Analyzing ChIP-Seq Data in Galaxy

Lauren Mills

RISS

ABSTRACT

Step-by-step guide to basic ChIP-Seq analysis using the Galaxy platform.

Table of Contents

Introduction

ChIP-seq is a popular technique for intergrading the chromatin state of a given genome. ChIP-seq can be used to identify the genomic locations of transcription factors, histone modifications and many other proteins that bind DNA. This tutorial is a guide for how to perform basic ChIP-seq analysis using the Galaxy platform at MSI.

Links to helpful information

BedTools: http://bedtools.readthedocs.org/en/latest/
Cistrome paper: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3245621/
Cistrome Galaxy instance: http://cistrome.org/ap/
SeqPos paper: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2932437/
UCSC Data File Formats: https://genome.ucsc.edu/FAQ/FAQformat.html

Text conventions

Throughout this document you will be asked to do tasks within the Galaxy web interface. Most of the tasks will involve working with Galaxy tools located in the left bar, data in your history located in the right bar or data from data libraries which can be accessed from the Shared Data tab at the top. To help clarify instructions the following colors will be used to signify the above data types:
Galaxy Tools
Items in your History
Items in Data Library

Starting Galaxy

There are several Galaxy instances on the web and each of them are slightly different. MSI hosts a Galaxy instance that is available to every MSI user. It is located at:
https://galaxy.msi.umn.edu

Get Tutorial Data

1. Go to https://galaxy.msi.umn.edu and login using your MSI account information
2. If needed create a new history using the gear icon in the top right corner of the page.
3. Name the history chip-tutorial.
4. Select Shared Data -> Data Libraries from the top blue bar.
5. Type Cistrome in the search bar and press enter to search.
6. Select the Cistrome data library
7. Open the CTCF folder using the blue arrow then select the check box next to G1E_CTCF.fastqsanger and G1E_input.fastqsanger
8. Select Go next to Import to current history at the bottom of the Data Library.

9. Select Analyze Data from the top blue bar to return to your current history.

Mapping Raw Reads

The first step in analyzing ChIP-Seq data, after initial QC and trimming, is the map the reads to the genome of interest. In our case we are using data taken from a ChIP-seq experiment done from murine G1E_ER4 cell line. This experiment used an antibody against mouse CTCF to try and identify where this transcription factors was bound in G1E_ER4 cells.

In this section we will map reads to the mouse genome (mm9), get mapping statics and remove duplicated reads caused by PCR errors.

Map Reads using Bowtie

1. Using the pencil icon next to G1E_CTCF.fastqsanger change the data type to fastqsanger under the Data Type tab.
2. Do the same for G1E_input.fastqsanger.
3. Under the NGS:Mapping header (or using the search bar) select Map with Bowtie for Illumina.
4. Change the reference genome to mm9_cononical
5. Select the G1E_CTCF as the FASTQ file
6. Leave the rest of the settings as default.
7. Select execute, this will create a new element in your history.
8. Select the name of the element in your history expanding the box. Select the recycle icon to reload the Bowtie settings used to map the CTCF FASTQ file.
9. Change the FASTQ file to G1E_input and select execute.
10. Under the NGS:Picard(beta) header or using the search bar select the SAM/BAM Alignment Summary Metrics tool.
11. Select the mapped reads from your history for the CTCF data (history element 3) as the dataset to generate statics for.
12. Rename the Title of the output file: aln metrics CTCF
13. Leave the other options as default and select execute.
14. Using the recycle icon run the SAM/BAM Alignment Summary Metrics tool using the input read mapping (history element 4).
15. Use the eye icon to view the output from the SAM/BAM Alignment Summary Metrics tool. What percent of reads aligned in each sample? (Hint: PCT_READS_ALIGNED)
16. Under the NGS: SAM Tools header select the rmdup tool.
17. Select Map with Bowtie for Illumina for the CTCF data as the BAM File and change the data to single end.
18. Run the rmdup tool on the input file as well.
19. Run SAM/BAM Alignment Summary Metrics on the outputs from rmdup. Did the read count go down? What is the percent of aligned reads?

Peak Calling

Now that we have aligned our de-duplicated reads to the genome we can use a peak caller to identify regions of the genome that have an enrichment of reads aka peaks. Peak callers tend to generate two type of files, discrete and continuous. Discrete files tend to be in BED format while continuous files tend to be in WIG or BigWIG format. BED files will tell you the discrete genomic location of your peaks i.e., chr12 100134 100308 while WIG/BigBWIG files are a summary of the signal (reads) across the entire genome. Both file types can be viewed in standard genome browsers.

Call Peaks using MACS

1. Rename the rmdup outputs so it is easier to know which mapping file is for the CTCF or Input file. Use the pencil icon.
2. Under the NGS: Peak Calling header select MACS
3. Rename the experiment MACS CTCF
4. Select the rmdup file from the CTCF sample as the ChIP-Seq Tag File and the rmdup file from the input sample as the Control File.
5. Set the Effective genome size to 1.87e9
6. Set the Tag size to 36, the length of the reads.
7. Select the check box next to Parse xls files into distinct interval files.
8. Select the check box next to Perform the new peak detection method (futurefdr)
9. Select execute.

Data from MACS

MACS creates 4 output files, peaks: bed, peaks: interval, negative peaks and a html report. The report is a log file that will give you information about your run. It also contains links that will let you download the other files in your Galaxy history as well as additional files that give you information about the model MACS calculated to identify the center of your peaks.

Peaks: bed and peaks: interval gives you information about the peaks MACS called from your data. Peaks: bed is a standard bed file giving the location of each peak called by MACS as well as the -10log(pvalue). Peaks: interval gives more information about the peaks such as length, where the summit of the peak is, number of reads in the peak, enrichment and FDR. This file will be the most useful when you want to select your strongest peaks.

Annotate Peaks

In this section we are going to select our top 100 most significant peaks then find any genes that these peaks may overlap with.

Find top 100 significant MACS peaks

1. Under the Filter and Sort header select the Sort tool.
2. Use the MACS output Peaks: interval as your dataset to sort
3. Select c7 (column 7) as your column to sort.
4. Leave the other options as default and select Execute.
5. View the sorted data using the eye icon in the history. Is the data sorted the way you wanted? How does p-value relate to enrichment, FDR and tags?
6. Under the Filter and Sort header select Filter.
7. Select the output from the sort to Filter
8. In the With the following condition: field type c1!='#' (Don't use any spaces).
9. This filtering step will result in a file that doesn't have the comment lines (#) but is still sorted by p-value.
10. Under the Text Manipulation header select the Select first tool.
11. Set Select first: to 100 and from to the filtered dataset.
12. Rename the resulting file Top 100 CTCF peaks using the pencil icon.

Identify genes that intersect with top 100 MACS peaks

1. From the Cistrome data library in the CTCF folder import mm9_refGene_chr19 into your current history. This is a BED file of genes on chromosome 19 of mm9.
2. With mm9_refGene_chr19 in your history under the BedTools header select Intersect interval files.
3. Select Top 100 CTCF peaks as the first BED/VCF/GFF/BAM file
4. Select mm9_refGene_chr19 as the overlap interval in this BED file.
5. Select "Write the original entry in B for each overlap" as what should be written to the output file. Then select execute.
6. The output from the intersection looks like two BED files lined up next to each other. The first 6 columns are from the peak file and column 7 – 18 are from mm9_refGene_chr19.
7. Under the Text Manipulation header select the Cut tool.
8. Cut columns c10 from the Intersection of Top 100 CTCF peaks and mm9_refGene_chr19. This will result in a list of genes that overlap MACS peaks.

Motif Analysis

There are several different motif identification software available both in Galaxy and at the command line. In this tutorial we use the SeqPos motif analysis tool from the Cistrome consortium to identify known motifs in the data.

Motif Analysis using SeqPos

1. Under the Cistrome header select SeqPos motif tool.
2. Select Top 100 CTCF peaks as the BED file
3. Select mm9 as the Cistrome Genome Assembly.

4. Select Homo Sapiens or Mus Musculus as species to filter the results
5. Leave the other options as default and select execute.
6. The HTML output from SeqPos gives clusters of motifs based on similarity. Motifs that are identified in the data and are similar to each other will be grouped together, similarity to top describes how similar each motif is to the top motif in the group. The zscores given for each motif is a good measure of the enrichment of that motif, the smaller the zscore the more enriched the data is for that motif.
7. Selecting the name of the motif will bring up another webpage with more information about the motif including what it looks like on both the positive and negative strands.