Final exercise ChIPseq analysis

Monica Sanchez Guixe

22/03/2020

1. Complete the Git exercise 1 (see hands-on).

Results:
http://github.com/msguixe/git_HandsOn

2. Complete the Docker exercise 1 (see hands-on). Make a Git repository containing the Dockerfile that you used to build your image and push it to GitHub. Add the URL that points to your image in DockerHub in the 'description' of the Dockerfile.

Results:
http://github.com/msguixe/Docker_exercise
http://hub.docker.com/msguixe/image_venn:final

Make another GitHub repository with the answer to the following points:

Results:
http://github.com/msguixe/Final_exercise_ChIPseq

3. Describe the workflow of a ChIP-seq experiment (experimental part)

The chromatin immunoprecipitation (ChIP) assay consists on the pull-down of DNA regions where a specific transcription factor binds or that contain a histone with a specific histone modification.

The first step of the experimental procedure is to cross-link proteins to DNA by adding formaldehyde to the cell media of cell culture plates (at 0.75%) or to tissue homogenates (at 1.5%). Crosslinking reaction is stopped using glycin.

The second step is to get the cell lysates of proteins and DNA. Cells or sample tissue are homogenized with a lysis buffer containing, among other elements, mild detergents (e.g. NP-40) to desintegrate the cell membrane; and protease-inhibitors (e.g. aprotinin, leupeptin, etc.) to avoid teh degradation of the proteins by the proteases present in the cells.
The third step is to shear the chromatin into smaller fragments by sonication. Ultrasounds will break the DNA into fragments between 200-1000bp (sonication conditions must be optimized for each sample type). After the sonication, samples are centrifuged to eliminate cell membrane fragments and conserve the DNA and protein part in the supernatant.

The fourth step is the immunoprecipitation (IP). In this step a selective antibody is used to bind specifically to our transcription factor (TF) or histone modification (HM) of interest. First, a portion of the lysate will be set apart and will be used as input control. Another portion of the lysate my be used as an IP control, using a non-specific antibody (IgG). The rest of the lysate is incubated with the specific antibody. If we are interested in more than one TF or HM, the lysate is splitted for each of the antobodies. In order to know how much sample we should put in each IP, DNA concentration must be calculated (an amount of 25 ug of DNA is recomended per IP). Once the antobody is mixed with the lysates, samples are incubated 1 hour at 4ºC in a rotating wheel. Then, samples are mixed either with protein A/G beads or magnetic beads, which will bind to the common fraction of the antibodies, and incubated over night at 4ºC. The next day, samples are centrifuged (for protein A/G beads) or magentized (for magnetic beads) to separate and remove the lysate from the beads. After several washing steps with increasing salt concentration, we should have the

DNA fragments, bound to the TP or histone containing the HM of interest, which are bound to the specific antibody and this to the the protein A/G of the beads.

The fifth step is to elute the DNA fragments and reverse the cross-linking following a DNA purification protocol such as phenol:clorophorm extraction.

The sixth step is the sequencing of the eluted DNA samples. Samples are processed by a massive sequencing platform in order to obtain all DNA fragments bound to the TF or histone containing the HM of interest. This step can be performed as single or paried-end, where single-end refer to sequencing performed only by 5' end only and paired-end to sequencing performed with both 5' and 3'-end.

4. Why do we need a control in ChIP-seq experiments?

We need a control because the IP procedure may not be optimal and some DNA regions may tend to precipitate with the beads or with the non-specific regions of the antibody (common fraction) even though they are not specifically targeted through the specific part of the antibody. For that, unsepcific IgGs may be incubated in a portion of the cell lysates and immunoprecipitated parallelly. If lacking an unspecific antibody to use as control for the IP, we could use only beads to incubate with teh lysate. In this case, the control will refer to all DNA/protein fragments bound to the beads, and not to the common regions of the antibody.

Another important control is the input. Conversely as the IP controls, this shows the initial proportion of DNA fragments in all the cell lysate previous to the IP. This means that in the input we are taking into account the relative initial amounts of DNA/protein in the samples but not those that could be enriched by non-specific bindings. However, input control reflects variations due to effective chromatin shearing, where we may have longer or shorter fragments due to open/condensed chromatin which differ effective sequencing, as shorter fragments are easier to sequence and may give false positives. Moreover, the use of different sequencing platforms may create a bias that generates certain variation in the samples.

**5. Explain at which stage of a ChIP-seq processing analysis and how we combine the information of ChIP and control reads.
Describe the output files we get if we are using the tool MACS2.**

we use the control samples to generate the fold-change peaks: the peak calling step. Once we have the sequenced fragments aligned in a genome (BAM files), we use MACS2 peak calling algorythm to generate the peaks.

The final peaks are generated by selecting those peaks in the ChIP sample that are enriched mfold times in respect to the control sample (input or unspecific IgGs/beads alone), in order to avoid non-specific enrichment of the given DNA sequence due to experimental conditions. Input DNA is often used as control rather than IP-related controls, which tend to have too little DNA material and, consequently, show a limited number of reads. The mfold value is set depending on the quality of the sample and the confidence of the obtained results. The signal of the final peaks are the fold-change of the ChIP peak in respect to the control peak.

The output files that we obtain after using MACS2 tool are the wiggle files (containing the position and the signal of the peaks) and the bed files (containing the chromosomes, the coordinates of the genes -start and end positions- and the presence of peaks). These are the uncompressed files, in a readable format. However, for easire manipulation of these files, these are compressed in a binary format: the bigWig and bigBed files, respectively. Besides the bigWig file with the fold-change of each peak, the pipeline generates another two bigWig files: one containing the p-values, and the other containing the pile-up reads. These data is given in 3 different bigWig files in order to reduce the amount of data in each file and facilitate the analysis (facilitate the compunting capacity).

**6. What is a pipeline?
In the case of the ENCODE ChIP-seq pipeline, which steps of the analysis of ChIP-seq data does it contain?**

A pipeline is a multistep process where experimental datasets are processed sequentially by different algorithms, software tools or fileformat manipulation.

In the case of the ENCODE ChIP-sep, the first step is the mapping of the FASTQ files: a genome of reference is used for genome indexing (through BWA tool). The indexed genome is then served as input, together with the experimental reads (FATSQ files), for a second step of alignment, fastq concatenation and filtering (through BWA, Samtools, Picard, Phantompeakqualtools and SPP tools). This creates two output files: the alignments and the unfiltered alignments.

The second step of the pipeline is peak calling. Here, the alignments are used as input for the peak calling and signal generation (through MACS2 and BEDtools). This creates 3 different outputs: the fold-change over control, the signal p-value and the peaks. The peaks are then processed differently depending if the experiment has replicates or not. In the case of having replicates, the peaks are processed for replicate concordance (through overlap_peaks.py), generating a file of replicated peaks; this file contains peaks found in both replicates or in two pseudoreplicates (random subsets of half of the sample). In the case of not having replicates, they are processed for partition concordance (through overlap_peaks.py), generating a file with the stable peaks; these peaks represent those that overlap at least 50% in both pseudoreplicates.

These files are then processed for file-format conversion: the pre-processed peaks, the replicated peaks (in the case of a replicated experiment) and the stable peaks (in the case of a unreplicated experiment), through bedToBigBed tool, generating bigBed files.

7. For the same EN-TEx donor that we have used in the hands-on session in class, use the Experiment Search Toolbar from the ENCODE portal to find all released experiments testing chromatin accessibility in stomach and sigmoid_colon (assembly GRCh38).

- **Paste here the filters you have applied**

Assay type: DNA accessibility
Status: released
Biosample term name: sigmoid colon, stomach

- **How many experiments are there?**

There are 4 experiments.

- **Paste here the link to download the corresponding metadata file.**

https://www.encodeproject.org/metadata/?type=Experiment&replicates.library.biosample.donor.uuid=d370683e-81e7-473f-8475-7716d027849b&status=released&status=submitted&status=in+progress&assay_slims=DNA+accessibility&biosample_ontology.term_name=sigmoid+colon&biosample_ontology.term_name=stomach

8. Download the metadata retrieved in point 5. Parse it to get:

- **File ID of bigWig file for fold-change over control in sigmoid_colon ATAC-seq experiment**
- **File ID of bigWig file for fold-change over control in stomach ATAC-seq experiment**
- **Paste the code used and the corresponding IDs.**

File ID for sigmoid colon:

```
grep -F ATAC-seq metadata.tsv | grep -F sigmoid_colon | grep -F fold_change_over_control | awk '{print $1}'
```

ENCFF997HHO

File ID for stomach:

```
grep -F ATAC-seq metadata.tsv | grep -F stomach | grep -F fold_change_over_control | awk ‘{print $1}’
ENCFF415RKU
```

9. What is an aggregation plot?

An aggregation plot is a plot that representents the distribution of continous ChIP-seq signals, such as fold-change, pile-ups or p-values, througout the whole sample in a given genomic coordinates.

- **Which tool do we use to generate one?**

We use the function *aggregation.plot.R.*

- **Which input data do we need?**

We need the tsv files with the aggregate signal of the most-expressed genes and the least-expressed genes, and the tissue type of our sample.

The files with the aggregate signal of the most and least-expressed genes are generated through the funtion *aggregate* from *bwtool*, using the genomic coordinates (in a bed file) specifying the start site by which we want to delimitate the genomic range of interest (e.g. promoter regions) and the bigWig files containing the continous signal (e.g. fold-change, p-value, pile-up signal).

- **Have a look at the aggregation plot done during the hands-on**

  - **Are the plots consistent between the two tissues?**

  The two plots generated in the hands-on (sigmoid colon and stomach samples) are consistent: they show a very similar distribution of the fold-change of the peaks in the promoter areas of protein-coding genes, in both least-expressed and most-expressed.

  - **Is this what you would expect, given the relationship between H3K4me3 and gene expression?**

  Yes, because this methilation is found more abundant in the most expressed and less to the least expressed genes.

  - **Why is it important to know the approximate location of a specific histone mark with respect to the gene?**

  It is important to know the approximate location of a specific histone mark with respect to the gene because we need to specify in the function *aggregate* from *bwtool* the genomic ranges of interest. This way, we can study the overall presence of the histone mark in the sample only in those areas where it is more present.

10. What type of plot are we using to visualize the correlation between two variables?

We use an x-y correlation plot or scatterplot. If there is a correlation between the variable in the x-axis and the y-axis, the distribution of the data can be adjusted in a function with a non-zero slope. The more the data points adjust to the function the more confident is the correlation, while if more dispersed, the less confident is the correlation.

- **Have a look at the plots generated during the hands-on to assess the correlation between expression and H3K4me3**

  - **Are these results consistent between the two tissues?**

The plots generated in the hands-on session are consistent between both tissues: none show a clear correlation between overall expression and the histone mark level.

- **Would you expect this degree of correlation? Formulate an interpretation of the results.**

The observed correlation between the histone mark and the gene expression is very low, and this is consistent with the fact that gene expression is multifactorial: there are multiple variables that can influence the expression of genes, such as the status of the chromatin (condensed/open), the regulation of transcription factors, other histone marks, etc. Moreover, the status of this histone mark (H3K4me3) can only explain 8 degrees of expression, given that each nucleosome has 4 histones that can be tri-methylated, and this is the same for both DNA copies (from the mother and the father).

11. During the hands-on session, we have checked the level of expression of genes with tissue-specific H3K4me3 marking.

- **Are these results consistent with the degree of correlation we have observed in point 10?**

The boxplot from the hands-on session show the distribution of the expression values in each of the grouped genes: marked in both tissues, not marked in both tissues, marked only in sigmoid colon and marked only in stomach. These plots show that the distribution of the gene expression levels is higher when the mark is present (either in both tissues or tissue-specific), whereas when it is not present the expression levels are lower. This representation of the data is a better option to show the differences in gene expression caused by the presence or absence of this mark than the scatterplot from the previous step, given that we are separating the datasets by the presence or absense of this histone mark specifically, whereas the scatterplot shows the overall expression across different degrees of the tri-methilation of the histone.

- **Do you observe any unexpected behavior?**

The only unexpected behaviour I can observe is that the expression levels observed in the tissue-specific marked genes is not as high as the expression levels observed from the genes marked in both tissues. As I would expect equal or higher level of expression for the tisse-specific marked genes.

- **How would you relate the presence of genes with tissue-specific marking with the GO terms obtained?**

The genes that contain the H3K4me3 histone mark specifically in sigmoid colon are significantly related to only 7 GO terms: 1) homophilic cell adhesion via plasma membrane adhesion molecules, 2) myelination, 3) anterior/posterior axis specification, 4) peptide ligand-binding, 5) adult locomotory behavior, 6) NABA ECM glycoproteins and 7) regulation of system processes. These pathways are difficult to relate to the context of sigmoid colon organ and its function. Due to the low number of GO terms obatined and the difficulty to relate those with the context of sigmoid colon, it is not optimal to drive conclusions over these results.

In the other hand, the genes that contain the H3K4me3 histone mark specifically in stomach are significantly related to up to 100 different GO terms, of which 8 present a -log10(P) superior to 10. Many of these pathways are related to immune system: lymphocyte activation, immunoregulatory interactions between a Lymphoid and non-Lymphoid cell, adaptive immune response, T-cell costimulation, etc. This could be due to all the published data about *Helicobacter pillori* and its relashionship with the immune system. As well, the GO term digestion appears among the most significant terms, which denotes a certain degree of reliability to these list of GO terms.

Overall, we can only make consistent speculations in stomach tissue due to the limited number of significant pathways and its poor correlation with its function in sigmid colon. The significant GO terms in stomach are more robust because they are higher in number and in significance, and the GO terms have relation to the function of the organ.

12. Have a look at the Venn diagram generated in the last task.

- **Comment on the number of peaks shared: is there more sharing between peaks of different type in the same tissue (e.g. H3K4me3 & POLR2A of stomach), or between peaks of the same type in different tissues (e.g. H3K4me3 of stomach and sigmoid colon)?**

The number of shared peaks of H3K4me3 peaks between both tissues is 14 889 peaks, and the number of shared peaks between both tissues for POL2RA peaks is 5 872 peaks.

When lookin at the shared peaks of both peak types in each of the tissue, we find that there are 6 332 shared peaks in stomach and 9 120 shared peaks in sigmoid colon.

Overall, the shared peaks between both tissues is higher (14 889 + 5872 = 20 761 shared peaks) than the shared peaks between both peak types (6 332 + 9 120 = 15 452 shared peaks).

However, the number of POLR2A peaks shared between both tissues is inferior to the number of shared peaks of both peak types in the two tissues, stomach and sigmoid colon (5 872 *vs.* 6 332 and 9 120 shared peaks, respectively).

We can conclude that the top shared peaks are between both tissues for H3K4me3 peak type.

13. (Extra question) Compute the percentage of genes with peaks of H3K4me3 and H3K27ac in the same donor and tissues we have used during the hands-on

There are 13 453 genes with peaks in both H3K4me3 and H3K28ac histone marks in sigmoid_colon tissue (97.6% of H3K27ac peaks and 89.8% of H3K4me3 peaks in sigmoid colon), and 13 725 genes with peaks in both histone marks in stomach tissue (97.7% of H3K27ac peaks and 88.4 of H3K4me3 peaks in stomach).

For both tissues, there are 12 967 genes with peaks of both histone marks (94.1% of H3K27ac and 86.5% of H3K4me3 in sigmoid colon; 92.4% of H3K27ac and 83.5% of H3K4me3 in stomach).

- **Provide the code**

The code is depicted in the Code_for_H3K27ac.txt file.

- **Provide the Venn Diagram of the intersection**

Venn diagram shown in the Venn.Diagram.H3K4me3.H3K27ac.png file.

References

https://www.abcam.com/protocols/cross-linking-chromatin-immunoprecipitation-x-chip-protocol
https://www.abcam.com/epigenetics/chromatin-preparation-from-tissues-for-chromatin-immunoprecipitation-chip