ChIP-seq hands-on tutorial using Chipster: STAT1 data set

Eija Korpelainen, CSC –IT Center for Science, Finland, chipster@csc.fi

In this section we call peaks, visualize them in genome browser, and filter them based on several criteria. We then retrieve the genes which are nearest to the peaks and perform pathway analysis for them. Finally, we also search for sequence motifs in the peaks and match the motifs to a database of known motifs.

Open example session course_ChIPseq_STAT1.zip. The session contains two samples of interferon-gamma stimulated HeLa cells: a treatment sample which was immunoprecipitated with STAT1 antibody, and a control sample of input DNA. The two BAM files were made by aligning reads to the human reference genome version hg18. For the interest of time only reads from chromosome 1 are included. Therefore we need to use the length of chromosome 1 in the following MACS2 runs.

-Open the treatment BAM file and check the length of chr1 in the header (2.47e8).

-Check also the length of reads by running FastQC

1. Detect binding locations for STAT1 with MACS2

Select the **treatment BAM** file and run the tool **ChIP, DNase, and Methyl-seq / Find peaks using MACS2** using the following parameters:
Mappable genome size = **user-specified**
User specified mappable genome size = 2.47e8

-How many peaks do you get (check the top of the file macs2-peaks.tsv)?

-How many reads were there and how long are they? Were any of them duplicates? To what length were the reads extended to the 3' direction (see macs2-log.txt)?

-Is the model plot smooth (see macs2_model.pdf)?

2. Detect binding locations for STAT1 with MACS2 using both treatment and control sample

Select both **BAM** files and run the tool **ChIP DNase and Methyl-seq / Find peaks using MACS2** so that you set the **Mappable genome size** as before and check at the bottom of the parameter panel that the files have been assigned correctly.

-How many peaks do you get now? How does the use of control sample affect the peak calling?

You can now delete the result files from exercise 1.

3. Visualize the peaks in the Chipster genome browser

-Open **macs2-summits.bed** as a spreadsheet, click **Detach** and put the new window aside.
-Select **both BAM files** and **macs2-summits.bed** and the visualization method Genome browser. **Maximize** the visualization panel, select **genome=Homo sapiens hg18**, and click **Go**. In the Settings tab, set the **coverage type = strand-specific** and **coverage scale = 250**.
-Sort the previously detached BED file by clicking on **Column4**. Click on the start positions of the peaks to navigate from one peak to another. Do the peaks have the bimodal shape expected?
-Zoom in and out with a mouse wheel.

4. Get the most significant peaks by filtering based on q-value

Select the file **macs2-peaks.tsv** and run the tool **Utilities / Filter table by column value** using the following parameter settings:

Column to filter by = neglog10qvalue
Cutoff = 10
Filtering criteria = larger than

-How many peaks pass the filter? **Rename** the result file to qfiltered.tsv

5. Filter out long peaks

Select the file **qfiltered.tsv** and run the tool **Utilities / Filter table by column value** using the following parameter settings:

Column to filter by = length
Cutoff = 1000
Filtering criteria = smaller than

-How many peaks do you have now? **Rename** the result file to length-filtered.tsv

6. Keep only high peaks

Select the file **length-filtered.tsv** and run the tool **Utilities / Filter table by column value** using the following parameter settings:

Column to filter by = pileup
Cutoff = 100
Filtering criteria = larger than

-How many peaks do you have now? **Rename** the result file to summit-filtered.tsv

7. Retrieve genes which are located closest to the peaks

Select the **summit-filtered.tsv** file and run the tool **ChIP, DNase, and Methyl-seq / Find the nearest genes for regions** so that you set the **genome** = **hg18**.

-Are all the peaks located upstream of genes (check the location column)?

8. Retrieve annotation for the nearby genes

Select **nearest-genes.tsv** and run the tool **ChIP, DNase, and Methyl-seq / Find unique and annotated genes**.

-How many unique and annotated genes did the list contain?

9. Pathway enrichment analysis using GO categories and ConsensusPathDB

Select **unique-genes.tsv** and run **ChIP, DNase, and Methyl-seq / GO enrichment for list of genes**.

-What is the second most significantly enriched GO category in our list of genes?

Select **unique-genes.tsv** and run the tool **RNA-seq / Hypergeometric test for ConsensusPathDB.** Check if any pathways involving STAT were found: Select the result file **cpdb-pathways.tsv** and run the tool **Utilities / Filter table by column term** using the following parameter settings:

Column to filter by = Pathway
Term to match = STAT

10. Find sequence motifs which are common in the detected peaks

Select **summit-filtered.tsv** and run the tool **ChIP, DNase, and Methyl-seq / Find motifs with GADEM and match to JASPAR** so that you set **Genome = hg18**.

-Did the peaks have the STAT1 binding motif (check logo-plot-1.pdf)?

11. Save session, get analysis history file, save a workflow

Save session: Select **File / Save local session**. Give a name to your session and save it.
Get a textual report: Select **hypergeo-go.tsv** and click on the history link in the visualization panel.
Save an automatic workflow: Select the file **macs2-peaks.tsv** and **Workflow / Save starting from selected**.