**Xiaoxu Na**

**ChIP-seq Exercises**

**https://usegalaxy.org/u/james/p/exercise-chip-seq**

For this exercise we will use a ChIP-seq dataset for CTCF in the murine G1E\_ER4 cell line. This dataset has been reduced to (mostly) contain only reads aligning to chr19:

[**Galaxy Dataset | G1E\_ER4 CTCF (chr9)**](https://usegalaxy.org/dataset/display_by_username_and_slug?username=james&slug=2979743ceed4f520)

A sample ChIP-seq dataset on CTCF in G1E\_ER4 cells, reads have been reduced to those mapping to chr9 for demonstration use.

Click the 'import this dataset' button above to add this dataset to your analysis history to being the analysis.

**Mapping reads and peak calling**

Step 1: First, for quality control, we will compute summary statistics on this dataset. Run the tool "NGS: QC and Manipulation > FASTQ Summary Statistics" on your dataset. When the job completes, inspect the results.

**Question 1:** How long are these reads? What is the median quality at the last position?

**Sequence length: 36**

**Median quality: 30.0000**

Step 2: Next we will map these reads to a reference genome. Use the "NGS: Mapping >  Bowtie.2 for Illumina" tool. You will need to change the reference genome build you are mapping against to "mm9". Otherwise you can leave the default mapping options.

Step 3: Once are reads are mapped, we will call peaks with MACS. Use the "NGS: Peak Calling > MACS " tool. **You should also change the tag size to the read length you observed in Step 1**. Otherwise the default values should be reasonable.

Step 4: Once MACS completes it will produce two datasets. One is a report on the peak calling process. The other contains the positions of the peaks.

**Question 2:** How many peaks were found? Click the link to "Display at UCSC main" and you will be able to see the positions of the peaks on the genome.

**616 peaks**

**Calling peaks with a control sample**

Next, we will incorporate an input DNA control, import the following dataset into your history:

[Galaxy Dataset | G1E\_ER4 input (chr19)](https://usegalaxy.org/dataset/display_by_username_and_slug?username=james&slug=fac8acfa217c9795)

Step 1: Map the input DNA control to mm9 using Bowtie.2

Step 2: Load the MACS tool again. Select your previous CTCF dataset for ChIP-seq tag file, but now select the mapped input DNA for "ChIP-seq control file".

**Question 3:** How many peaks are called this time? What is the effect of using the input control?

**802 peaks are called this time. Using the input control highly increase the number of peaks called.**

**Create a workflow and reuse**

Step 1: At the top of the History panel, click "Options" and select "Extract Workflow". Here you have the chance to select which jobs will be included in the workflow. Click "Uncheck all" and the select the two "Map with Bowtie.2" jobs and the last "MACS" job.

Step 2: Import the following datasets -- CTCF ChIP and control for the G1E line:

[Galaxy Dataset | G1E CTCF](https://usegalaxy.org/dataset/display_by_username_and_slug?username=james&slug=d78ba454458040fd)

[Galaxy Dataset | G1E input](https://usegalaxy.org/dataset/display_by_username_and_slug?username=james&slug=5aa51f1c13683c33)

Step 3: At the bottom of the tools menu, select "Workflows > All Workflows", this will show the workflow list. Select the workflow you just created. You will be able to select input datasets for the two Bowtie steps, select the CTCF and input datasets. Click "Run Workflow".

**Question 4:** How many peaks are called?

**674 peaks.**

**Identify differential binding sites**

[G1E is a model for erythropoiesis](http://stokes.chop.edu/web/weiss/G1Eindex.html), the G1E line is a GATA1 null derived line which can be induced to differentiate by estradiol treatment (thus G1E-ER4). Here we will use Galaxy to identify sites that have differential binding across the two developmental stages.

Step 1: Select the "Operate on Genomic Intervals > Subtract" tool. For the first input ("Subtract") select your second set of peaks (Peaks from G1E), for the second input ("from") select your first set of peaks (Peaks from G1E-ER4) and run the tool. The resulting dataset contains peaks that are only present in the differentiated line.

**Question 5.** How many are there?

227

Step 2: Perform the subtract operation, switching the input datasets, to find peaks that are unique to the undifferentiated line.

**Question 6.** How many are there?

83

Step 3: Use the "Operate on Genomic Intervals > Intersect" tool to find peaks that are common to both datasets.

**Question 7.** How many are there?

591

Step 4: Finally, load the "Graph / Display Data > Build Custom Track" tool. Add each of your three tracks by clicking "Add Track" and give them descriptive names. Run the tool, and inspect the resulting dataset when complete. Click "Display at UCSC main" and all three tracks will be displayed in the UCSC browser. You can now inspect differential CTCF binding sites between two differentiation time points.

**Question 8.**  Save a screenshot of your UCSC display.

