Chip – Seq Peak Calling in Galaxy

Chris Seward

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: http://compgen.knoweng.org/galaxy

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 the following
http://compgen.knoweng.org/galaxy/u/shounakbhogale/h/sbchipseq

Click **Import history** and the next page click **Import**.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>4: G1E_input.fastqsanger</td>
<td></td>
</tr>
<tr>
<td>3: G1E_CTCF.fastqsanger</td>
<td></td>
</tr>
<tr>
<td>2: G1E_ER4_input_(chr19).fastqsanger</td>
<td></td>
</tr>
<tr>
<td>1: G1E_ER4_CTCF_(chr19).fastqsanger</td>
<td></td>
</tr>
<tr>
<td>sb_chipseq</td>
<td>103.92 MB</td>
</tr>
</tbody>
</table>

search datasets
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_(chr19).fastqsanger** is selected.

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

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

Click \( \text{\[ \text{icon} \]} \) to display the file in the Main pane.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>#column</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>
<td></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>
<td></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>
<td></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>
<td></td>
</tr>
<tr>
<td>6</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>
<td></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:

<table>
<thead>
<tr>
<th>CHROM</th>
<th>POS</th>
<th>NAME</th>
<th>QUALITY</th>
<th>SCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr1</td>
<td>1</td>
<td>ENCODE</td>
<td>20</td>
<td>100</td>
</tr>
<tr>
<td>chr2</td>
<td>2</td>
<td>ENCODE</td>
<td>20</td>
<td>100</td>
</tr>
<tr>
<td>chr3</td>
<td>3</td>
<td>ENCODE</td>
<td>20</td>
<td>100</td>
</tr>
<tr>
<td>chr4</td>
<td>4</td>
<td>ENCODE</td>
<td>20</td>
<td>100</td>
</tr>
<tr>
<td>chr5</td>
<td>5</td>
<td>ENCODE</td>
<td>20</td>
<td>100</td>
</tr>
<tr>
<td>chr6</td>
<td>6</td>
<td>ENCODE</td>
<td>20</td>
<td>100</td>
</tr>
<tr>
<td>chr7</td>
<td>7</td>
<td>ENCODE</td>
<td>20</td>
<td>100</td>
</tr>
<tr>
<td>chr8</td>
<td>8</td>
<td>ENCODE</td>
<td>20</td>
<td>100</td>
</tr>
<tr>
<td>chr9</td>
<td>9</td>
<td>ENCODE</td>
<td>20</td>
<td>100</td>
</tr>
<tr>
<td>chr10</td>
<td>10</td>
<td>ENCODE</td>
<td>20</td>
<td>100</td>
</tr>
</tbody>
</table>

Chip-Seq Peak Calling in Galaxy | Lisa Stubbs | 2019
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 read...

ChIP-Seq Control File
Nothing selected

Effective genome size
2700000000.0
default: 2.7e+9

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 No

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 No
up to 9X more time consuming

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)
up to 9X more time consuming

Perform the new peak detection method (futurefdr)
Yes No
The default method only consider the peak location, 1k, 5k, and 10k regions in the control data; whereas the new future method also consider the 5k, 10k regions in treatment data to calculate local bias.

Execute
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 **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:

![Chip-Seq Peak Calling in Galaxy | Lisa Stubbs | 2019](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 \texttt{G1E\_ER4\_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:
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) Options](image)
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?
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 2\textsuperscript{nd} dataset and your G1E-ER4 peaks as your 1\textsuperscript{st} 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 Step 1 only **SWITCH** the input order.

Choose your G1E MACS peaks as your 1\textsuperscript{st} dataset and your G1E-ER4 peaks as your 2\textsuperscript{nd} 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**.

The number of peaks should be 72.