Chip – Seq Peak Calling in Galaxy

Lisa Stubbs

PowerPoint by Pei-Chen Peng
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 0B: Logging into Galaxy

Go to: galaxy.knowhub.org

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:

<table>
<thead>
<tr>
<th>Filename</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1E_ER4_CTCF_(chr19).fastqsanger</td>
<td>A sample ChIP-seq dataset on CTCF in G1E_ER4 cells, reads have been reduced to those mapping to chr19 for demonstration use.</td>
</tr>
<tr>
<td>G1E_ER4_input_(chr19).fastqsanger</td>
<td>Control DNA taken from chr19.</td>
</tr>
<tr>
<td>G1E_CTCF.fastqsanger</td>
<td>CTCF Chip for G1E line.</td>
</tr>
<tr>
<td>G1E_input.fastqsanger</td>
<td>Control for G1E line.</td>
</tr>
</tbody>
</table>

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.
Step 2B: Import Data in Galaxy

Go to http://galaxy.knowhub.org/u/pei-chen/h/compgenchip1-2

Click Import history
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 Import.
Step 2C: Import Data into Galaxy

Your Galaxy page should look like the following now:
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_(chr9).fastqsanger` is selected.

Press **Execute**.
Step 3C: FASTQ Summary Statistics

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

Click \[\text{眼}\] to display the file in the Main pane.

<p>| | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>#column</td>
<td>count</td>
<td>min</td>
<td>max</td>
<td>sum</td>
<td>mean</td>
<td>Q1</td>
<td>med</td>
<td>Q3</td>
<td>IQR</td>
<td>IW</td>
</tr>
<tr>
<td>-------</td>
<td>-------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>------</td>
<td>----</td>
<td>-----</td>
<td>----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>1</td>
<td>270631</td>
<td>2</td>
<td>33</td>
<td>8498504</td>
<td>31.4025518141</td>
<td>32.0</td>
<td>33.0</td>
<td>33.0</td>
<td>1.0</td>
<td>31</td>
</tr>
<tr>
<td>2</td>
<td>270631</td>
<td>2</td>
<td>34</td>
<td>8324960</td>
<td>30.7612948997</td>
<td>30.0</td>
<td>33.0</td>
<td>33.0</td>
<td>3.0</td>
<td>26</td>
</tr>
<tr>
<td>3</td>
<td>270631</td>
<td>2</td>
<td>34</td>
<td>8181664</td>
<td>30.231806408</td>
<td>29.0</td>
<td>32.0</td>
<td>33.0</td>
<td>4.0</td>
<td>23</td>
</tr>
<tr>
<td>4</td>
<td>270631</td>
<td>2</td>
<td>34</td>
<td>8184981</td>
<td>30.2440629492</td>
<td>29.0</td>
<td>32.0</td>
<td>33.0</td>
<td>4.0</td>
<td>23</td>
</tr>
<tr>
<td>5</td>
<td>270631</td>
<td>2</td>
<td>34</td>
<td>8161333</td>
<td>30.1566819766</td>
<td>29.0</td>
<td>32.0</td>
<td>33.0</td>
<td>4.0</td>
<td>23</td>
</tr>
<tr>
<td>36</td>
<td>270631</td>
<td>2</td>
<td>34</td>
<td>7244057</td>
<td>26.7672846052</td>
<td>25.0</td>
<td>30.0</td>
<td>32.0</td>
<td>7.0</td>
<td>17</td>
</tr>
</tbody>
</table>

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_(chr9).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 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.

**MACS Model-based Analysis of ChIP-Seq (Galaxy Version 1.0.1)**

- **Experiment Name**: MACS in Galaxy
- **Paired End Sequencing**: Single End
- **ChIP-Seq Tag File**: 6: Map with Bowtie for Illumina on data 1: mapped reads
- **ChIP-Seq Control File**: Nothing selected
- **Effective genome size**: 2700000000.0
- **Tag size**: 25
- **Band width**: 300
- **P-value cutoff for peak detection**: 1e-05

**Select the regions with MFOLD high-confidence enrichment ratio against background to build model**

- 32

**Parse xls files into into distinct interval files**

- Yes

**Save shifted raw tag count at every bp into a wiggle file**

- Do not create wig file (faster)

**Use fixed background lambda as local lambda for every peak region**

- Yes

**3 levels of regions around the peak region to calculate the maximum lambda as local lambda**

- 1000, 5000, 10000

**Build Model**

- Build the shifting model

**Diagnosis report**

- Do not produce report (faster)

**Perform the new peak detection method (futurefdr)**

- Yes

The default method only consider the peak location, 1k, 5k, and 10k regions in the control data; whereas the new method also consider the 5k, 10k regions in treatment data to calculate local bias.
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.
Step 5D: Calling Peaks with MACs

In the **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:

![Diagram of MACs output](image)

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_(chr19).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 should look like the following:

```
<table>
<thead>
<tr>
<th>QNAME</th>
<th>FLAG</th>
<th>RNAME</th>
<th>POS</th>
<th>MAPQ</th>
<th>CIGAR</th>
<th>MRNM</th>
<th>MPOS</th>
<th>ISIZE</th>
<th>SEQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>@HD VN 1.0 SO:unsorted</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>@SQ SN chr1 LN:1097195432</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>@SQ SN chr10 LN:129993235</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>@SQ SN chr11 LN:121843856</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>@SQ SN chr12 LN:121257330</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>@SQ SN chr13 LN:130248312</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>@SQ SN chr13_random LN:400311</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>@SQ SN chr14 LN:125194864</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>@SQ SN chr15 LN:103494974</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>@SQ SN chr16 LN:94319150</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>@SQ SN chr16_random LN:3994</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>@SQ SN chr17 LN:95172651</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>@SQ SN chr17_random LN:628739</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>@SQ SN chr18 LN:90772031</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>@SQ SN chr18_random LN:61342430</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>@SQ SN chr1_random LN:1231697</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>@SQ SN chr2 LN:181748087</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>@SQ SN chr3 LN:139999783</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>@SQ SN chr4_random LN:41899</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>@SQ SN chr4 LN:155630120</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>@SQ SN chr5_random LN:160594</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>@SQ SN chr5 LN:152537259</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>@SQ SN chr6_random LN:357350</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>@SQ SN chr6 LN:149517037</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>@SQ SN chr7 LN:152524553</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>@SQ SN chr7_random LN:362490</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>@SQ SN chr8 LN:117338871</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>@SQ SN chr8_random LN:849593</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>@SQ SN chr9 LN:124076172</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>@SQ SN chr9_random LN:449403</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>@SQ SN chrM LN:16299</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>@SQ SN chrM_random LN:5900358</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>@SQ SN chX LN:166650296</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>@SQ SN chX_random LN:1785075</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>@SQ SN chX LN:15902555</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>@SQ SN chY_random LN:15662461</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```
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 6: Map with Bowtie for Illumina on data 1 (the experimental aligned reads) for the Chip-Seq Tag File.

Select 9: Map with Bowtie for Illumina on data 2 (the control aligned reads) for the Chip-Seq Control File.

Click Execute

MACS Model-based Analysis of ChiP-Seq (Galaxy Version 1.0.1)

Experiment Name
MACS in Galaxy

Paired End Sequencing
Single End

ChiP-Seq Tag File
Select 6: Map with Bowtie for Illumina on data 1: mapped reads

ChiP-Seq Control File
Select 9: Map with Bowtie for Illumina on data 2: mapped reads

Effective genome size
2700000000.0

default: 2.7e+9

Tag size
25

Band width
300
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 name a new name.

Click **Uncheck All**.

Check **these boxes**.

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

Step 1: Input dataset
- Input Dataset
  - 3: G1E_CTCF.fastqsanger

Step 2: Input dataset
- Input Dataset
  - 4: G1E_input.fastqsanger

Step 3: Map with Bowtie for Illumina (version 1.1.2)
- Will you select a reference genome from your history or use a built-in index?
  - Use a built-in index
- Select a reference genome
  - mm9 (value not yet validated)
- Is this library mate-paired?
  - Single-end
- FASTQ file
  - 3: G1E_CTCF.fastqsanger

Step 4: Map with Bowtie for Illumina (version 1.1.2)
- Will you select a reference genome from your history or use a built-in index?
  - Use a built-in index
- Select a reference genome
  - mm9 (value not yet validated)
- Is this library mate-paired?
  - Single-end
- FASTQ file
  - 4: G1E_input.fastqsanger

Click when done
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 1\(^{st}\) 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 1\textsuperscript{st} dataset and your G1E peaks as your 2\textsuperscript{nd} dataset.

Click **Execute**.