Epigenomics Task 3

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

- Treatment of cells/tissues of interest with formaldehyde to glue proteins to DNA

- Cell disruption and sonication to get chromatin fragments of 100-300 bp

- Selection of the fragments bound to the protein of interest by immunoprecipitation (using a specific antibody for our protein and then a secondary antibody)

- Separation of proteins from DNA fragments by heating the samples and washing to get the DNA fragments only

- Sequencing or microarray analysis of DNA fragments

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

We need control samples because:

- the DNA fragmentation step is not uniform and accessible chromatin regions are more represented which may lead to false-positive enrichment in these regions

- there are platform-specific sequencing biases which could yield non-uniform 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.

Control reads are used in the peak calling step of the ChIP-seq processing analysis.

Information of ChIP and control reads is combined to calculate the fold-enrichment of the signal. The fold enrichment is the amount of reads in a ChIP sample relative to a control sample.

The files obtained from the MACS2 peak caller algorithm are BED files. BED files are compact and provide information of the peak localization and statistics. In particular, MACS2 BED files provide: the chromosome, the start and end region of the peak, a peak id, a score for display in the USCS genome browser, NA for strand information, fold-change at peak summit, unadjusted p-values, empirical FDR adjusted p-values, and the summit position relative to the peak start. In addition to the this information, MACS2 also provides 3 types of signals at the nucleotide level: the pile-up aligned reads, the fold-enrichment, and the p-value. These files are provided as more compressed files called bigwig.

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 way of automatizing a number of steps. It has inputs that go through the different steps and some outputs are produced during the pipeline and/or at the end of it.

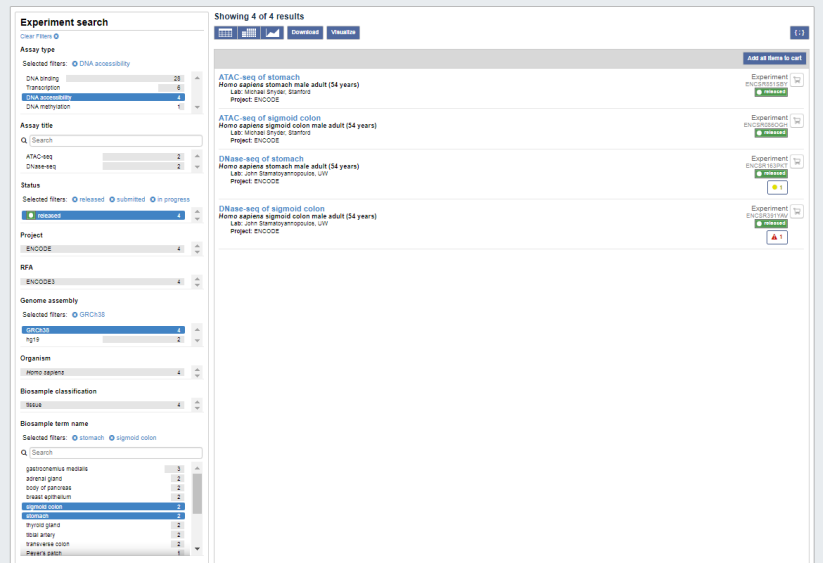
A pipeline is a workflow that concatenates different tasks and allows us to perform in a continuous manner these tasks.

The ENCODE ChIP-seq pipeline contains the steps of:

- Mapping
- Peak calling, which includes peak statistics, and peaks obtained from the analysis of biological replicates or pseudo-replicates, for either histone marks or transcription factors.

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



● How many experiments are there? 4

● 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+acces

sibility&biosample_ontology.term_name=stomach&biosample_ontology.term_name=sigmoid+colon&assembly=GRCh38

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.**

```
../bin/download.metadata.sh
"https://www.encodeproject.org/metadata/?type=Experiment&replicates.library.biosample.donor.uuid=d370683e-81e7-473f-8475-7716d027849b&status=released&status=submitted&status=in+progress&assembly=GRCh38&biosample_ontology.term_name=sigmoid+colon&biosample_ontology.term_name=stomach&assay_slims=DNA+accessibility"
```

```
grep -F ATAC-seq metadata.tsv | grep -F sigmoid_colon | awk 'BEGIN{FS="\t"}$2=="bigWig"{n++}END{print n}'
```

```
grep -F ATAC-seq metadata.tsv | grep -F stomach | awk 'BEGIN{FS="\t"}$2=="bigWig"{n++}END{print n}'
```

There were 2 bigWig files for each tissue because one corresponded to the signal p-value and the other to the fold change. To select just the fold change files I did:

```
grep -F ATAC-seq metadata.tsv | grep –F stomach |awk 'BEGIN{FS=OFS="\t"; print "file_id\ttissue\ttarget" } $2=="bigWig" && $3=="fold_change_over_control" && $44!="hg19" {print $1, $7, $19}' > bigWig.FC.s.ids.txt

grep -F ATAC-seq metadata.tsv | grep –F sigmoid_colon |awk 'BEGIN{FS=OFS="\t"; print "file_id\ttissue\ttarget" } $2=="bigWig" && $3=="fold_change_over_control" && $44!="hg19" {print $1, $7, $19}' > bigWig.FC.sc.ids.txt
```

- ENCFF997HHO is the file ID of the bigwig file for FC over control in the sigmoid colon ATAC-seq experiment

- ENCFF415RKU is the file ID of the bigwig file for FC over control in the stomach ATAC-seq experiment

9. What is an aggregation plot? Is a plot that is used to examine the distribution of continuous ChIP-seq signals over a set of genomic coordinates.

● **Which tool do we use to generate one?** bwtool to compute the aggregated FC signal and R to draw the plot

● **Which input data do we need?** bigwig fold change files, a Gencode annotation file, and expression matrices of the tissues we are working with

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

○ **Are the plots consistent between the two tissues?** results are quite consistent between tissues. Peaks are in regions of about 500 bp from the TSS. The shape of the signal in highly and lowly expressed genes is similar in both tissues. The most expressed genes have a FC signal peak around the TSS much larger than the least expressed genes. However, results are not exactly the same. In stomach, the FC signal peak is between 25-30 for the most expressed genes and between 5-10 for the least expressed. In sigmoid colon, the FC signal peak is between 30-35 for the most expressed genes and around 5 for the least expressed genes. These differences indicate that there is some degree of tissue-specificity of the regulatory regions that control TSS.

○ **Is this what you would expect, given the relationship between H3K4me3 and gene expression?** H3K4me3 is associated with transcription activation of nearby genes, so I would expect a much higher peak in highly expressed compared to lowly expressed genes as it is observed.

○ **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 as we know the association with gene expression of some histone marks, we can define whether close genes are expressed/repressed in different tissues, after certain condition, etc.

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

● **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?** Results are very consistent between tissues. The distribution of observations is quite similar between sigmoid colon and stomach. Most observations have a low-medium level of H3K4me3 and a low-medium level of expression. There are, however, a number of genes with medium-high expression, particularly in sigmoid colon. The correlation coefficients were 0.51 (Pearson) and 0.74 (Spearman) for sigmoind

colon, and 0.47 (Pearson) and 0.67 (Spearman) for stomach. The correlation coefficient values showed a moderate correlation between H3K4me3 and expression. The plots showed that at methylation levels below 10-15, expression levels were mostly below 2.5. While at methylation levels between 25-60 expression levels ranged mostly between 0 and 3-3.5.

○ **Would you expect this degree of correlation? Formulate an interpretation of the results.** I expected a higher degree of correlation, such as low levels of methylation associated with low levels of expression, moderate levels of methylation associated with moderate levels of expression, and high levels of correlation associated with high levels of expression. However, while low levels of methylation are associated with low levels pf expression, medium-high levels of methylation are associated with low-moderate-high levels of expression, mainly low-moderate. This correlation that is not the same for all genes could mean that H2K4 trimethylation affects particularly a subset of the genes expressed in sigmoid colon and stomach.

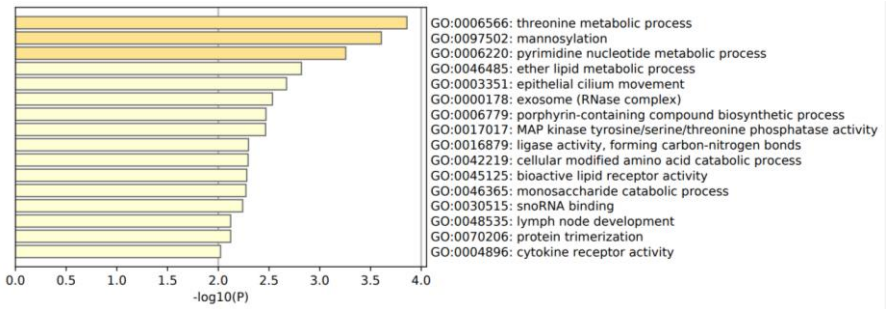
11. During the hands-on session, we have checked the level of expression of genes with tissue-specific H3K4me3 marking.

● **Are these results consistent with the degree of correlation we have observed in point 8?** I think so. The correlation between methylation of H3K4 and gene expression was moderate showing not only genes with low methylation and low expression, and genes with high methylation and high expression, but also, genes with low methylation and high expression and genes with high methylation and low expression. In this results of specific H3K4me3 marking, we see that the median expression of not marked genes (close to 0) is lower that the median expression of marked genes (close to 1). However, in both cases there are highly expressed genes, particularly for marked genes. In addition, the tissue specific marked genes are more expressed in the specific tissue than in the other. Although the difference is lower than when comparing expression between marked and non-marked genes.

● **Do you observe any unexpected behavior?** I did not expect to find non-marked genes so highly expressed genes. However, as there are other marks, it makes sense that this mark only is not explaining the whole expression pattern. I would also expect a greater difference in the expression of tissue specific marked genes.

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

I did only obtain GO terms for the stomach gene set (figure below). Due to the GO terms obtained it appears that the stomach H3K4me3 marked genes are not related to specific stomach functions but to general cell functions such as exosomes, MAPK signaling, nucleotide and amino acid processes, etc.



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 sharing between different peaks is 9120 in sigmoid colon and 6332 in stomach. On the other side, the sharing in both tissues of H3K4me3 peaks is 14889 and of POLR2A peaks is 5872. So there is more sharing between peaks of the same origin in different tissues if we consider H3K4me3, but less if we consider POLR2A.