Students should write a short report which includes the performed analysis and the contributions of each student. Students have to provide the code. The report should contain a part written in groups and a part written by each student individually.
The maximum is 10 points: 3 points for code, 4 points for description, 3 points for the equal students input.

DNAme in iPSC-derived and Post-mortem neurons

Students:
Aleksandra Gorbunova
Veronika Karpushenkova
Bamiwo Adelere

TA:
Anna Kononkova

Background of the study

iPSC-derived (human induced pluripotent stem cell) neurons differ from post-mortem ones in several features such as transcription, epigenetic and chromatin. DNA methylation is a characteristic of neuronal cells. It is stated in the literature [https://doi.org/10.1093/hmg/ddv637] that iPSC-derived cortical neurons potentially present a powerful new model to understand corticogenesis and neurological disease.
The question arises: does DNA methylation in CpG and non-CpG context differentiate iPSC-derived neurons from post-mortem ones?

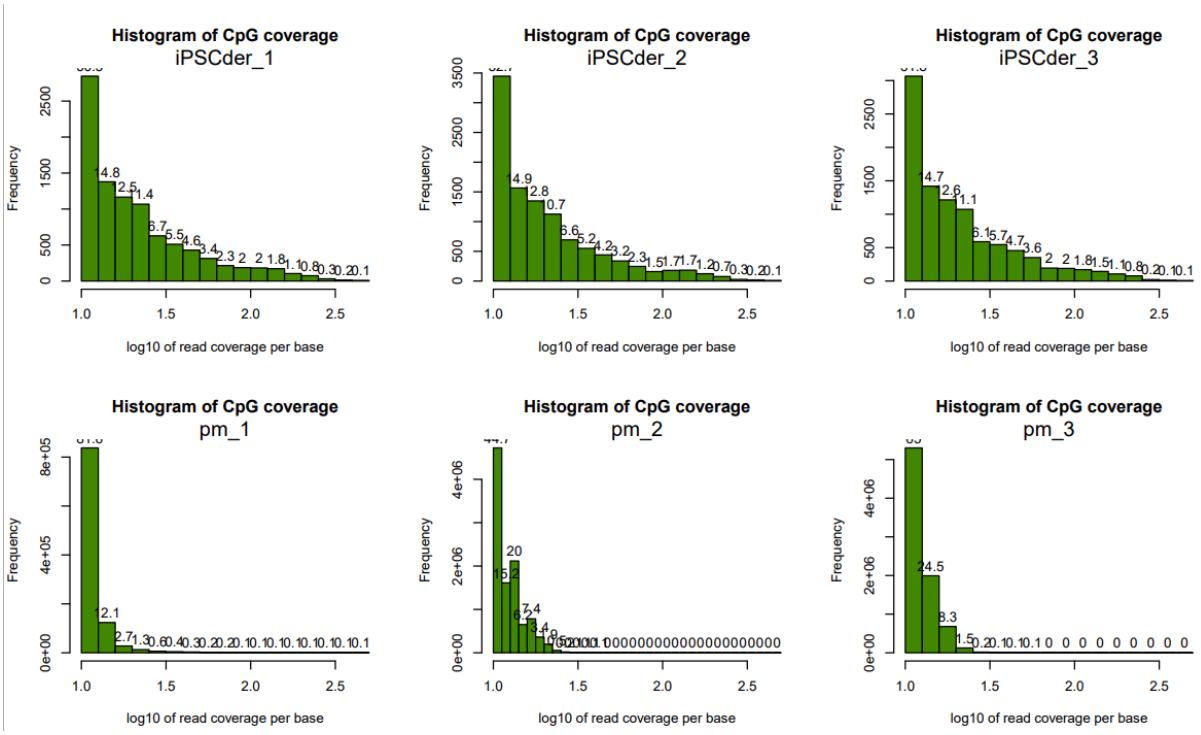
Goal of the study

The goal is to compare DNAme profiles in iPSC-derived and post-mortem neurons, and establish interconnection and differences in transcription, H3K9me3 modification and chromatin compartments.

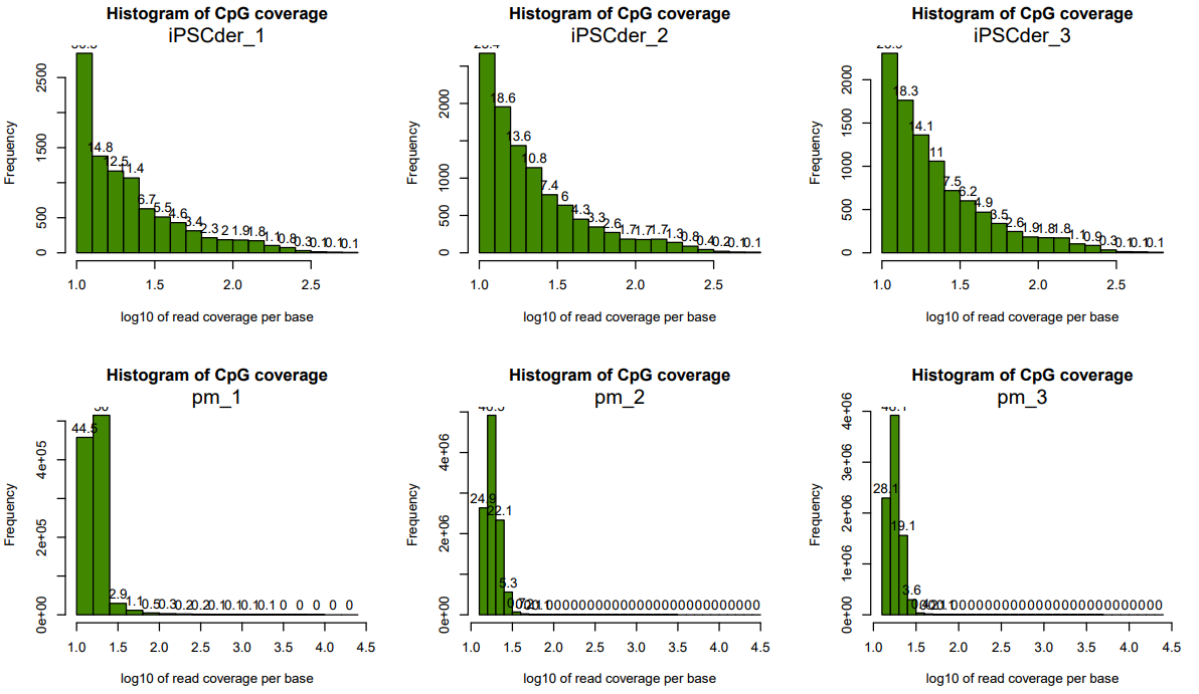
Methylation part
This part was done by Aleksandra

code source: https://github.com/GorAleks/omics_fin_proj.git

After mapping the reads, we obtained quite divergent results. For iPSC samples (induced neurons), the mapping efficiency was on average around 10%, whereas for postmortem samples, it was about 70%. Therefore, it was decided to analyze the average coverage for each bin to determine whether we can perform methylation analysis, where the minimum required coverage for the algorithm to work is 10.

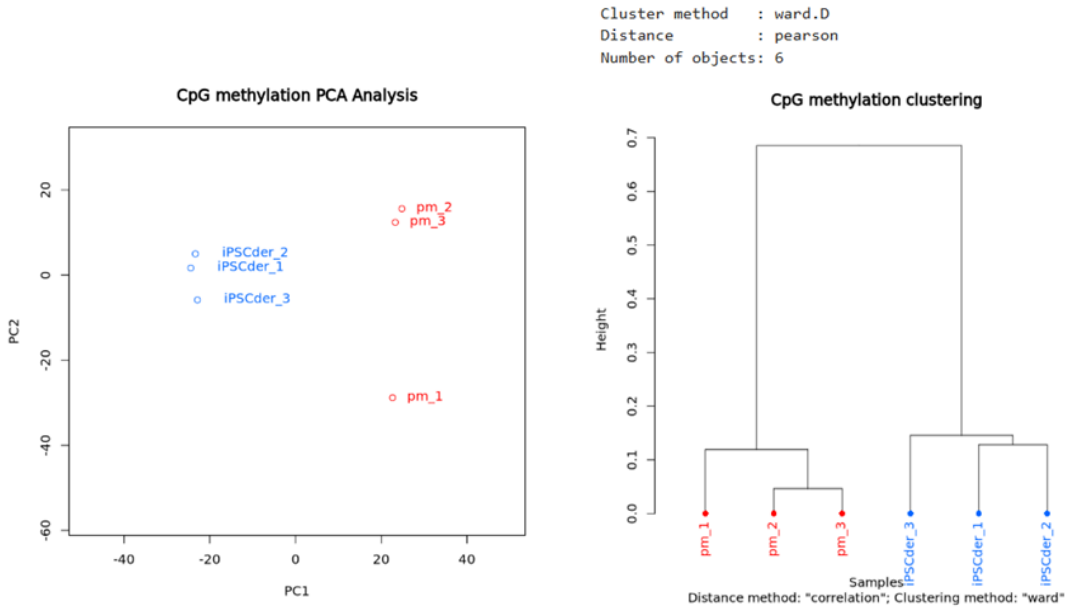


On the coverage distribution plot for CpG, we can see that for the post-mortem (pm) samples, the situation is quite poor - coverage is mostly near the minimum, somewhere just above 10 reads per region (the same situation was for non-CpG sites).

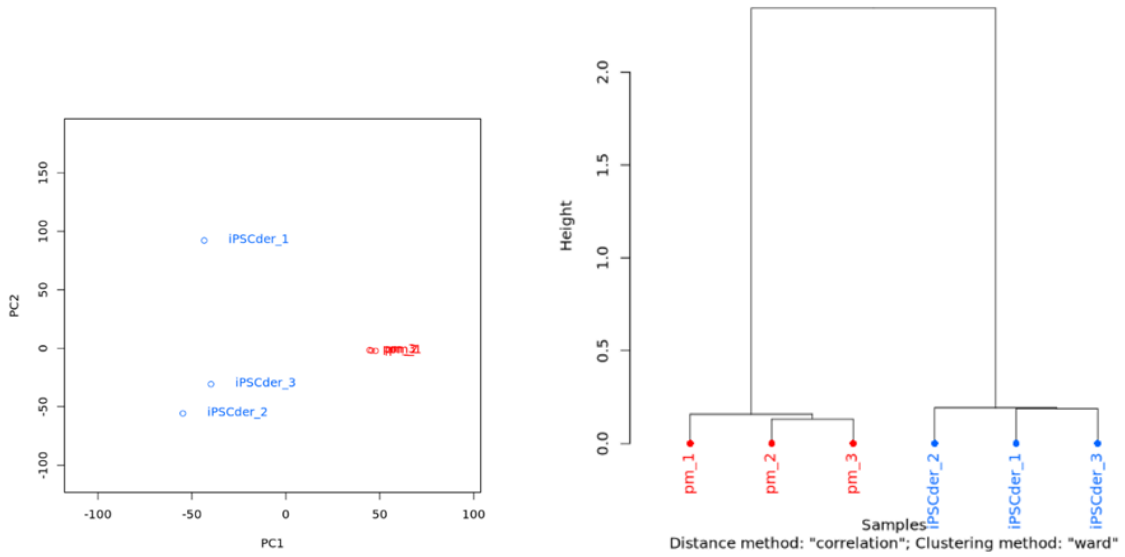


For CpG sites, I performed normalization (graphs for these results are above); overall, this did not significantly change the situation, but about 10 sites were filtered out in the differential methylation results because of this. Therefore, I decided to use the unnormalized data for both CpG and non-CpG going forward.

The clustering for both CpG and non-CpG sites looks relatively good in the samples; based on the first principal component, they can be clearly separated without overlaps.

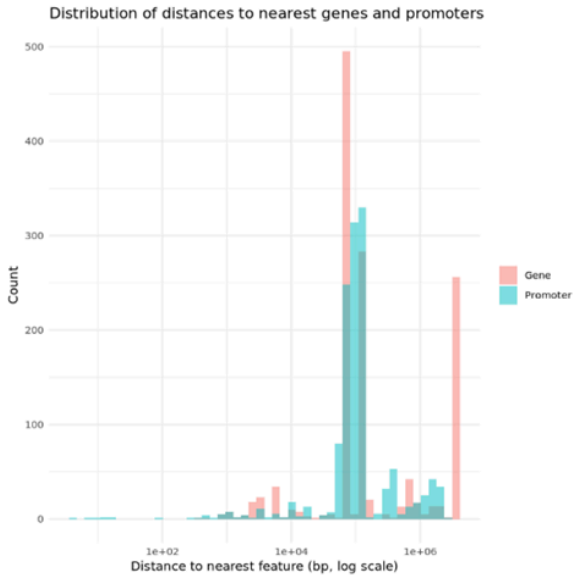


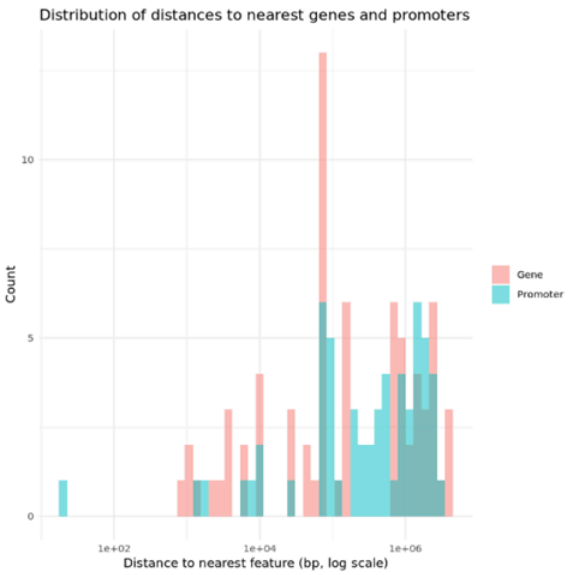
Results of clusterization for CpGs. (above)



Results of clusterization for non-CpGs.(above)

Annotation was performed using the transcriptome to identify not only genes but also promoters, since using only genes give us very few results.
As a result, this is the distribution of distances obtained:

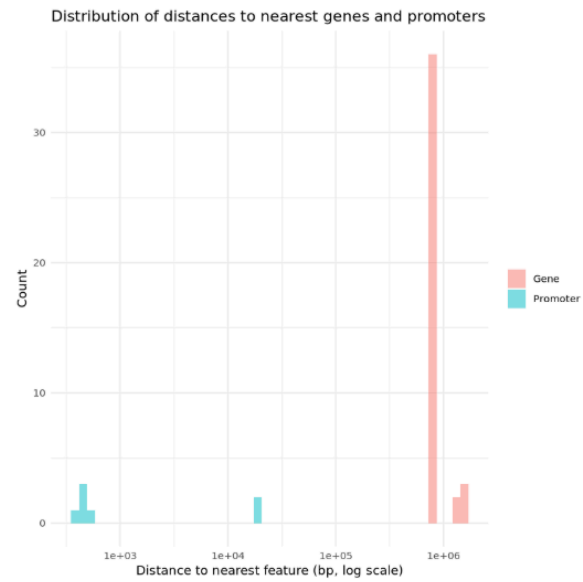


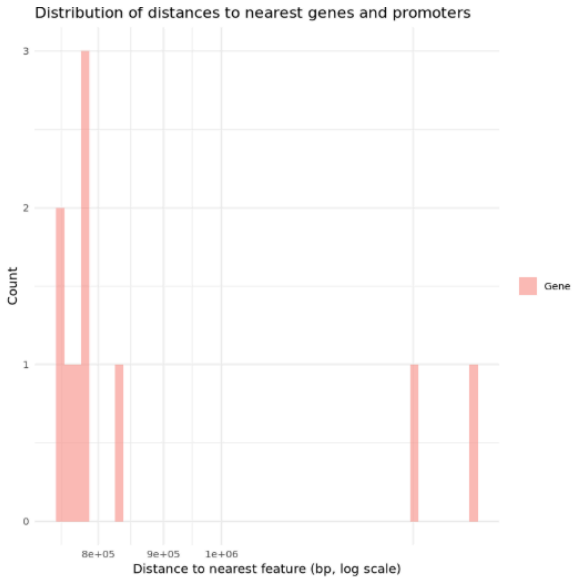


On the left is the plot for all differentially methylated sites for CpG, and on the right for the significant ones.
In both cases, we observe main peaks at distances around 10^5–10^6 base pairs, which seems quite large and logically should mostly influence chromatin structure in that region or affect gene regulation through loops and other types of tertiary structure.

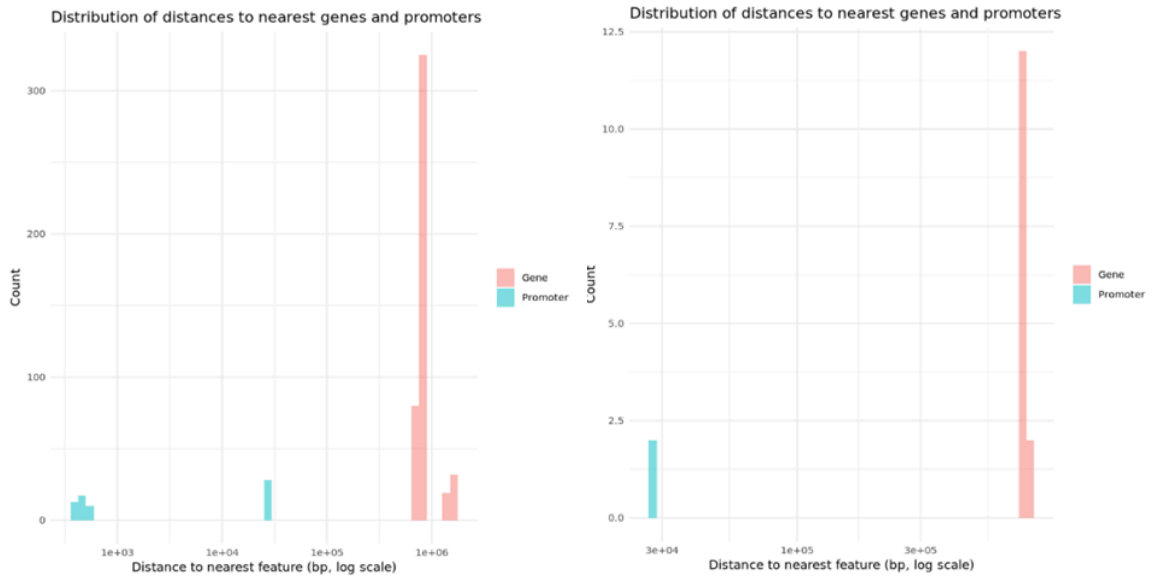
But due to some problems with memory on the server, we couldn't analyse results for non-CpG regions for all chromosomes, so further investigation and association of different types of data will be done with just some chromosomes to estimate, if there are any patterns.

For the 5th chromosome CpG distribution is like that:





On the left is the plot for all differentially methylated sites, and on the right for the significant ones. As you can see, patterns for this chromosome differ from the whole genome pattern and also we have a lack of data.



But we can see a correlation of pattern for all methylation sites between CpG and non-CpG (graphs are above, all meth sites on the left and significant one on the right).

We also tried to conduct GO enrichment analysis, but it could not be performed successfully; no statistically significant results were found. So, we can say that there are no significant differences in some pathways for post mortem and induced neurons.

ChIP-seq part

This part was done by Veronika. Full analysis with code can be found in pdf file from the repository: https://github.com/GorAleks/omics_fin_proj/blob/main/ChIPseq.pdf.

Materials and Methods

The raw data and its description is provided below in a table

H3K9me3 and H3K27me3 ChIP-seq in .broadPeaks format for iPSC and post-mortem samples. Signal and input files for merged replicates in .bw format.	iPSC-derived: GSE196109 (2 replicates) Post-mortem: GSE211871 (3 replicates)

H3K9me3 is tightly linked with DNAme, therefore, we will use the following tools to analyze this correlation:
- deeptools for merging and visualization of enrichment H3K9me3 and DNAme in TADs
- computeMatrix and plotHeatmap to plot signal (.bw) around genes

In detail, we compared all the data between each other with plotCorrelation (https://deeptools.readthedocs.io/en/develop/content/tools/plotCorrelation.html) firstly. Then, normalization of signal to input with bigwigCompare was done (https://deeptools.readthedocs.io/en/develop/content/tools/bigwigCompare.html). Followed by merging replicates within groups using bigwigAverage (https://deeptools.readthedocs.io/en/develop/content/tools/bigwigAverage.html): post-mortem with post-mortem and iPSC-derived with iPSC-derived.
The control samples were used for normalization, not the main comparison. Having normalized data, we will compared two cell types using computeMatrix (https://deeptools.readthedocs.io/en/develop/content/tools/computeMatrix.html) and plotHeatmap (https://deeptools.readthedocs.io/en/develop/content/tools/plotHeatmap.html).

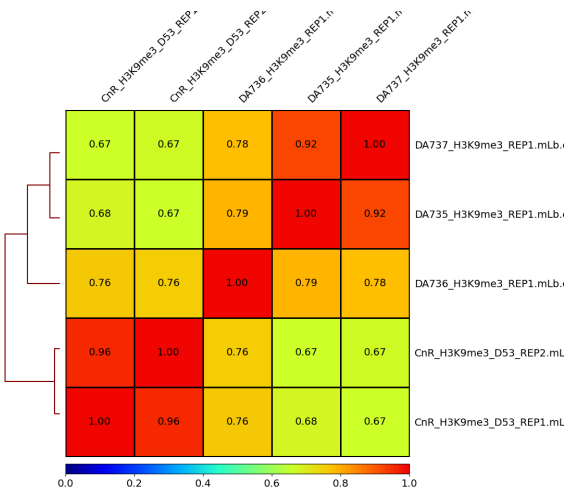
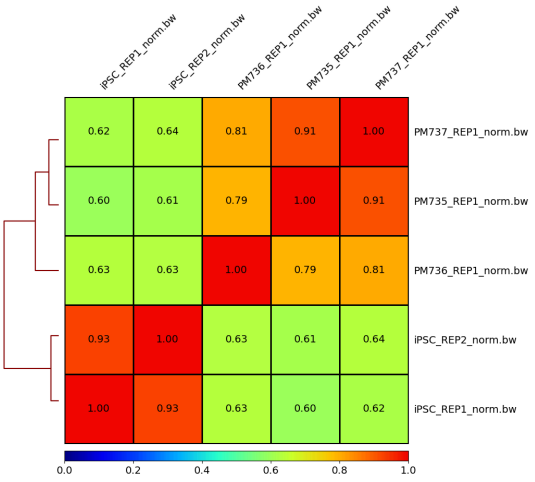
For bw visualization in IGV following .fa and .fa.fai files were used: Homo_sapiens.GRCh38.dna.toplevel.fa.gz.fai (https://ftp.ensembl.org/pub/release-114/fasta/homo_sapiens/dna_index/Homo_sapiens.GRCh38.dna.toplevel.fa.gz.fai), and Homo_sapiens.GRCh38.dna.toplevel.fa.gz (https://ftp.ensembl.org/pub/release-114/fasta/homo_sapiens/dna_index/Homo_sapiens.GRCh38.dna.toplevel.fa.gz).

Abbreviations

DAXXX - are post-mortem samples.
Every sample having IgG in name is an input.
Files started from CnR are from iPSC-derived data.

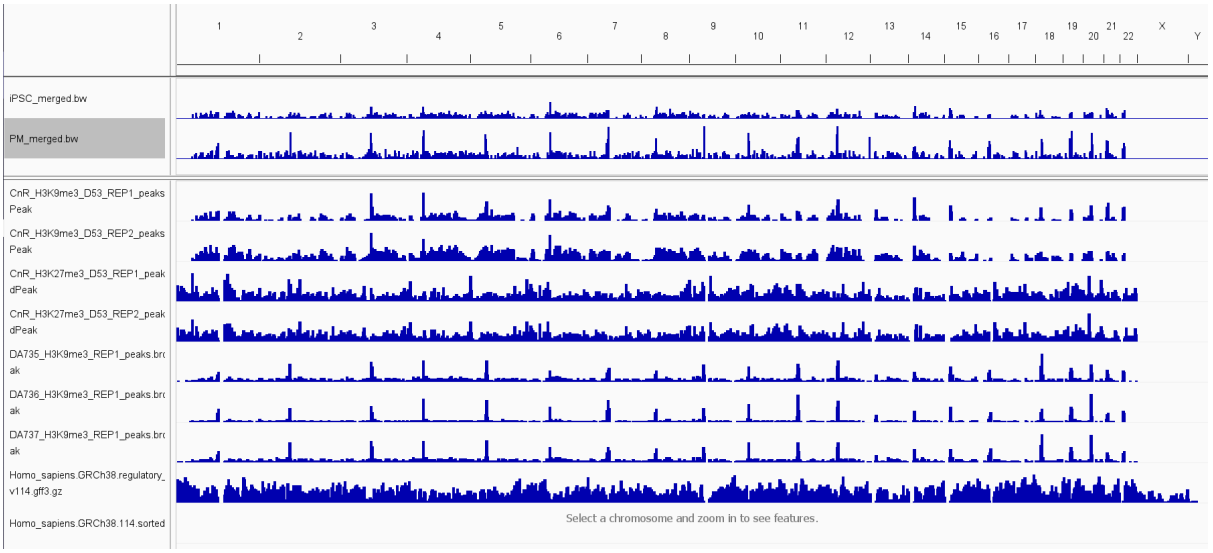
Results

- Correlation of raw data vs normalized data

Before normalization	After normalization
	
We can see that after normalization, questionable PM replicate (DA736) now correlates better with other PM samples. And all coefficients in general between replicates and samples increased.	

- Visualization in IGV

On two merged bigwig files the same range was set: from 0 to 0.20.



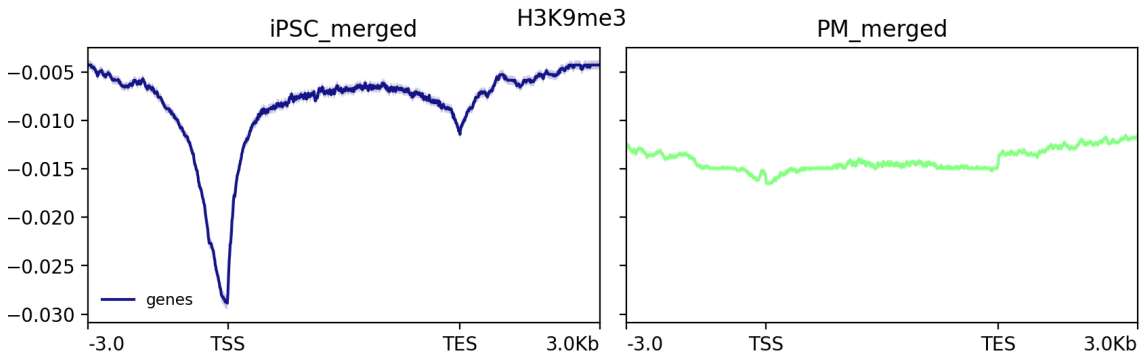
First two tracks are ChIP-seq signal intensity for H3K9me3 in iPSC and in post-mortem cells. Then various broadPeak files are placed, and finally a gff3 file with regulatory elements, and sorted with igvtools gff3 file. Latest two were downloaded from Ensembl. On the overall view we can see that enrichment between replicates is consistent which is a sign of high reproducibility, especially in case of post-mortem (DA) samples. We also can see that our H3K9me3 peaks do not overlap with H3K27me3 which is a sign of high specificity.
Let's take a closer look at chr1 region:

First of all, iPSC are less enriched with H3K9me3 compared to post-mortem cells. Some little peaks are present in PM but absent in iPSC and vice versa. Those might be cell specific modifications or some noise as enrichment is not really high for those peaks.

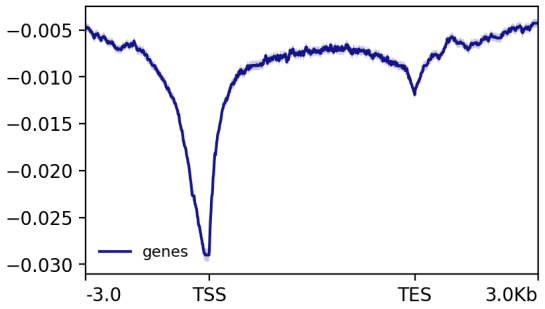
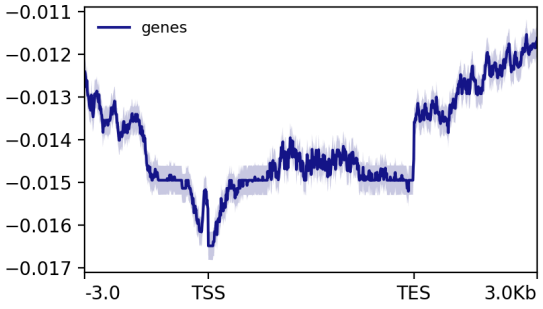
H3K27me3 modification is more present than H3K9me3. They both are absent at regions around 125-145 mb, so most probably this is an actively transcribed region. From biological

interpretation H3K9me3 corresponds to heterochromatin and gene silencing, and H3K27me3 corresponds to facultative heterochromatin.

- Plotting enrichment around genes



TSS and TES are transcription start and end sites respectively. Methylation plot shows that upstream of TSS the enrichment is low as it is usually an active regulatory region. Between TSS and TES we see an almost baseline level of enrichment. Downstream of TES the enrichment returns to baseline level as we get to non-transcribed areas. Generally, H3K9me3 is an inhibitory modification of transcription. On the same scale PM enrichment has much smaller amplitude in enrichment.
Let's plot them separately with different scales:

iPSC	Post-mortem
	

On a different scale we can more clearly see the enrichment changes for post-mortem cells. The same patterns are seen as for iPSC but enrichment for PM samples is more noisy which can be explained by higher heterogeneity of the post-mortem samples and replicates compared to cultural iPSC samples.

RNA-seq part

This part was done by Veronika. Full analysis with code can be found here: https://github.com/GorAleks/omics_fin_proj/blob/main/rnaseq.ipynb

Materials and methods

The baseline for the code was taken from hw4 description on: https://cosmoskaluga.github.io/OmicsDataAnalysis/input-dataset.html.

The RNA-seq data is used to find a correlation between the level of gene expression and methylation patterns. We use DESeq2 (https://bioconductor.org/packages/release/bioc/html/DESeq2.html) to analyze differential expression, and a basic method of normalisation is used (median-of-ratios). Since we have salmon files ready, we matched the transcript with the corresponding gene. For clusterization and further visualisation we have done regularized log normalisation to minimise the noise and stabilize dispersion.

Our RNA-seq files:
iPSC-derived: GSE212252 (3 replicates)
Post-mortem: GSE96615 (3 replicates)
Also the following .gtf file was used for analysis:
Homo_sapiens.GRCh38.113.gtf.gz
(https://ftp.ensembl.org/pub/release-114/gtf/homo_sapiens/Homo_sapiens.GRCh38.114.gtf.gz)

Abbreviations

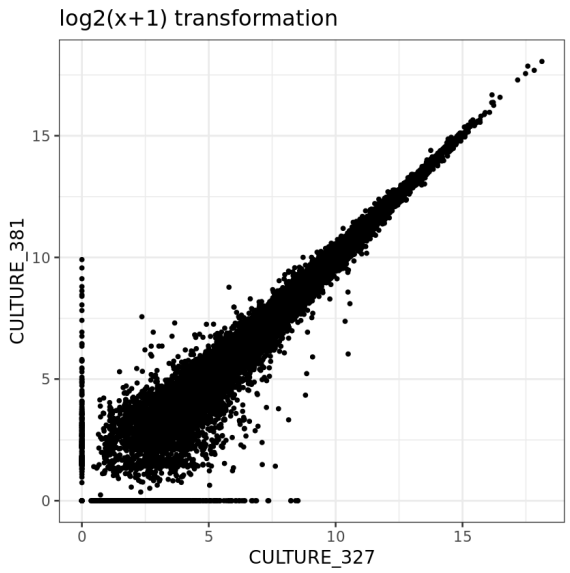
PM corresponds to post-mortem, and CULTURE corresponds to iPSC-derived

Results

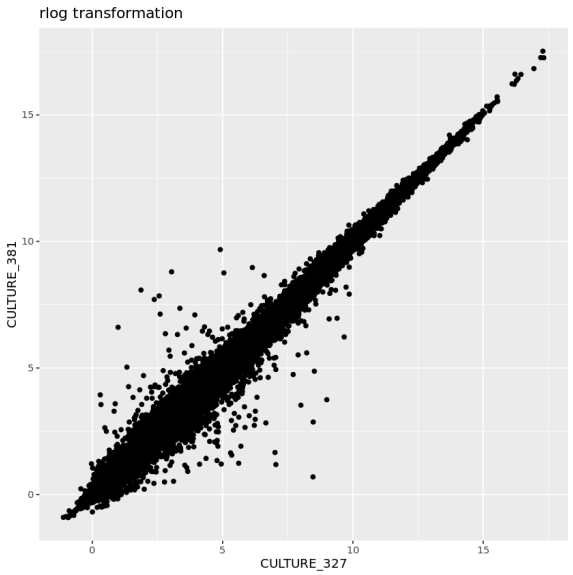
- Normalization and quality control

We performed QC checks on the count data to help us ensure that the samples/replicates look good after normalization.

Before rlog transformations	After rlog transformation



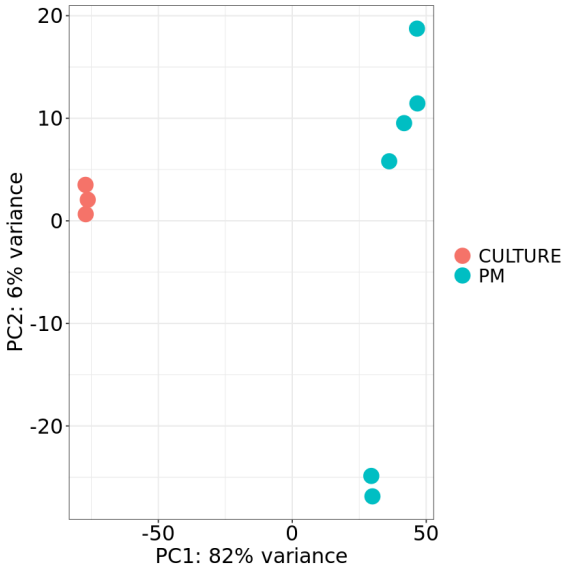
Above we can see heteroscedasticity in the data, illustrating the dependence of the variance on the mean count value.



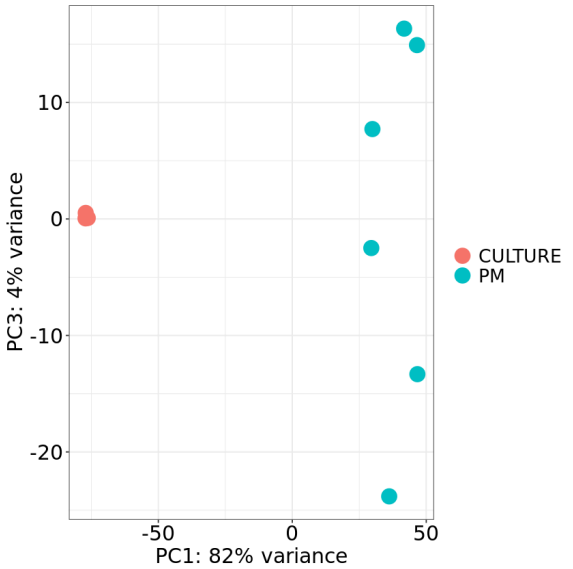
After rlog transformation we can see a more clear picture with less dispersed data.

- Principal Component Analysis (PCA)

PC1 vs PC2 plot



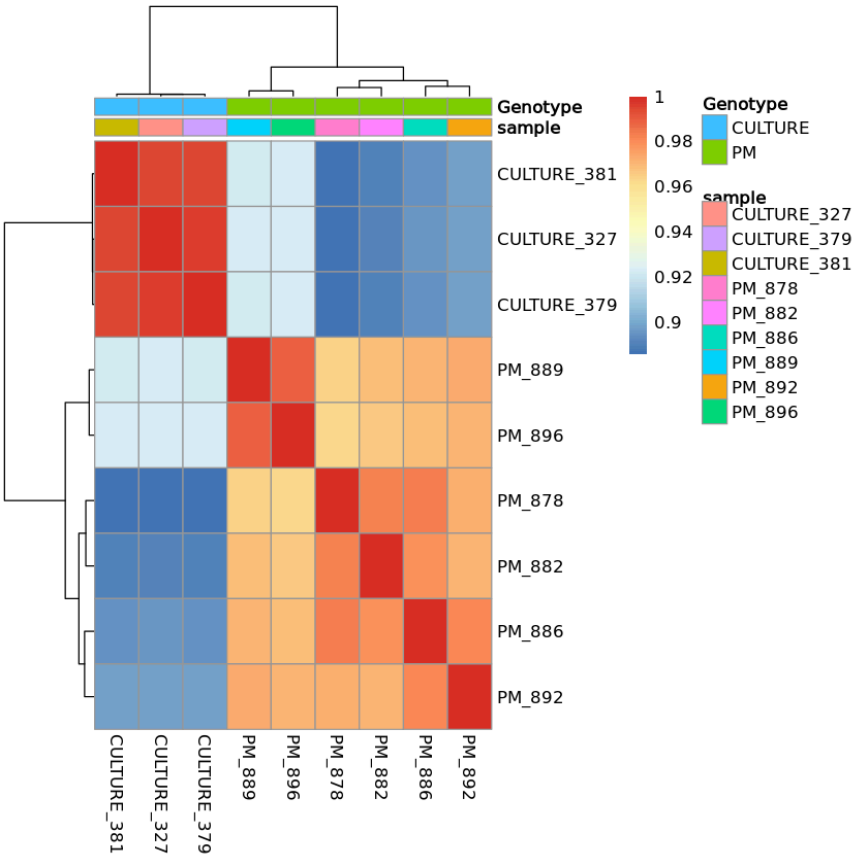
PC1 vs PC3 plot



Those above are PCA plots which might show and emphasize variations and strong patterns. We can see that clustering is quite obvious, and we can see that biological replicates cluster together on PC1 (82% of variance explained), and the samples from different treatment groups cluster separately.The percentage of variance corresponds to the amount of variation captured by the principal

component. Post-mortem samples have bigger variance on PC2 and PC3, than culture samples which is due to biological heterogeneity (time after death, tissue degradation, and individual biological variability), and technical heterogeneity (degradation, differences in sample handling, or batch effects) of post-mortem samples.

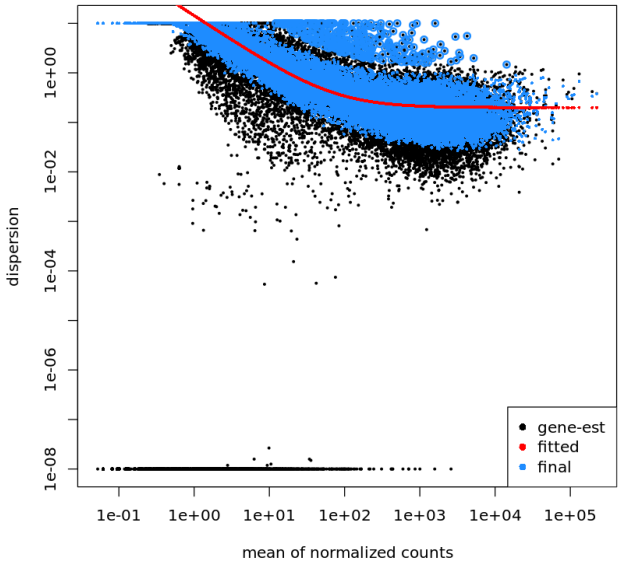
- Correlation Heatmap



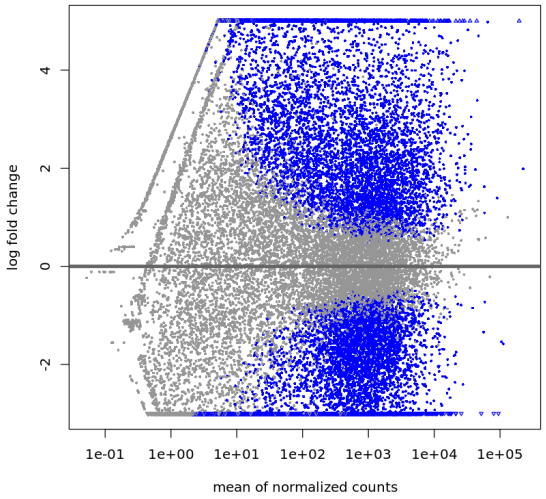
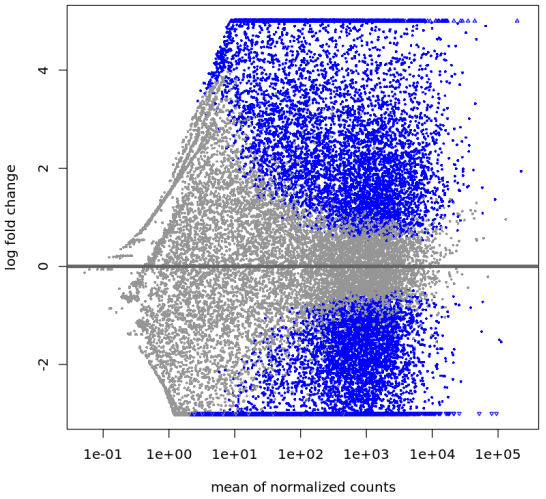
As expected, correlation between culture samples is really high. PM samples also correlate really well, and bigger correlation is seen between replicates.

- DE analysis

Dispersion estimates plot:

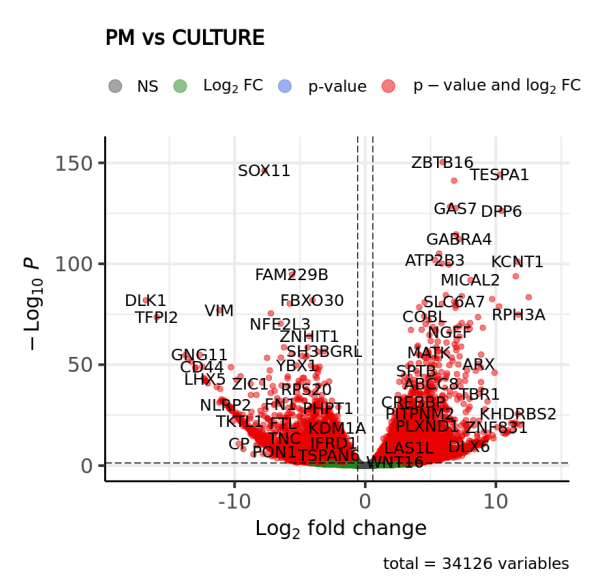
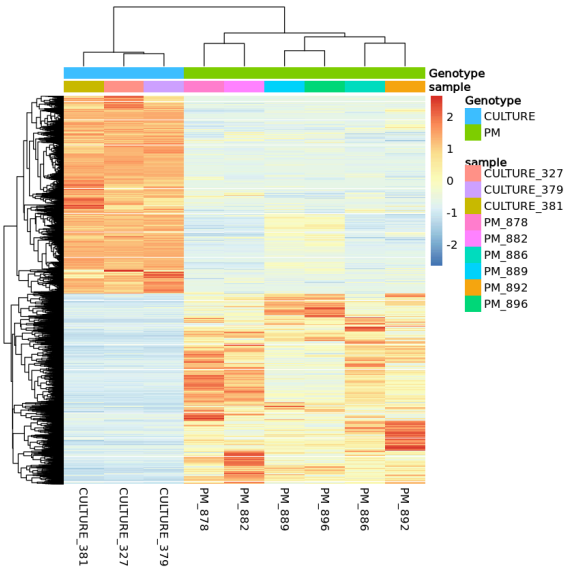


Dispersion is quite high. We will apply shrinkage to lower it.

Before shrinkage	After shrinkage
	
Here unshrunken data is a little bit more dispersed, and noisy.	

- Visualization

Volcano plot	Heatmap of significant genes

	
Each point on the volcano plot is a DE gene. On x-axis negative values correspond to genes more highly expressed in culture, and positive values correspond to genes more highly expressed in post-mortem samples. On y-axis higher value corresponds to more statistically significant result (lower p-value). Red dots correspond to genes statistically significant (by p-value) and have a large fold change. Grey are not significant results, blues are significant only by p-value, and greens are significant only by fold change. So, SOX11, FAM229B, VIM, DLK1, etc. are significant DE genes for culture samples. Meanwhile, ZBTB16M TESPA1, GAS7, DPP6, GABRA4, etc. are significant DE genes for post-mortem samples.	Here also we can see quite high and clear clusterization between culture samples. Same for post-mortem samples with a little bit less correlation, and between replicates correlation is higher than between samples.

- Comparison with cortical layer markers

Let's analyze cortical layer marker expression in iPSC-derived neurons

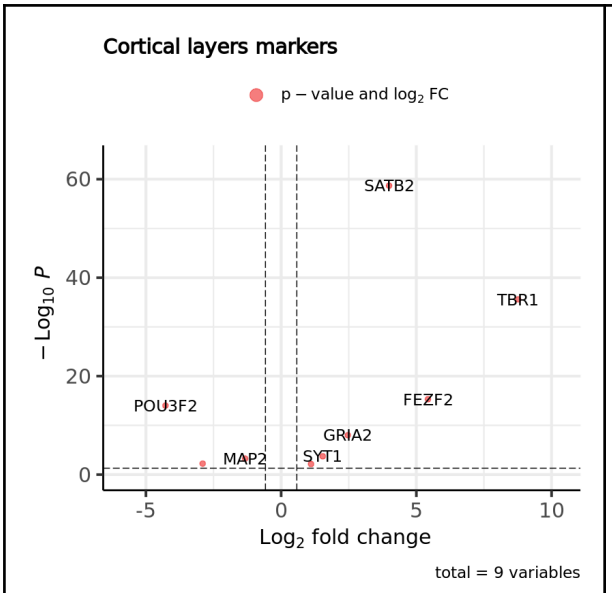
According to:
https://www.pnas.org/doi/10.1073/pnas.1216793109
https://pubmed.ncbi.nlm.nih.gov/21228164/
Cortical layer markers: Tbr1, Ctip2, and Satb2 are expressed in postmitotic neurons, Fezf2 is expressed in cycling cortical progenitors from very early stages of corticogenesis. Fezf2 blocks corticothalamic fate in layer 5 by reducing Tbr1 expression in subcerebral neurons
According to:
https://www.pnas.org/doi/10.1073/pnas.1216793109
Tbr1, EphA4, and Unc5H3 are critical downstream targets of Satb2 in callosal fate specification. This represents a unique role for Tbr1, implicated previously in specifying corticothalamic projections. We further show that Tbr1 expression is dually regulated by Satb2 and Ctip2 in layers 2–5.
According to:
https://www.bio-connect.nl/news/cortical-layers/
LHX2 and PAX6 together play a crucial role in the specification of neo-cortical progenitors which give rise to the projection neurons. MEF2C is another example of transcription factor essential for normal neural development and spatial distribution in the neocortex.
Upper layers neurons can be identified by expression of CUX1 and POU3F2 (BRN2), the neurons of layer V – by expression of BCL11B (CTIP2) and neurons of layer VI – by FOXP2 expression

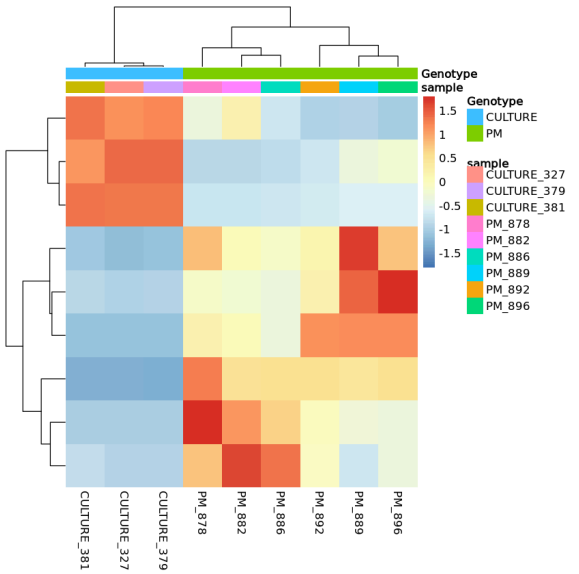
Summary:
-Deep layers (V–VI): Tbr1, Fezf2, Ctip2

-Upper layers (II–IV): Satb2, Cux1, Pou3f2 (Brn2)

-Pan-neuronal markers: Map2 (maturation), Syt1 (synaptic), Gria2 (glutamatergic)

Volcano plot of cortical layers markers	Heatmap of cortical layers markers





On volcano plots: We can see that the bigger the y-value is, the more probable that the post-mortem result is different from culture. The bigger the x-value is, the bigger the contribution of this gene to expression. Let's remind markers:
-Deep layers (V–VI): Tbr1, Fezf2, Ctip2

-Upper layers (II–IV): Satb2, Cux1, Pou3f2 (Brn2)

-Pan-neuronal markers: Map2 (maturation), Syt1 (synaptic), Gria2 (glutamatergic)

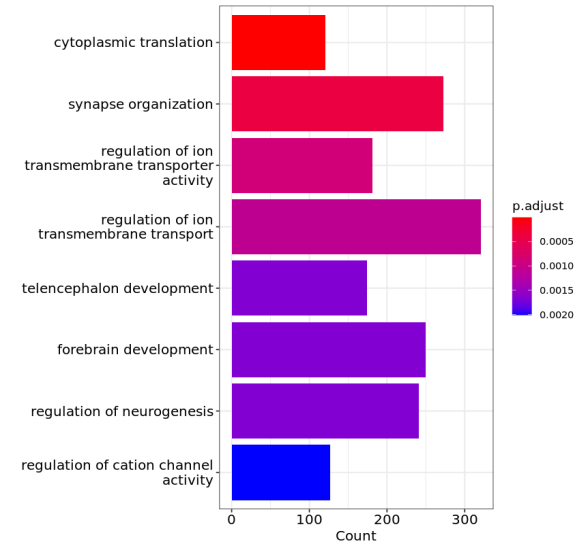
Pou3f2 and Map2 are seen as significant in culture samples. These results correspond to upper layer and maturation.

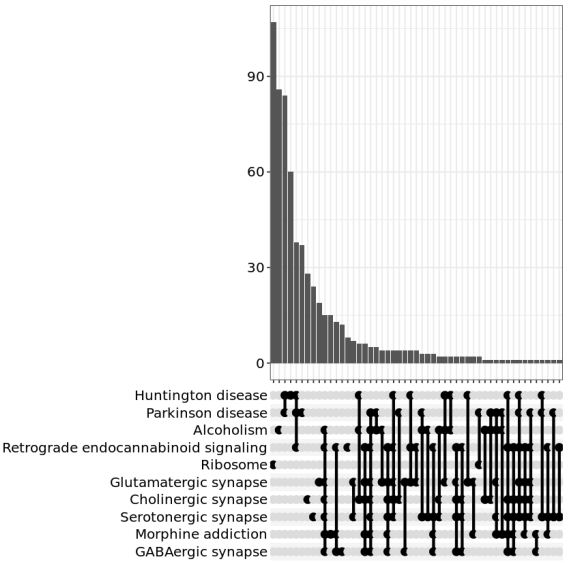
Satb2, Tbr1, Fezf2, Gria2, and Syt1 are seen as significant in post-mortem samples. These results correspond to each of the categories, and probably it is because of a higher heterogeneity in post-mortem samples.

On a heatmap we can clearly see the difference in expression of genes between post-mortem and culture groups.

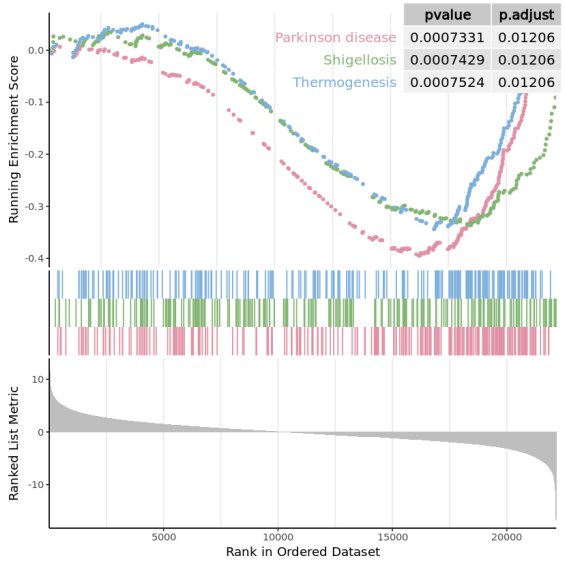
- Functional annotation

Gene Ontology	Pathway analysis

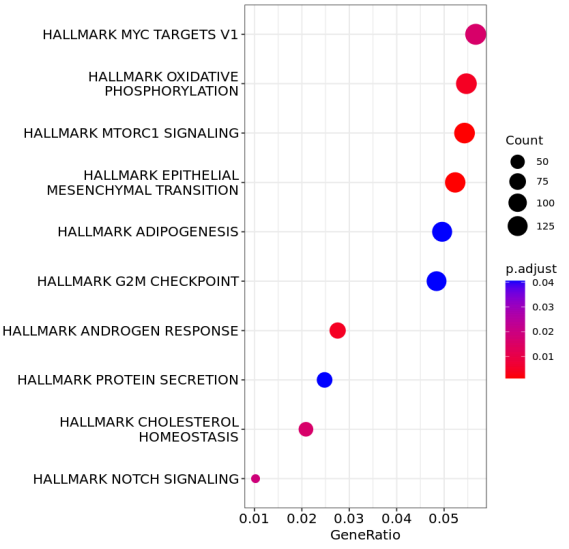




Gene-Set Enrichment Analysis (GSEA) for 3 manually selected pathways

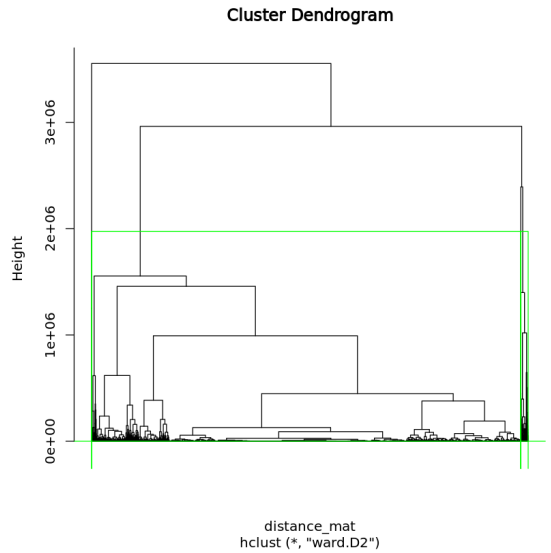


Gene Set Enrichment Analysis with mSigDB



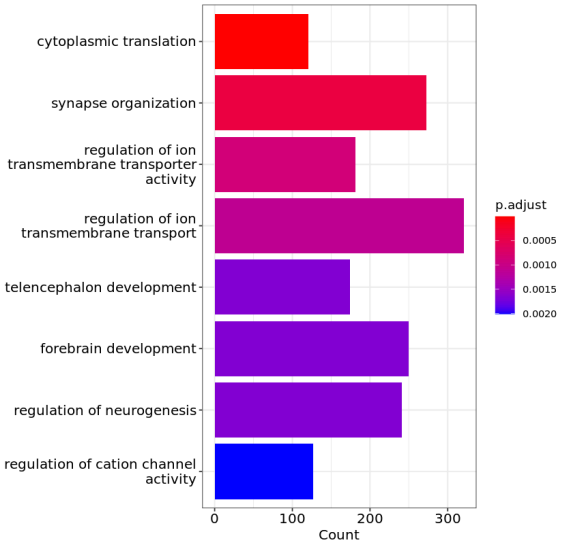
- Hierarchical clustering:

Let's perform clustering of DE genes to identify groups of genes with similar expression trends, and perform the enrichment analysis (enrichGO/enrichKEGG) on the groups of our genes.

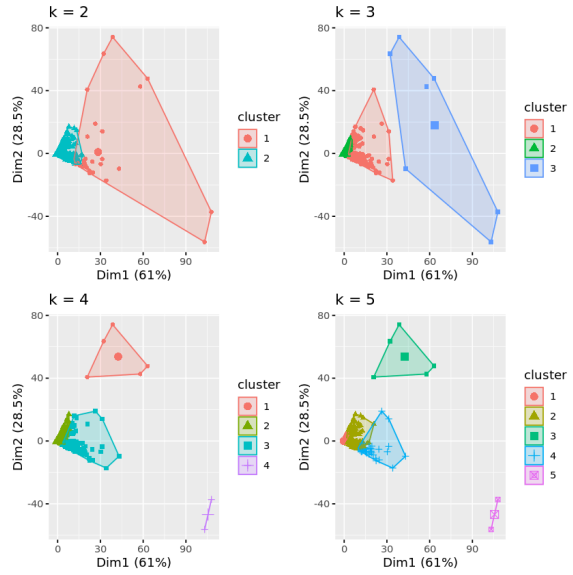


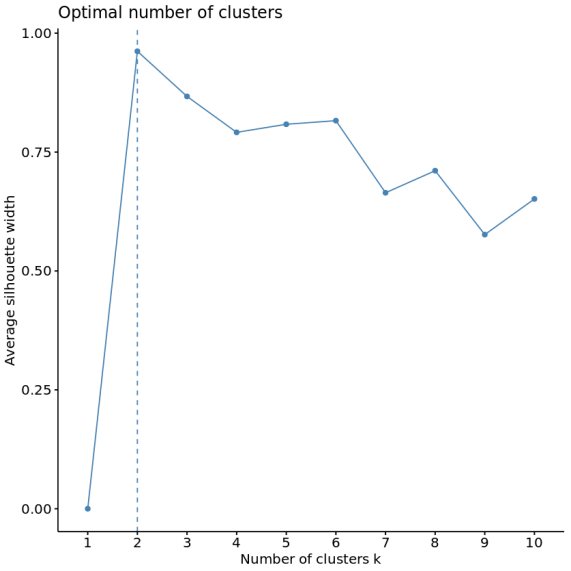
So, probably we can see 3 distinct clusters at least. Let's check with k-means

enrichGO



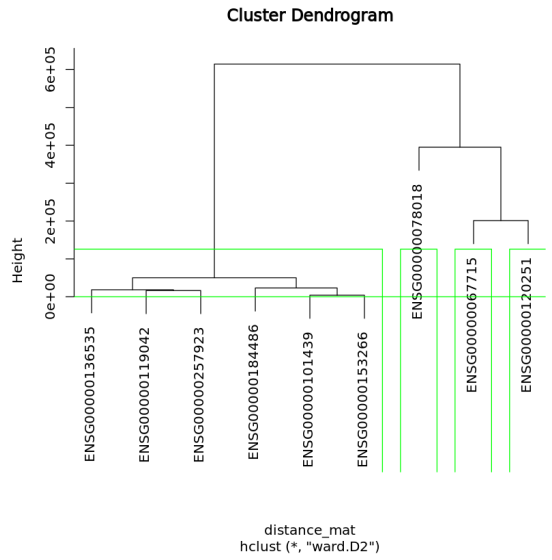
k-means clustering



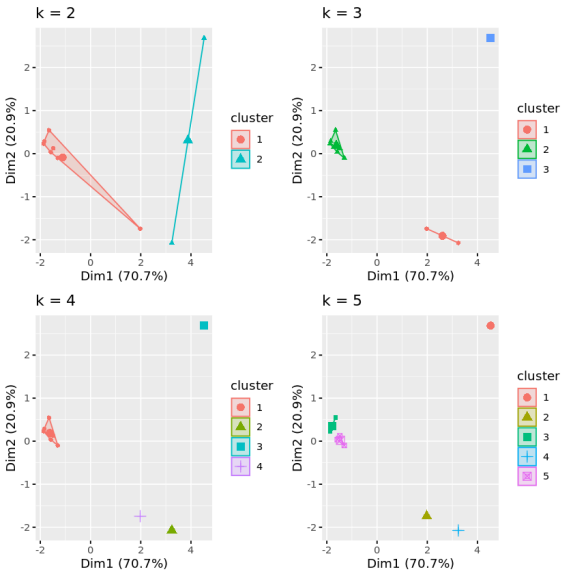


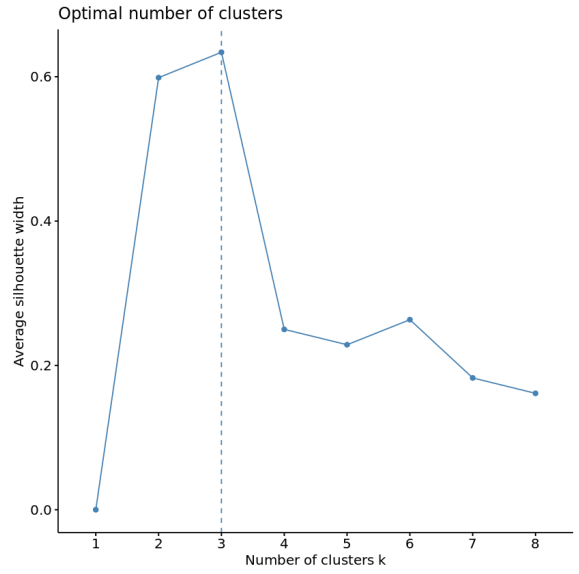
So, according to silhouette the most probable number of clusters is 2.

- same analysis for cortical markers:



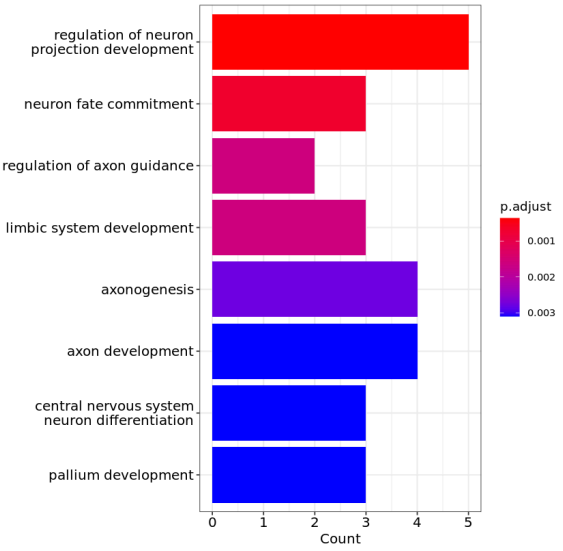
k-means clustering





Probable number of clusters is 3.

The enriched categories for cortical layer markers (sanity check):

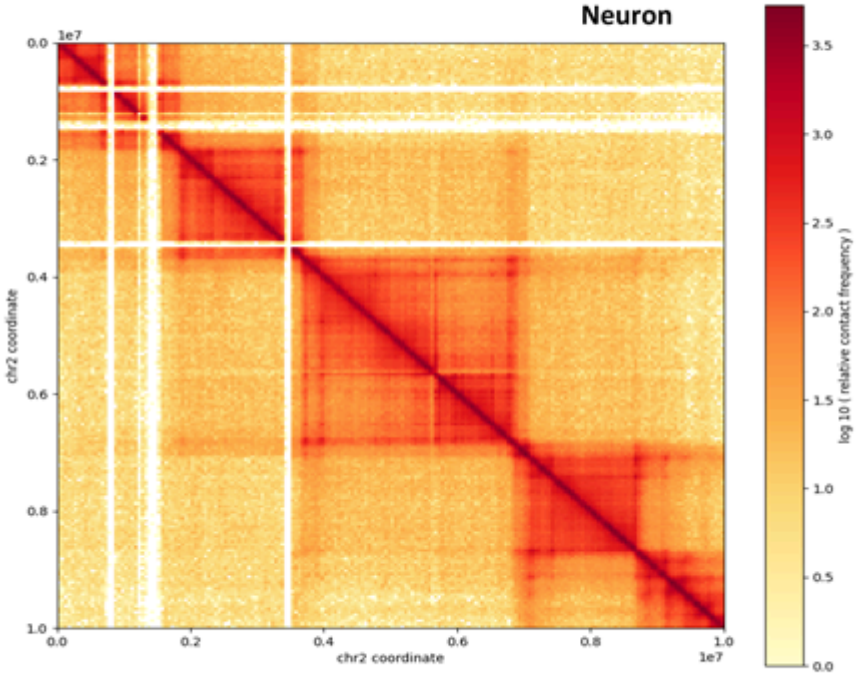


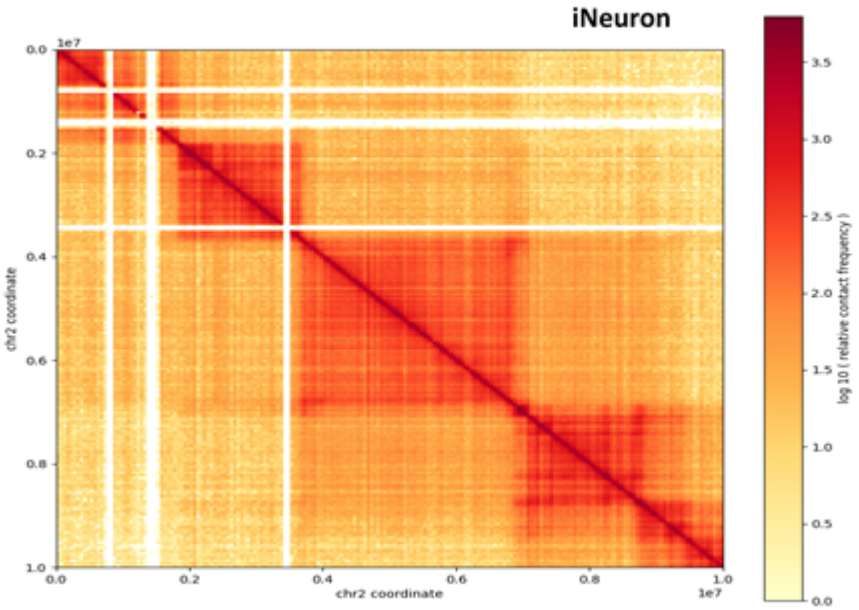
HiC Report

Assigned Files: Two cool files (Neurons and iPSc-derived neuron (iNeuron) cool files)

Task 1

Created mcool files with 10kb, 25kb, 50kb and 100kb resolutions for each cell type (Neuron and iNeuron). To view the genome HiC, I used ‘cooler show’ tool to view chr2:0-1000000 region at 50kb resolution, shown below:

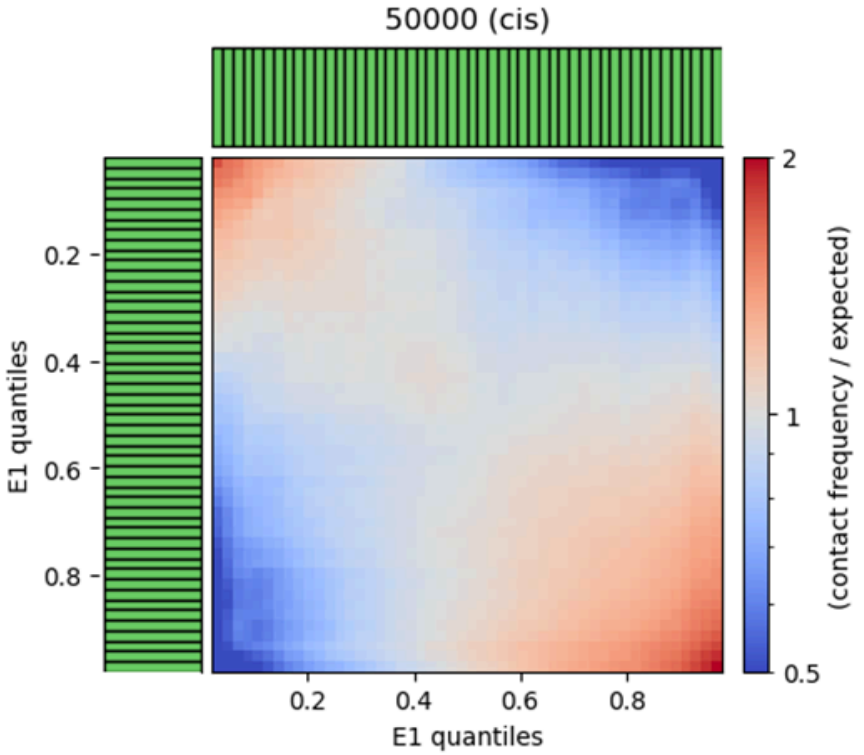
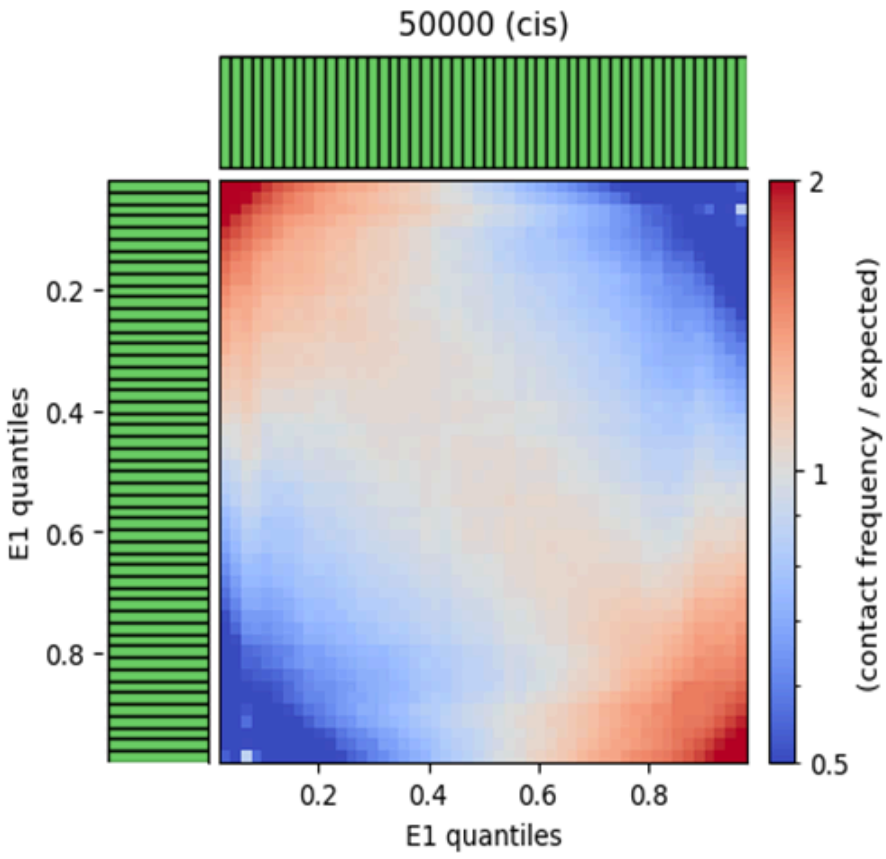




The maps above show some differences in the structural chromatin organisation in the selected region (chr2:0-10000000),for Neuron and iNeuron cell type. Also, some features (like TADs) can also be seen in the maps.

Task 2

With cooltools, I calculated the expected interaction, performed eigenvector decomposition, then compartment calling. The output compartment and expected tsv file were used to generate the saddle plot. This was repeated for all HiC resolutions.
The saddle plot for 50kb resolution is shown:

	iNeuron saddle plot
	Neuron saddle plot

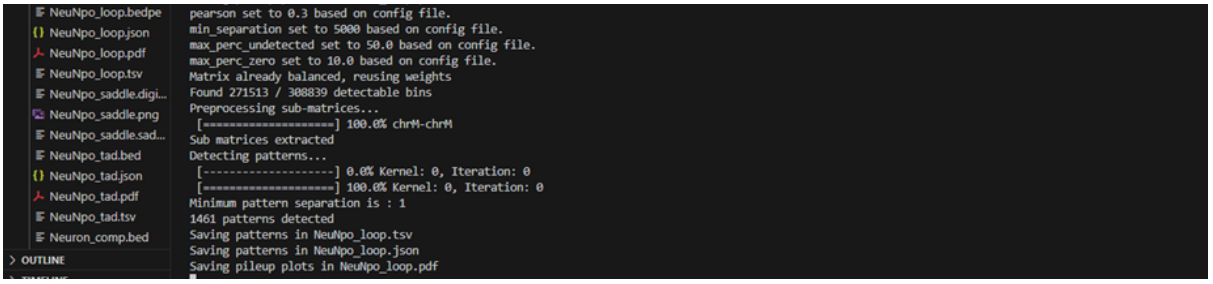
The saddle plots show more interaction within the neurons compartments than in iPSc-derived neurons. The plot for other resolutions are in the submitted file.

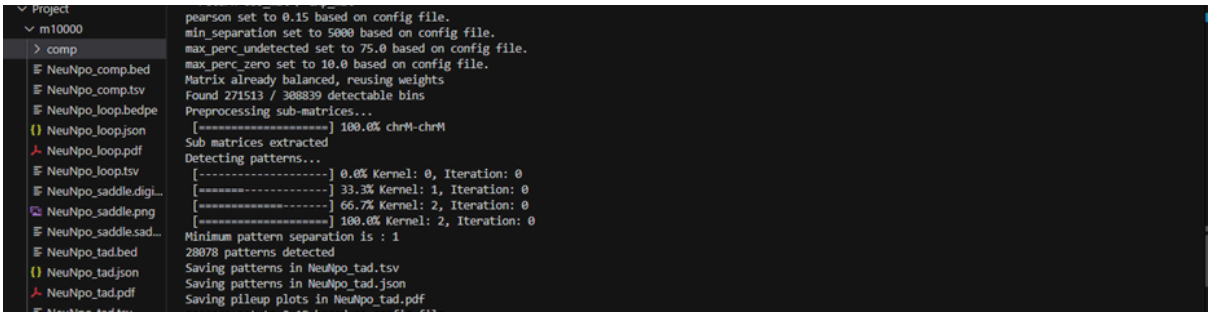
Task 3

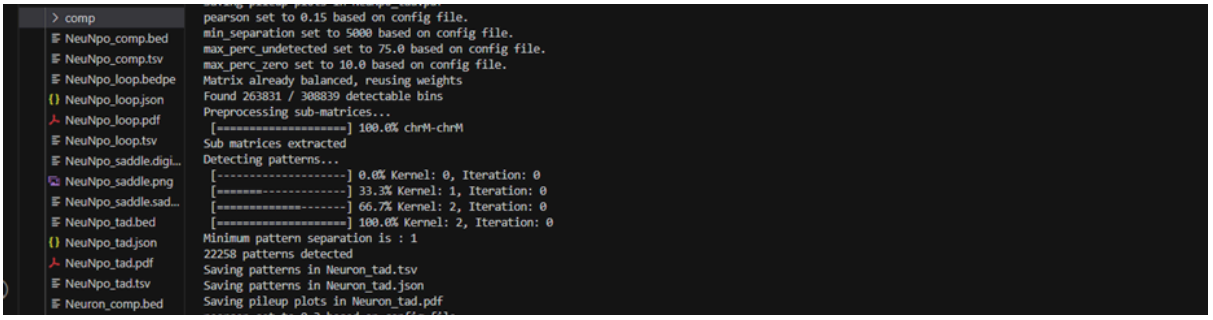
The files with pattern Neuron correspond to neurons. The files with name NeuNpo - iPSC-derived.

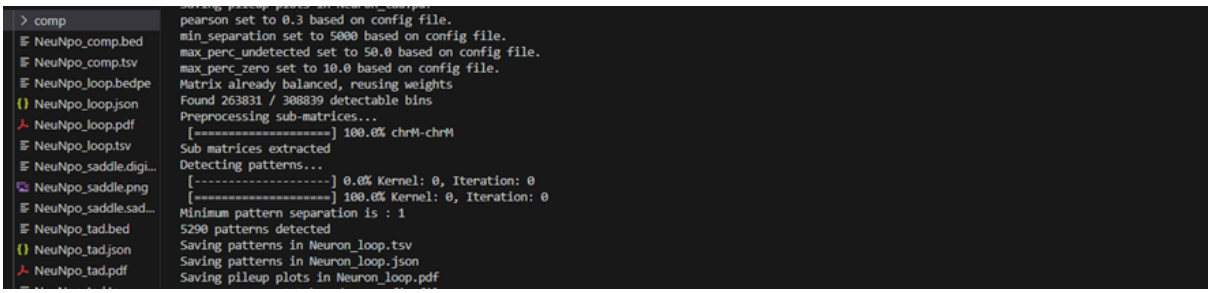
With chromosight tool, I carried out TAD and Loop calling, setting - -pattern parameter to border and loop respectively. I generated the average border plot and average loop plot for the datasets, at several resolution mentioned previously.

The figures below show the report on Tad and Loop calling with chromosight detect, at 10kb resolution. The outputs plot of other resolutions are in the submitted files.

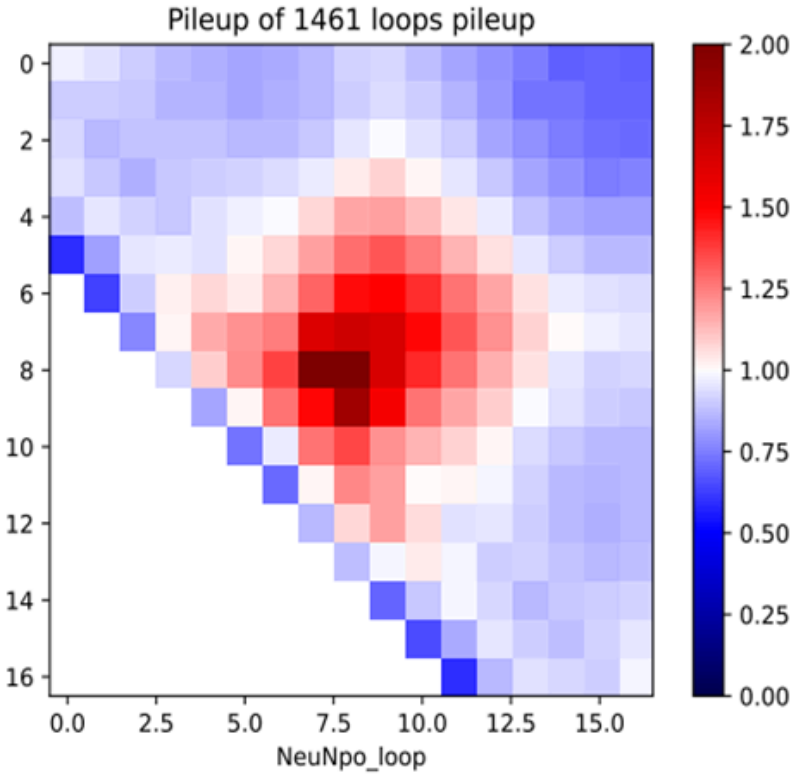


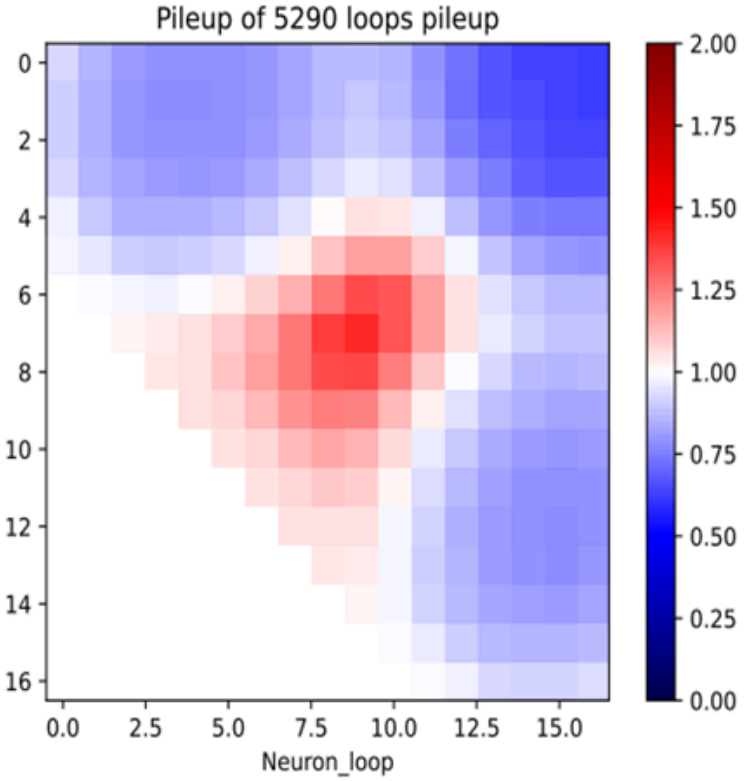


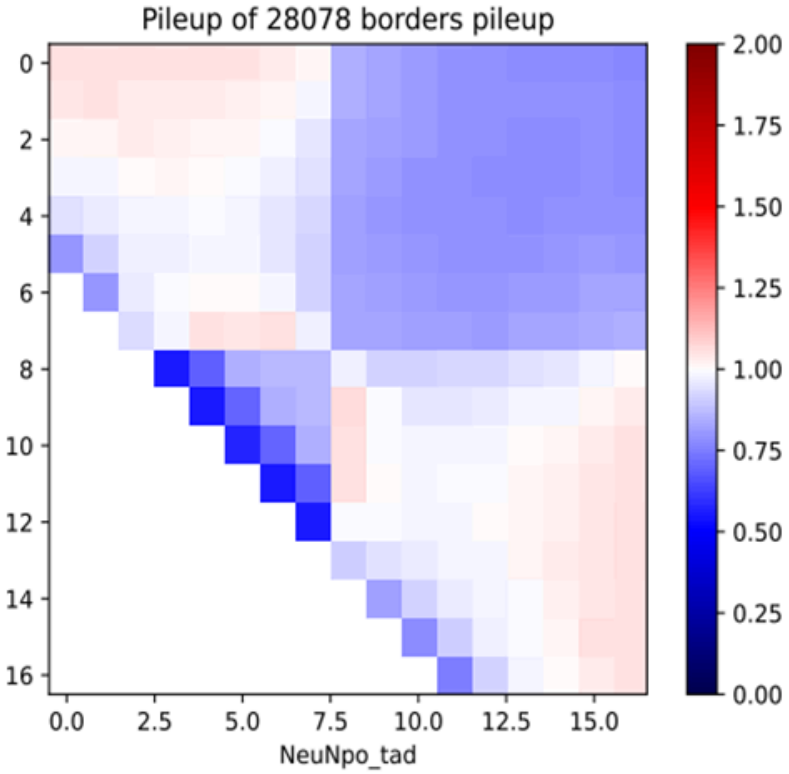


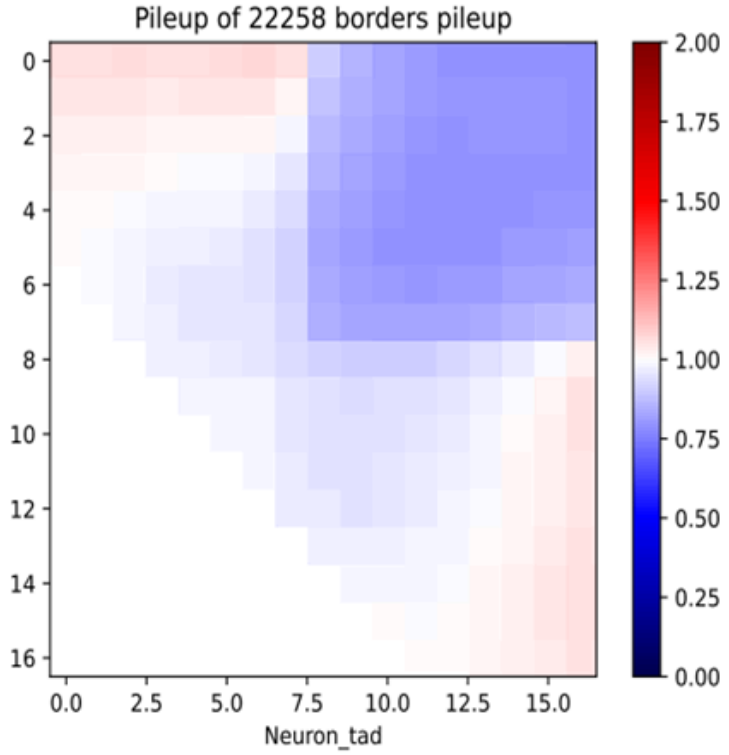


The output average border and anchor plots for 10kb resolution are shown below:









For further analysis, I used the tsv file for each of the chromatin features (compartment, TAD and Loops) to create a bed file for compartment and tad, while bedpe for loops. For subsequent integration of all datasets, TADs, loops and compartments generated from 25kb, 10kb and 50kb resolution respectively.

Overall results and further plans

Ultimately, after processing this set of results, we can conclude the following:

- Based on RNA-seq, ChIP-seq, and methylation analysis data, induced neuron cell lines differ very little from those obtained from postmortem patient samples. Overall, we observe only minor differences in both methylation patterns and gene expression levels.
- Due to the low level of differences, it is possible to use induced cell lines as model samples. But all this is taking into account the fact that small differences are observed due to the greater heterogeneity of postmortem samples compared with a monotonous cell line.
- From the perspective of genome organization and Hi-C analysis, it becomes a little more difficult to draw the same conclusion about the suitability of induced neurons as model systems. Although all indicators show that induced samples demonstrate a clearer and more defined organization (all graphs are sharper), the overall patterns of TADs (topologically associating domains) and loops do not differ significantly between induced and postmortem samples. This is likely due to the heterogeneity of postmortem samples, which prevents obtaining the clear and distinct patterns observed in induced neurons.

Further plans:

- To integrate the results obtained for differential methylation and genome organization analysis in order to understand whether there is a relationship between differentially methylated sites and the tertiary structure of the genome.
- to analyze whether the data on changes in the expression level of certain genes obtained during RNA-sec analysis converge with changes in methylation and the genes associated with these positions.
- to use insulation_and_boundaries tool (https://cooltools.readthedocs.io/en/latest/notebooks/insulation_and_boundaries.html) to study more thoroughly differences and similarities of cell types HiC data