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
```

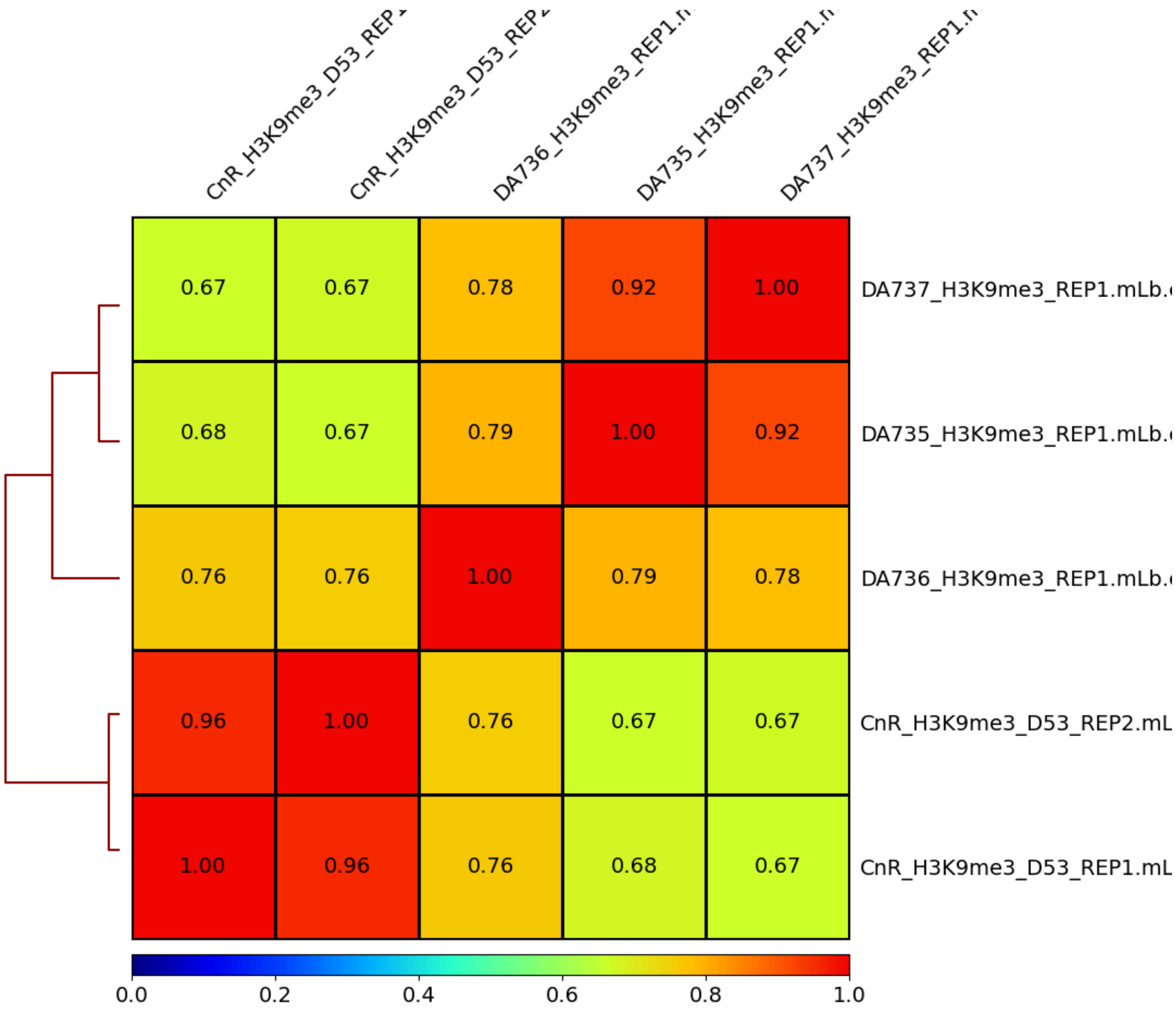
For bw visualization in IGV I downloaded fa and fa.fai files

```
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
```

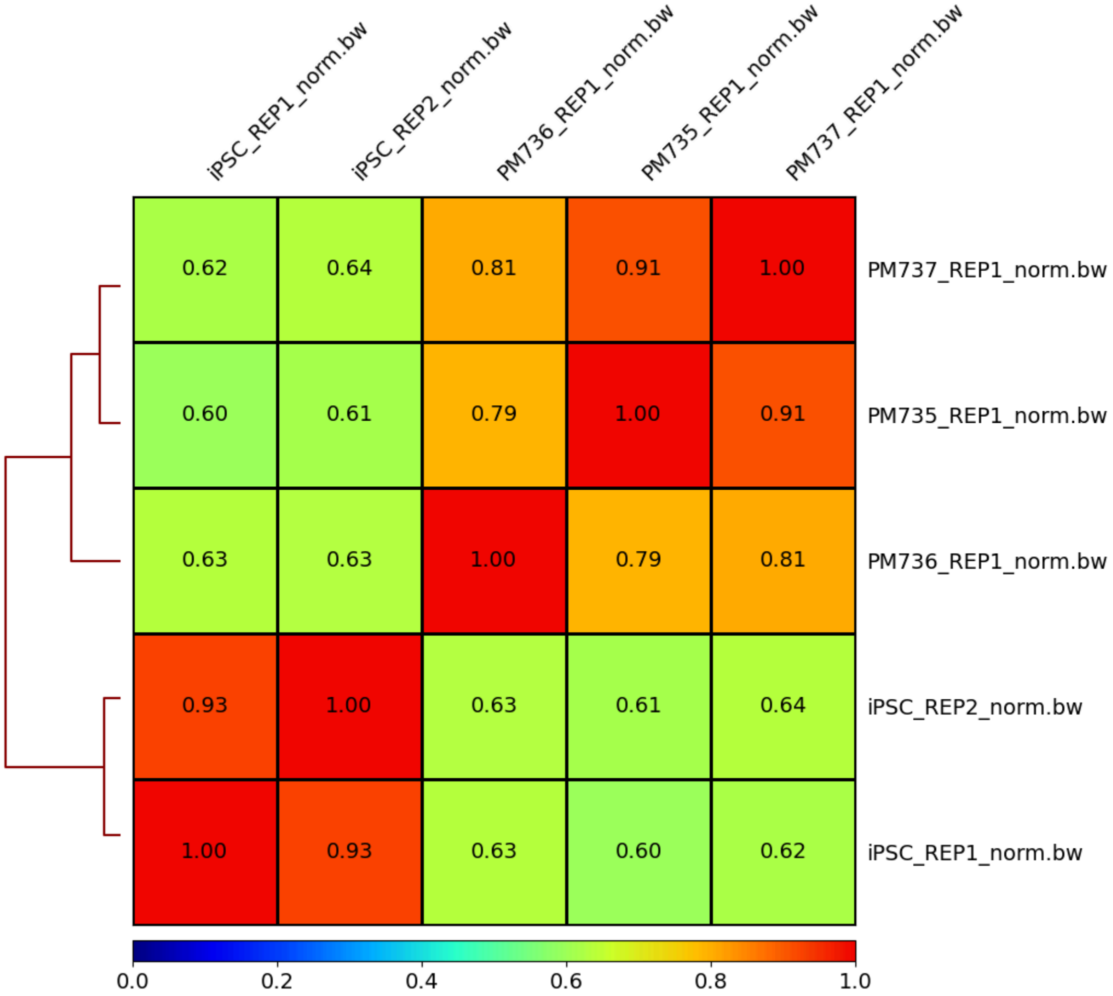
```
bigwigCompare -b1 ./bigwigs/DA736_H3K9me3_REP1.mLb.clN.bigWig  -b2
./bigwigs/DA736_Ig_REP1.mLb.clN.bigWig --operation log2 -o PM736_REP1_norm.bw
```

```
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
```



	iPSC_REP1_norm.bw	iPSC_REP2_norm.bw	PM736_REP1_norm.bw	PM735_REP1_norm.bw	PM737_REP1_norm.bw
PM737_REP1_norm.bw	0.62	0.64	0.81	0.91	1.00
PM735_REP1_norm.bw	0.60	0.61	0.79	1.00	0.91
PM736_REP1_norm.bw	0.63	0.63	1.00	0.79	0.81
iPSC_REP2_norm.bw	0.93	1.00	0.63	0.61	0.64
iPSC_REP1_norm.bw	1.00	0.93	0.63	0.60	0.62

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
```

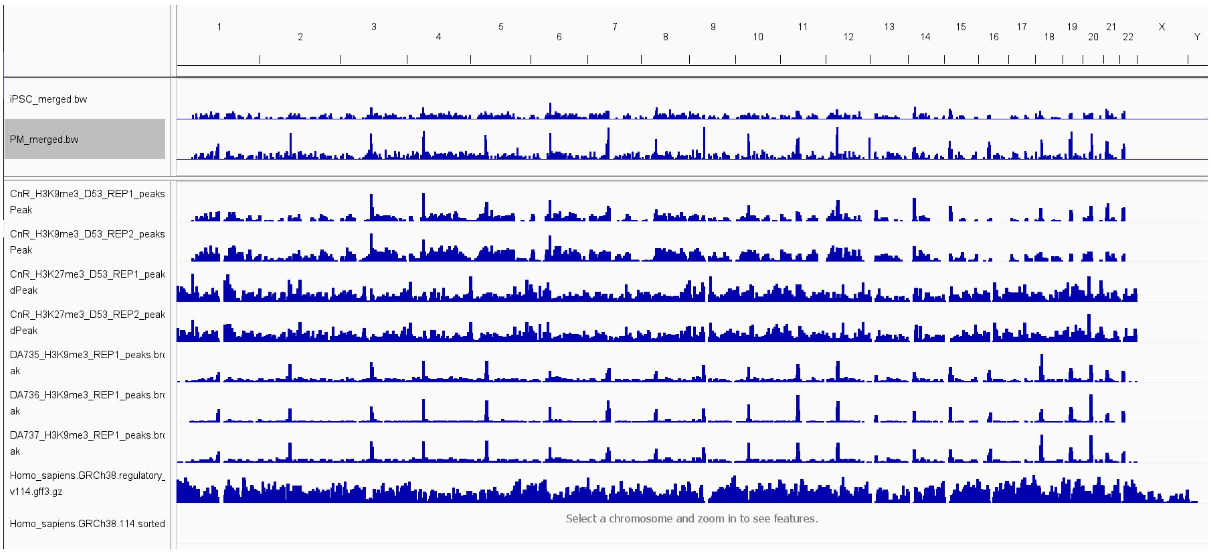
```
bigwigAverage -b iPSC_REP1_norm.bw iPSC_REP2_norm.bw -o iPSC_merged.bw
bigwigAverage -b PM735_REP1_norm.bw PM736_REP1_norm.bw -o PM_1_merged.bw -p 4
bigwigAverage -b PM_1_merged.bw PM737_REP1_norm.bw -o PM_merged.bw -p 4
```

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
```

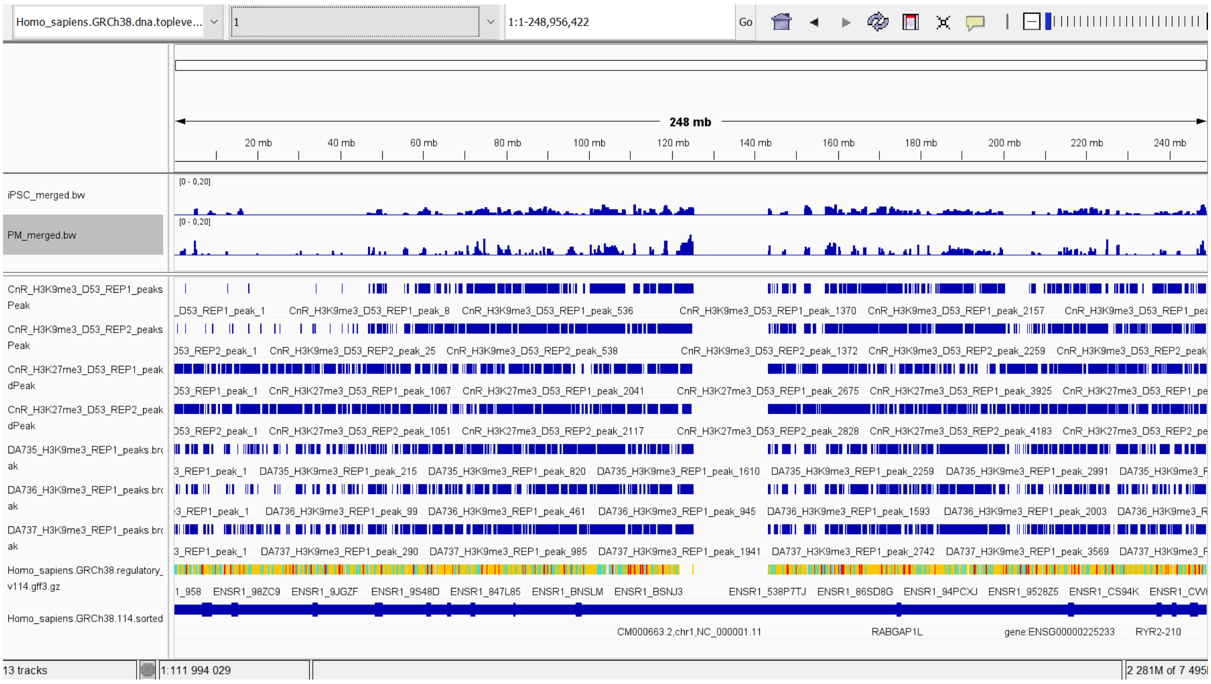
Visualization in IGV

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
Homo_sapiens.GRCh38.regulatory_ v114.gff3.gz
Homo_sapiens.GRCh38.114.sorted
CM000663.2,chr1,NC_000001.11
RABGAP1L
gene:ENSG00000225233
RYR2-210
13 tracks
1:111 994 029
2 281M of 7 495

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.

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

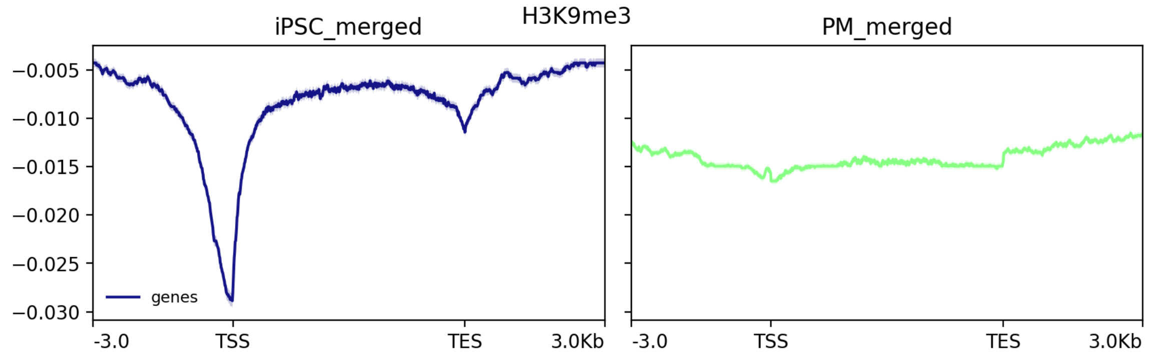
computeMatrix scale-regions -S PM_merged.bw -R genes.bed --beforeRegionStartLength 3000 --regionBodyLength 5000 --afterRegionStartLength 3000 --skipZeros -o matrix_PM_H3K9me3.gz --smartLabels

Plotting enrichment around genes

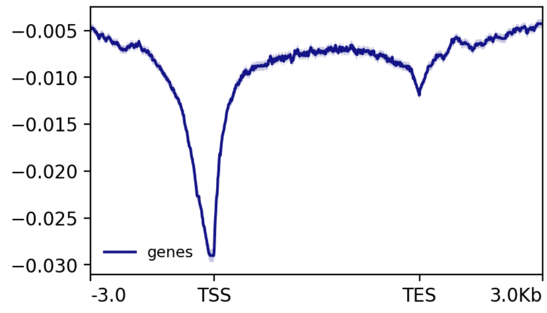
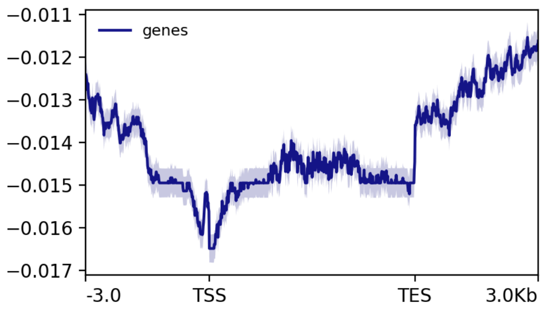
```
plotProfile -m matrix_H3K9me3.gz -out PLOT_final.png --plotTitle "H3K9me3" --averageType
median --plotType se

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
```



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.