# ChIP-seq hands-on practical using Galaxy

In this exercise we will cover some of the basic NGS analysis steps for ChIP-seq using the *Galaxy* framework:

- Quality control
- Mapping of reads using Bowtie2
- Peak-calling using MACS
- Assign peaks to genes

We will work with two cell lines, each with two replicated sample-control experiments.

## Startup / Login to Lifeportal Galaxy

We will use the UiO Lifeportal instance of Galaxy. Open a web browser and navigate to [https://lifeportal.uio.no](https://lifeportal.uio.no)

Select "User|Login" from the Galaxy front-page and log in with your Feide or course credentials (received via email invite).

The CPU hours requires for compute on Lifeportal are tallied in a project that you own.
**For this course, use the project name "lp51" for all calculations**

## Obtain example data

Obtain data. When working with your own data, you typically upload files via the `Get Data | Upload File` function in the left menu. Today, for convenience and to save time, I have made the example data available as shared histories on the Lifeportal server.

- Click the `Shared data` menu option and choose `Published histories`.
- Choose the set "INFBIOx121_ChIP_0_Fastq_raw_seq" for the next exercise.

For manual loading (another time):

```
* Specify "fastqsanger" as file format and "Homo sapiens, hg19" as database
* Tip: Close the "Get Data" sub menu again by clicking it's name
* Rename your fastq file element in the history to shorter meaningful names
revealing cell-line, TF/Ctrl name and replicate no (Tip: Click the pencil
icon to edit the item in the history).
```

## Quality Control

First we'll will use the FastQC tool to get some statistics on the reads in the *fastq* raw data files for the ChIP-seq experiments.

1. Rename your history to "QC on H1hsec" by clicking on the history name to edit it
2. Open the "NGS: QC and manipulation" sub menu in the left menu and select the "FastQC: Read QC" tool
3. Check that your fastq file is suggested correctly as the input and start the tool
4. This should produce a new element in your history, inspect it and review the results of the tool in the centre panel.
5. Note the read length of the data (needed later)
6. Try the same procedure for a second fastq file, the one matching to your previous file in this sample-control pair.

# Mapping

Due to time constraints we will only do "live" mapping of 1 fastq file and import pre-aligned bam files for the others later on.

1. In the left menu, open the "NGS: Mapping" sub-menu and select the "Bowtie2" tool.
2. Select your "H1hsec Egr1" sample fastq file, select "chr19" as your reference genome, and use defualt settings for the rest.
3. Execute your mapping job, this can take some time (15-20 min depending on server load/your computers capacity), there will come a new data item in your history that should turn from grey via yellow to green when the job is done.
4. Rename your new bam file entry in the history to something meaningful containing cell-line, TF/Ctrl name and rep no.
5. For visual inspection of the results, either
   - try to visualize your bam file using Trackster (horizontal bar chart icon in the detailed view of the bam file entry in the history)
   - or download the bam file and inspect it in IGV (if available)

ChIP-seq does not have the same clear expectations for read distribution across the genome as RNA-seq data does. But for curiosity we might ask "how many of the reads binds within 10kb upstream of the genes?".

1. The "Get data | UCSC Main" tool let us query and download interval data in bed and gtf format easily.
   - Navigate to the hg19 genome, select "Genes and Genes Predicions", only for "chr19:1-59128983", tick "Send output to Galaxy", specify BED format and click the "get output" button.
   - In the next page, tick and fill in "Upstream by 10000 bases" and click "Send query to Galaxy" button.
2. You will then get a new BED file item in your history.
3. Use the "NGS:Something | Slice bam file" tool to cut out the reads from your bam file that overlap the BED file. (Find the tool by searching by name in the top form input on the left side menu!)
4. The number of reads in the resulting bam file now tells you how many out of the total mapped reads that lie within 10kb upstream of the genes on chr19. How big is this percentage?

# Peak-calling

First, import the complete set of mapped reads to your Galaxy history. Again,

- Click the `Shared data` menu option and choose `Published histories`.
- Choose the set "INFBIOx121_ChIP_1_Mapped_reads" for this exercise.

Now we have several bam files: for each cell line, there are two biological replicates for TF bound DNA (ChIP) and non-enriched control ("input DNA").

The next task is to call peaks of reads that are significantly enriched in each sample replicate over the matched control. For this we use the MACS tool.

1. Open the MACS tool under "NGS: Peak-calling | MACS".
2. Name the experiment "MACS on H1hsec rep 1" (or other sample)
3. Select your TF sample as the "ChIP-Seq Tag File"
4. Select your Control as the "ChIP-Seq Control File"
5. Enter the effective genome size as the length of chr19: 59128983
6. Tag size is the read length that you found, for that sample, under the FastQC step above :)
7. Since we have downsampled the original data, we also lower the minimum MFOLD high-confidence ratio to 10.
8. Leave other parameters at default.
9. Execute

# Visualize your data in the UCSC browser

Let's then examine our results visually. Either in Trackster inside Galaxy (although this function at Lifeportal never finished preparing the tracks during my testing) - click "Visualize" in the top Galaxy menu

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

1. Switch over to chromosome 19 in the UCSC broswer.
2. The called peaks are named "MACS_no", so by using the position/search field, try to inspect the first peaks and their relation to gene annotation and functional annotation from the Encode project.
3. Can you find and activate the coverage plot for the data that we have just analyzed (H1esc Egr1) among all the Encode data available in the UCSC browser?

# Assign peaks to genes

The next step will be to assign peaks to the "closest" gene. To do this we need to combine our bed interval file for the peaks, with one file representing genes. Galaxy provides by default many tools that works with combinations of genome interval files. We will use two of them that:

- Find genes with Trancription Start Site (TSS) that overlaps with our MACS peak calling results.
- Find the closest non-overlapping TSS for all peaks.

Then we first need a bed file encoding the TSS in an interval file like bed foramt. This we have prepared for you for chr19, and is available at the course download webpage. We will use the tools under the **Operate on Genomic Intervals** sub menu of the tools menu to the left. So, let's get started.

1) Create transcription start sites, from your UCSC gene annotation 2) Use the **Intersect** tool to see how many of our called peaks overlap with a TSS. 3) Then we'll use the **Fetch closest non-overlapping feature** tool to find the nearest TSS not overlapping with our peaks.

# Next replicate

Since we have two replicated experiments for our TF ChIP-seq data, start from the two bam files from the second replicate do the peak-calling and assign genes to the peaks afterwards.

1. How many peaks were called in each of the replicate experiments?
2. How many of these peaks are overlapping between the replicates?

# Next cell line

We also have a second cell line Mcf7 with two replicated ChHIP-seq experiments for the same TF.

1) Create a new history and call it "Mcf7 MACS" 2) Start from bam files and do peak calling and assign to genes 3) How good is the overlap between the replicates in this cell line?

# Between cell line comparison

This exercise you can either do in Galaxy (get your bed files holding peaks called in both replicates of a cell-line into a new history) or using bed-tools from the commandline (tip: intersectBed commandline tool)

1) How many of the peaks called by both replicates in a cell-line is present in both cell lines? 2) And how many peaks are unique for each cell-line

# Retrieving the peak sequences

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

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.

Once the box become green in the History frame, click on the pencil icon and rename the data set (for example H1hesc Egr1 peak sequences).

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 H1hesc_Egr1_MACS_peak_sequences.fasta ).

# Try to identify over represented motifs

Goal: Use the sequences under the peaks to identify an Egr1 specific binding motif

Go to the peak motif website [http://rsat.ulb.ac.be/peak-motifs_form.cgi](http://rsat.ulb.ac.be/peak-motifs_form.cgi)

Start a new analysis (enter a meaningful title)

Paste in the URL of the sequences that we have extracted (you find it at the little disk symbol in the history pane of galaxy)

Start the analysis (select display as the output) and wait a bit Check if the known Oct4 motifs were found HINTS to refine the analysis: Use only highly significant peaks (column 5 of the bed file output of macs contains -10*log(P-value) ) You can get histograms and summary statistics from galaxy to decide on a threshold Find the 75 percentile of the score distribution Filter the bed file to retain only peaks with score > 75th percentile Repeat the sequence retrieval and motif analysis with this set Do you find an Oct4 motif now?