Chip-seq Analysis with Galaxy: from reads to peaks (and motifs)

Contents

---

1 - Introduction

This tutorial explains how to access a public dataset of ChIP?-seq data and calculate the peaks.
We will use a publicly accessible server that provides tools for Next Generation Sequencing (NGS) analyses: The **Galaxy server** (http://main.g2.bx.psu.edu/)
The goal of this tutorial is to perform the successive steps to obtain a list of peaks. We will first **retrieve the raw reads**, get basic information on this dataset, then perform the **mapping of the reads** on the reference genome to obtain their coordinates, and finally perform the **peak-calling** step, to look for clusters of reads forming peaks.
For this exercice, we will use a dataset produced by a study of transcription factors involved in the differenciation of stem cells. For time reasons, we will focus on one factor: **Oct4**. The ChIP?-seq experiment was conducted on **mouse** cells, on an **Illumina Genome Analyzer sequencer**. These two information are necessary before starting analyzing these data.

---

2 - Obtaining the raw data: Accessing ChIP-seq reads from ArrayExpress database

Goal: Identify the dataset corresponding to the article by Chen et al., 2008 (Pubmed ID: 18555785) and Retrieve the data for the **Oct4** experiment.
The easiest way is to achieve this goal is to use the GEO/SRA database at NCBI (USA). As of February 2011, the NCBI will slowly stop to be a repository for NGS data, due to the cost that it represents. The EBI in Europe has decided to continue their repository, so we will explain below how to get the sequences from the EBI.

1. Open another browser window to the ArrayExpress database.
2. In the *Experiment Archive* box, enter the title of the article:
3. And click on the **Query** button.
4. This should give a single result. Click on the **+** icon on the left side to see the information for this dataset.
5. On the top left side, click on the **small graphic icon** under the column **Raw**, a new page appears.
6. Unfortunatly, this representation does not list the name of the transcription factors in the headers (the only solution is to open each box to look for the name). For the sake of time, directly look for the box with **SRA Run: SRR002012**, that correspond to the Oct4 transcription factor.
7. In the lower table, under the column *FTP*, there is a link to the sequences in **FASTQ** format. Do not download the data, just copy the link location of the first dataset (right click on the "download"link -> copy link location): **ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR002/SRR002012/SRR002012.fastq.gz**. This will allow to directly transfer the data from the EBI to Galaxy, without transiting on your computer.

---

3 - Upload the reads in the Galaxy server

Goal: Connect to the Galaxy server and upload the dataset of raw reads

1. Open a new page on the Galaxy server (http://main.g2.bx.psu.edu/).
2. In the menu at the left of the window, click **Get Data > Upload File**.
3. For this exercise, we will upload the FASTQ read file from ENA. In the **URL/Text** box, paste the URL of the Oct4 sample:
   ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR002/SRR002012/SRR002012.fastq.gz
4. In *File format*, choose **fastqsanger**, **not** fastq**cs**sanger.
5. In the **Genome** menu, select *Mouse July 2007 (NCBI37/mm9)*. Tip: he genome is selected if you simply type *mm9* when the menu is selected.
6. Leave the other options to their default value, and click **Execute**. The upload may take several minutes. When the file will be uploaded, the yellow box on the right side will turn to green.
   - **Note:** the upload speed depends on the availability of the two

servers. Indeed, the ChIP?-seq reads were directly transferred from ENA to Galaxy, without transiting by your computer. .

1. Once the right box is green, click on this box and make sure that the format is fastqsanger and the genome is mm9. How big is this file (in Mb) ?

---

4 - Some statistics on the raw data

Goal: Get some basic information on the data (read length, number of reads, quality of dataset)

1. On the left side, there is a menu with all the tools available in Galaxy. There is a section **NGS Toolbox Beta**. Click on **NGS: QC and manipulation**, where QC means *Quality Check*
2. The various tools are ordered by sequencer types, with at the end some more generic tools to deal with the FASTQ sequences. In the **FASTX-Toolkit for FASTQ data** section, click on **Compute quality statistics**, and then click on the **execute** button. Wait until the job is finished.
3. Analyse your results by clicking on the *eye* icon in the green box: how many lines are there in the file ? Each line correspond to one read position. The number of lines is thus the read length.
4. How many reads are there in the file (check the column *count*).
5. The scale of quality values goes from *0* to *40*. In the column *mean*, this is the mean quality for each position of the read. This values decreases when getting to the end of the reads, because the Illumina sequencer is known to produce more errors at the end of the reads.
6. Let's check this visually: in the left menu, click on **Draw quality score boxplot**. Look at the produced plot by clicking on the *eye* in the green box. Our dataset is of relatively good quality (but not very good !), as the quality values only drops towards the end of the reads.

---

5 - Mapping the reads with Bowtie

Goal: Obtain the coordinates of each read on the reference genome

1. There are multiple programs to perform the mapping step. For reads produced by an Illumina machine, the currently "standard" programs are *BWA* and *Bowtie*. We will use Bowtie for this exercice. There is a section **NGS Toolbox Beta**. Click on **NGS: Mapping**, and then **Map with Bowtie for Illumina**
2. For the reference genome, keep *Use a built-in index* and select the mouse assembly **mm9** (Full)

3. Keep *single-end* for the library
4. The FASTQ file should be your read file (which is in FASTQ format)
5. In the *Bowtie settings*, choose **Full parameter list**. As you can see, this program has many parameters !!!. We will only change few ones:
6. Change the **Maximum permitted total of quality values at mismatched read positions (-e):** to **40**.
7. Change the **Number of mismatches for SOAP-like alignment policy (-v):** to **2**, which will allow two mismatches anywhere in the read, when aligning the read to the genome sequence.
8. Change the **Suppress all alignments for a read if more than n reportable alignments exist (-m):** to **1**, which will exclude the reads that do not map uniquely to the genome.
9. Click on the **execute** button to launch the mapping. This is the longest step of this protocol, wait until the job is finished (it usually take few minutes, but this is a good time to take a break !!).
10. The output is SAM format, which contains all reads (mapped and not mapped), along with flags indicating whether there are mapped or not, their quality values and their genomic coordinates (only for the mapped ones)
11. For the following steps, we are only interested in the mapped reads. We are going to filter out these reads: click on **NGS: SAM Tools**, and then **Filter SAM**
12. Click on **add a new flag** button, then in the *Type* menu, select **read is unmapped**, and then select **No**. Indeed, we do not want the unmapped reads (= we want the mapped ones).
13. Click on the **execute** button.
14. How many lines are there in this final file ? This represent the number of mapped reads. Calculate the percentage of mapped reads for this experiment.

---

6 - Peak calling with MACS

Goal: Define the peaks, i.e. the region with a high density of reads, where the studied factor was bound.

1. There are multiple programs to perform the peak-calling step. One of the currently "standard" programs is *MACS*. In the section **NGS Toolbox Beta**. Click on **NGS: Peak Calling**, and then **MACS**
2. Enter an **Experiment Name** (e.g. *OCT4 Chen-2008*).
3. For the **ChIP-seq tag file**, select the **filtered SAM file** you created in the previous step.
    - **Note:**For this exercice, we dispose of a single set of reads, so we will run the peak-calling without providing any control. For a *real analysis, you would need to provide the control dataset* !
4. **Effective genome size**: this is the size of the genome considered

"usable" for peak calling. This value is given by the MACS developpers on their website. It is smaller than the complete genome because many regions are excluded (telomeres, highly repeated regions...). The default value is for human (2700000000.0), as we work on mouse, enter **1870000000.0**

5. Set the **Tag size** to 26bp (the default is 25).
6. Leave all other options to their default values and click **Execute**.
7. While the program is running, two yellow boxes should appear in the "History" frame at the right of the Galaxy Window. After completion of the job, the boxes will be colored in green. The first box contains an HTML page with links to the results in various formats. The second box contain a BED file with the coordinates of the peaks. How many peaks ("regions") were detected by MACS ?

---

7 - Retrieving the peak sequences

Goal: Retrieve the sequences from the peak coordinate file (BED)

1. In the left menu of Galaxy, click on **Fetch Sequences > Extract Genomic DNA**. Your peak dataset (bed) should be selected. click on the ***Execute*** button.
2. Once the box become green in the History frame, click on the **pencil icon** and rename the data set (for example **Oct4 peaks sequences**).
3. If you wish to download the sequences, open the green box and click on the **disk icon** to store the result on your computer (for example in a file *Oct4_MAC_peak_sequences.fasta*).

---

8 - Visualize the peak regions in UCSC genome browser

1. In the green box of the MACS results, simply click on the link **display at UCSC main**.
2. A new page opens. Your peak regions are displayed in the first *track*. You will need to zoom on one peak to better see its gene environment.

---

9 - Try to identify over represented motifs

Goal: Use the sequences under the peaks to identify an Oct4 specific binding motif

1. Go to the peak motif website http://rsat.ulb.ac.be/rsat//RSAT_home.cgi
2. Start a new analysis (enter a meaningful title)
3. Paste the URL of the sequences that we have extracted (you find it at

the little disk symbol in the history pane of galaxy)
4. Start the analysis (select display as the output) and wait a bit
5. Check if the known Oct4 motifs were found

HINTS to refine the analysis:

1. Use only highly significant peaks (column 5 of the bed file output of macs contains -10*log(P-value))
2. You can get histograms and summary statistics from galaxy to decide on a threshold
3. Find the 75 percentile of the score distribution
4. Filter the bed file to retain only peaks with score > 75th percentile
5. Repeat the sequence retrieval and motif analysis with this set
6. Do you find an Oct4 motif now?

---

10 - Annotate peaks with genes

Goal: Assign the closest gene to each of the top scoring peaks Obtain the mouse gene annotation from UCSC

1. Go to ucsc table browser (http://genome.ucsc.edu/cgi-bin/hgTables)
2. Select mm9 known genes
3. Tick the checkbox "Send to galaxy" and get the output directly into your galaxy session
4. Now we need to transform the text into bed format
5. Text manipulation: cut (c2,c4,c5,c1,c13,c3)
6. Use the pencil in the history pane to change the format to bed
   Finally we can assign peaks to genes
7. Operate on intervals
8. Fetch closest non-overlapping feature
9. Select filtered regions and gene annotation
   Now we have genes for each peak
10. Compute the distance between peak and gene (use text manipulation, compute expression)
11. Plot a histogram of the distance distribution

---

11 - Functional enrichment analysis of the peaks

Goal: Find functional categories over-represented in Oct4 targets

1. Remove the score and name columns from the bed file (cut columns c1,c2,c3)
2. Save the complete peak list as bed file on your computer (disk symbol)
3. Go to http://great.stanford.edu

4. Upload the file
5. Select mouse genome
6. Run the analysis
7. What biological process is enriched for Oct4 targets?

---

12 - Shared results on Galaxy

http://main.g2.bx.psu.edu/u/morgane/h/fromreadstopeaks

References

- Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., Wong, E., Orlov, Y. L., Zhang, W., Jiang, J., Loh, Y. H., Yeo, H. C., Yeo, Z. X., Narang, V., Govindarajan, K. R., Leong, B., Shahab, A., Ruan, Y., Bourque, G., Sung, W. K., Clarke, N. D., Wei, C. L. and Ng, H. H. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell 133, 1106-17.

For suggestions or information request, please contact *Morgane Thomas-Chollier* **thomas-c[at]molgen.mpg.de** or *Matthias Heinig* **matthias.heinig[at]molgen.mpg.de**
