Analysis of human and mouse expressed genes enriched by H3K9me3 histone modification

Sidorov Svyatoslav

Scientific advisor:

Alexander Predeus, Bioinformatics Institute

11 May 2016

Outline

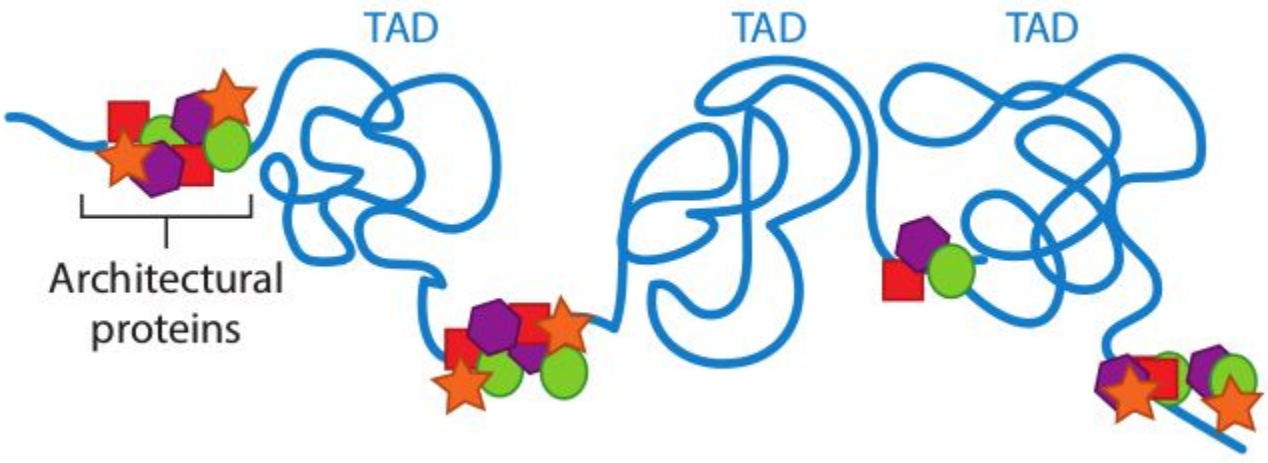
1. Introduction.
2. My master thesis in brief.
3. The current project.
4. Current results.
5. Perspective.

Outline

1. Introduction.
2. My master thesis in brief.
3. The current project.
4. Current results.
5. Perspective.

Topologically associating domains

Topologically associating domains (TADs) are such regions of chromatin that frequency of intra-TAD interactions is higher than frequency of inter-TAD interactions.



Adapted from *Nguyen H. Q.* and *Bosco G.*, Annual Review of Genetics. 2015, 49:627-646.

Topologically associating domains

TADs as functional domains in mammals:

- TAD borders are enriched in active transcription, housekeeping genes, tRNA genes and SINE repeats, as well as binding sites for the architectural proteins CTCF and cohesin.

See *Dekker J. and Heard E.,* FEBS Letters. 2015, 589(20 Pt A):2877-2884.

Topologically associating domains

TADs as functional domains in mammals:

- TAD borders are enriched in active transcription, housekeeping genes, tRNA genes and SINE repeats, as well as binding sites for the architectural proteins CTCF and cohesin.
- TADs are units of coordinated gene expression.

See *Dekker J. and Heard E.,* FEBS Letters. 2015, 589(20 Pt A):2877-2884.

Topologically associating domains

TADs as functional domains in mammals:

- TAD borders are enriched in active transcription, housekeeping genes, tRNA genes and SINE repeats, as well as binding sites for the architectural proteins CTCF and cohesin.
- TADs are units of coordinated gene expression.
- Series of adjacent TADs correspond to replication domains.

See *Dekker J. and Heard E.,* FEBS Letters. 2015, 589(20 Pt A):2877-2884.

Topologically associating domains

TADs as functional domains in mammals:

- TAD borders are enriched in active transcription, housekeeping genes, tRNA genes and SINE repeats, as well as binding sites for the architectural proteins CTCF and cohesin.
- TADs are units of coordinated gene expression.
- Series of adjacent TADs correspond to replication domains.
- TAD borders are to a significant extent conserved between different cell types, and even between mouse and human.

See *Dekker J. and Heard E.,* FEBS Letters. 2015, 589(20 Pt A):2877-2884.

Chromatin modifications

Histone proteins can obtain various post-translational modifications (acetylation, methylation, phosphorylation and others) which regulates transcription.

Portela A., Esteller M., Nature Biotechnology, 2010, 28:1057-1068.

Chromatin modifications

Histone proteins can obtain various post-translational modifications (acetylation, methylation, phosphorylation and others) which regulates transcription.

Portela A., Esteller M., Nature Biotechnology, 2010, 28:1057-1068.

Some of these modifications are well established markers of specific genomic loci, e. g.,

- **H3K36me3** and **H3K79me2** mark actively transcribed genes;
- **H3K9me3** and **H3K27me3** mark heterochromatin;
- **H3K4me3** marks promoters of transcribed genes.

Outline

1. Introduction.
2. My master thesis in brief.
3. The current project.
4. Current results.
5. Perspective.

Goal and the main steps

The main goal of the master thesis project was to perform a joint analysis of TADs and chromatin modifications in human embryonic stem cells (hESC-H1 cell line).

Goal and the main steps

The main goal of the master thesis project was to perform a joint analysis of TADs and chromatin modifications in human embryonic stem cells (hESC-H1 cell line).

The main steps were as follows:

- Process ChiP-seq data for several histone modifications and some of the main hESC-H1 transcription factors to find genome loci with such modifications and TFs.

Goal and the main steps

The main goal of the master thesis project was to perform a joint analysis of TADs and chromatin modifications in human embryonic stem cells (hESC-H1 cell line).

The main steps were as follows:

- Process ChiP-seq data for several histone modifications and some of the main hESC-H1 transcription factors to find genome loci with such modifications and TFs.
- Find stable combinations of such modifications and TF binding sites along the genome.

Goal and the main steps

The main goal of the master thesis project was to perform a joint analysis of TADs and chromatin modifications in human embryonic stem cells (hESC-H1 cell line).

The main steps were as follows:

- Process ChiP-seq data for several histone modifications and some of the main hESC-H1 transcription factors to find genome loci with such modifications and TFs.
- Find stable combinations of such modifications and TF binding sites along the genome.
- Process Hi-C data to obtain genome segmentation into TADs and TAD borders.

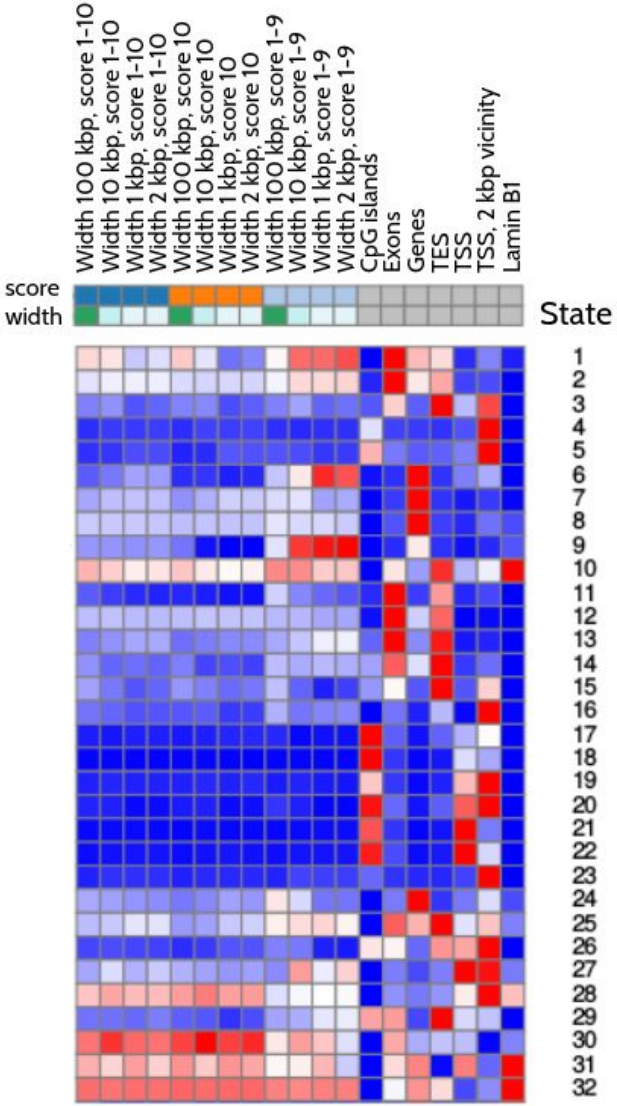
Goal and the main steps

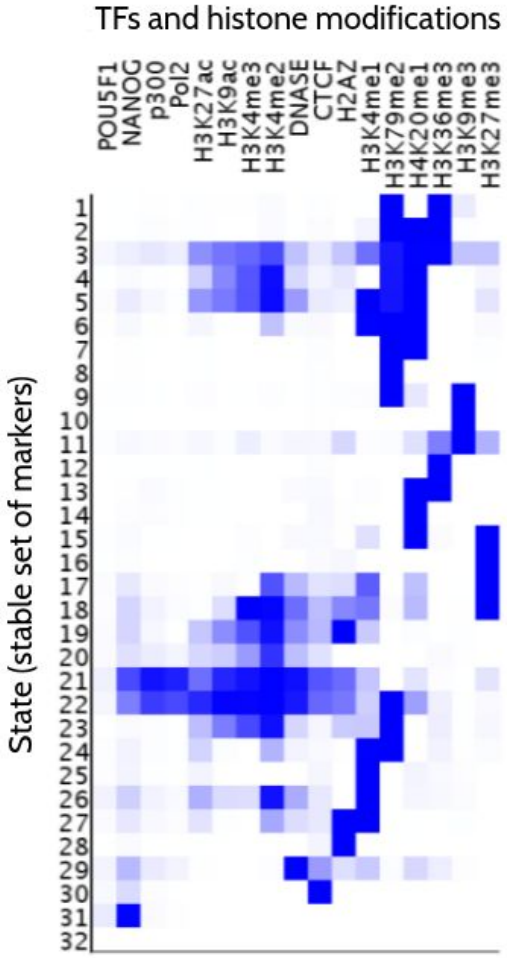
The main goal of the master thesis project was to perform a joint analysis of TADs and chromatin modifications in human embryonic stem cells (hESC-H1 cell line).

The main steps were as follows:

- Process ChiP-seq data for several histone modifications and some of the main hESC-H1 transcription factors to find genome loci with such modifications and TFs.
- Find stable combinations of such modifications and TF binding sites along the genome.
- Process Hi-C data to obtain genome segmentation into TADs and TAD borders.
- Find TAD border enrichments by the stable combinations and some genomic features.

The main result

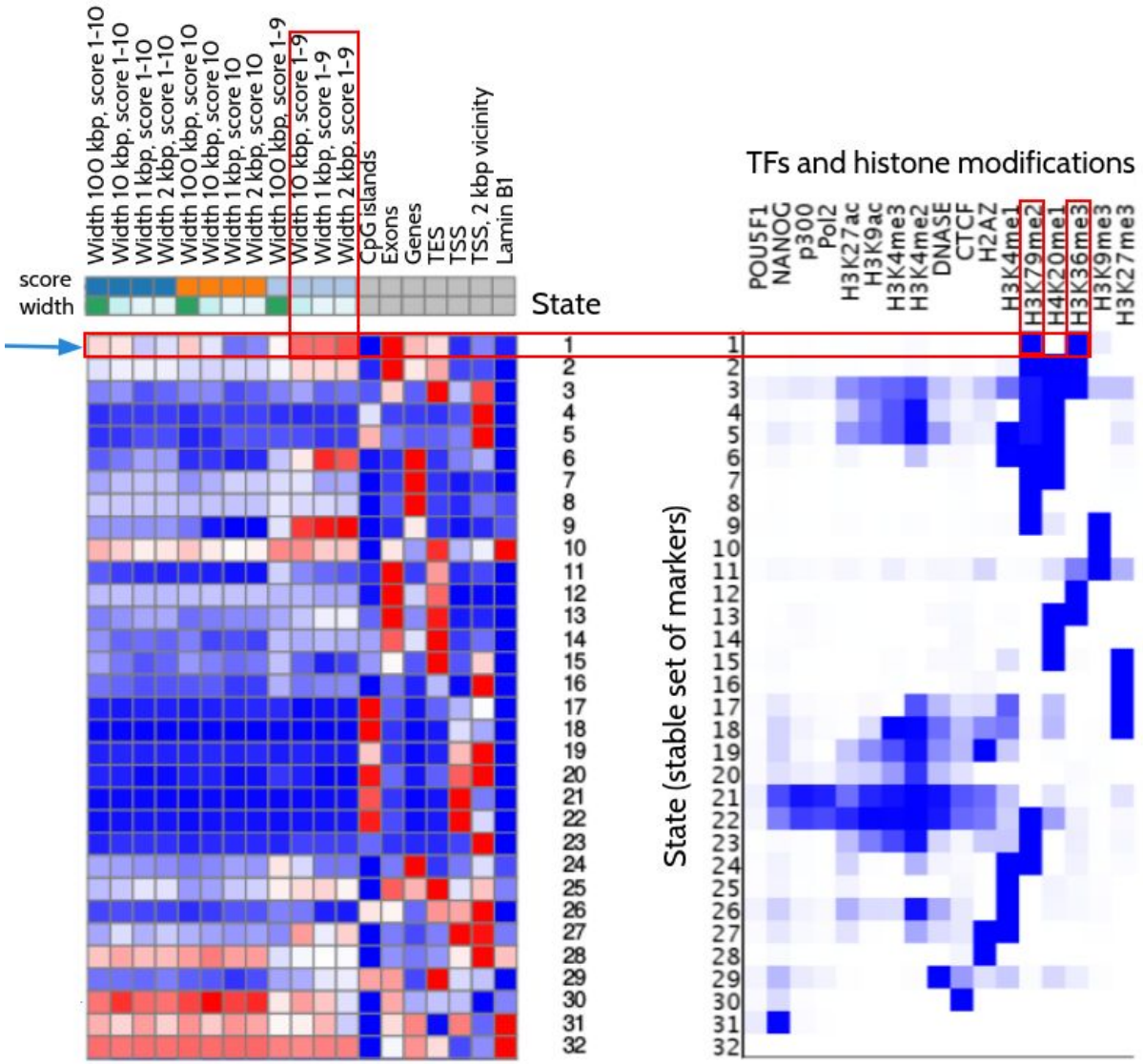




TAD border is strong if it nearly prohibits contacts over itself, and weak otherwise. Border strength is assessed with a score from 1 (the weakest) to 10 (the strongest).

Here we call borders with score 10 ‘strong’ and all other borders - ‘weak’.

The main result

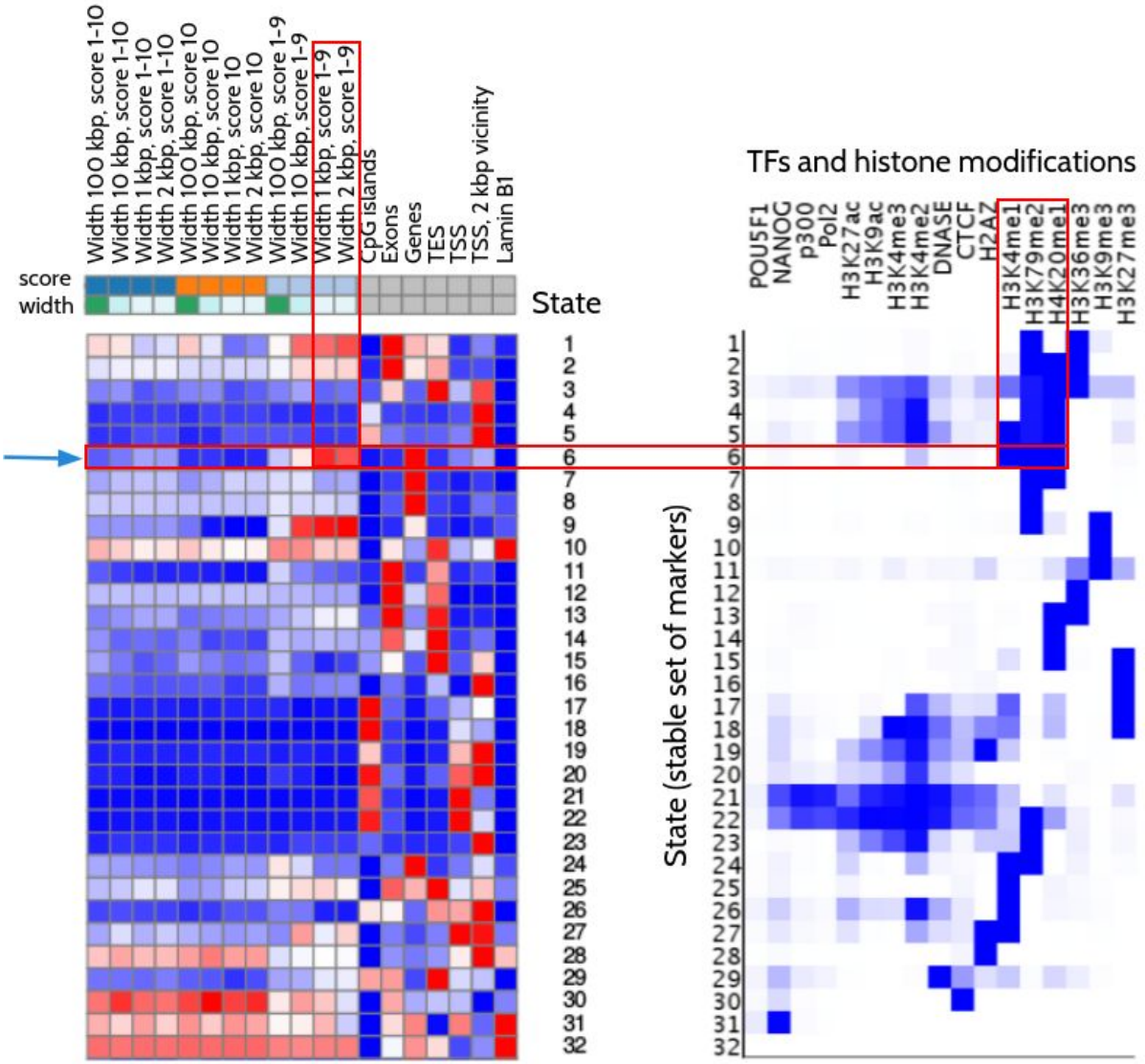


TAD border is strong if it nearly prohibits contacts over itself, and weak otherwise. Border strength is assessed with a score from 1 (the weakest) to 10 (the strongest).

Here we call borders with score 10 'strong' and all other borders - 'weak'.

Weak borders are enriched in states 1 and 6 (corresponding to transcribed genes).

The main result

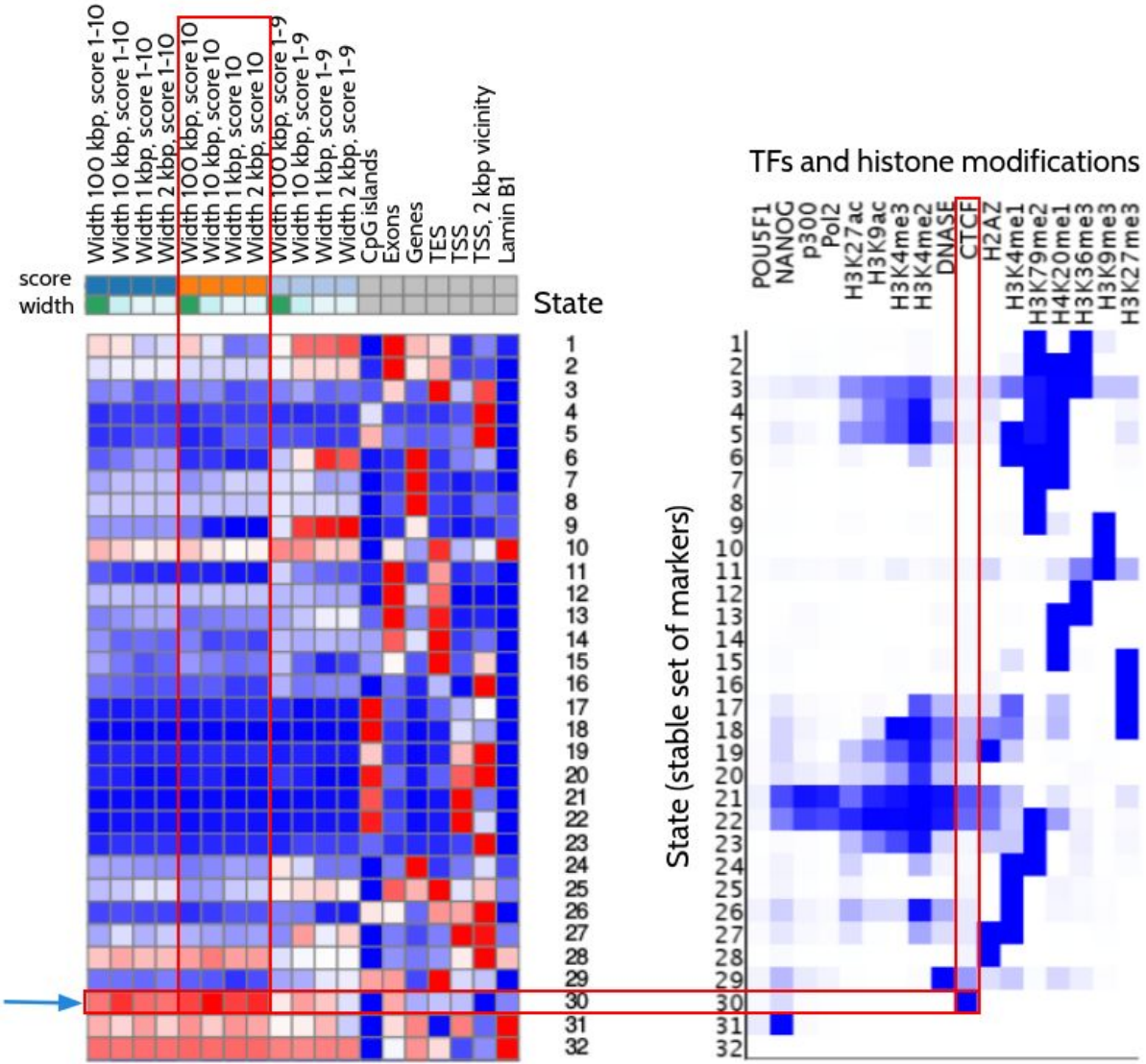


TAD border is strong if it nearly prohibits contacts over itself, and weak otherwise. Border strength is assessed with a score from 1 (the weakest) to 10 (the strongest).

Here we call borders with score 10 'strong' and all other borders - 'weak'.

Weak borders are enriched in states 1 and 6 (corresponding to transcribed genes).

The main result

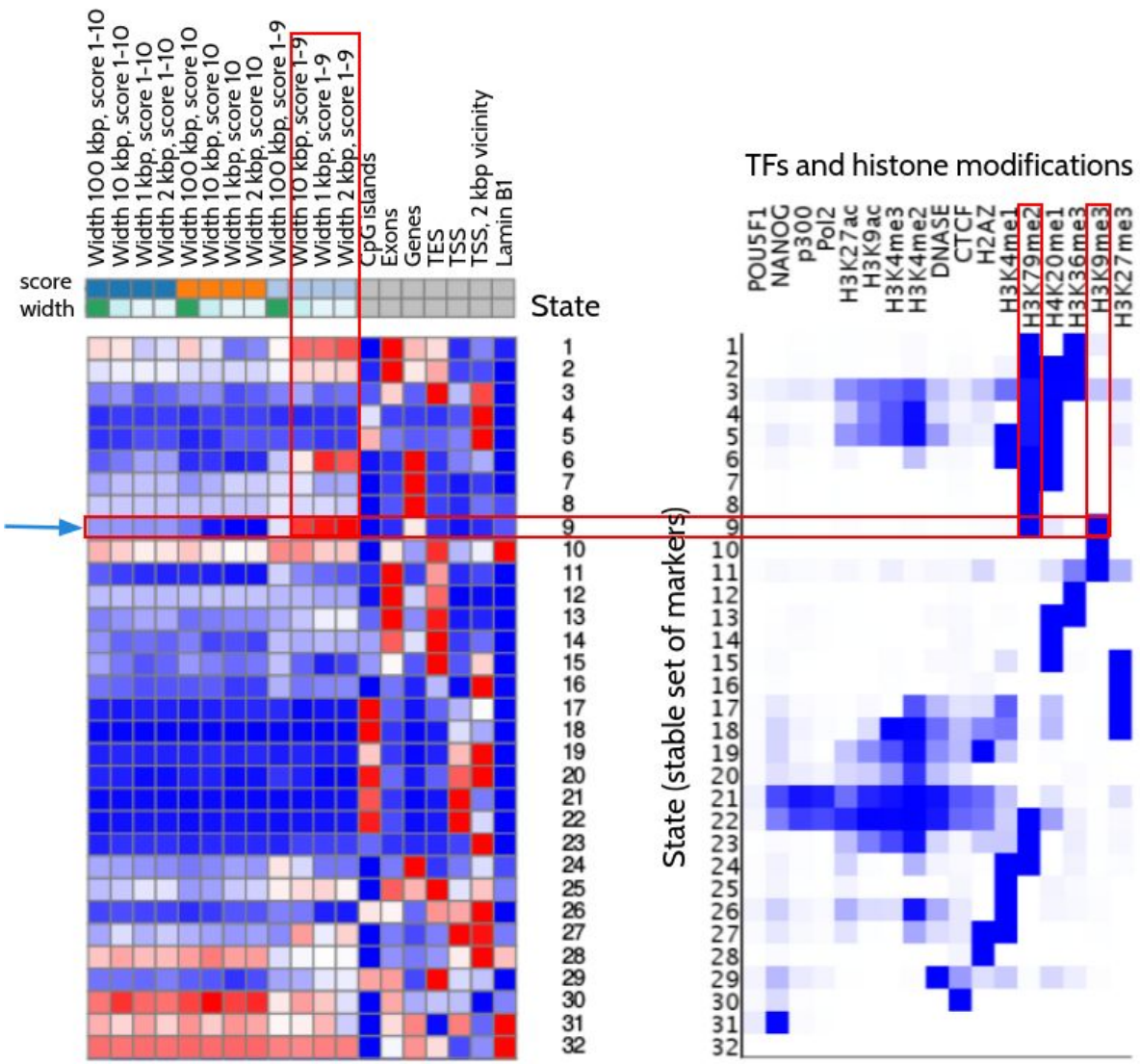


TAD border is strong if it nearly prohibits contacts over itself, and weak otherwise. Border strength is assessed with a score from 1 (the weakest) to 10 (the strongest). **Here we call borders with score 10 ‘strong’ and all other borders - ‘weak’.**

Weak borders are enriched in states 1 and 6 (corresponding to transcribed genes).

Strong borders are enriched in state 30 (CTCF alone).

The main result



TAD border is strong if it nearly prohibits contacts over itself, and weak otherwise. Border strength is assessed with a score from 1 (the weakest) to 10 (the strongest).

Here we call borders with score 10 'strong' and all other borders - 'weak'.

Weak borders are enriched by states 1 and 6 (corresponding to transcribed genes).

Strong borders are enriched by state 30 (CTCF alone).

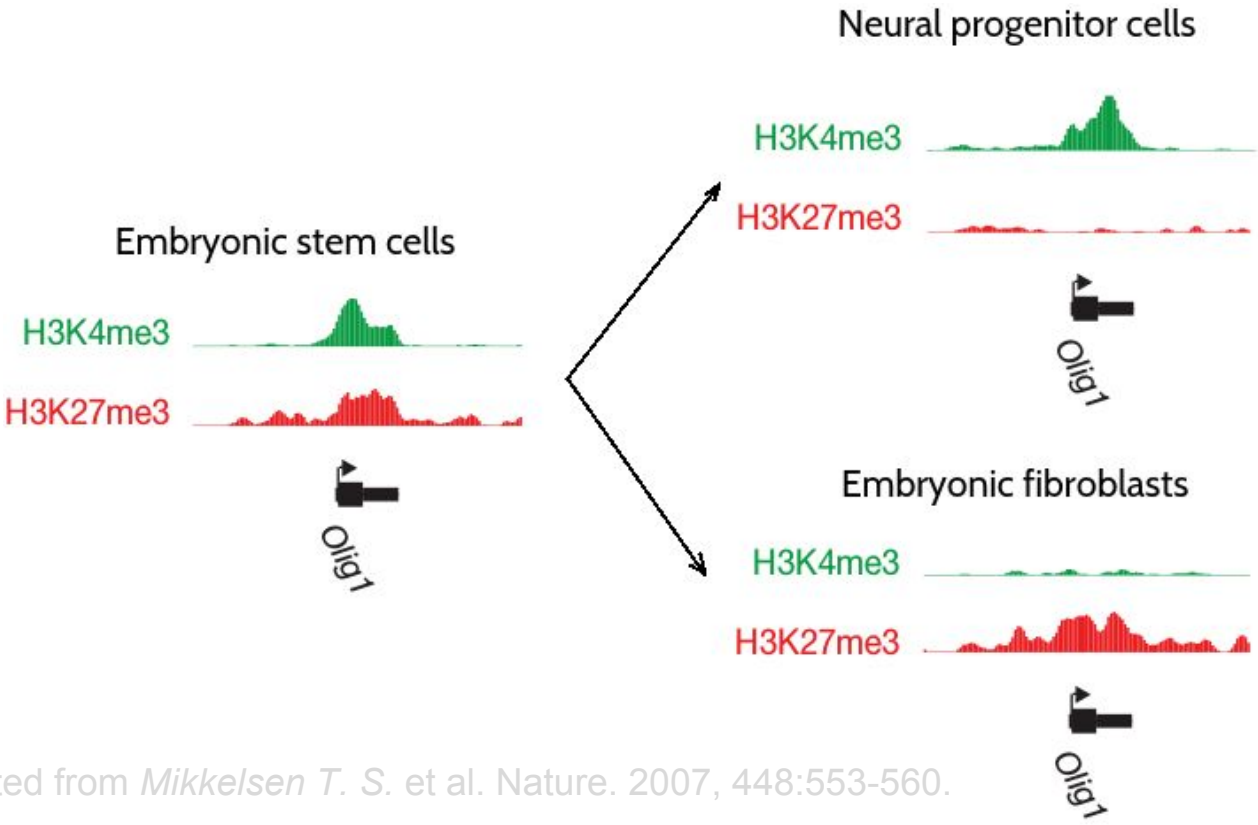
Also, weak borders are enriched by state 9 that contains H3K9me3 and H3K79me2 at the same time.

Outline

1. Introduction.
2. My master thesis in brief.
3. The current project.
4. Current results.
5. Perspective.

First hypothesis

Since H3K79me2 is an active modification, and H3K9me3 is a repressive modification, we considered the possibility that the combination **H3K79me2 + H3K9me3 marks a kind of bivalent promoters** (similar to the well known combination H3K4me3 + H3K27me3):



Adapted from *Mikkelsen T. S.* et al. Nature. 2007, 448:553-560.

Test of the first hypothesis

If H3K79me2 + H3K9me3 marks bivalent promoters similar to that of H3K4me3 + H3K27me3, then it should disappear in cells differentiated from embryonic stem cells.

Test of the first hypothesis

If H3K79me2 + H3K9me3 marks bivalent promoters similar to that of H3K4me3 + H3K27me3, then it should disappear in cells differentiated from embryonic stem cells.

Embryonic stem cells differentiate into three primary germ layers: **ectoderm**, **mesoderm**, and **endoderm**. They, in turn, differentiate into various cell types forming tissues.

Test of the first hypothesis

If H3K79me2 + H3K9me3 marks bivalent promoters similar to that of H3K4me3 + H3K27me3, then it should disappear in cells differentiated from embryonic stem cells.

Embryonic stem cells differentiate into three primary germ layers: **ectoderm**, **mesoderm**, and **endoderm**. They, in turn, differentiate into various cell types forming tissues.

So, we selected one cell line (with ChiP-seq and Hi-C data) from each of the primary germ layers:

- **HeLa-S3** (a clonal derivative of the parent HeLa line; cervix adenocarcinoma) from ectoderm;
- **HUVEC** (umbilical vein) from mesoderm;
- **IMR90** (fetal lung fibroblasts) from endoderm.

Test of the first hypothesis

If H3K79me2 + H3K9me3 marks bivalent promoters similar to that of H3K4me3 + H3K27me3, then it should disappear in cells differentiated from embryonic stem cells.

Embryonic stem cells differentiate into three primary germ layers: **ectoderm**, **mesoderm**, and **endoderm**. They, in turn, differentiate into various cell types forming tissues.

So, we selected one cell line (with ChiP-seq and Hi-C data) from each of the primary germ layers:

- **HeLa-S3** (a clonal derivative of the parent HeLa line; cervix adenocarcinoma) from ectoderm;
- **HUVEC** (umbilical vein) from mesoderm;
- **IMR90** (fetal lung fibroblasts) from endoderm.

Also we found data for murine embryonic stem cells (**mESC**) to compare them with human embryonic stem cells (**hESC**) regarding H3K79me2 + H3K9me3 combination.

Test of the first hypothesis

ChIP-seq data for hESC-H1, HeLa-S3, HUVEC, IMR90: H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac, H3K36me3, H3K79me2, H3K27me3, H3K9me3, RNA Pol II, H2A.Z, DNase, P300, CTCF. For mESC: the same except H3K4me2 and H2A.Z.

Test of the first hypothesis

ChIP-seq data for hESC-H1, HeLa-S3, HUVEC, IMR90: H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac, H3K36me3, H3K79me2, H3K27me3, H3K9me3, RNA Pol II, H2A.Z, DNase, P300, CTCF. For mESC: the same except H3K4me2 and H2A.Z.

ChIP-seq data processing:

1. Map single-end reads to hg19 human reference genome with bowtie2, process BAM files with SAMtools and BEDtools.

Test of the first hypothesis

ChIP-seq data for hESC-H1, HeLa-S3, HUVEC, IMR90: H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac, H3K36me3, H3K79me2, H3K27me3, H3K9me3, RNA Pol II, H2A.Z, DNase, P300, CTCF. For mESC: the same except H3K4me2 and H2A.Z.

ChIP-seq data processing:

1. Map single-end reads to hg19 human reference genome with bowtie2, process BAM files with SAMtools and BEDtools.
2. Produce TDF files for visualization purpose with IGVtools.

Test of the first hypothesis

ChIP-seq data for hESC-H1, HeLa-S3, HUVEC, IMR90: H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac, H3K36me3, H3K79me2, H3K27me3, H3K9me3, RNA Pol II, H2A.Z, DNase, P300, CTCF. For mESC: the same except H3K4me2 and H2A.Z.

ChIP-seq data processing:

1. Map single-end reads to hg19 human reference genome with bowtie2, process BAM files with SAMtools and BEDtools.
2. Produce TDF files for visualization purpose with IGVtools.
3. Call peaks with MACS2 for RNA Pol II, P300, and CTCF (assuming their peaks to be narrow) and with SICER for all other marks (whose peaks are likely to be wide).

Test of the first hypothesis

ChIP-seq data for hESC-H1, HeLa-S3, HUVEC, IMR90: H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac, H3K36me3, H3K79me2, H3K27me3, H3K9me3, RNA Pol II, H2A.Z, DNase, P300, CTCF. For mESC: the same except H3K4me2 and H2A.Z.

ChIP-seq data processing:

1. Map single-end reads to hg19 human reference genome with bowtie2, process BAM files with SAMtools and BEDtools.
2. Produce TDF files for visualization purpose with IGVtools.
3. Call peaks with MACS2 for RNA Pol II, P300, and CTCF (assuming their peaks to be narrow) and with SICER for all other marks (whose peaks are likely to be wide).
4. Filter peaks by their score with awk.

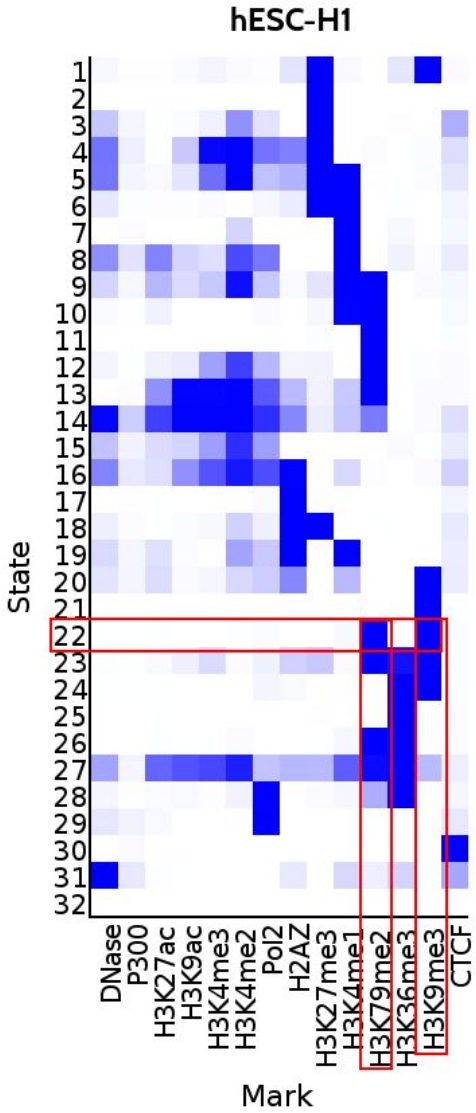
Test of the first hypothesis

ChIP-seq data for hESC-H1, HeLa-S3, HUVEC, IMR90: H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac, H3K36me3, H3K79me2, H3K27me3, H3K9me3, RNA Pol II, H2A.Z, DNase, P300, CTCF. For mESC: the same except H3K4me2 and H2A.Z.

ChIP-seq data processing:

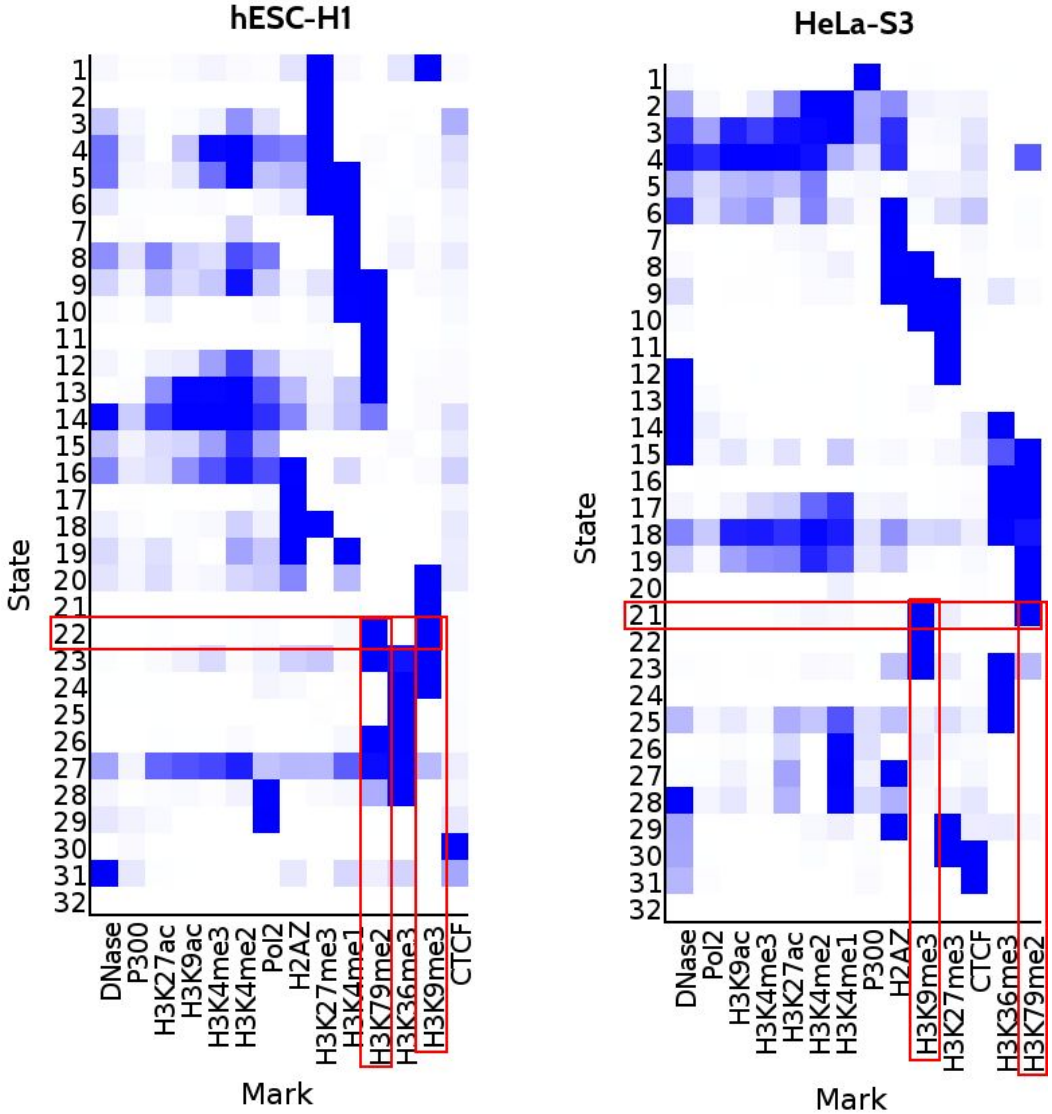
1. Map single-end reads to hg19 human reference genome with bowtie2, process BAM files with SAMtools and BEDtools.
2. Produce TDF files for visualization purpose with IGVtools.
3. Call peaks with MACS2 for RNA Pol II, P300, and CTCF (assuming their peaks to be narrow) and with SICER for all other marks (whose peaks are likely to be wide).
4. Filter peaks by their score with awk.
5. Find stable sets of marks (states) with ChromHMM.

Test of the first hypothesis



Combination H3K79me2 + H3K9me3 was reproduced in hESC-H1.

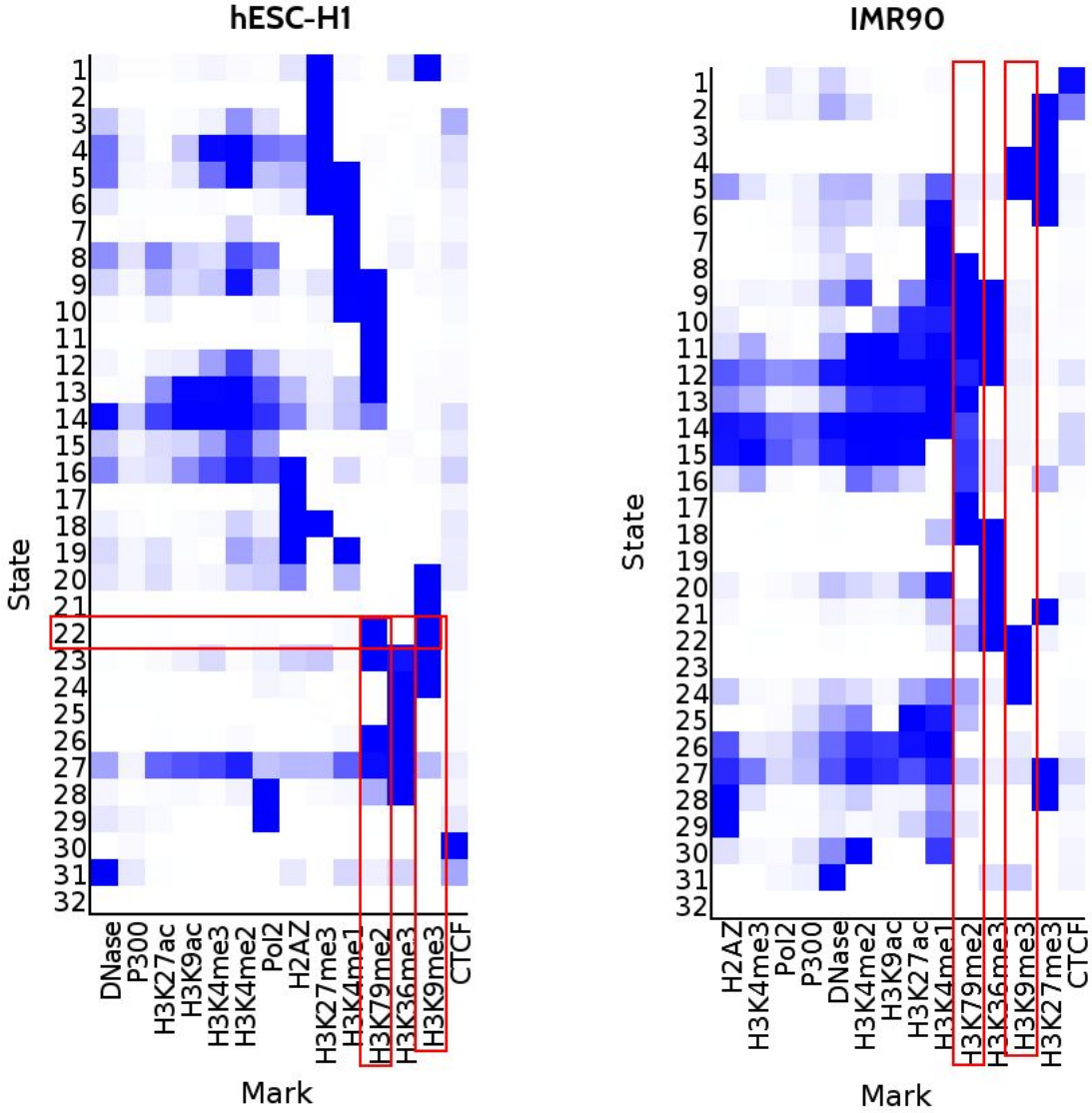
Test of the first hypothesis



Combination H3K79me2 + H3K9me3 was reproduced in hESC-H1.

But this combination also was reproduced in HeLa-S3.

Test of the first hypothesis

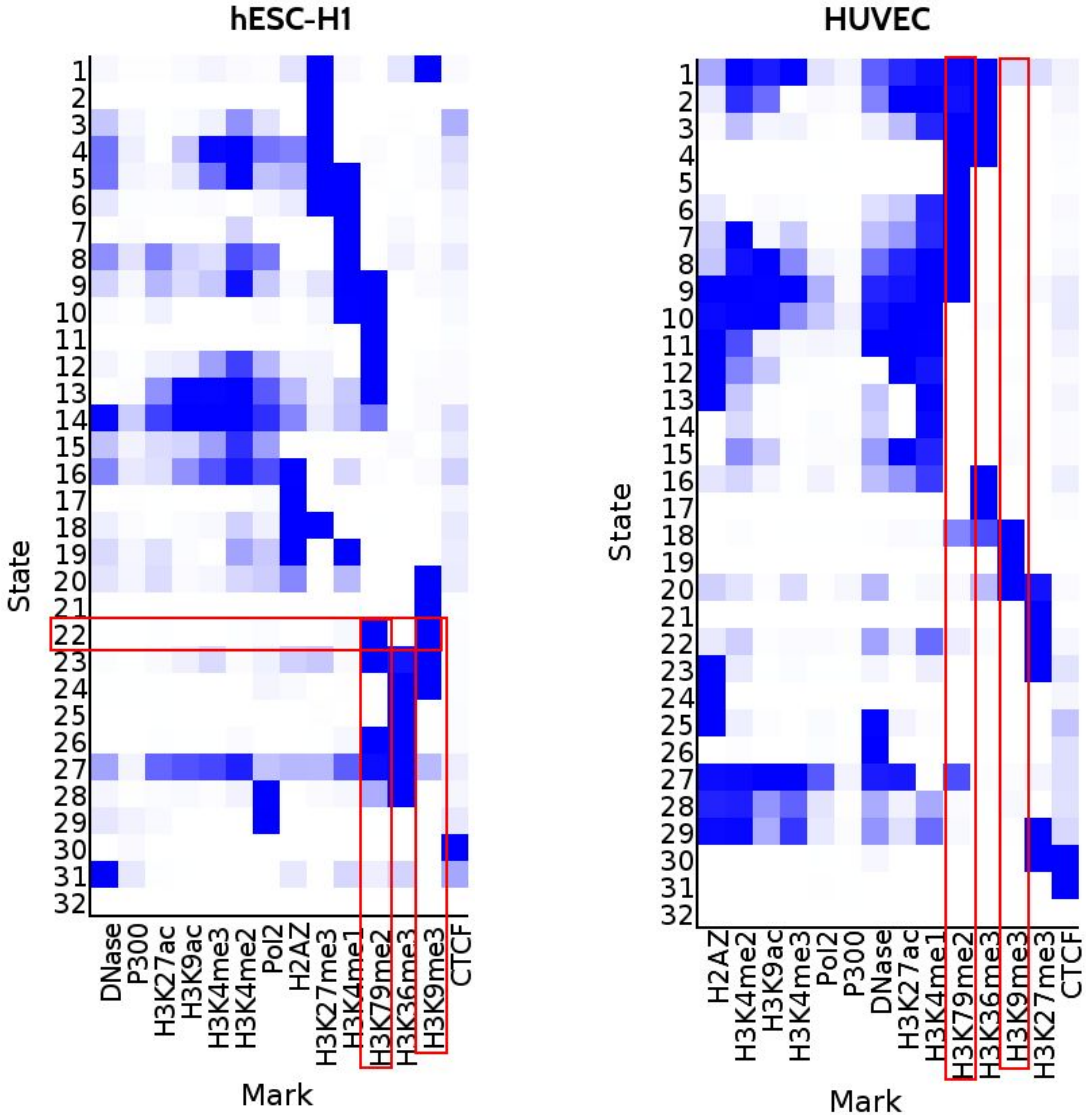


Combination H3K79me2 + H3K9me3 was reproduced in hESC-H1.

But this combination also was reproduced in HeLa-S3.

In IMR90, HUVEC, and mESC there are no clear states for H3K79me2 + H3K9me3 combination.

Test of the first hypothesis

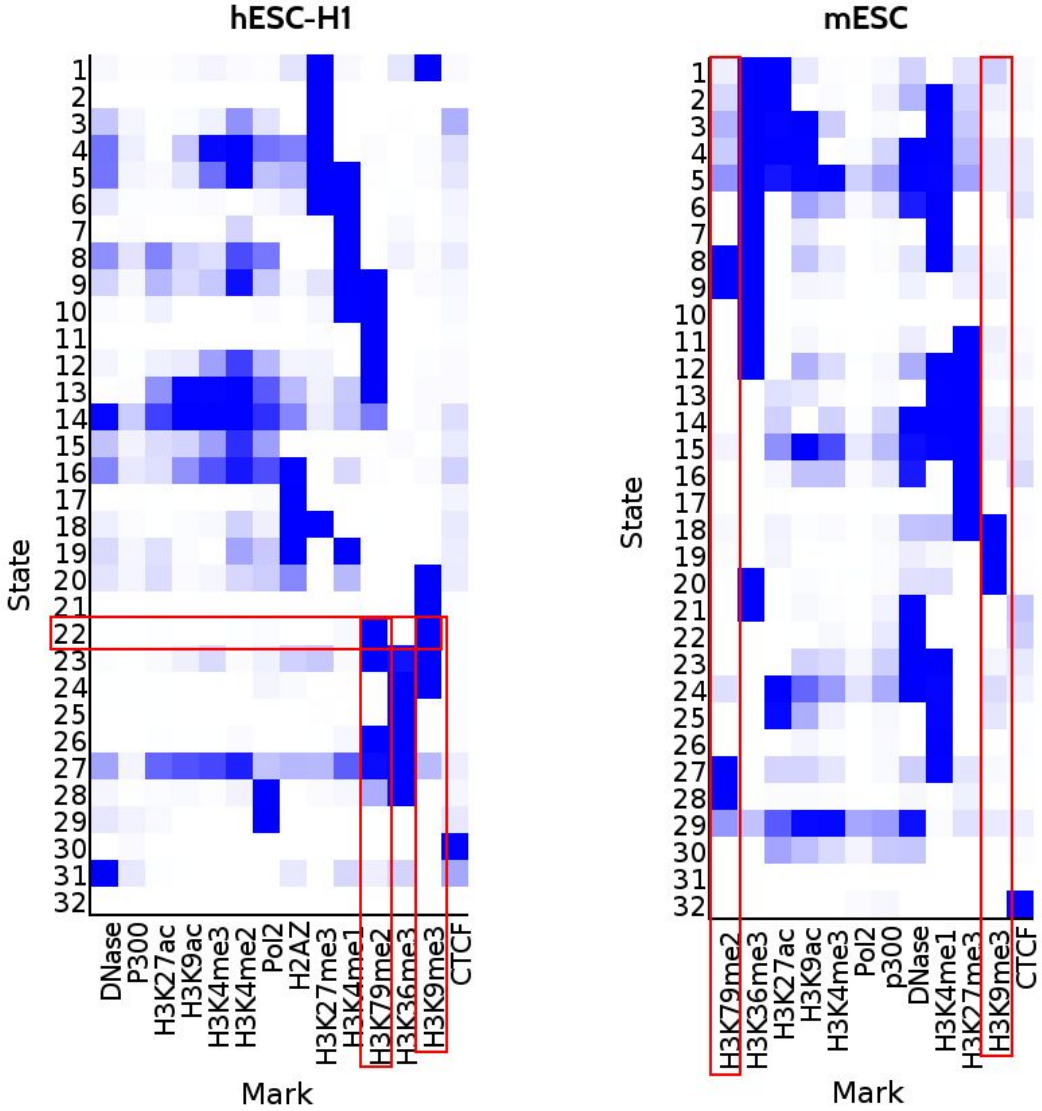


Combination H3K79me2 + H3K9me3 was reproduced in hESC-H1.

But this combination also was reproduced in HeLa-S3.

In IMR90, HUVEC, and mESC there are no clear states for H3K79me2 + H3K9me3 combination.

Test of the first hypothesis

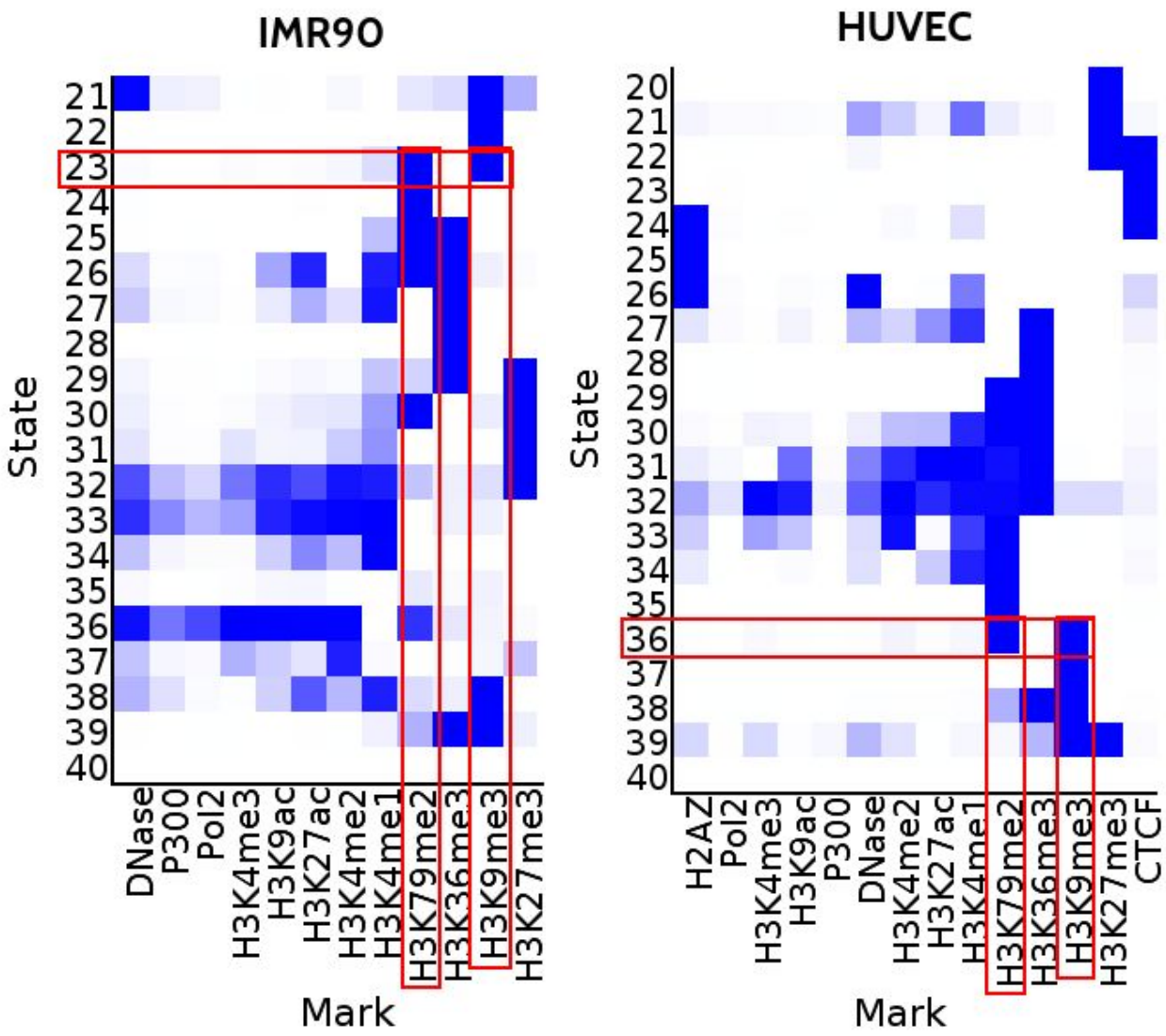


Combination H3K79me2 + H3K9me3 was reproduced in hESC-H1.

But this combination also was reproduced in HeLa-S3.

In IMR90, HUVEC, and mESC there are no clear states for H3K79me2 + H3K9me3 combination.

Test of the first hypothesis



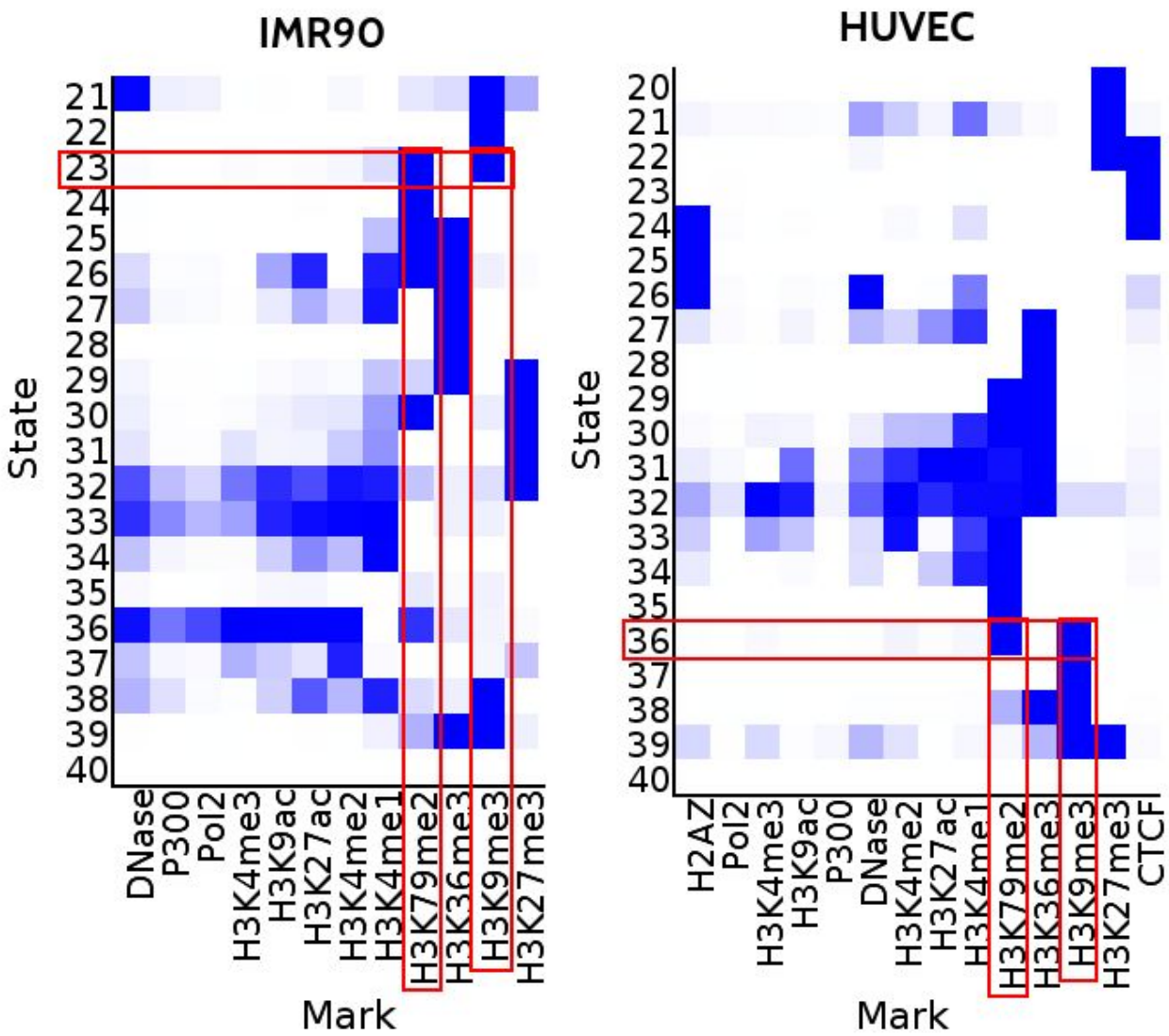
Combination H3K79me2 + H3K9me3 was reproduced in hESC-H1.

But this combination also was reproduced in HeLa-S3.

In IMR90, HUVEC, and mESC there are no clear states for H3K79me2 + H3K9me3 combination.

But it was reproduced for IMR90 and HUVEC in 40-state segmentation.

Test of the first hypothesis



Combination H3K79me2 + H3K9me3 was reproduced in hESC-H1.

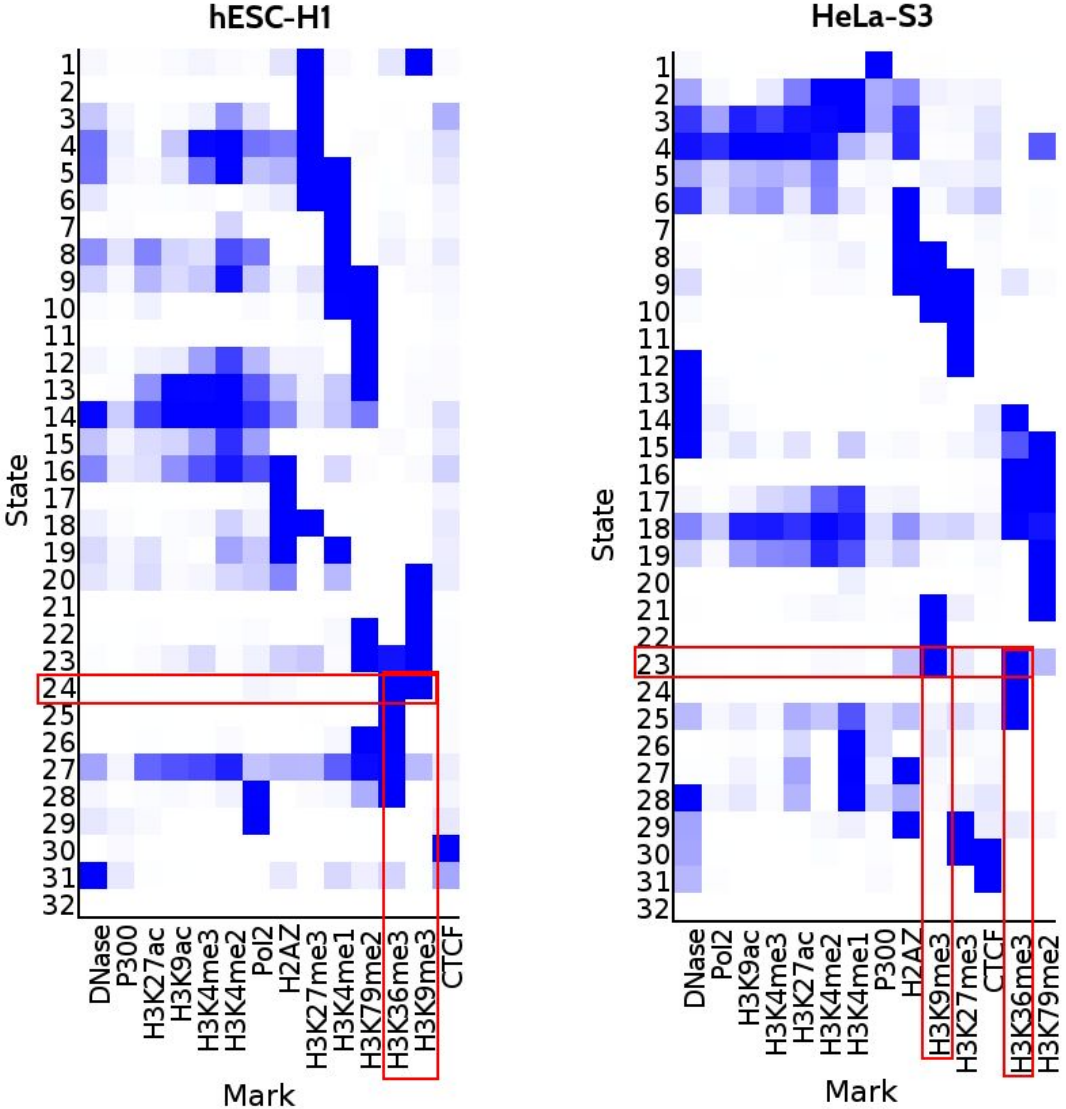
But this combination also was reproduced in HeLa-S3.

In IMR90, HUVEC, and mESC there are no clear states for H3K79me2 + H3K9me3 combination.

But it was reproduced for IMR90 and HUVEC in 40-state segmentation.

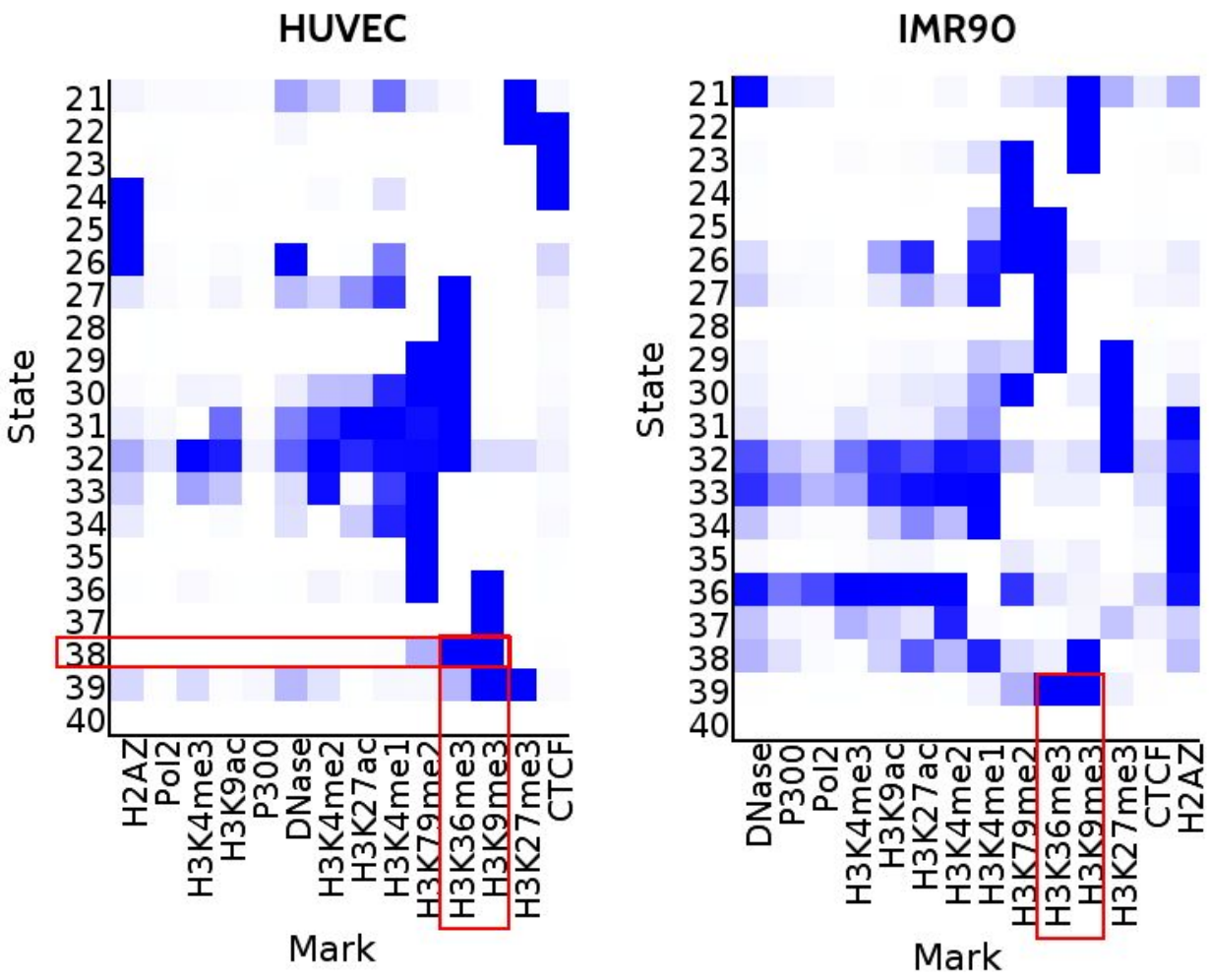
Thus, it can't be a mechanism like that of bivalent promoters.

Two observations



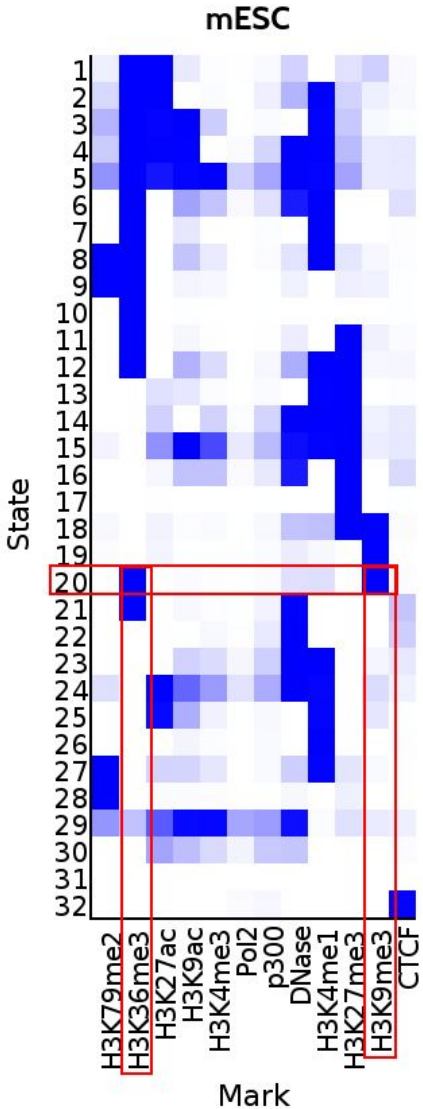
(1) A combination **H3K36me3 + H3K9me3** is presented as a clear separate state in all considered cell lines: hESC-H1, HeLa-S3, HUVEC, IMR90, and mESC.

Two observations



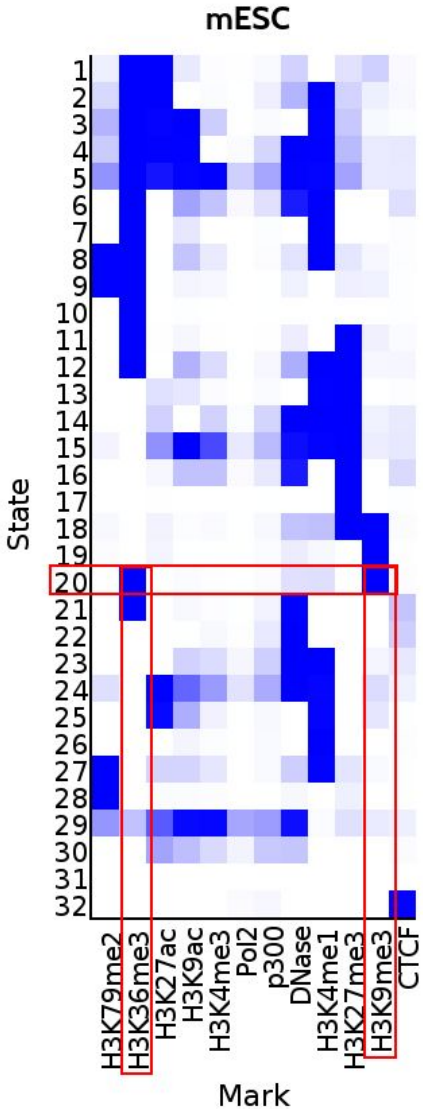
(1) A combination **H3K36me3 + H3K9me3** is presented as a clear separate state in all considered cell lines: hESC-H1, HeLa-S3, HUVEC, IMR90, and mESC.

Two observations



(1) A combination **H3K36me3 + H3K9me3** is presented as a clear separate state in all considered cell lines: hESC-H1, HeLa-S3, HUVEC, IMR90, and mESC.

Two observations



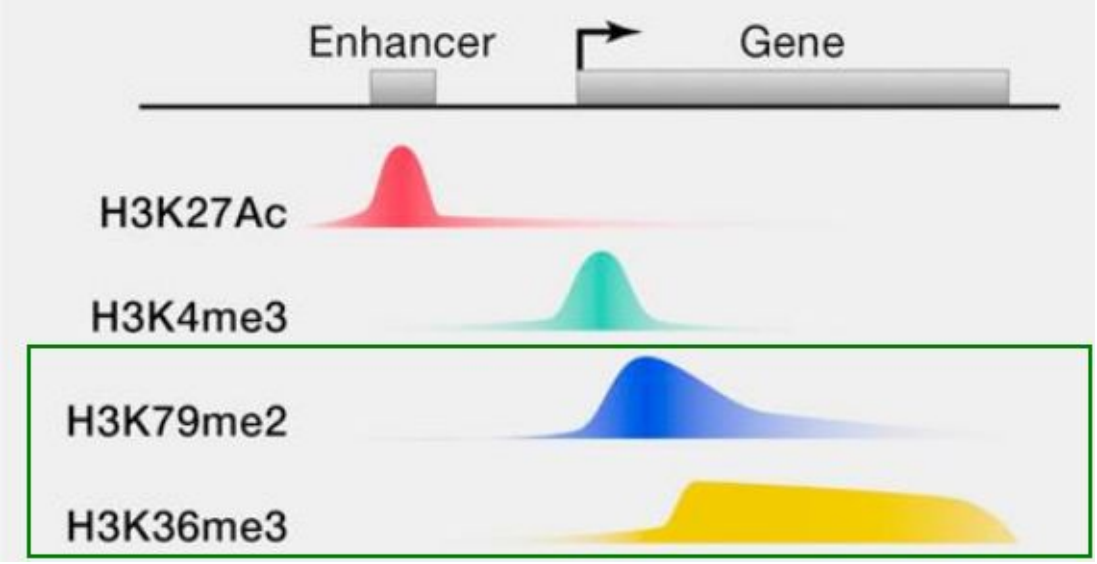
(1) A combination **H3K36me3 + H3K9me3** is presented as a clear separate state in all considered cell lines: hESC-H1, HeLa-S3, HUVEC, IMR90, and mESC.

(2) Combinations H3K79me2 + H3K9me3 and H3K36me3 + H3K9me3 are situated predominantly within genes.

Two observations

(2) Combinations H3K79me2 + H3K9me3 and H3K36me3 + H3K9me3 are situated predominantly within genes.

But H3K79me2 and H3K36me3 are well known as marks of transcribed genes:



Adapted from *Lee T. I.* and *Young R. A.* Cell. 2013, 152(6):1237-1251.

Two observations

(2) Combinations H3K79me2 + H3K9me3 and H3K36me3 + H3K9me3 are situated predominantly within genes.

But H3K79me2 and H3K36me3 are well known as marks of transcribed genes.

And H3K9me3 is one of the typical marks of heterochromatin.

Two observations

(2) Combinations H3K79me2 + H3K9me3 and H3K36me3 + H3K9me3 are situated predominantly within genes.

But H3K79me2 and H3K36me3 are well known as marks of transcribed genes.

And H3K9me3 is one of the typical marks of heterochromatin.

Thus, **the most interesting thing is** not an existence of H3K79me2 + H3K9me3 and H3K36me3 + H3K9me3 combinations themselves, but **the presence of H3K9me3 marks in supposedly transcribed genes!**

Study of H3K9me3-enriched genes

First, we checked whether genes with H3K36me3, H3K79me2, and H3K9me3 are really actively transcribed.

Study of H3K9me3-enriched genes

First, we checked whether genes with H3K36me3, H3K79me2, and H3K9me3 are really actively transcribed.

RNA-seq data processing for hESC-H1, HeLa-S3, HUVEC, IMR90, mESC:

1. Map reads to hg19 reference genome (for human) or mm10 reference genome (for mouse) with STAR using GENCODE human annotation.

Study of H3K9me3-enriched genes

First, we checked whether genes with H3K36me3, H3K79me2, and H3K9me3 are really actively transcribed.

RNA-seq data processing for hESC-H1, HeLa-S3, HUVEC, IMR90, mESC:

1. Map reads to hg19 reference genome (for human) or mm10 reference genome (for mouse) with STAR using GENCODE annotation.
2. Produce TDF files for visualization purpose with IGVtools.

Study of H3K9me3-enriched genes

First, we checked whether genes with H3K36me3, H3K79me2, and H3K9me3 are really actively transcribed.

RNA-seq data processing for hESC-H1, HeLa-S3, HUVEC, IMR90, mESC:

1. Map reads to hg19 reference genome (for human) or mm10 reference genome (for mouse) with STAR using GENCODE human annotation.
2. Produce TDF files for visualization purpose with IGVtools.
3. Quantify reads per gene and per transcript with RSEM. Read count, TPM, and FPKM are calculated.

Study of H3K9me3-enriched genes

First, we checked whether genes enriched by H3K36me3, H3K79me2, and H3K9me3 are really actively transcribed.

RNA-seq data processing for hESC-H1, HeLa-S3, HUVEC, IMR90, mESC:

1. Map reads to hg19 reference genome (for human) or mm10 reference genome (for mouse) with STAR using GENCODE human annotation.
2. Produce TDF files for visualization purpose with IGVtools.
3. Quantify reads per gene and per transcript with RSEM. Read count, TPM, and FPKM are calculated.
4. Select genes with TPM >= 5 (they are considered to be transcribed).

Study of H3K9me3-enriched genes

First, we checked whether genes enriched by H3K36me3, H3K79me2, and H3K9me3 are really actively transcribed.

To be sure that a gene is really actively transcribed, we selected genes that overlap with H3K36me3 and H3K79me2 peaks from genes with TPM >= 5.

Study of H3K9me3-enriched genes

First, we checked whether genes enriched by H3K36me3, H3K79me2, and H3K9me3 are really actively transcribed.

To be sure that a gene is really actively transcribed, we selected genes that overlap with H3K36me3 and H3K79me2 peaks from genes with TPM >= 5.

Finally, we took actively transcribed genes that overlapped with H3K9me3 peaks. **And these sets of genes were not empty.**

Study of H3K9me3-enriched genes

Finally, we took actively transcribed genes that overlapped with H3K9me3 peaks. **And these sets of genes were not empty.**

Gene counts within these sets are as follows:

hESC-H1	635
HeLa-S3	1273
HUVEC	110
IMR90	1686
mESC	158

Study of H3K9me3-enriched genes

By visual inspection of actively transcribed genes overlapped with H3K9me3 peaks we found the following frequent cases:

Study of H3K9me3-enriched genes

By visual inspection of actively transcribed genes overlapped with H3K9me3 peaks we found the following frequent cases:

1. H3K9me3 peak is situated inside an intron.

Study of H3K9me3-enriched genes

By visual inspection of actively transcribed genes overlapped with H3K9me3 peaks we found the following frequent cases:

1. H3K9me3 peak is situated inside an intron.
2. HeK9me3 peak overlaps with several introns and exons but is situated within gene body.

Study of H3K9me3-enriched genes

By visual inspection of actively transcribed genes overlapped with H3K9me3 peaks we found the following frequent cases:

1. H3K9me3 peak is situated inside an intron.
2. HeK9me3 peak overlaps with several introns and exons but is situated within gene body.
3. H3K9me3 peak overlaps with the last long exon (it's typical for zinc-fingers).

Study of H3K9me3-enriched genes

By visual inspection of actively transcribed genes overlapped with H3K9me3 peaks we found the following frequent cases:

1. H3K9me3 peak is situated inside an intron.
2. HeK9me3 peak overlaps with several introns and exons but is situated within gene body.
3. H3K9me3 peak overlaps with the last long exon (it's typical for zinc-fingers).

To study this cases more formally, we produced classification tables.

Classification tables

Such table was produced for each cell line. It has the following columns:

1. `chr` - chromosome name (chr1, chr2, …, chr22, chrX, chrY);

Classification tables

Such table was produced for each cell line. It has the following columns:

1. `chr` - chromosome name (chr1, chr2, …, chr22, chrX, chrY);
2. `gene_id` - ID of the gene in GENCODE annotation;
3. `gene_name` - the name of the gene in GENCODE annotation;

Classification tables

Such table was produced for each cell line. It has the following columns:

1. `chr` - chromosome name (chr1, chr2, …, chr22, chrX, chrY);
2. `gene_id` - ID of the gene in GENCODE annotation;
3. `gene_name` - the name of the gene in GENCODE annotation;
4. `tpm` - level of expression by TPM;
5. `fpkm` - level of expression by FPKM;

Classification tables

Such table was produced for each cell line. It has the following columns:

1. `chr` - chromosome name (chr1, chr2, …, chr22, chrX, chrY);
2. `gene_id` - ID of the gene in GENCODE annotation;
3. `gene_name` - the name of the gene in GENCODE annotation;
4. `tpm` - level of expression by TPM;
5. `fpkm` - level of expression by FPKM;
6. `left_H3K9me3_density` - total length (in bp) of H3K9me3 peaks in 500 kbp vicinity to the left of the gene divided by 500,000;
7. `right_H3K9me3_density` - the same to the right of the gene;

Classification tables

Such table was produced for each cell line. It has the following columns:

1. `chr` - chromosome name (chr1, chr2, …, chr22, chrX, chrY);
2. `gene_id` - ID of the gene in GENCODE annotation;
3. `gene_name` - the name of the gene in GENCODE annotation;
4. `tpm` - level of expression by TPM;
5. `fpkm` - level of expression by FPKM;
6. `left_H3K9me3_density` - total length (in bp) of H3K9me3 peaks in 500 kbp vicinity to the left of the gene divided by 500,000;
7. `right_H3K9me3_density` - the same to the right of the gene;
8. `dist_to_cent` - distance from the middle of the gene to the centromere (**-1** for murine genes, as its chromosomes have no typical centromeres);

Classification tables

Such table was produced for each cell line. It has the following columns:

1. `chr` - chromosome name (chr1, chr2, …, chr22, chrX, chrY);
2. `gene_id` - ID of the gene in GENCODE annotation;
3. `gene_name` - the name of the gene in GENCODE annotation;
4. `tpm` - level of expression by TPM;
5. `fpkm` - level of expression by FPKM;
6. `left_H3K9me3_density` - total length (in bp) of H3K9me3 peaks in 500 kbp vicinity to the left of the gene divided by 500,000;
7. `right_H3K9me3_density` - the same to the right of the gene;
8. `dist_to_cent` - distance from the middle of the gene to the centromere (**-1** for murine genes, as its chromosomes have no typical centromeres);
9. `H3K9me3_peak_start` - start coordinate of H3K9me3 peak [start, stop);
10. `H3K9me3_peak_stop` - stop coordinate of H3K9me3 peak [start, stop);

Classification tables

Such table was produced for each cell line. It has the following columns:

1. `chr` - chromosome name (chr1, chr2, …, chr22, chrX, chrY);
2. `gene_id` - ID of the gene in GENCODE annotation;
3. `gene_name` - the name of the gene in GENCODE annotation;
4. `tpm` - level of expression by TPM;
5. `fpkm` - level of expression by FPKM;
6. `left_H3K9me3_density` - total length (in bp) of H3K9me3 peaks in 500 kbp vicinity to the left of the gene divided by 500,000;
7. `right_H3K9me3_density` - the same to the right of the gene;
8. `dist_to_cent` - distance from the middle of the gene to the centromere (**-1** for murine genes, as its chromosomes have no typical centromeres);
9. `H3K9me3_peak_start` - start coordinate of H3K9me3 peak [start, stop);
10. `H3K9me3_peak_stop` - stop coordinate of H3K9me3 peak [start, stop);
11. `if_overlaps_exon` - **1** if H3K9me3 peak overlaps with an exon, **0** otherwise;
12. `max_exon_overlapped` - length of the longest exon overlapped with the peak;

Classification tables

Such table was produced for each cell line. It has the following columns:

1. `chr` - chromosome name (chr1, chr2, …, chr22, chrX, chrY);
2. `gene_id` - ID of the gene in GENCODE annotation;
3. `gene_name` - the name of the gene in GENCODE annotation;
4. `tpm` - level of expression by TPM;
5. `fpkm` - level of expression by FPKM;
6. `left_H3K9me3_density` - total length (in bp) of H3K9me3 peaks in 500 kbp vicinity to the left of the gene divided by 500,000;
7. `right_H3K9me3_density` - the same to the right of the gene;
8. `dist_to_cent` - distance from the middle of the gene to the centromere (**-1** for murine genes, as its chromosomes have no typical centromeres);
9. `H3K9me3_peak_start` - start coordinate of H3K9me3 peak [start, stop);
10. `H3K9me3_peak_stop` - stop coordinate of H3K9me3 peak [start, stop);
11. `if_overlaps_exon` - **1** if H3K9me3 peak overlaps with an exon, **0** otherwise;
12. `max_exon_overlapped` - length of the longest exon overlapped with the peak;
13. `if_overlaps_intron` - **1** if H3K9me3 peak overlaps with an intron, **0** otherwise;
14. `max_intron_overlapped` - length of the longest intron overlapped with the peak.

Active genes enriched by H3K9me3

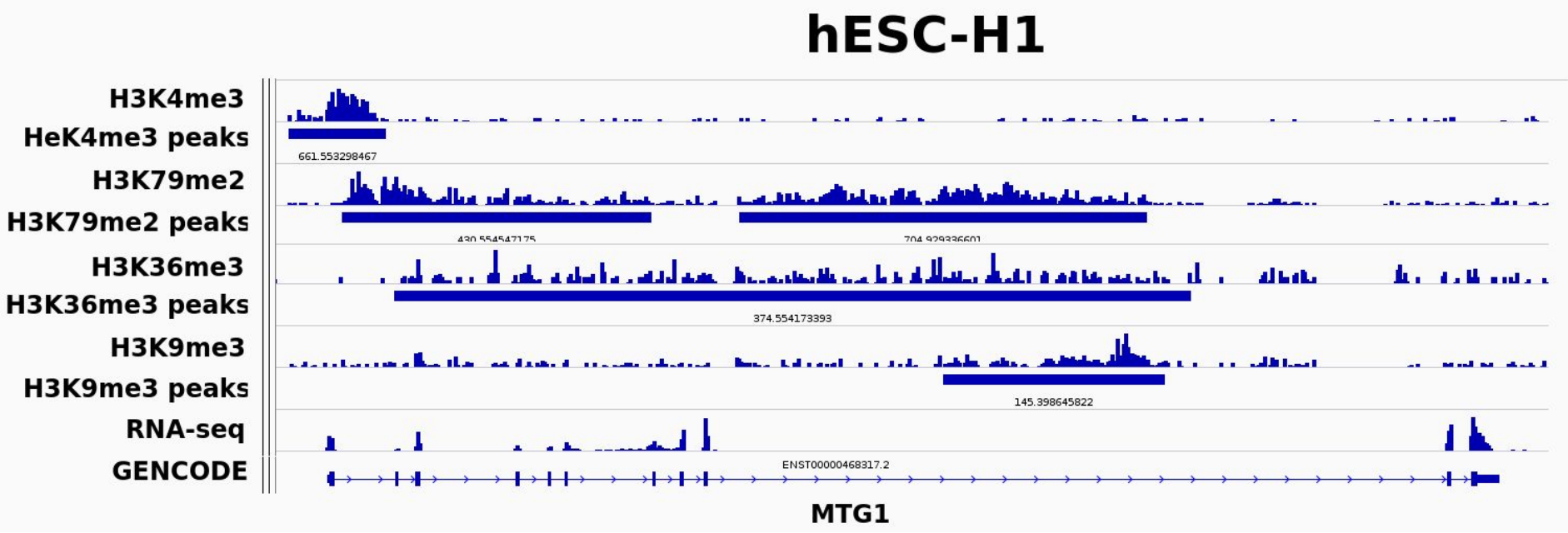
Let’s find transcribed genes with H3K9me3 peaks strictly within introns:

```
if_overlaps_exon == 0 && if_overlaps_intron == 1
```

Active genes enriched by H3K9me3

Let's find transcribed genes with H3K9me3 peaks strictly within introns:

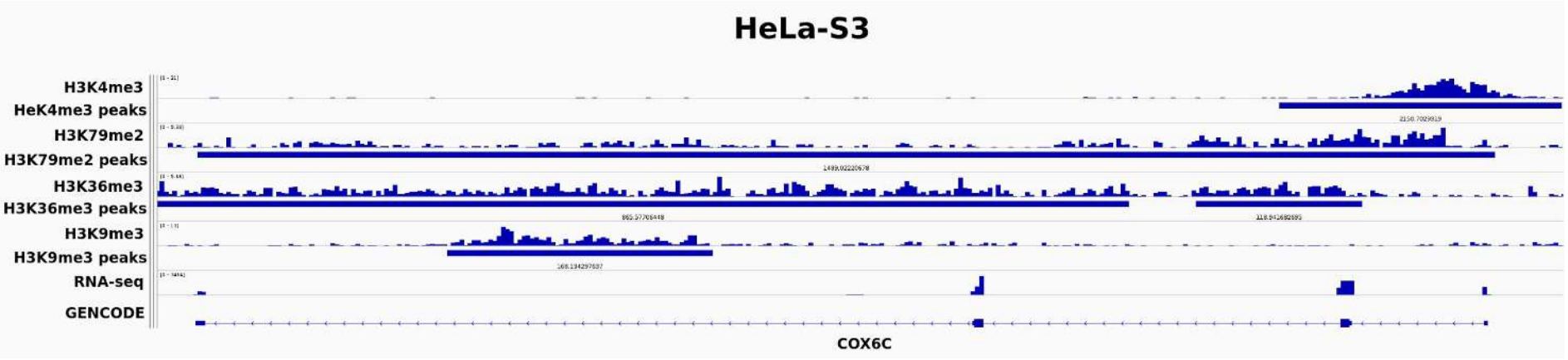
```
if_overlaps_exon == 0 && if_overlaps_intron == 1
```



Active genes enriched by H3K9me3

Let's find transcribed genes with H3K9me3 peaks strictly within introns:

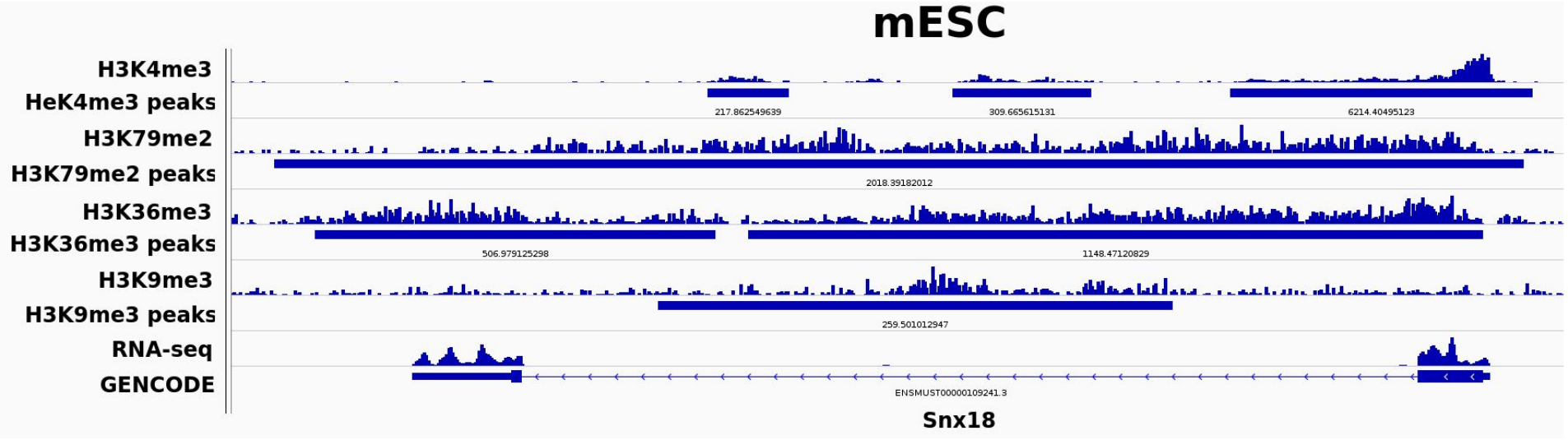
```
if_overlaps_exon == 0 && if_overlaps_intron == 1
```



Active genes enriched by H3K9me3

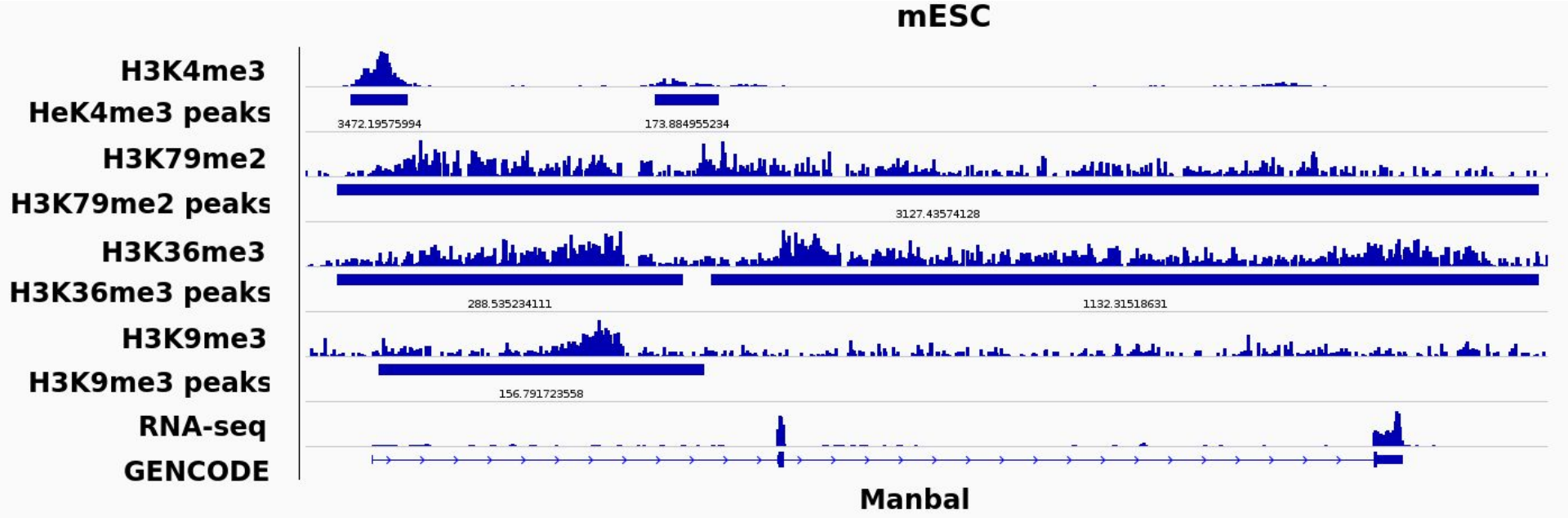
Let's find transcribed genes with H3K9me3 peaks strictly within introns:

```
if_overlaps_exon == 0 && if_overlaps_intron == 1
```



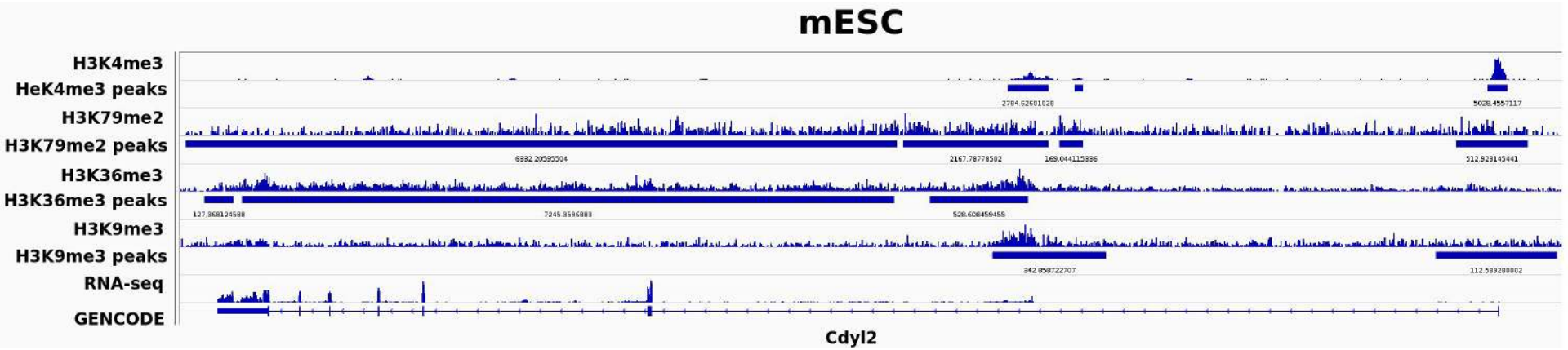
Active genes enriched by H3K9me3

Also, for some genes we could see the following:



Active genes enriched by H3K9me3

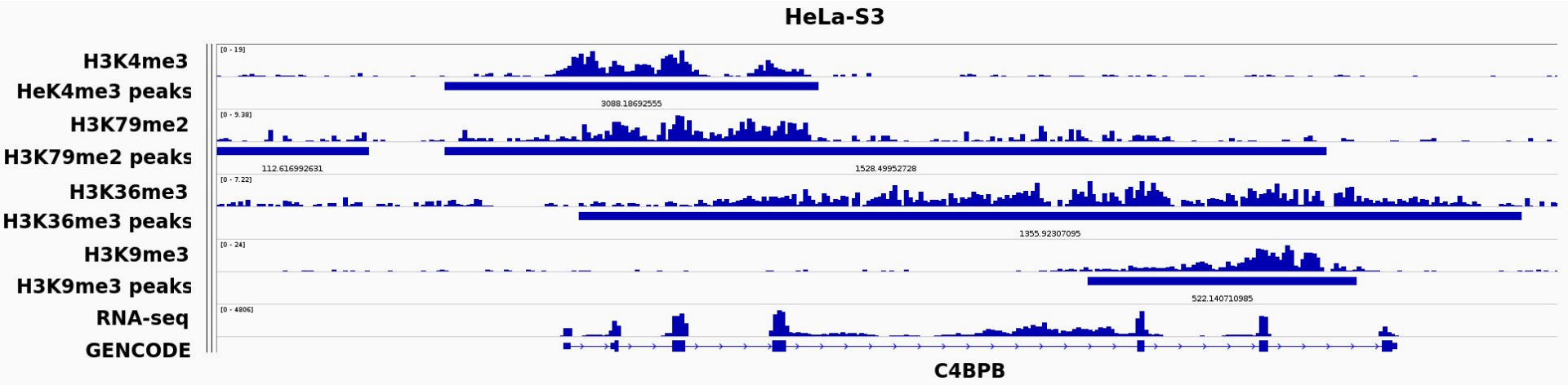
Also, for some genes we could see the following:



So, we suggest that such H3K9me3 peaks may mark alternative promoters.

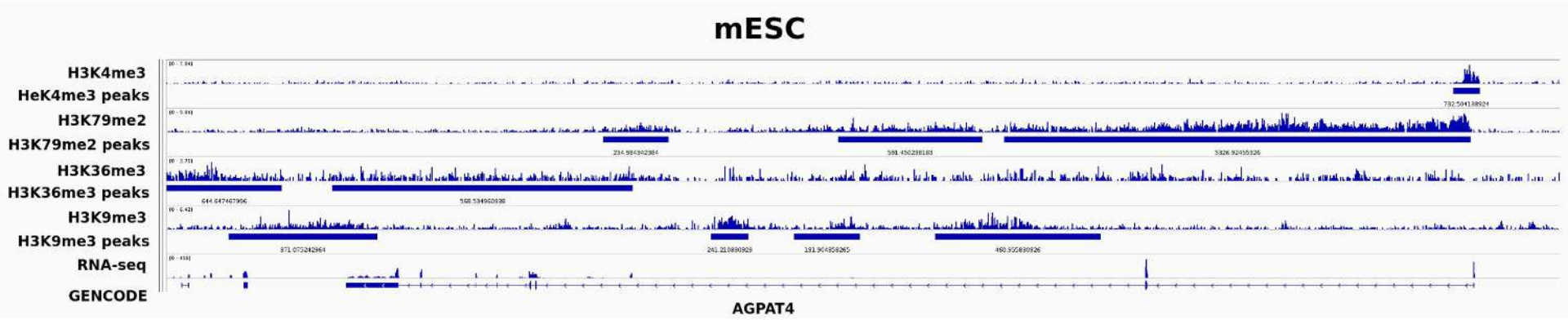
Active genes enriched by H3K9me3

Additionally, we can observe through visual inspection H3K9me3 peaks that overlap exons and introns but are situated within active genes:



Active genes enriched by H3K9me3

Additionally, we can observe through visual inspection H3K9me3 peaks that overlap several exons and introns but are situated within active genes or overlap with active genes from intergenic regions:

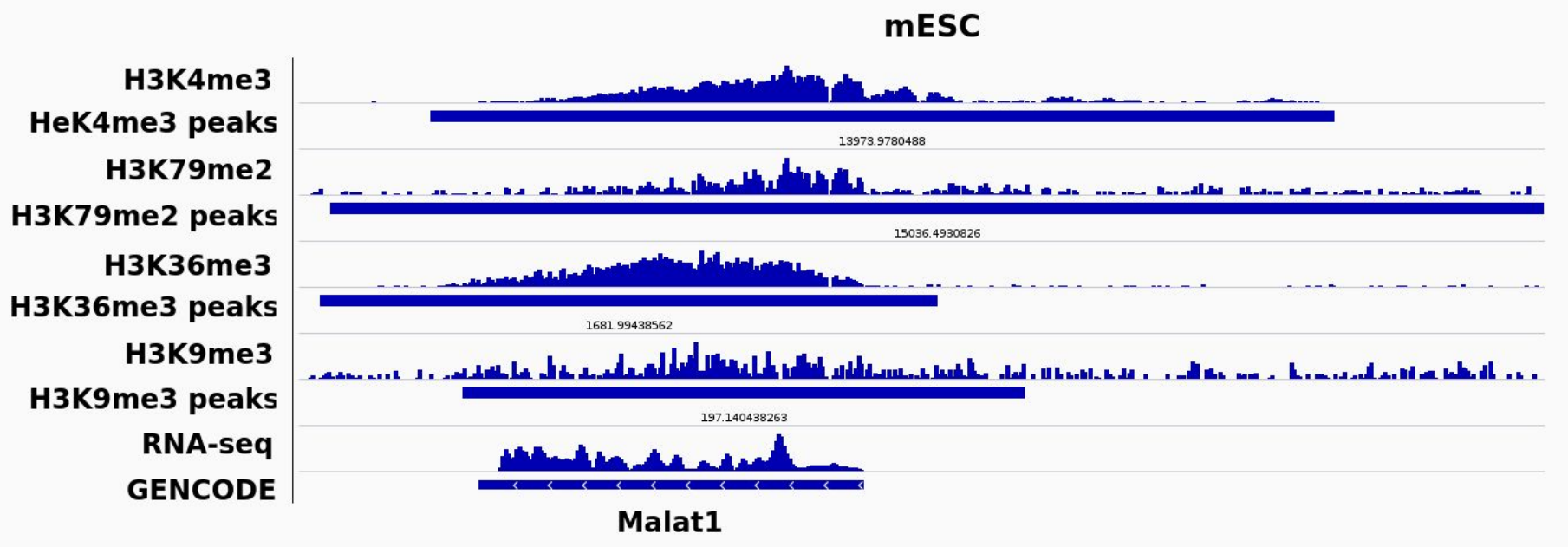


Active genes enriched by H3K9me3

Let's select active genes that overlap with HeK9me3 peaks only with their exons:

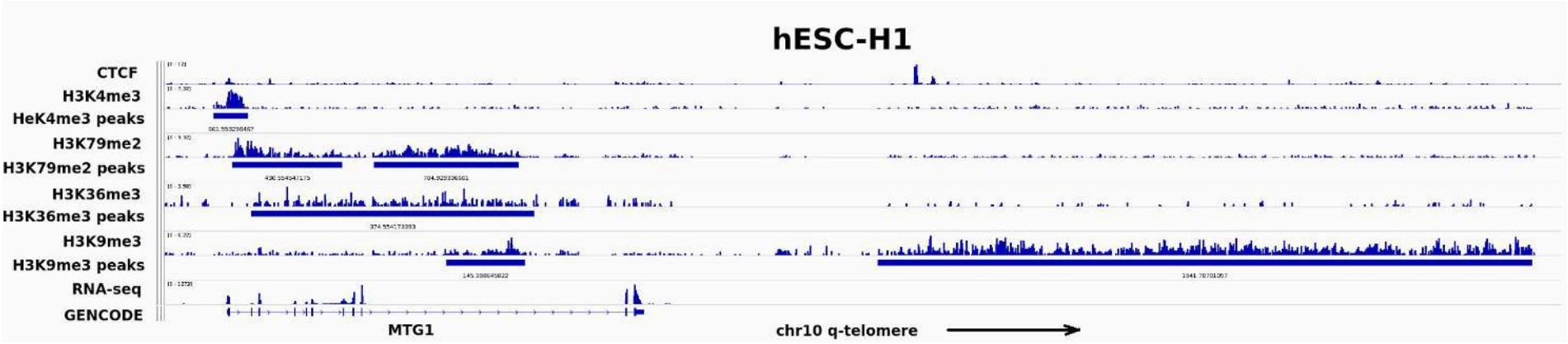
```
if_overlaps_exon == 1 && if_overlaps_intron == 0
```

Among them we could see genes with a long last exons overlapped with H3K9me3, but not always:



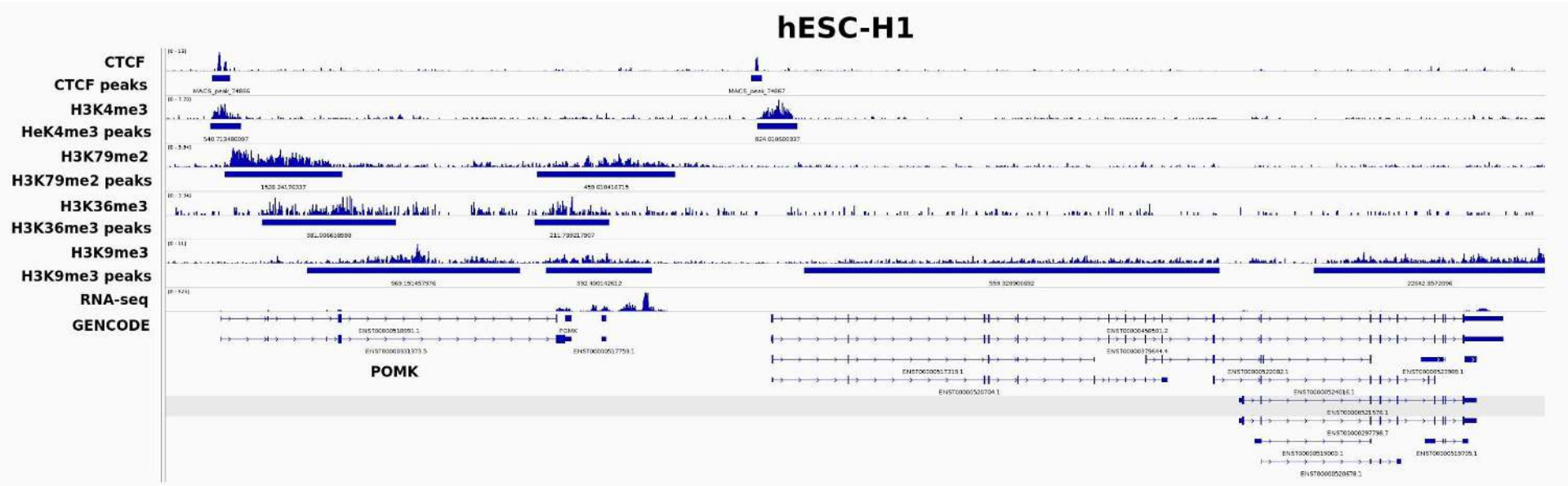
Active genes enriched by H3K9me3

Finally, we were able to observe some active genes overlapped with H3K9me3 peaks and situated at the edge of long H3K9me3-enriched regions (near telomeres and centromeres, as well as within chromosome arms):



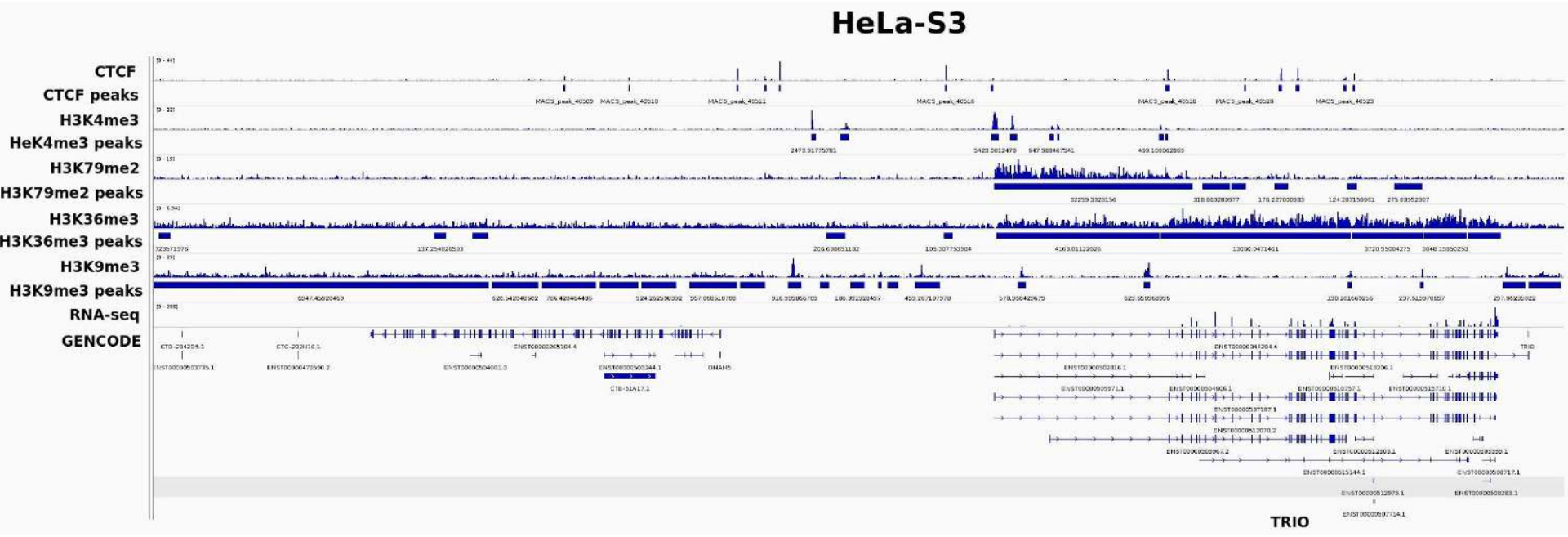
Active genes enriched by H3K9me3

Finally, we were able to observe some active genes overlapped with H3K9me3 peaks and situated at the edge of long H3K9me3-enriched regions (near telomeres and centromeres, as well as within chromosome arms):



Active genes enriched by H3K9me3

Finally, we were able to observe some active genes overlapped with H3K9me3 peaks and situated at the edge of long H3K9me3-enriched regions (near telomeres and centromeres, as well as within chromosome arms):



Enrichments

Next we decided to calculate enrichments for H3K9me3 + H3K79me2/H3K36me3 combinations (approximately marking H3K9me3 regions in active genes) in the following genome elements:

1. Genes, exons, promoters, TSS, TES.
2. CpG-islands, lamina-associated domains.
3. Origins of replication.
4. High-occupancy target (HOT) regions, super-enhancers.
5. Known repeat families.
6. Strong and weak TAD borders 1 kbp, 2 kbp, 10 kbp, 100 kbp wide.

TAD calling

We took Hi-C data for hESC-H1, HeLa-S3, HUVEC, IMR90, and mESC and processed them using TADbit:

1. Mapped paired-end reads to hg19 or mm10 reference genome with GEM-mapper.
2. Produced contact matrices.
3. Found TADs (back-to-back segmentation) and scored borders between them (from 1 that is the weakest border possible to 10 that is the strongest one).
4. Generated BED files with borders of various width: 1 kbp, 2 kbp, 10 kbp, 100 kbp.

TAD calling

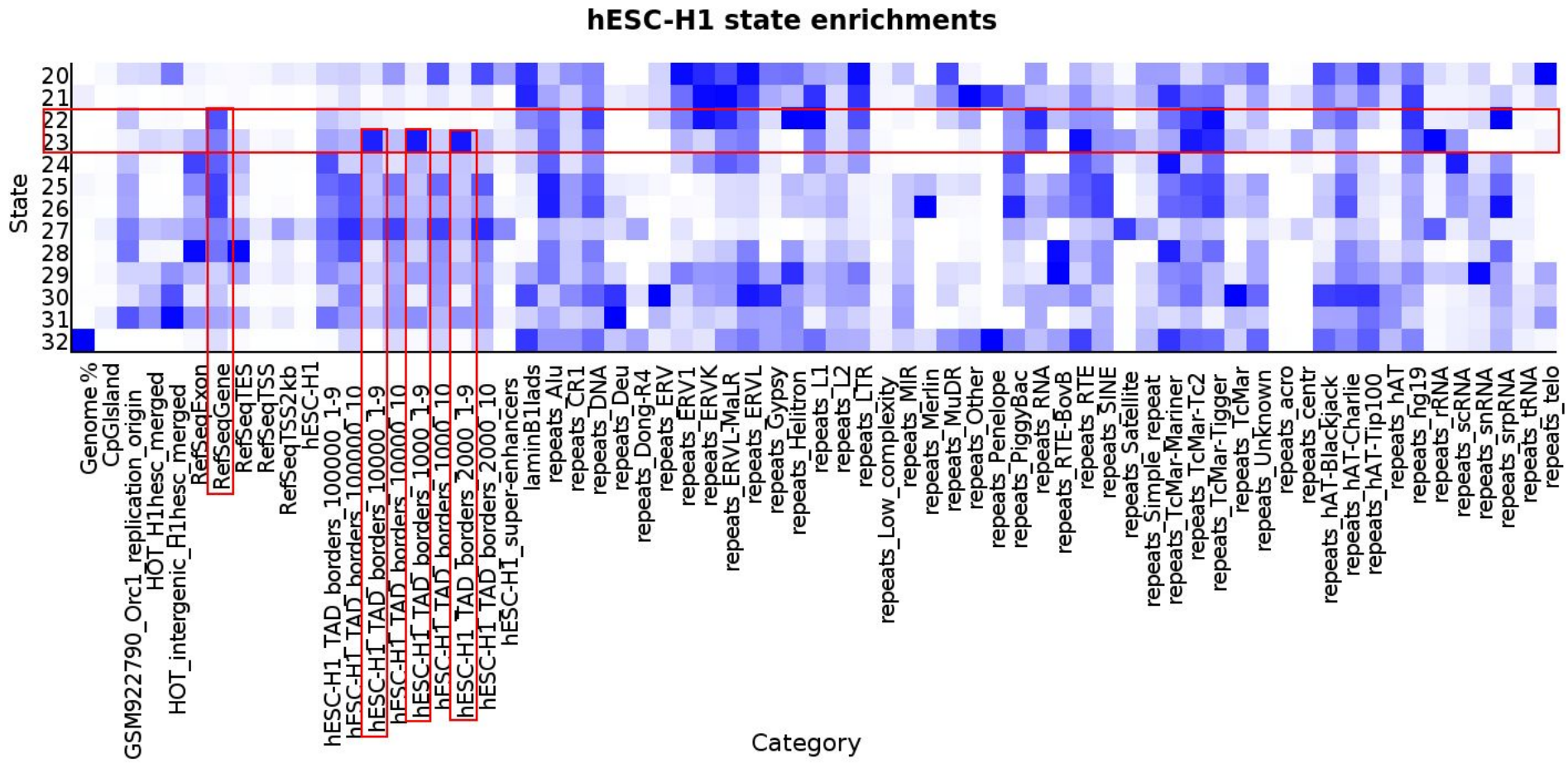
We took Hi-C data for hESC-H1, HeLa-S3, HUVEC, IMR90, and mESC and processed them using TADbit:

1. Mapped paired-end reads to hg19 or mm10 reference genome with GEM-mapper.
2. Produced contact matrices.
3. Found TADs (back-to-back segmentation) and scored borders between them (from 1 that is the weakest border possible to 10 that is the strongest one).
4. Generated BED files with borders of various width: 1 kbp, 2 kbp, 10 kbp, 100 kbp.

We obtained 2627 borders for hESC-H1, 2014 borders for HeLa-S3, 2353 borders for HUVEC, 2244 borders for IMR90, and 1922 borders for mESC.

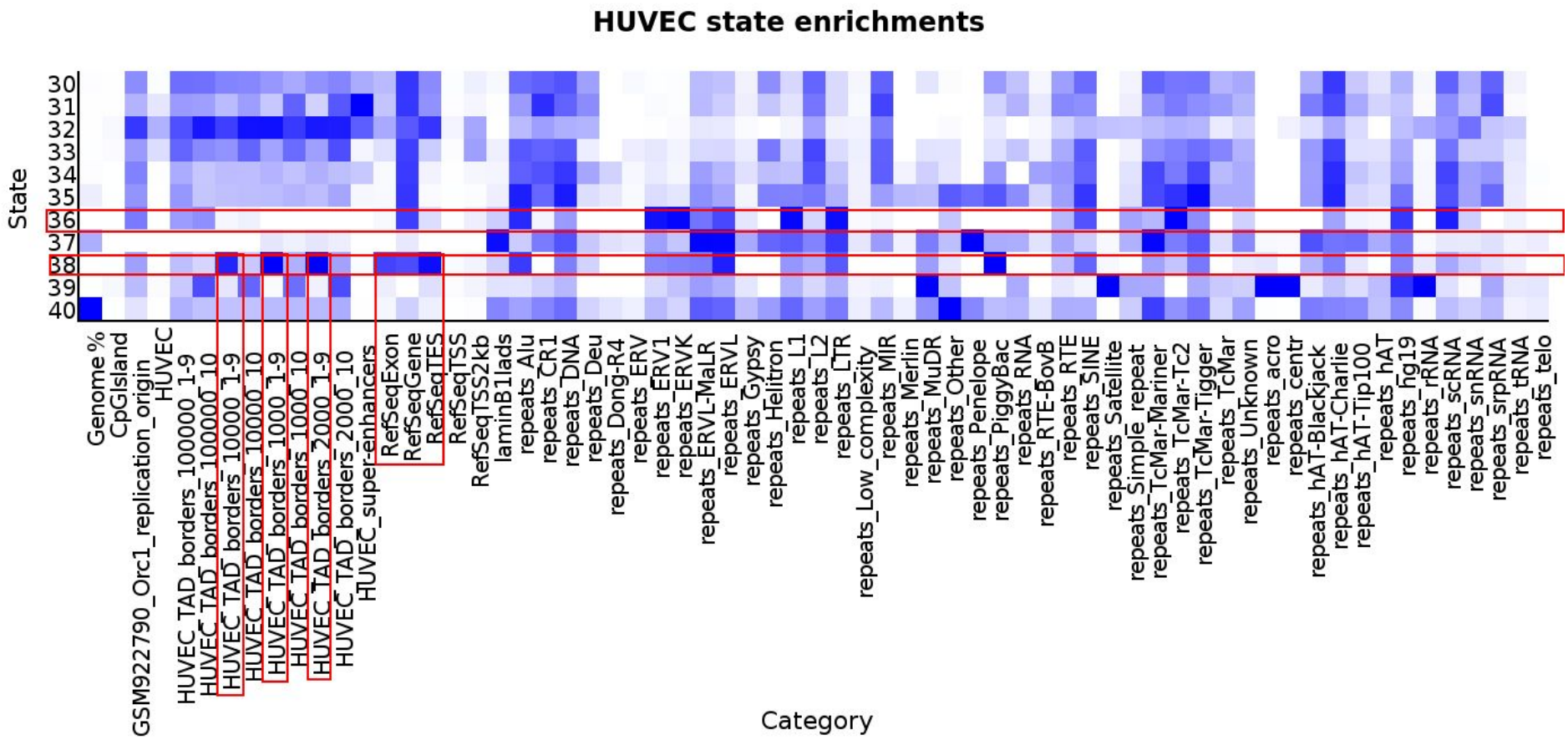
Enrichments in human cell-lines

Weak TAD borders of **hESC-H1** and **HUVEC** are enriched by H3K9me3 + H3K79m2/H3K36me3:



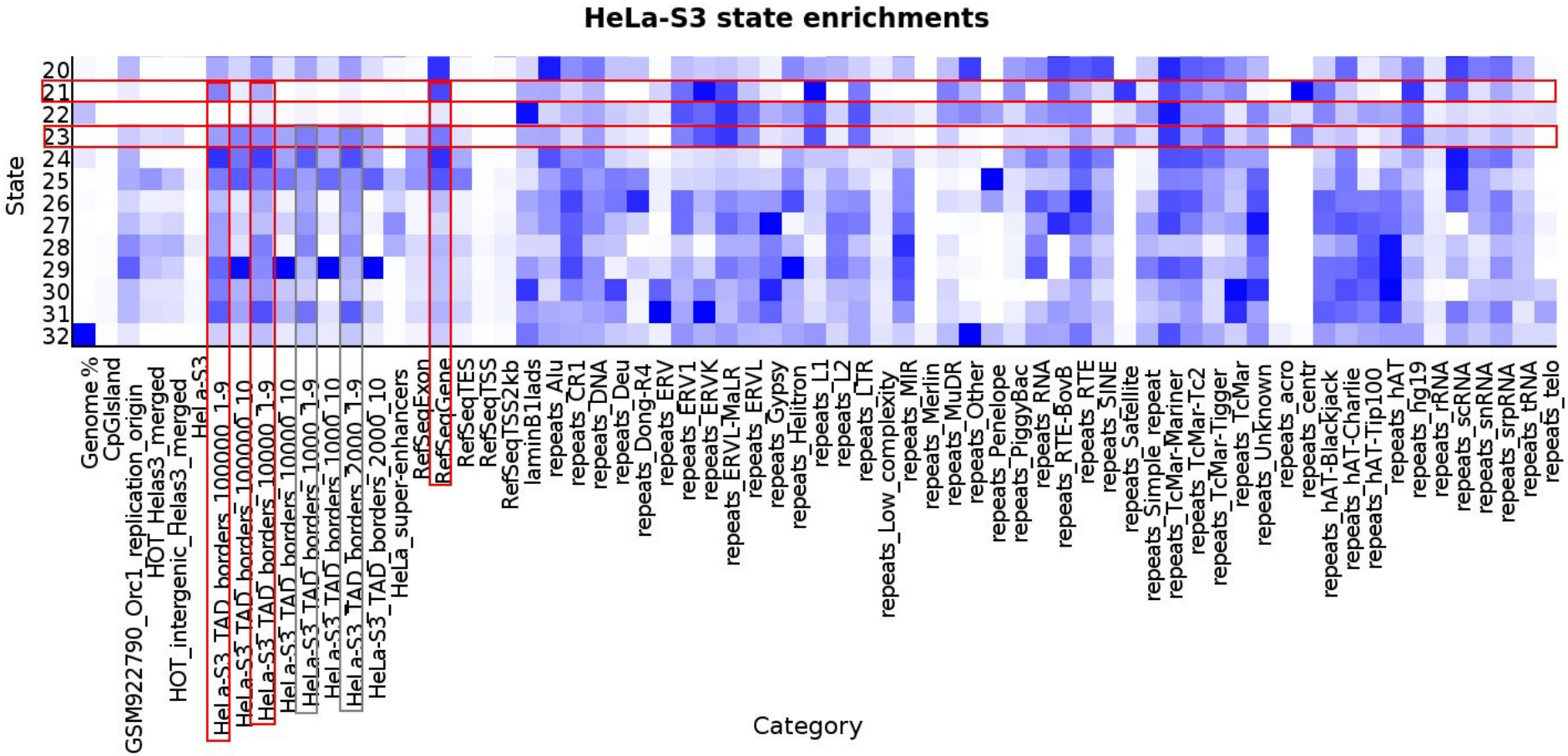
Enrichments in human cell-lines

Weak TAD borders of **hESC-H1** and **HUVEC** are enriched by H3K9me3 + H3K79m2/H3K36me3:



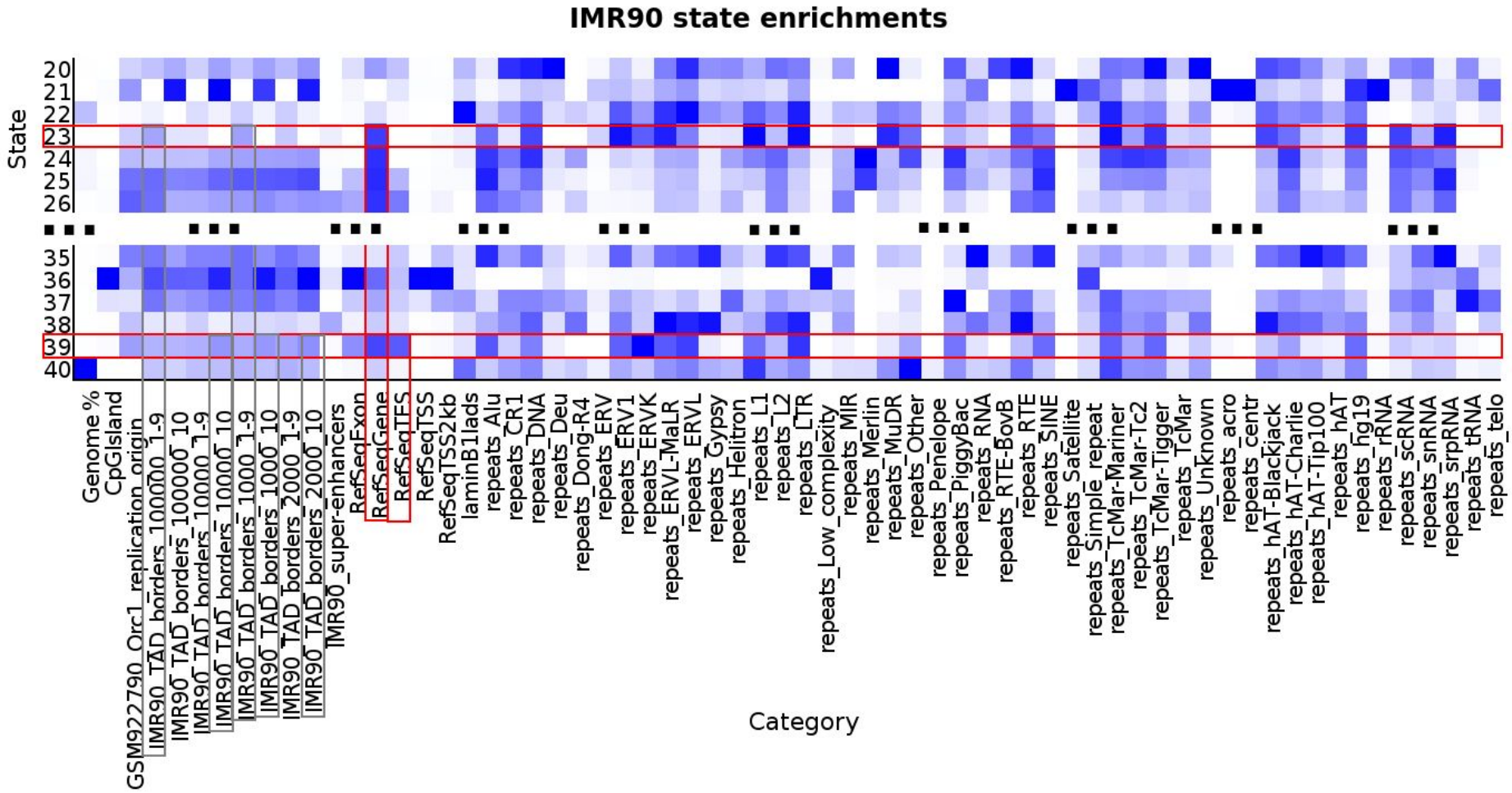
Enrichments in human cell-lines

Weak TAD borders of **HeLa-S3** and **IMR90** are not enriched considerably by H3K9me3 + H3K79m2/H3K36me3:



Enrichments in human cell-lines

Weak TAD borders of **HeLa-S3** and **IMR90** are not enriched considerably by H3K9me3 + H3K79m2/H3K36me3:



Outline

1. Introduction.
2. My master thesis in brief.
3. The current project.
4. Current results.
5. Perspective.

Current results

1. We found H3K9me3-enriched active genes in hESC-H1, HeLa-S3, HUVEC, IMR90, and mESC cell lines.

Current results

1. We found H3K9me3-enriched active genes in hESC-H1, HeLa-S3, HUVEC, IMR90, and mESC cell lines.
2. We classified this genes into several groups:
   a. H3K9me3 peak is strictly within intron (including the cases of putative alternative promoters).
   b. H3K9me3 peak overlaps transcribed exons, but is situated within gene body.
   c. H3K9me3 peak overlaps with gene from an intergenic region.
   d. H3K9me3 peak covers an active gene completely (predominantly, such genes are mono-exonic).
   e. Active ‘barrier’ genes at the edge of long H3K9me3-enriched regions with no transcription.

Current results

1. We found H3K9me3-enriched active genes in hESC-H1, HeLa-S3, HUVEC, IMR90, and mESC cell lines.
2. We classified this genes into several groups:
   a. H3K9me3 peak is strictly within intron (including the cases of putative alternative promoters).
   b. H3K9me3 peak overlaps transcribed exons, but is situated within gene body.
   c. H3K9me3 peak overlaps with gene from an intergenic region.
   d. H3K9me3 peak covers an active gene completely (predominantly, such genes are mono-exonic).
   e. Active ‘barrier’ genes at the edge of long H3K9me3-enriched regions with no transcription.
3. We calculated enrichments of various genome elements by H3K9me3 + H3K79me2/H3K36me3 combination (which approximately marks H3K9me3 regions within active genes).

Outline

1. Introduction.
2. My master thesis in brief.
3. The current project.
4. Current results.
5. Perspective.

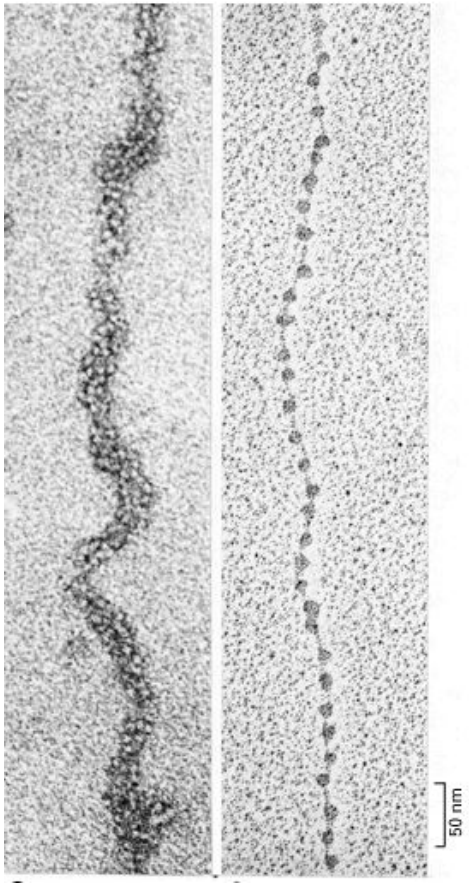
Perspective

1. Finish enrichment analysis.
2. Analyse the sequences of the H3K9me3-enriched regions within active genes.
3. Analyse literature on H3K9me3 in active genes.
4. Select a subset of such genes and check proteins from the chromatin of their H3K9me3-enriched regions.

Thank you!

TAGGTGCAGGAGCCTGCGTGACCGGTGCACTAACGGTGGGCAGCGCCAGCACATCTTGAGTCTCATGCCAGAGCTGTGGT
CGCTGGGCTAGGGCGTGCAGGAGTTGGCAATAACACTGGAACATCTGATCCATTAGCGCCGGTTCAAACAATTGTTCCAC
CACGTCCCAATTAAAGTGCAATTCCCCGTTACGCTCCATTACCTGATGATCGAGCCACACCTGCGGTGTCTGTGAAATAC
TGTACAGATCATCCTGCGGTTGGTTGAACAGCGTATCATCCGTTTCCCAGCTCGCCTCATATTGATTGAGCAAGCTGGTG
AACACGATAGGAATCACCGATTTCGCCGGATCGCGGTGAGTTTGTACCAGCAGTCGAGAGACCTGGATACCGCCAAACAA
GCGGTTATCCAAATCGCGCCACAGTTGAGAGTGGATCACGTTTGCCCGCGCCTGCAAGGTTTCTCCCTGCGACATATCAA
TCTCCAGCAACGTCAGTGCAGTAAAATCGCCAATTAAGTGATTGATATCTGCGTGCAACGGCAGACGATTAAACAGCGTC
AGGTTAAGGCTAAAGTGCGGCGAGGAACTAAAGCGCGCCAACACCTGAGCAAACCCCGTGAGCAGCAGTGTAGTAGGCGT
GATACTCATCTGACCCGCCTGCGTTTTCAAGCGCTGCCAGATAGCACGAGATAGACAATAGCTACGGCGCACGAATGTGG
GGTTTTCCAGCTTCGTGGGGTCGGTACGCAGCGGCAATCGTGGCCCCGGCGGCAGAGTTTCGATGCGGGCGCGCCAGTAG
TCACGTGCCCGTTCAAACTGCGGCGTTTTCTCGTAGGATTTTAAGGTCAGCACATAGTCGCGGAAACTGTACGTCAATGT
AGGCAATGCCGCCTGCGGATGGCGGTAGTGATAAGCTAGTTCCCGAATGAACAACTCCAGACTCCATGCATCGGCTATCA
GCAGGTCGATAGATACATGTAGCCGGGCTTTTTGCTGCGGATCAACCACGACCTCCATCTGAAACAGAGGCCAGCGACTA
CAATCAATCATCTGATGTGAGAGGGTGTCACGCAGCTCACGGCAACGCTGTTGAAACGCATCCTCGCACTGGGTGGTATA
GAAGGCAACGTGGTATTCCGGCACGTTTGGCAATATCTGCTGCATACCGTCATCGTTGACAATTGCACGCAACATGCTGT
GCCGCGCAATCACCGCATTGAGCGCGCGGTTGAACGCTGTTTCATCTAGCCCCTGTAGTTCATATTCCACATAGATATGG
GTGGAAATATTGCCCAATCCCAAGGATTTTCGCCGCCCAACCCAATAGGCCTGTTGCACGTCAGTCAGCGGAAAAGGTAG
GTGCTGCTGGTCAGGGGCAAAGACCGTATCAGGCAATGCTGAAACCGCTGTCTCACTGCCGCCTTGTTCGAGCAGATAGC
AACTTAATCCCTCGATGGTGGGATCGGCAAAGAATGCGCTTAGCGTACAGGTGATGCCGTAACGTTGCTGAATATCATAT
AAAATGGTTGACGCTAACAGCGAATGCCCACCCAGTTCGAAGAAGCTGGCGTGGTGATCATCAATGCTGACACCAAGCGT

What a bioinformatics specialist thinks the genome is like.



And what it’s really like...