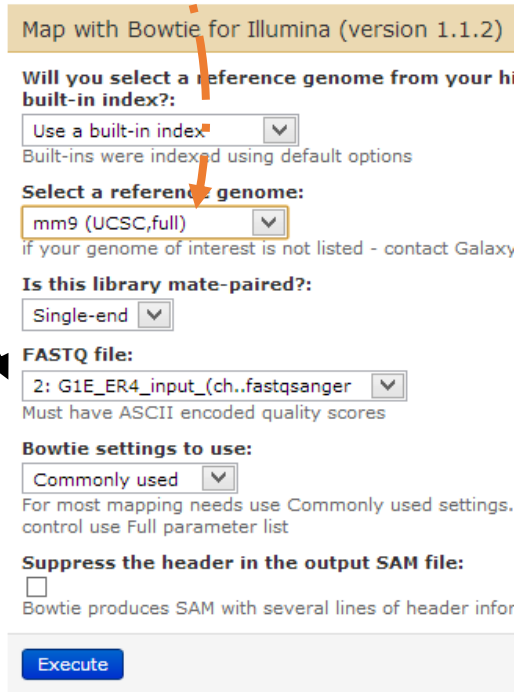
Make sure to select **mm9** as the reference genome.

Make sure **2: G1E_ER4_input_(chr..fastqsanger)** is selected.

Click **Execute**.

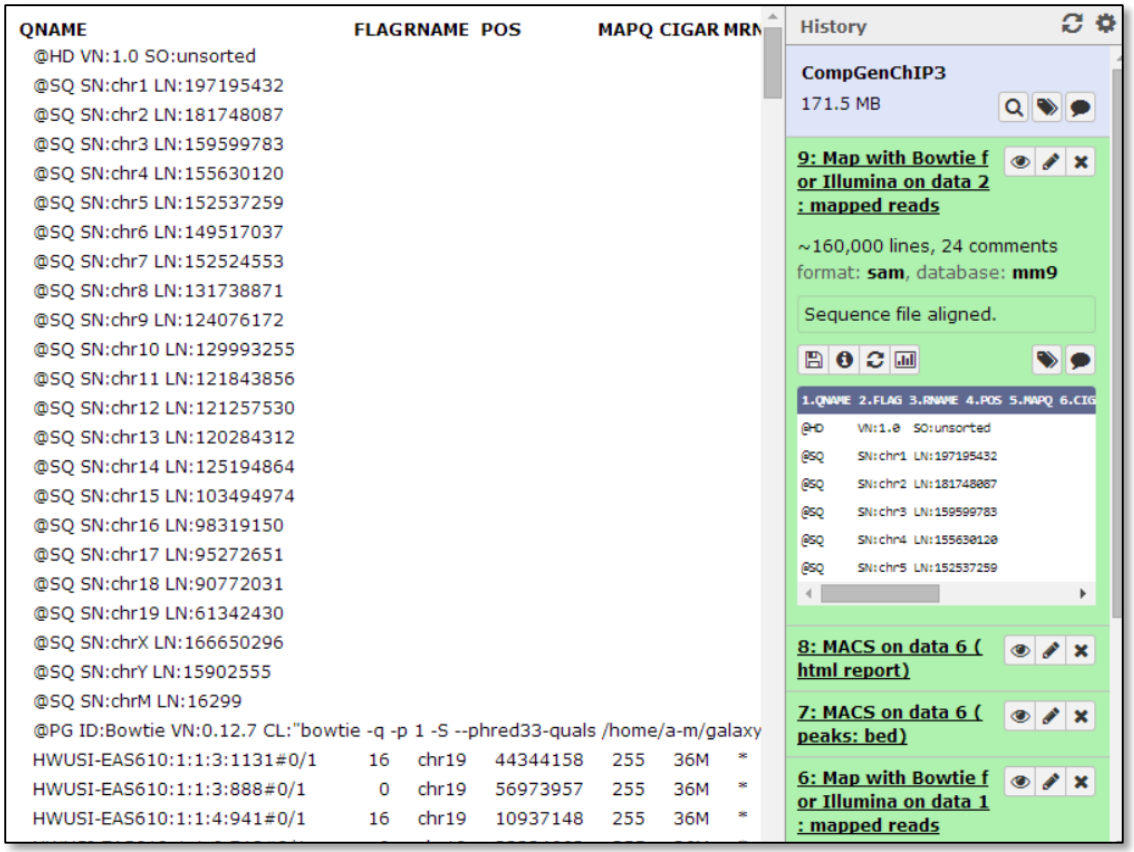
It will take a few moments to complete.

When done, click the view icon. 



Step 6C: Map Control ChIP-Seq Reads to MM9 Genome

Your **Main** pane and **History** pane should look like the following:

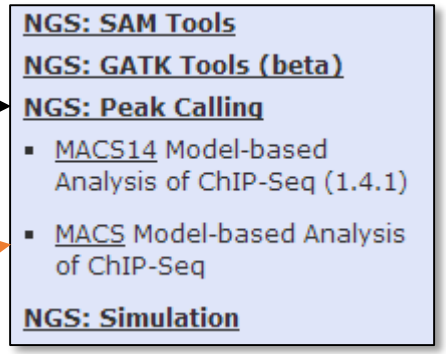


Step 7A: Calling Peaks with MACs on Control Chip-Seq Reads

Like before, we want to call peaks in our mapped Control ChiP-Seq reads.

Select **NGS: Peak Calling**

Select **MACS**

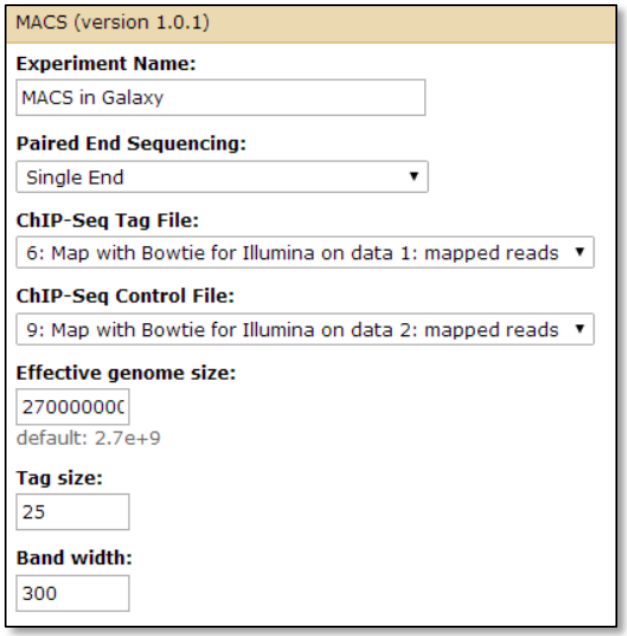


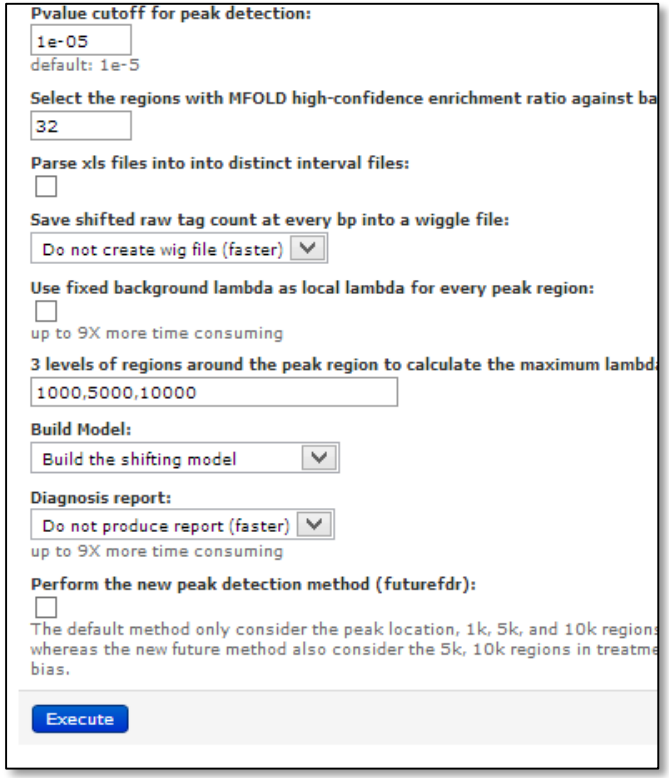
Step 7B: Calling Peaks with MACs on Control Chip-Seq Reads

Select **Map with Bowtie for Illumina on data 1** (the experimental aligned reads) for the Chip-Seq Tag File.

Select **Map with Bowtie for Illumina on data 2** (the control aligned reads) for the Chip-Seq Control File.

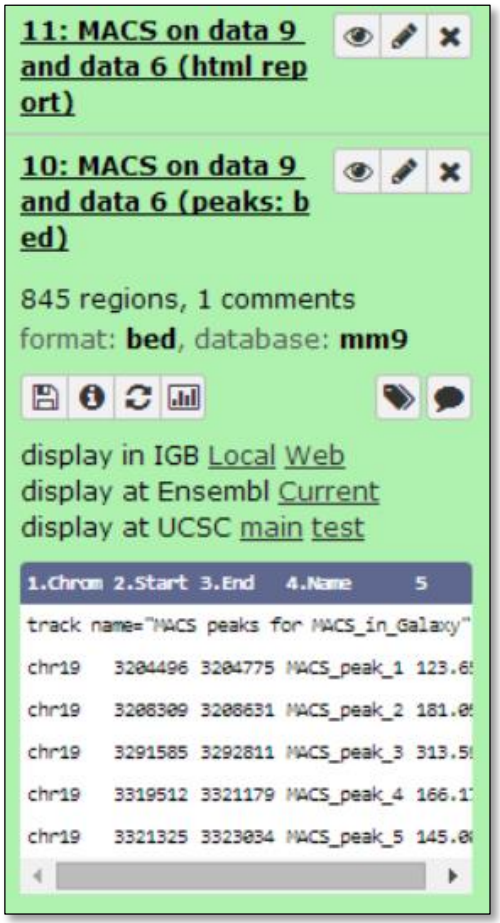
Click Execute





Step 7C: Calling Peaks with MACs on Control Chip-Seq Reads

Once again, MACs creates a **BED** file containing the peak coordinates and an **HTML** file containing information on the peak calling process.



Discussion

1. Examine the **BED** track.

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?

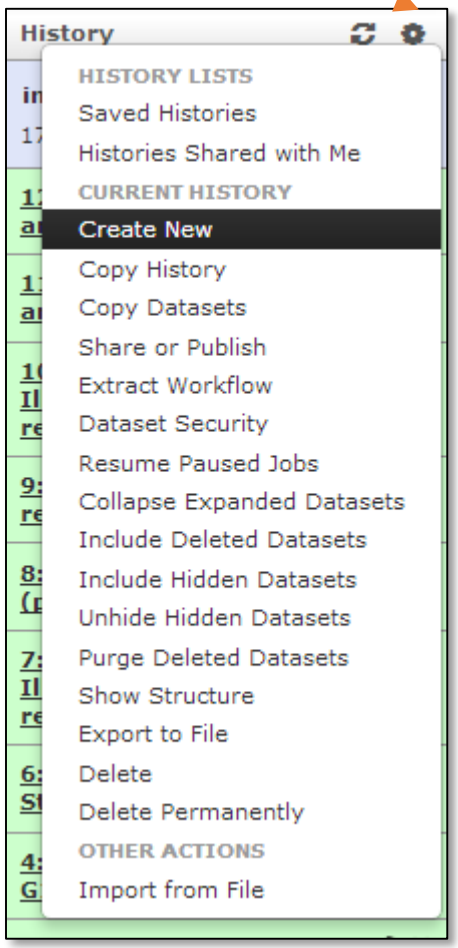
Workflow Extraction

In this exercise, we will automate the Bowtie runs of our other 2 datasets.

Step 1A: Workflow Extraction

Click the  located in the **History** pane.

From the drop down menu, select **Extract Workflow.**

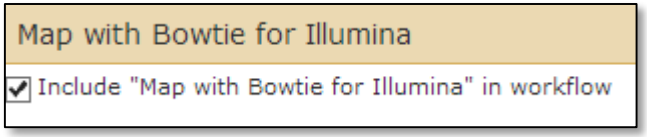


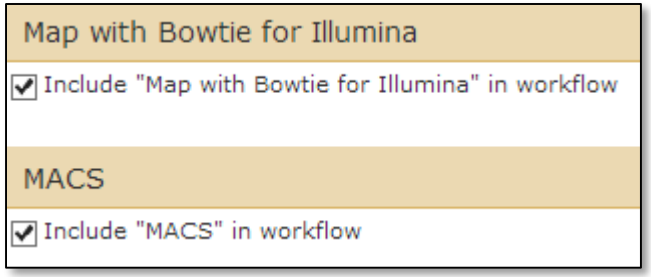
Step 1B: Workflow Extraction

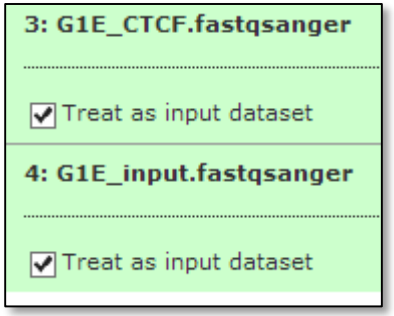
In the resulting window, give the workflow a new name.

Click **Uncheck All**.

Ensure the following are checked and press **Create Workflow.**



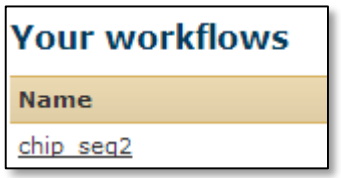




Step 1C: Workflow Extraction

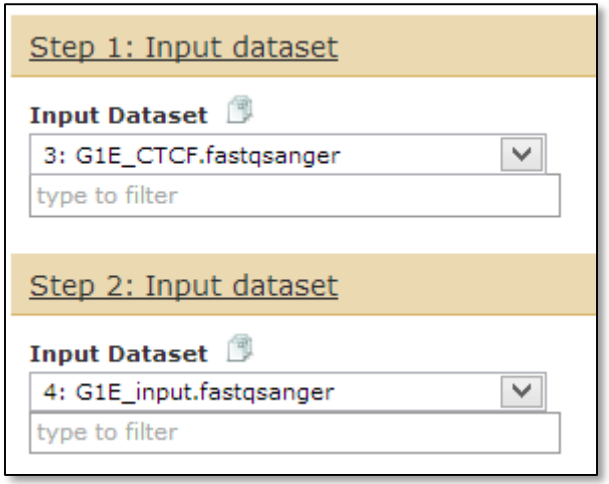
Select **All Workflows** at the bottom of the **Tool** pane.

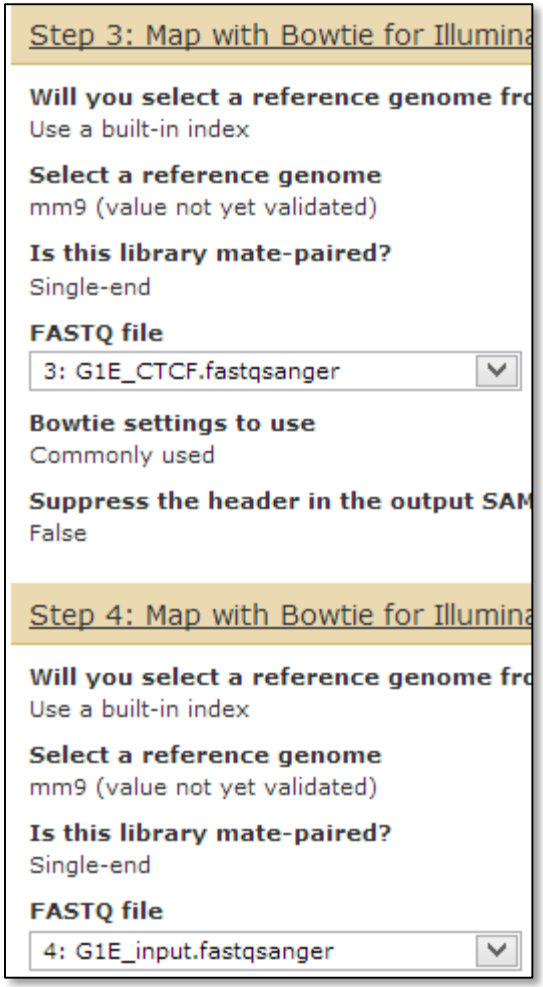
Select **chipseq2**

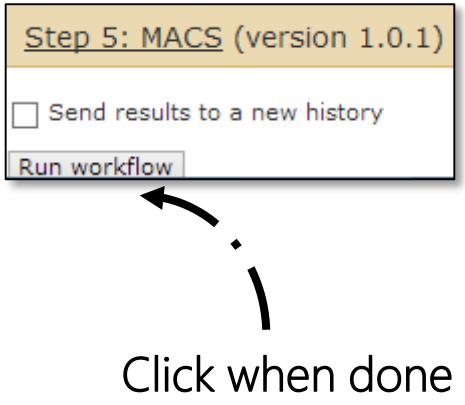


On the next page, ensure the following settings and click **Run**.

Step 1D: Workflow Extraction





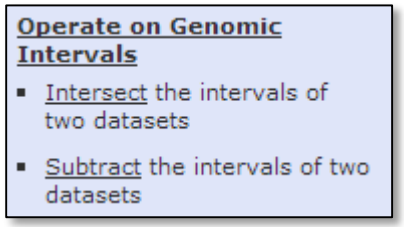


Identifying Differential Binding Sites

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

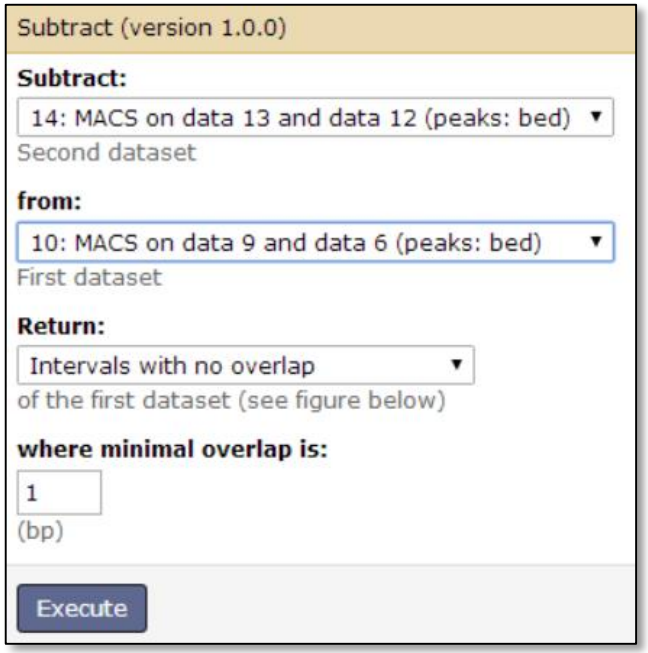
Step 1: Subtract Peaks Between Cell Lines

Select **Operate on Genomic Intervals** and **Subtract.**



Choose your G1E MACS peaks as your 2nd dataset and your G1E-ER4 peaks as your 1st dataset.

Click **Execute**.



Step 2: Subtract Peaks Between Cell Lines.

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

Discussion

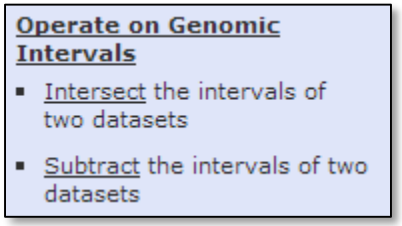
1. How many peaks are exclusive to G1E-ER4?

Redo Step1 only **SWITCH** the input order.

Choose your G1E MACS peaks as your 1st dataset and your G1E-ER4 peaks as your 2nd dataset.

Step 3: Intersect Peaks Between Cell Lines

Select **Operate on Genomic Intervals** and **Intersect.**



Choose your G1E-ER4 MACS peaks as your 1st dataset and your G1E peaks as your 2nd dataset.

Click **Execute**.