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

<https://github.com/Gerardts9/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. Make another GitHub repository with the answer to the following points:

Git with Dockerfile: <https://github.com/Gerardts9/Dockerfile-Hands-On>

DockerHub: <https://hub.docker.com/r/gerardts9/handson>

Git with answers: <https://github.com/Gerardts9/Epigenomics_Uvic>

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

In order to locate a protein of interest in the genome, such as a transcription factor or a modified histone, the protocol starts by covalently bind all these proteins to DNA using formaldehyde normally. Then, the DNA is fragmented into pieces measuring between 100 and 300bp and the sample is purified using a specific antibody against the protein of interest. After immunoprecipitation, the junction between the proteins and the DNA is broken and the DNA is prepared for sequencing. In Chip-seq, sequencing is done only from the 5 'end, which is called single-end.

The sequencing results are obtained in a FASTQ format file that contains the sequences of the cluster readings accompanied by a parameter that measures their quality.

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

There are two main reasons to use a control. First one is that DNA fragmentation may not be uniform throughout the genome, for example, the more compacted regions (heterochromatin), will be less fragmented than those that are more open (euchromatin), which can lead to false positives. The second is that there may be biases in the sequencing process that augment the non-uniformity of the sample. The most widely used method is the input method where the sample is treated like the test sample, but without immunoprecipitation. This control is used in the subsequent peak calling phase.

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.

The stage is known as peak calling, in it, the results are compared throughout the genome between the ChIP and control reads obtaining a relative result such as fold-changes (FC). This reduces the number of errors in the results and those truly significant regions are obtained.

This tool gives two different types of files: the BigBed, that includes, among other data, the chromosome, the start and end position of the peak and the fold-change (FC) and the BigWig where a signal quantity is assigned to each position in the genome. This signal can be measured in three different ways, with the FC, p-value or pile-up signal. These two files in binary format can be unzipped to BED files and wiggle files allowing them to be read by humans.

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 method of processing serial data, where the result of one element is used as input to the next element in the pipeline.

The first stage in this pipeline consists in the mapping of FASTQ files obtained in the sequencing. It compares the reads in the file with those of a reference genome and obtains alignments and unfiltered alignments.

In a histone experiment MACS2 algorithm is used for peak calling. We obtain the values of the fold-change with respect to the control, the p-value of the signal and the peaks. The method of obtaining peaks varies depending on whether it is an experiment with replicas or not and is obtained in a BED format file.

In a transcription factors (TF) experiment, MACS2 is used to generate the BigWig file and, to carry out the peak calling, SPP and Phantompeakqualtools. There are also certain differences in the process depending on whether there are replicas or not.

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.

Genome assembly: GRCh38.

Bio sample term name: sigmoid colon and stomach.

● How many experiments are there? 4 in total, 2 for stomach and 2 for sigmoid colon.

● Paste here the link to download the corresponding metadata file: [Link 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&assembly=GRCh38&biosample_ontology.term_name=stomach&biosample_ontology.term_name=sigmoid+colon).

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.

1. For sigmoid colon: ID: ENCFF997HHO.

grep -F ATAC metadata.tsv | grep -F sigmoid | grep -F fold | awk 'BEGIN{FS=OFS="\t"; print "file\_id\tfile\_format\toutput\_type\texperiment\_type\tsample"}$2=="bigWig"{print $1, $2, $3, $4,$5, $7}'

2. For stomach: ID: ENCFF415RKU.

grep -F ATAC metadata.tsv | grep -F stomach | grep -F fold | awk 'BEGIN{FS=OFS="\t"; print "file\_id\tfile\_format\toutput\_type\texperiment\_type\tsample"}$2=="bigWig"{print $1, $2, $3,$4, $5,$7}'

9. What is an aggregation plot?

An aggregation plot is a continuous representation of a ChIP-seq signal, either pile-up, FC or p-value, over a set of genomic coordinates.

● Which tool do we use to generate one?

First, we use function aggregate of bwtool to compute the aggregate ChIP-seq signal over a genome region and then we use the R script aggregation.plot.R to create the plot.

● Which input data do we need?

A bigWig file, for instance fold-change, a gencode annotation file that provides, for example, a list of protein coding genes or the coordinates of all genes in the genome and an expression matrix.

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

○ Are the plots consistent between the two tissues?

Yes, plots show similar levels of expression. They present a peak before position 0 of the TSS and a higher one afterwards, in both cases with signals of similar value.

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

As H3K4me3 is highly enriched at active promoters near transcription start sites (TSS) and positively correlated with transcription, yes, you expect a higher methylation in high expressed genes than in least expressed genes.

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

On the one hand, it can be useful to facilitate the creation of the aggregation plot, since it would be more complicated to have to represent the whole gene instead of only a part, in our case the promoter.

On the other hand, although it is not always reliable, the position of the modification with respect to the gene, may give some clue about the function it is performing, in our case, it is known that modifications in the promoter regions of genes usually affect its expression.

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

A scatterplot with expression levels at the X axis and H3K4me3 methylation levels on the Y one.

● 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?

Yes, the scatterplots are similar between the two tissues.

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

We probably would have expected a greater correlation between the presence of methylation and the amount of gene expression. This is probably not observed due to the difference in the scale of values that the two factors can take. Methylation can occur a limited number of times due to the limited presence of histones in the region (4, two for each copy of the gene), whereas gene expression can be raised to much higher parameters. On the other hand, it is not possible to always associate the same effect for the same mark in the genome.

Lastly, it must be considered that in the case of the aggregation plot we were comparing the most expressed and the least expressed genes, in the correlation we are using the entire set of genes, so the result is not as clear as before.

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?

Yes, the expression levels of the genes are quite similar between tissues with lower levels of expression for sigmoid colon specific marking genes. Furthermore, not marked genes in any tissue show the lowest expression levels.

● Do you observe any unexpected behaviour?

A surprising fact when looking at the number of genes that belong to each group, is the large number of labelled genes shared by both tissues, about 15,000 genes. Whereas the number of genes that are not marked is around 4000 and the genes marked exclusively in each tissue are few: 600 for stomach and 100 for colon.

This is explained by the fact that the two tissues that we are using come from the same system, the digestive system, which causes the expression, in both tissues, of the same genes.

The levels of expression of the colon-specific genes are also surprising, since many genes appeared on the scatter plot with higher levels for this tissue. This could be explained by the large number of genes that are shared between the two tissues with high levels of expression.

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

The clearest results obtained when performing the enrichment come from the stomach genes since there were quite a few more. Many genes appear related to the immune response such as activation of lymphocytes or the adaptive immune response, these results could be due to the fact that the stomach tissue is in contact with the outside and a strong immune response is necessary in it to avoid infections. Furthermore, this fact could also be associated with the presence in the stomach of the *H. pylori* bacteria, the control of which also requires a strong immune response.

Gene families related to digestion also appear, which makes a lot of sense since we are working with samples of stomach tissue. There are other families also present in other tissues such as receptors associated with G proteins (Rhodopsin like receptors) or cell adhesion molecules (CAM).

The results with the colon samples are very little informative since we had a very limited number of genes. Only 5 families of genes appear and with p-values lower than those obtained with stomach samples. These enriched families are cell adhesion, myelination, anterior / posterior axis specification genes, polysaccharide synthesis and adult locomotory behaviour. As we can see, these are very general results, for example, myelination genes can be related to regions of the peripheral nervous system present throughout the body, in the same way that cell adhesion genes are.

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 highest number of peaks are shared by those of the H3K4me3 type between the two tissues, whereas, among those of the POLR2 type, they are the group that shares the fewest peaks. On the other hand, if we compare between the same tissue type, the sigmoidal colon tissue samples occupy the second place and the stomach samples remain as the third ones, globally. The point that presents the greatest number of peaks is the union between the 4 different situations.