Done by Veronika Karpushenkova
H3K9me3 ChIP-seq analysis

Task

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 details

Comparison of all the data between each other with plotCorrelation (https://deeptools.readthedocs.io/en/develop/content/tools/plotCorrelation.html).

Then, normalization of signal to input with bigwigCompare (https://deeptools.readthedocs.io/en/develop/content/tools/bigwigCompare.html).

Then 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 are used for normalization, not the main comparison. Having normalized data, we will compare two cell types.

The data

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

Abbreviations

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

Files

```
(base) [Veronika.Karpushenko@srv-khrameeva-01 chipseq]$ ls bigwigs/

CnR_H3K9me3_D53_REP1.mLb.clN.bigWig
DA735_H3K9me3_REP1.mLb.clN.bigWig
DA736_Ig_REP1.mLb.clN.bigWig
CnR_H3K9me3_D53_REP2.mLb.clN.bigWig
DA735_Ig_REP1.mLb.clN.bigWig
```

DA737_H3K9me3_REP1.mLb.clN.bigWig
CnR_IgG_D53_REP1.mLb.clN.bigWig
DA736_H3K9me3_REP1.mLb.clN.bigWig
DA737_Ig_REP1.mLb.clN.bigWig

```
(base) [Veronika.Karpushenko@srv-khrameeva-01 chipseq]$ ls peaks/
```

CnR_H3K27me3_D53_REP1_peaks.broadPeak
CnR_H3K9me3_D53_REP1_peaks.broadPeak
DA735_H3K9me3_REP1_peaks.broadPeak
DA737_H3K9me3_REP1_peaks.broadPeak
CnR_H3K27me3_D53_REP2_peaks.broadPeak
CnR_H3K9me3_D53_REP2_peaks.broadPeak
DA736_H3K9me3_REP1_peaks.broadPeak

Working directory is /home/Veronika.Karpushenko/final_project/chipseq.

```
export PATH="/home/i.zhegalova/anaconda3/bin/:$PATH"
source activate chipseq_add
```

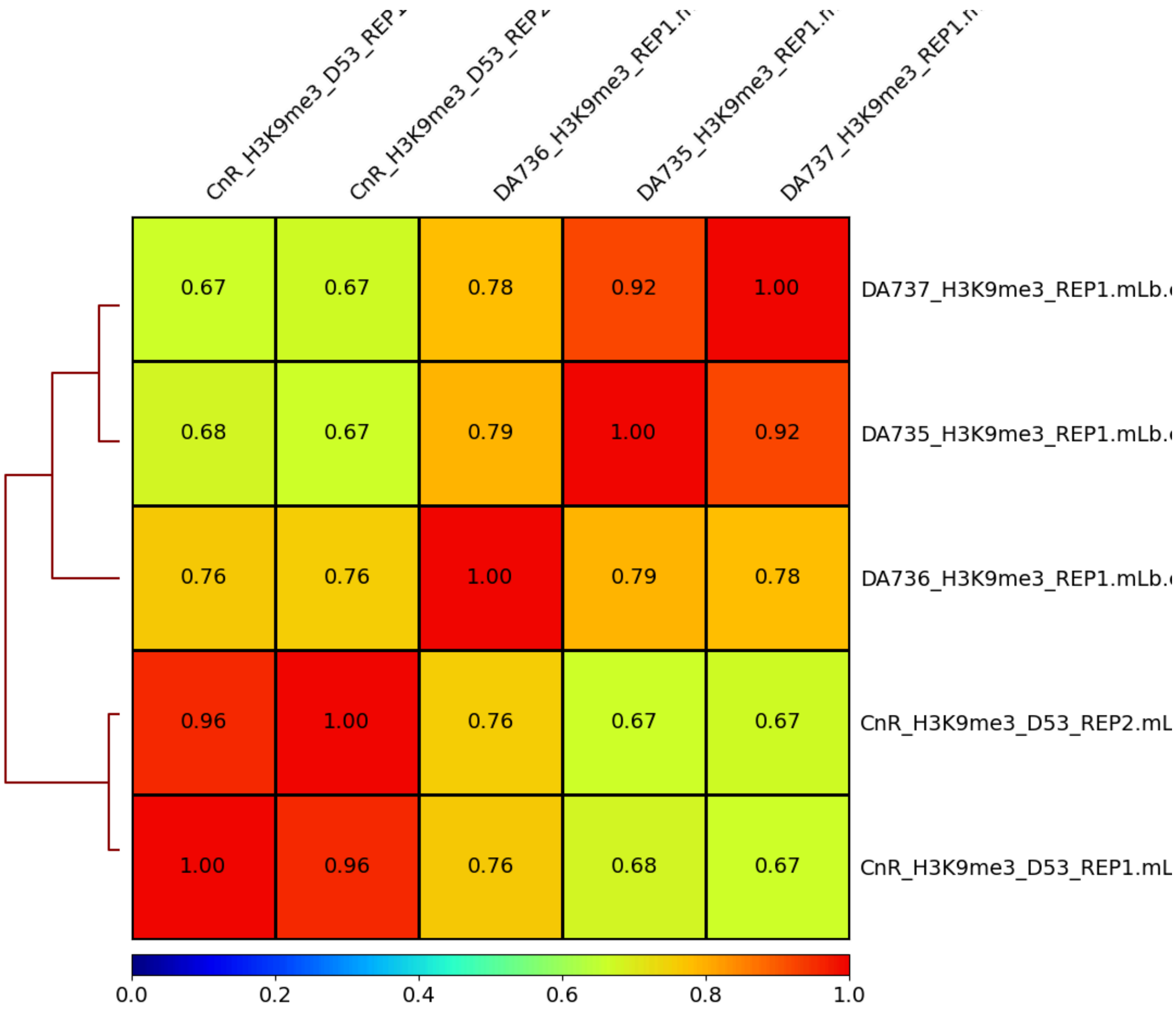
For bw visualization in IGV I downloaded fa and fa.fai files

```
wget https://ftp.ensembl.org/pub/release-114/fasta/homo_sapiens/dna_index/Homo_sapiens.GRCh38.dna.toplevel.fa.gz.fai
wget https://ftp.ensembl.org/pub/release-114/fasta/homo_sapiens/dna_index/Homo_sapiens.GRCh38.dna.toplevel.fa.gz
```

Correlation of raw H3K9me3 signals:

```
multiBigwigSummary bins -b ./bigwigs/*_H3K9me3_* -o raw_signal.npz --outRawCounts raw_signal.tab
```

```
plotCorrelation -in raw_signal.npz --corMethod pearson --whatToPlot heatmap --plotNumbers -o raw_signal_correlation_heatmap.png
```



iPSC replicates (CnR) correlate nicely, and two post-mortem (DA) replicates also correlate well except for DA736 which correlates almost to the same extent with iPSC and post-mortem replicates.

Normalization

Redirecting temporary files:

```
export TMPDIR=/home/Veronika.Karpushenko/final_project/chipseq/mytmp
```

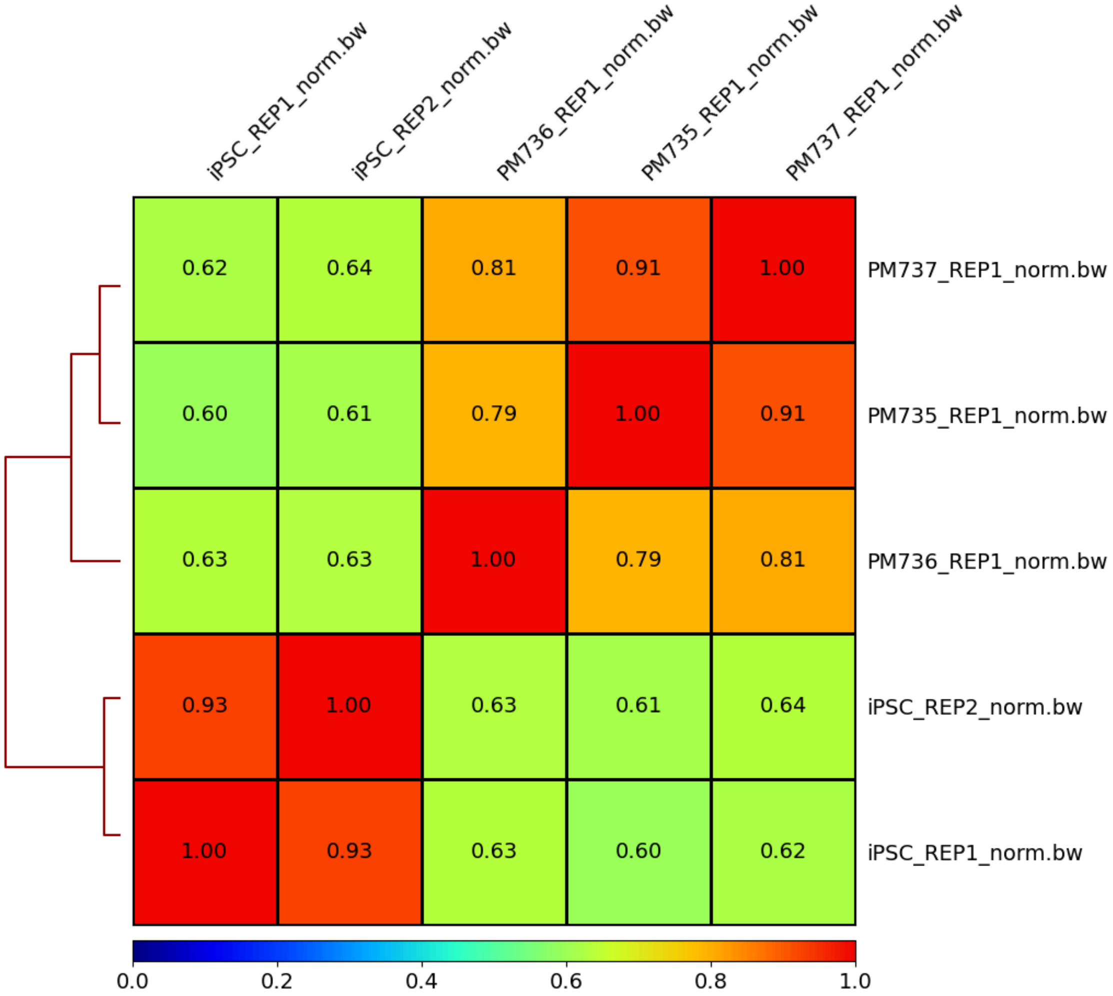
```
bigwigCompare -b1 ./bigwigs/CnR_H3K9me3_D53_REP1.mLb.clN.bigWig -b2
./bigwigs/CnR_IgG_D53_REP1.mLb.clN.bigWig --operation log2 -o iPSC_REP1_norm.bw

bigwigCompare -b1 ./bigwigs/CnR_H3K9me3_D53_REP2.mLb.clN.bigWig -b2
./bigwigs/CnR_IgG_D53_REP1.mLb.clN.bigWig --operation log2 -o iPSC_REP2_norm.bw

bigwigCompare -b1 ./bigwigs/DA735_H3K9me3_REP1.mLb.clN.bigWig  -b2
./bigwigs/DA735_Ig_REP1.mLb.clN.bigWig --operation log2 -o PM735_REP1_norm.bw

bigwigCompare -b1 ./bigwigs/DA736_H3K9me3_REP1.mLb.clN.bigWig  -b2
./bigwigs/DA736_Ig_REP1.mLb.clN.bigWig --operation log2 -o PM736_REP1_norm.bw
```

```
bigwigCompare -b1 ./bigwigs/DA737_H3K9me3_REP1.mLb.clN.bigWig  -b2
./bigwigs/DA737_Ig_REP1.mLb.clN.bigWig --operation log2 -o PM737_REP1_norm.bw
```

Correlation after normalization

```
multiBigwigSummary bins -b *_norm.bw -o norm_signal.npz --outRawCounts
norm_signal.tab
```

```
plotCorrelation -in norm_signal.npz --corMethod pearson --whatToPlot heatmap
--plotNumbers -o norm_signal_correlation_heatmap.png
```



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

Merging replicates within groups

```
export PATH="/home/i.zhegalova/anaconda3/bin/:$PATH"
source activate atacseq_dicty
```

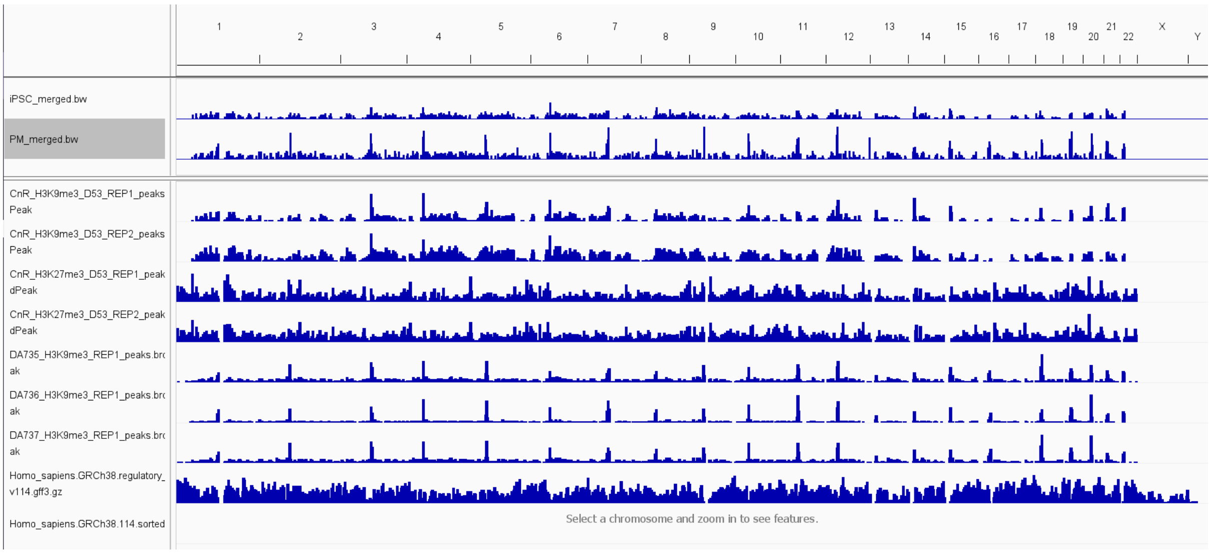
```
bigwigAverage -b iPSC_REP1_norm.bw iPSC_REP2_norm.bw -o iPSC_merged.bw
bigwigAverage -b PM735_REP1_norm.bw PM736_REP1_norm.bw -o PM_1_merged.bw -p 4
bigwigAverage -b PM_1_merged.bw PM737_REP1_norm.bw -o PM_merged.bw -p 4
```

Getting gtf file and bed file

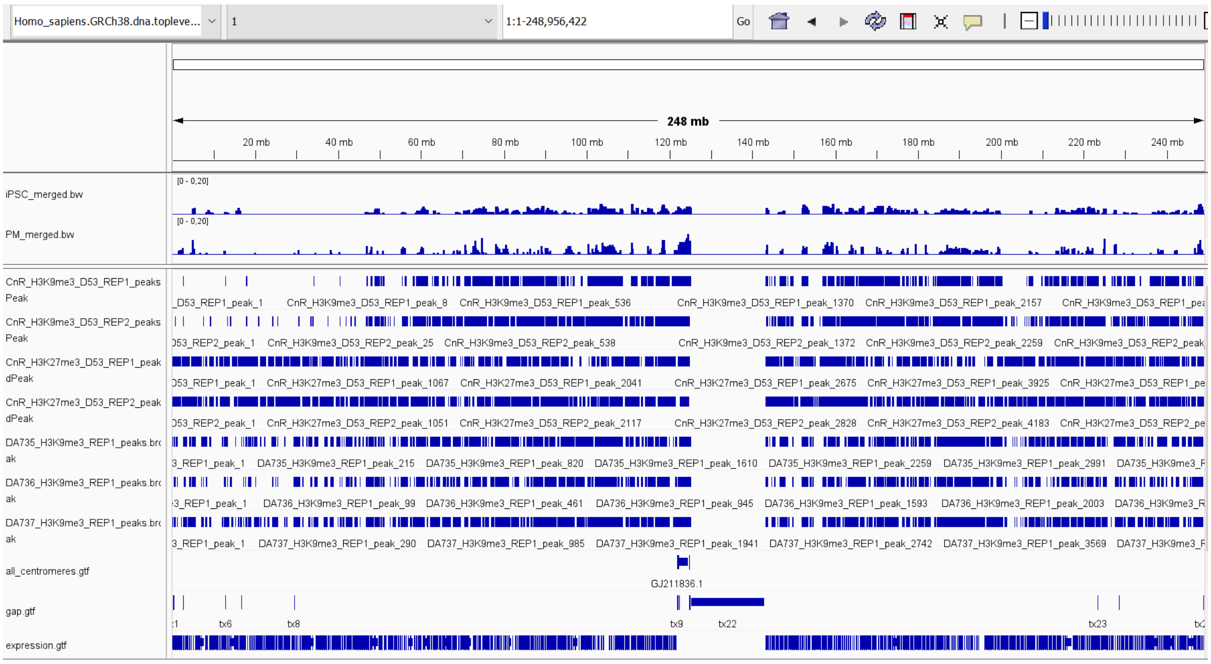
```
wget https://ftp.ensembl.org/pub/release-114/gtf/homo_sapiens/Homo_sapiens.GRCh38.114.gtf.gz
zcat Homo_sapiens.GRCh38.114.gtf.gz | \
awk 'BEGIN{OFS="\t"} $3=="gene" {print $1, $4-1, $5, $10, ".", $7}' > genes.bed
```

Visualization in IGV

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


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
iPSC_merged.bw
PM_merged.bw
CnR_H3K9me3_D53_REP1_peaks Peak
CnR_H3K9me3_D53_REP2_peaks Peak
CnR_H3K27me3_D53_REP1_peak dPeak
CnR_H3K27me3_D53_REP2_peak dPeak
DA735_H3K9me3_REP1_peaks.brc ak
DA736_H3K9me3_REP1_peaks.brc ak
DA737_H3K9me3_REP1_peaks.brc ak
Homo_sapiens.GRCh38.regulatory_ v114.gff3.gz
Homo_sapiens.GRCh38.114.sorted
Select a chromosome and zoom in to see features.

The 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 gff3 file with regulatory elements, and sorted with igvtools gff3 file. The 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:


Homo_sapiens.GRCh38.dna.topleve...
1
1:1-248,956,422
Go
248 mb
20 mb
40 mb
60 mb
80 mb
100 mb
120 mb
140 mb
160 mb
180 mb
200 mb
220 mb
240 mb
iPSC_merged.bw
[0 - 0.20]
PM_merged.bw
[0 - 0.20]
CnR_H3K9me3_D53_REP1_peaks Peak
CnR_H3K9me3_D53_REP2_peaks Peak
CnR_H3K27me3_D53_REP1_peak dPeak
CnR_H3K27me3_D53_REP2_peak dPeak
DA735_H3K9me3_REP1_peaks.bro ak
DA736_H3K9me3_REP1_peaks.bro ak
DA737_H3K9me3_REP1_peaks.bro ak
all_centromeres.gtf
gap.gtf
expression.gtf
CnR_H3K9me3_D53_REP1_peak_8
CnR_H3K9me3_D53_REP1_peak_536
CnR_H3K9me3_D53_REP1_peak_1370
CnR_H3K9me3_D53_REP1_peak_2157
CnR_H3K9me3_D53_REP2_peak_25
CnR_H3K9me3_D53_REP2_peak_538
CnR_H3K9me3_D53_REP2_peak_1372
CnR_H3K9me3_D53_REP2_peak_2259
CnR_H3K27me3_D53_REP1_peak_1067
CnR_H3K27me3_D53_REP1_peak_2041
CnR_H3K27me3_D53_REP1_peak_2675
CnR_H3K27me3_D53_REP1_peak_3925
CnR_H3K27me3_D53_REP2_peak_1051
CnR_H3K27me3_D53_REP2_peak_2117
CnR_H3K27me3_D53_REP2_peak_2828
CnR_H3K27me3_D53_REP2_peak_4183
DA735_H3K9me3_REP1_peak_215
DA735_H3K9me3_REP1_peak_820
DA735_H3K9me3_REP1_peak_1610
DA735_H3K9me3_REP1_peak_2259
DA735_H3K9me3_REP1_peak_2991
DA736_H3K9me3_REP1_peak_99
DA736_H3K9me3_REP1_peak_461
DA736_H3K9me3_REP1_peak_945
DA736_H3K9me3_REP1_peak_1593
DA736_H3K9me3_REP1_peak_2003
DA737_H3K9me3_REP1_peak_290
DA737_H3K9me3_REP1_peak_985
DA737_H3K9me3_REP1_peak_1941
DA737_H3K9me3_REP1_peak_2742
DA737_H3K9me3_REP1_peak_3569
GJ211836.1
tx6
tx8
tx9
tx22
tx23

Here the last three tracks correspond to centromeres, gapped, and expression regions downloaded from: https://genome.ucsc.edu/cgi-bin/hgTables (human genome, Dec.2013 GRCh38/hg38 assembly).
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, and this region corresponds to a gapped region with a peak corresponding to centromere at the very beginning, and with a gap in expression track, so this is probably centromere region. From biological interpretation H3K9me3 corresponds to heterochromatin and gene silencing, and H3K27me3 corresponds to facultative heterochromatin.

Computing matrix for visualization

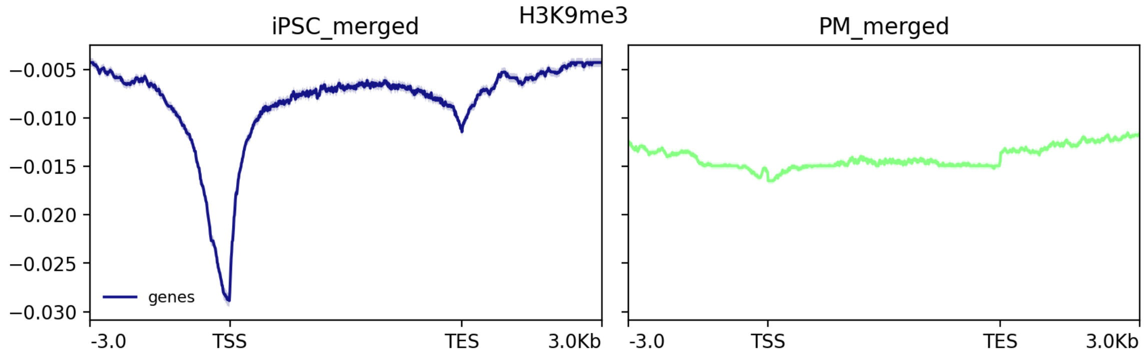
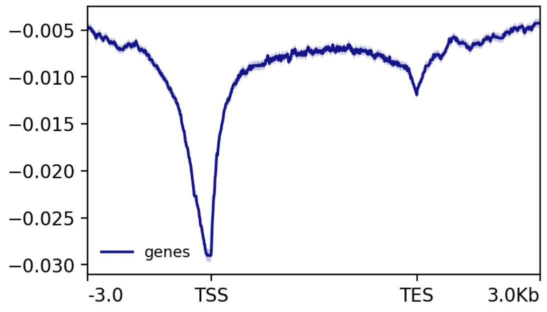
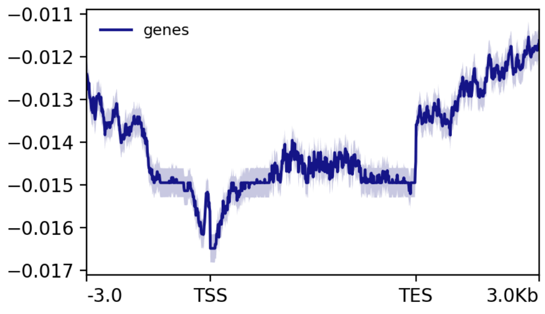
```
computeMatrix scale-regions -S iPSC_merged.bw PM_merged.bw -R genes.bed
--beforeRegionStartLength
3000 --regionBodyLength 5000 --afterRegionStartLength 3000 --skipZeros -o
matrix_H3K9me3.gz --smartLabels
```

```
computeMatrix scale-regions -S iPSC_merged.bw -R genes.bed --beforeRegionStartLength
3000 --region
BodyLength 5000 --afterRegionStartLength 3000 --skipZeros -o matrix_iPSC_H3K9me3.gz
--smartLabels -p 4
```

```
computeMatrix scale-regions -S PM_merged.bw -R genes.bed --beforeRegionStartLength
3000 --regionBodyLength 5000 --afterRegionStartLength 3000 --skipZeros -o
matrix_PM_H3K9me3.gz --smartLabels
```

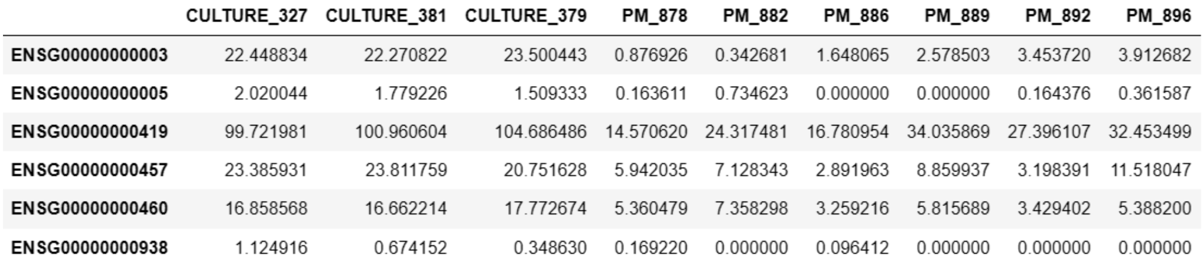
Plotting enrichment around genes

```
plotProfile -m matrix_H3K9me3.gz -out PLOT_final.png --plotTitle "H3K9me3" --averageType
median --plotType se

plotProfile -m matrix_iPSC_H3K9me3.gz -out PLOT_iPSC.png --plotTitle "H3K9me3_iPSC"
--averageType median --plotType se

plotProfile -m matrix_PM_H3K9me3.gz -out PLOT_PM.png --plotTitle "H3K9me3_PM"
--averageType median --plotType se
```

<table>
<tr><td colspan="2">
</td></tr>
<tr><td colspan="2">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:</td></tr>
<tr><td>iPSC</td><td>Post-mortem</td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td colspan="2">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.</td></tr>
</table>

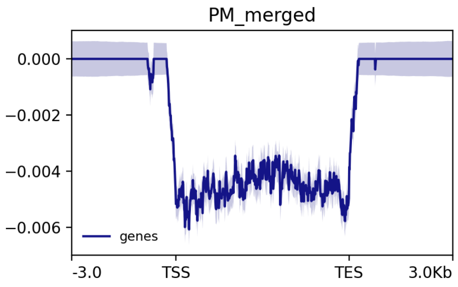
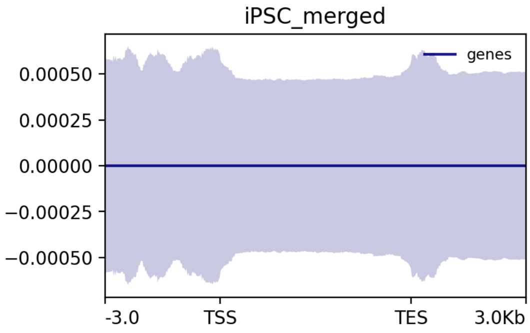
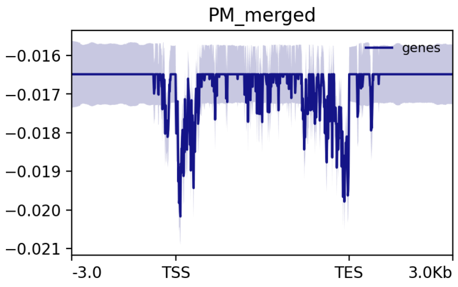
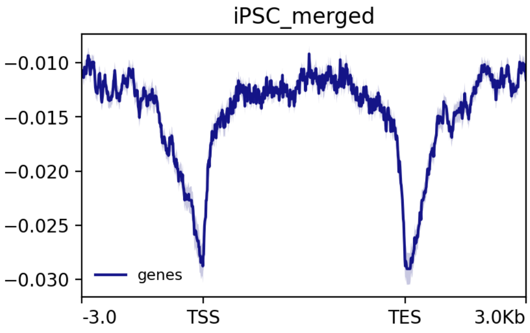
Plotting with TMP filtering

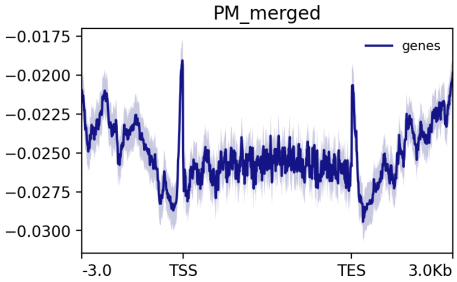
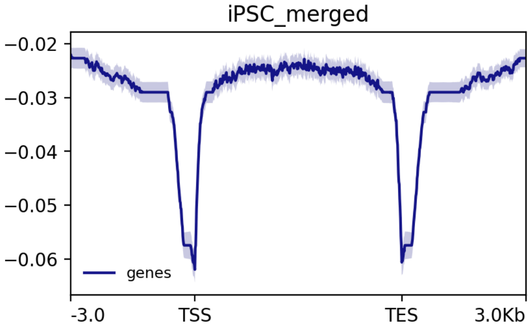
The code for this part can be found in rnaseq.ipynb in the repository.
Firstly, from RNA-seq data I extracted TPM (in transcripts per million) matrix:



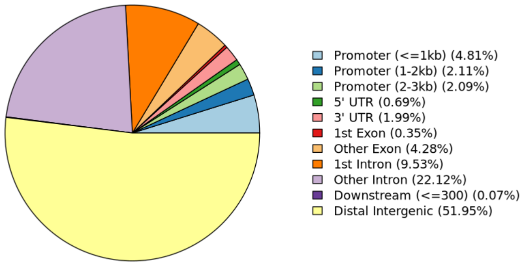
	CULTURE_327	CULTURE_381	CULTURE_379	PM_878	PM_882	PM_886	PM_889	PM_892	PM_896
ENSG00000000003	22.448834	22.270822	23.500443	0.876926	0.342681	1.648065	2.578503	3.453720	3.912682
ENSG00000000005	2.020044	1.779226	1.509333	0.163611	0.734623	0.000000	0.000000	0.164376	0.361587
ENSG00000000419	99.721981	100.960604	104.686486	14.570620	24.317481	16.780954	34.035869	27.396107	32.453499
ENSG00000000457	23.385931	23.811759	20.751628	5.942035	7.128343	2.891963	8.859937	3.198391	11.518047
ENSG00000000460	16.858568	16.662214	17.772674	5.360479	7.358298	3.259216	5.815689	3.429402	5.388200
ENSG00000000938	1.124916	0.674152	0.348630	0.169220	0.000000	0.096412	0.000000	0.000000	0.000000

Now we can divide genes into 3 groups according to gene expression level in transcripts per million: TPM < 1, 1 < TPM < 10, TPM >10.

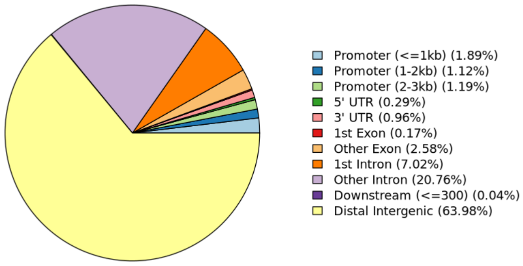
TPM	post-mortem	i-PSC
< 1	 	
	Not really strong signal in terms of enrichment but between TSS and TES we see a characteristic decrease inChIP-seq signal as it is active region of transcription.	Baseline signal is only present for genes with low expression level.
from 1 to 10	 	
	The pictures above and below are	Those enrichment levels above and below

<table>
<tr><td></td><td>really noisy. Below we can see enrichment level around -0.03 which is twice bigger in absolute value from the averaged on all genes for post-mortem samples. So probably in post-mortem samples there is quite a high amount of genes with TMP < 1 affecting the whole picture. The pattern in the picture above is comparable to iPSC but in the picture below at TSS and TES we have got positive peaks because baseline level is just different. Compared to zero those peaks would have been negative</td><td>mostly correspond to the picture with all genes regardless of TMP so for iPSC most of the genes have TMP &gt; 1</td></tr>
<tr><td>&gt; 10</td><td>
</td><td>
</td></tr>
<tr><td colspan="3">Summarising results in this table, baseline levels in iPSC, and especially in post-mortem samples are lower than 0 for middle or highly expressed genes (genes with TPM > 1). iPSC samples have a good consistency in pattern and low noise. PM samples are noisy and inconsistent between genes with different TPM. More or less this behaviour of PM samples could be explained by their heterogeneity, batch effect, technical procedures.</td></tr>
</table>

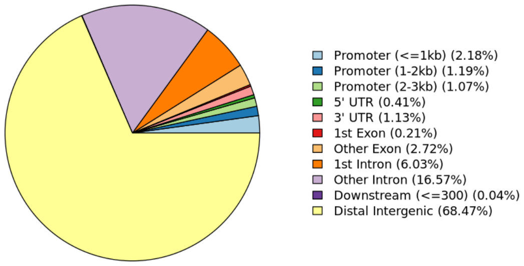
ChIP-seq peaks annotation with ChIPseeker



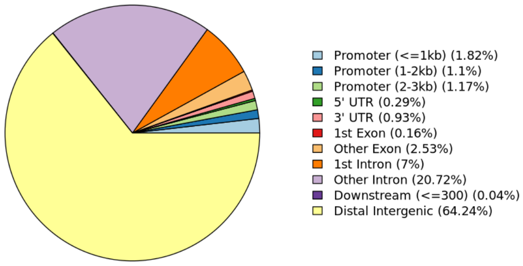
Annotation of unique peaks present in both iPSC replicates with H3K27me3 peaks



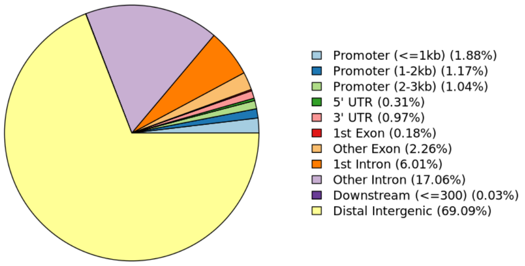
Annotation of unique peaks present in both iPSC replicates with H3K9me3 peaks



Annotation of unique peaks present in both PM replicates with H3K9me3 peaks



Annotation of unique peaks present in both iPSC H3K9me3 and H3K27me3 peaks



Annotation of unique peaks present in both PM and iPSC H3K9me3 peaks

—---

NeuNpos relates to post-mortem neurons,
but Neurons - to iPSC-derived

```
tail -n +2 ./Project/m25000/NeuNpo_tad.bed > ./Project/m25000/NeuNpo_tad_noheader.bed
tail -n +2 ./Project/m25000/Neuron_tad.bed > ./Project/m25000/Neuron_tad_noheader.bed

computeMatrix scale-regions -S PM_merged.bw -R
./Project/m25000/NeuNpo_tad_noheader.bed --beforeRegionStartLength 50000
--regionBodyLength 100000 --afterRegionStartLength 50000 -o matrix_PM_chip_tads.gz

computeMatrix scale-regions -S iPSC_merged.bw -R
./Project/m25000/Neuron_tad_noheader.bed --beforeRegionStartLength 50000
--regionBodyLength 100000 --afterRegionStartLength 50000 -o matrix_iPSC_chip_tads.gz
```

—----------------

Cleaned and modifies gene annotation files:

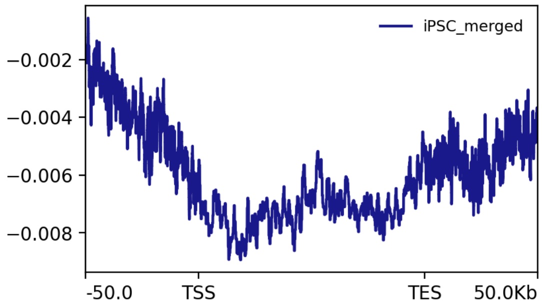
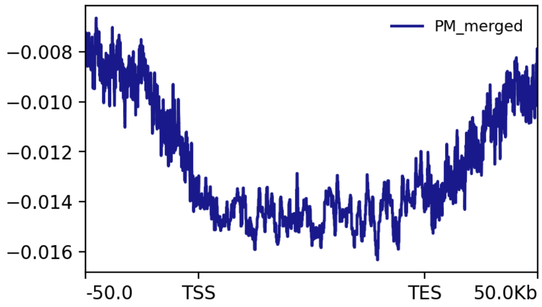
```
awk 'BEGIN{OFS="\t"} {gsub(/[";]/,"",$4); print}' genes.bed > genes_clean.bed
awk 'BEGIN{OFS="\t"} {if($1 !~ /^chr/) $1="chr"$1; print}' genes_clean.bed >
genes_clean_chr.bed
```

Finding genes within TADs

```
 bedtools intersect -a genes_clean_chr.bed -b ./Project/m25000/Neuron_tad_noheader.bed
-wo > Neuron_genes_in_TADs.bed

bedtools intersect -a genes_clean_chr.bed -b ./Project/m25000/NeuNpo_tad_noheader.bed
-wo > NeuNpo_genes_in_TADs.bed
```

Plotting ChIP-seq signal profiles across genes in TADs

```
computeMatrix scale-regions -S PM_merged.bw -R
./Project/m25000/NeuNpo_tad_noheader.bed --beforeRegionStartLength 50000
--regionBodyLength 100000 --afterRegionStartLength 50000 -o matrix_PM_chip_tads.gz

computeMatrix scale-regions -S iPSC_merged.bw -R
./Project/m25000/Neuron_tad_noheader.bed --beforeRegionStartLength 50000
--regionBodyLength 100000 --afterRegionStartLength 50000 -o matrix_iPSC_chip_tads.gz

plotProfile -m matrix_iPSC_chip_tads.gz -out iPSC_chip_tads.png --perGroup
plotProfile -m matrix_PM_chip_tads.gz -out PM_chip_tads.png --perGroup
```

<table>
<tr><td>ChIP-seq signal profiles across genes of iPSC</td><td>ChIP-seq signal profiles across genes of post-mortem cells</td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td colspan="2">As we can see, the characteristic shape for H3K9me3 is the same for two cell types, yet the difference from basal level of enrichment is twice bigger for post-mortem samples.</td></tr>
</table>

Plotting ChIP-seq signal profiles across genes in compartments

```
awk '$4 > 0 {print $1"\t"$2"\t"$3}' Neuron_comp.bed > Neuron_A.bed
tail -n +2 ./Project/m50000/Neuron_A.bed > ./Project/m50000/Neuron_A_noheader.bed
awk '$4 <= 0 {print $1"\t"$2"\t"$3}' Neuron_comp.bed > Neuron_B.bed
tail -n +2 ./Project/m50000/Neuron_B.bed > ./Project/m50000/Neuron_B_noheader.bed

awk '$4 > 0 {print $1"\t"$2"\t"$3}' NeuNpo_comp.bed > NeuNpo_A.bed
tail -n +2 ./Project/m50000/NeuNpo_A.bed > ./Project/m50000/NeuNpo_A_noheader.bed
awk '$4 <= 0 {print $1"\t"$2"\t"$3}' NeuNpo_comp.bed > NeuNpo_B.bed
tail -n +2 ./Project/m50000/NeuNpo_B.bed > ./Project/m50000/NeuNpo_B_noheader.bed

multiBigwigSummary BED-file --bwfiles iPSC_merged.bw --BED
./Project/m50000/Neuron_A_noheader.bed -out Neuron_A.npz --outRawCounts
Neuron_A_signal.tab

multiBigwigSummary BED-file --bwfiles iPSC_merged.bw --BED
./Project/m50000/Neuron_B_noheader.bed -out Neuron_B.npz --outRawCounts
Neuron_B_signal.tab
```

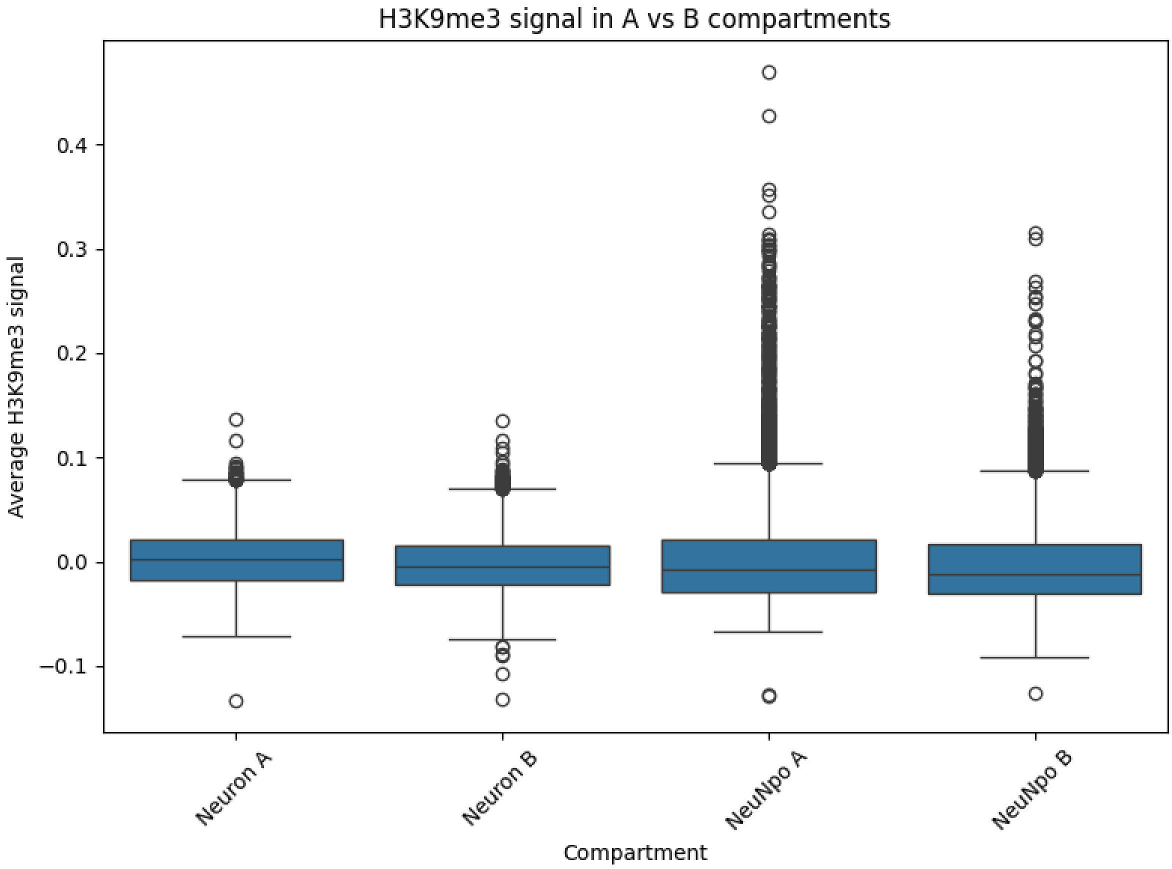
```
multiBigwigSummary BED-file --bwfiles PM_merged.bw --BED
./Project/m50000/NeuNpo_A_noheader.bed -out NeuNpo_A.npz --outRawCounts
NeuNpo_A_signal.tab
```

```
multiBigwigSummary BED-file --bwfiles PM_merged.bw --BED
./Project/m50000/NeuNpo_B_noheader.bed -out NeuNpo_B.npz --outRawCounts
NeuNpo_B_signal.tab
```

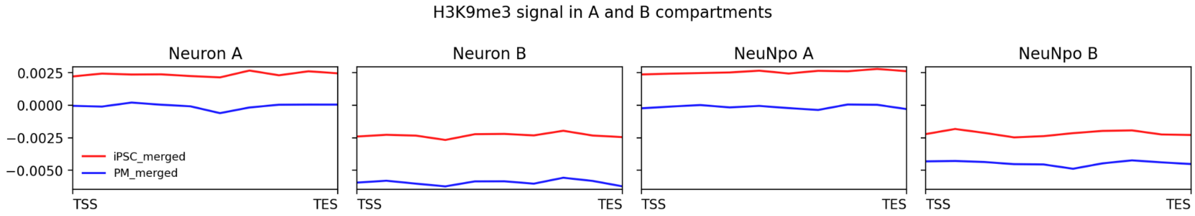
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

def load_signal(file, label):
    df = pd.read_csv(file, sep='\t', comment='#', header=None)
    df = df.iloc[:,3]
    return pd.DataFrame({'Signal': df, 'Compartment': label})

neuron_A = load_signal('Neuron_A_signal.tab', 'Neuron A')
neuron_B = load_signal('Neuron_B_signal.tab', 'Neuron B')
neunpo_A = load_signal('NeuNpo_A_signal.tab', 'NeuNpo A')
neunpo_B = load_signal('NeuNpo_B_signal.tab', 'NeuNpo B')

all_data = pd.concat([neuron_A, neuron_B, neunpo_A, neunpo_B])

plt.figure(figsize=(8,6))
sns.boxplot(x='Compartment', y='Signal', data=all_data)
plt.title('H3K9me3 signal in A vs B compartments')
plt.ylabel('Average H3K9me3 signal')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
plt.savefig(‘chip_in_comp.png’)
```

```
computeMatrix scale-regions   -S iPSC_merged.bw PM_merged.bw   -R
./Project/m50000/Neuron_A_noheader.bed ./Project/m50000/Neuron_B_noheader.bed
./Project/m50000/NeuNpo_A_noheader.bed ./Project/m50000/NeuNpo_B_noheader.bed
--regionBodyLength 50000   --binSize 5000   -out matrix_compartments.npz   --skipZeros
--missingDataAsZero
```

```
plotProfile -m matrix_compartments.npz \
  -out H3K9me3_compartments_profile.png \
  --perGroup \
  --plotTitle "H3K9me3 signal in A and B compartments" \
  --regionsLabel "Neuron A" "Neuron B" "NeuNpo A" "NeuNpo B" \
  --colors red blue \
  --plotHeight 6 --plotWidth 8
```



As we can see, there is no significant difference between two cell types and between two compartments. Post-mortem cells (NeuNpo) have bigger dispersion compared to iPSC (Neuron), which is due to bigger heterogeneity of post-mortem samples, batch effect, and technical errors.



Gene expression per compartments

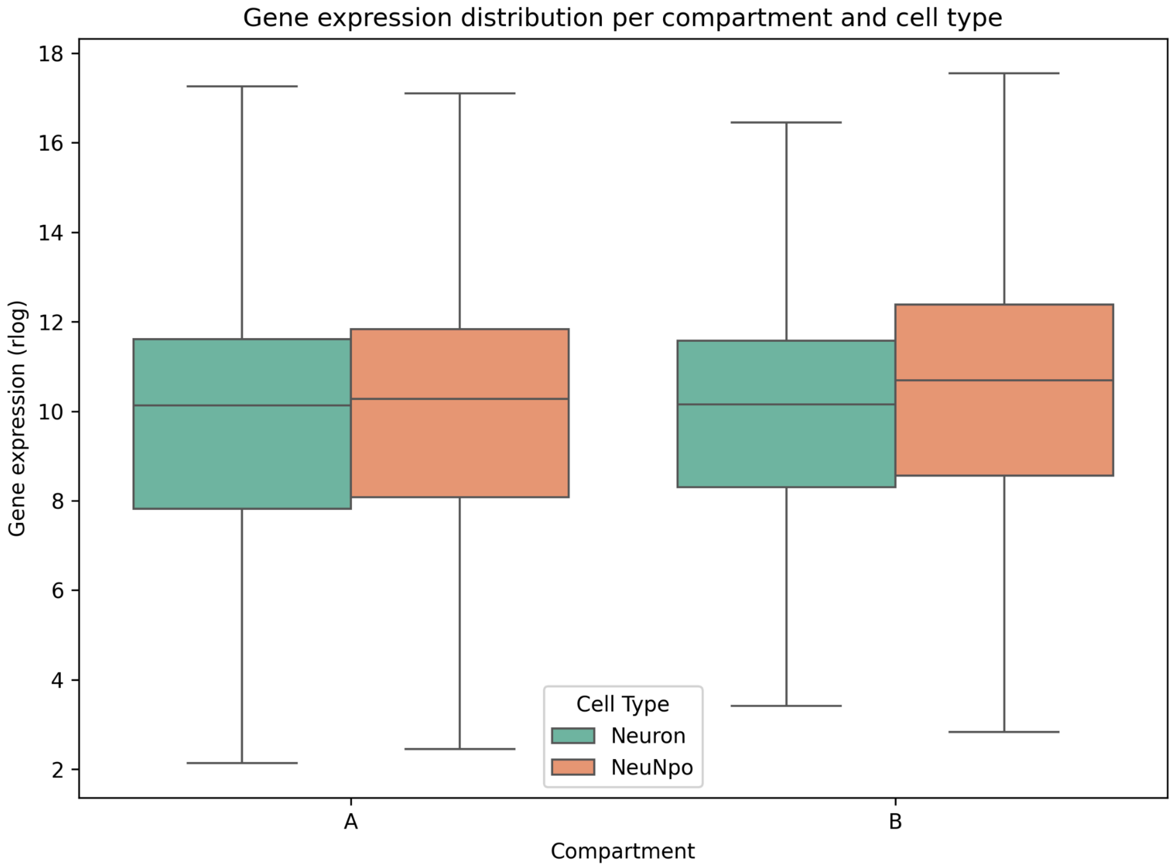
```
awk '{gsub(/"/,"",$4); gsub(/;/,"",$4); print "chr"$1"\t"$2"\t"$3"\t"$4}' genes.bed >
genes_fixed.bed
```

```
bedtools intersect -a genes_fixed.bed -b ./Project/m50000/Neuron_comp.bed -wa -wb >
gene2NeuronComp.txt
```

```
 bedtools intersect -a genes_fixed.bed -b ./Project/m50000/NeuNpo_comp.bed -wa -wb >
gene2NeuNpoComp.txt
```

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

expression_df = pd.read_csv('rlog_gene_expression_matrix.csv')
neuron_map = pd.read_csv('gene2NeuronComp.txt', sep='\t', header=None)
neuron_map = neuron_map[[3, 7]].rename(columns={3: 'gene_id', 7: 'Neuron_E1'})
neuron_map['Neuron_compartment'] = neuron_map['Neuron_E1'].apply(lambda x: 'A' if
float(x) > 0 else 'B')

neunpo_map = pd.read_csv('gene2NeuNpoComp.txt', sep='\t', header=None)
neunpo_map = neunpo_map[[3, 7]].rename(columns={3: 'gene_id', 7: 'NeuNpo_E1'})
neunpo_map['NeuNpo_compartment'] = neunpo_map['NeuNpo_E1'].apply(lambda x: 'A' if
float(x) > 0 else 'B')

compartment_df = pd.merge(
    neuron_map[['gene_id', 'Neuron_compartment']],
    neunpo_map[['gene_id', 'NeuNpo_compartment']],
    on='gene_id', how='outer'
)

merged_df = pd.merge(expression_df, compartment_df, on='gene_id')
neuron_cols = [col for col in merged_df.columns if col.startswith('CULTURE_')]
neunpo_cols = [col for col in merged_df.columns if col.startswith('PM_')]

neuron_long = merged_df.melt(
    id_vars=['gene_id', 'Neuron_compartment'],
    value_vars=neuron_cols,
    var_name='Sample',
    value_name='Expression'
)
neuron_long['CellType'] = 'Neuron'
neuron_long['Compartment'] = neuron_long['Neuron_compartment']

neunpo_long = merged_df.melt(
    id_vars=['gene_id', 'NeuNpo_compartment'],
    value_vars=neunpo_cols,
    var_name='Sample',
    value_name='Expression'
)
neunpo_long['CellType'] = 'NeuNpo'
neunpo_long['Compartment'] = neunpo_long['NeuNpo_compartment']

# Combine for plotting
plot_df = pd.concat([neuron_long, neunpo_long], ignore_index=True)

plot_df = plot_df.dropna(subset=['Compartment'])
```

```
plt.figure(figsize=(8,6))
sns.boxplot(
    data=plot_df,
    x='Compartment',
    y='Expression',
    hue='CellType',
    showfliers=False,
    palette='Set2'
)
plt.ylabel('Gene expression (rlog)')
plt.xlabel('Compartment')
plt.title('Gene expression distribution per compartment and cell type')
plt.legend(title='Cell Type')
plt.tight_layout()
plt.savefig('expression_per_compartment_boxplot.png', dpi=300)
plt.show()
```



As we can see, there is no significant difference between two cell types and between two compartments in gene expression. Post-mortem cells (NeuNpo) have bigger dispersion compared to iPSC (Neuron), which is due to bigger heterogeneity of post-mortem samples, batch effect, and technical errors.