Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Report for Softwareproject

Integrative analysis of next generation sequencing data: binding events of pluripotent transcription factors and histone modification pattern across the genome in MEF and ESC

Quang Vu Nhu, Sonja Batke and Sofya Laskina

AG Functional Genomics, Max plank Institut for molecular genetics,Ihnestrasse 63-73, 14195 Berlin, Germany.

* Alena von Bömmel and Robert Schöpflin.

Abstract

Motivation: Results: Availability: The data from the ATAC-seq, Chip-seq and RNA-seq experiments can be found in the GEO database under accession number GSE90895 **Contact:** name@bio.com **Supplementary information:** Supplementary data are available at *Bioinformatics* online.

1 Introduction

fgdfgwgrgwgrg

2 Approach

3 Methods

3.1 Getting data and quality analysis

Experiment data of ChIP-Seq, ATAC-Seq and RNA-Seq of the associated cell line was downloaded in .sra format from NCBI web site[1]. Locally stored files were then rewritten into .fastq files using *fastq-dump* function in SRA-Toolkit[2]. Quality of obtained read files was checked with FastQC software (Andrews, 2010).

3.2 Univariate analysis

3.2.1 Aligning to reference genome

Reads from ChIP-Seq and ATAC-Seq experiments were mapped to mouse genome (mm9) with Bowtie2 software (B.Langmead, 2007). Bowtie2 is memory efficient tool for aligning Next Generation Sequencing data using FM-Index. Only those mapped reads were retained, that had a mapping Quality of 10 or more and less than three mismatches. Duplications in read files were eliminated. Mapped Reads were translated into coverage track files (bigWig) with bamCoverage tool of deeptools software (Ramírez *et al.*, 2016). Reads were extended by 200bp. Number of reads was calculated for 25bp bins and then normalized by Reads Per Kilobase per Million mapped reads (RPKM).

RNA-Seq data was aligned to mm9 with STAR 2.4.0.1 software (Dobin *et al.*, 2013) obtaining unique reads with 2 or less mismatches and normalizing signal by Reads Per Million. For each read file of RNA-seq experiment a table with read counts per gene was produced with STAR also. Wiggle files were converted to bigWig files with *wigToBigWig* of UCSC Tools. As the chromosome names of UCSC differ from those of STAR, an own Python code was written to ignore rows with appearances of noncanonical chromosomes.

For the quality control of our .bam files we were using deepTools plotFingerprint and deeptools plotCorrelation. One plot with fingerprints was done for all transcription factors and one for all histone modifications in both stages. Before using plotCorrelation we generated a matrix from all .bam files using multiBamSummary. With plotCorrelation we produced a heatmap with the pearson correlation method and with removed outliers for both stages.

3.2.2 Peak Calling with MACS2

We used MACS2 software v.2.1.2 (Zhang *et al.*, 2008) to call peaks of our ChIP-Seq and ATAC-Seq data using a bandwidth of 150bp. We set the q-val cutoff to <0.005, the same as in Chronis *et al.* (2017), but we didn't change the default Mfold range [5-50], since the results only differed minimally. As control data we used the WCE ChIP-Seq data for the TFs, epigenetic regulators and for H3K79me2, H3K9me3, H3.3 and H3. MNase ChIP-Seq data was used as control for the remaining histone modifications. For the ChIP-Seq data of H3K9me3, we used broad peak calling because this histone modification doesn't produce narrow peaks.

[1] https://www.ncbi.nlm.nih.gov/sra/

[2] https://www.ncbi.nlm.nih.gov/sra/docs/toolkitsoft/

We have later visualized the peaks, mapping and RNA-Seq data in IGV software v.2.5.0 (Robinson *et al.*, 2011; Thorvaldsdóttir *et al.*, 2013) to look at target genes of different transcription factors such as the gene Ccnb1, also known as Cyclin B1, which is a target gene for cMyc(Menssen and Hermeking, 2002).

3.2.3 Analysis of TF peaks

In this step we used the .summits file of our peak calling results to create unified peaks that are 200bp long. With the sub-command *intersect* of the BEDTools software v.2.27.1 (Quinlan and Hall, 2010) and the sub-command *merge* of the BEDOPS software v.2.4.35 (Reynolds *et al.*, 2012), we were able to create files, which were used for the analysis of common peaks between MEFs and ESCs and common peaks between OSKM.
Using promoter regions[3], which we defined as 2000bp upstream of a annotated transcription start site (TSS) of RefSeq genes, we analysed how many peaks of OSKM were located in promoters.
The results of the analysis were visualized with R (R Development Core Team, 2008) and the package *VennDiagramms* (Chen and Boutros, 2011).

3.3 Differential analysis

3.3.1 Correlation between different histone modifications and transcription factors

For the quality control of our .bam files we were using deepTools plotFingerprint and deeptools plotCorrelation. One plot with fingerprints was done for all transcription factors and one for all histone modifications in both stages. Before using plotCorrelation we generated a matrix from all .bam files using multiBamSummary. With plotCorrelation we produced a heatmap with the pearson correlation method and with removed outliers for both stages.

3.3.2 Visualization

The number of significant peaks for both mouse embryonic fibroblast and embryonic stem cells were visualized with own R script.
Converted bigWig files and bed files with peaks were visualized with IGV. Also for both cell lines heatmaps were produced with deepTools software. First, a table with scores per genome was produced with *computeMatrix*, then a plot itself with *plotHeatmap*. As region files we used a file containing gene scores.This was obtained by putting together a promoter file containing gene names and their position with read table of paired-end RNA-seq data. This file then was devided into files with score above and below 500 and also separated by strand. Signal distribution was calculated relative to center of genomic region, which was set to 2kb up- and downstream, bigWig files of MEF with associated histone modifications were used as scores. Only modifications with signal were retained in plot.

3.3.3 Prediction model

We wrote a R-script, which built a multivariate linear regression that models the gene expression of protein coding genes in dependency to the histone modifications. We started by extracting all protein coding genes and their promoters, here defined as 2000bp upstream and 500bp downstream of the TSS, from the ensembl archive of may 2012. Then we counted the number of reads in the promoter regions of each histone modification. The resulting pseudocounts and the genecounts of the RNA-Seq data were log-normalized. To evaluate our model we need to split our data in half into a training and test set. Given these sets we now build the model from our training data and evaluate it with the test data by correlating the predicted expression to the expression measured in the test data.
For this script we used the following packages: biomaRT [Durinck *et al.* (2009);Brazma *et al.* (2005)], GenomicRanges (Lawrence *et al.*, 2013), GenomicFeatures (Lawrence *et al.*, 2013), bamsignals (Mammana and Helmuth, 2016) and DESeq2 (Love *et al.*, 2014).

[3] `http://hgdownload.soe.ucsc.edu/goldenPath/mm9/bigZips/`

4 Results

4.1 Data quality

Downloaded .fastq files were overall of a good quality with some exceptions, which had poor per sequence and per base quality, some overrepresented sequences, duplicates,etc. As reads were filtered when aligning to the genome, we skipped further formatting of .fastq files.

4.2 Univariate analysis

4.2.1 Aligning to reference genome

Reads were mapped with high quality of $>$ 70% with a few exceptions of about 50%. The table with percentages of mapping is shown in Figure 1.

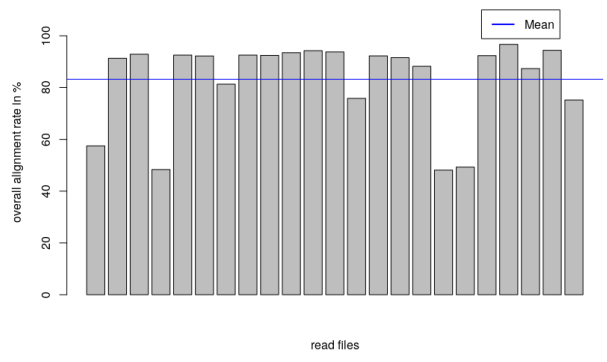

Fig. 1. Percentages of aligned reads for ATAC-seq and ChIP-seq experiments of mouse embryonic fibroblats.

4.2.2 Peak Calling

Figure 2shows that the majority of significant peaks, with an average length of 432bp, are histone modifications. This is not only because there are more histone modifications than transcription factors, but the absolute numbers are higher in general (Supplementary table **??**). In ESCs it is apparent that the amount of significant peaks for the transcription factor cMyc is very low in comparison to the other transcription factors.

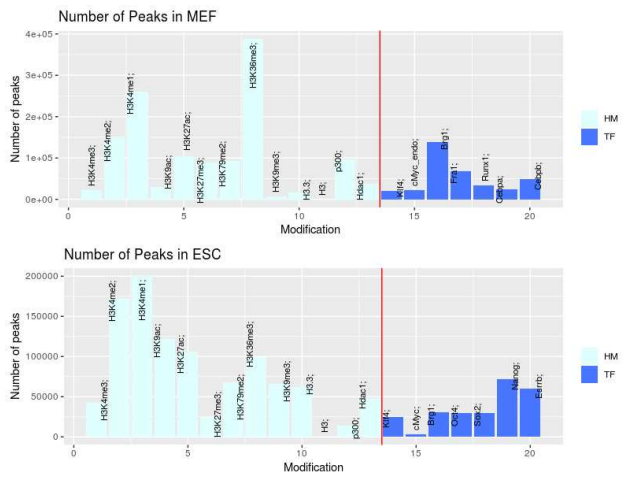

Fig. 2. Number of significant Peaks in mouse embryonic fibroblasts and embryonic stem cells with corresponding histone modifications or transcription factors.

The venndiagram in Figure 3 illustrates that Oct4 and Sox2 share many binding sites in ESCs with each other. Klf4 shares a few binding sites with Oct4 and Sox2 and cMyc barely shares any binding site with another transcription factor, but this may be a result of the small amount of significant peaks for cMyc (supplementary figure 8).

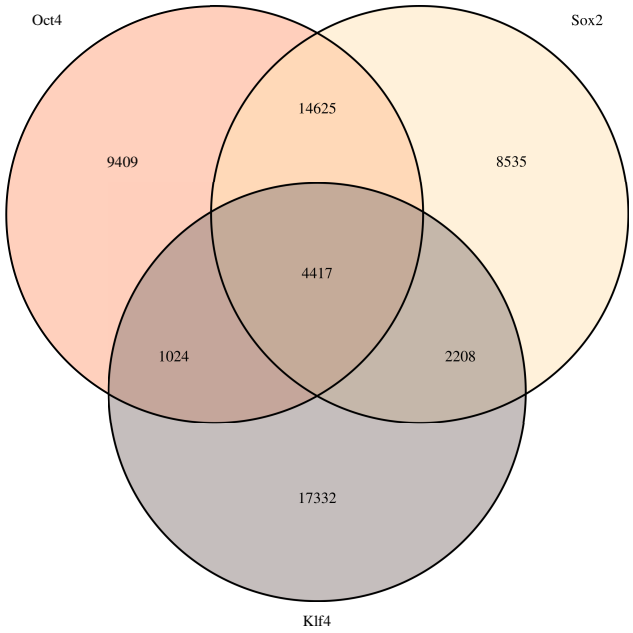

Fig. 3. Number of unique and shared Oct4, Klf4 and Sox2 peaks in embryonic stem cells

cMyc binding sites are located primarily in promoter regions for both stages. We assume that the other binding sites are located in enhancers. The binding sites for Klf4 however change between the stages. In MEF there are more binding sites in promoters than in enhancers, but in ESC the majority of binding sites are located in enhancers. Distribution of those signals according to their locus is shown in Figure 4.

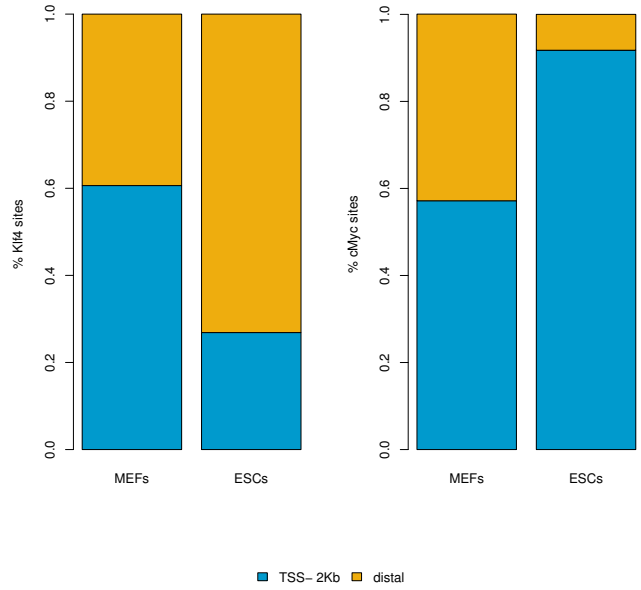

Fig. 4. Klf4 and cMyc binding rates in promoter region and outside in MEF and ESC

In cMyc binds in both MEF and ESC to the promoter region of the target gene. The peaks however differ in size. In ESC the signal is stronger and wider than the signal in MEF. This is also reflected in the RNA-Seq data. The signal in ESC is stronger than the signal in MEF (Supplementary Figure 6).

4.3 Differential analysis

4.3.1 Correlation between different histone modifications and transcription factors

Correlation between the different histone modifications and transcription factors from the Chip-seq and ATAC-seq experiments in the MEF's and ESC's is shown in Apendix Figure 5 . In the MEF heatmap all experiments are very high correlated, except the ATAC-seq experiments and H3K79me2. In the ESC KLF4, cMyc, p300 and H3K9me3 are very high correlated, as well as H3K9ac and H3K4me3. Also Sox2, Oct4, H3K27ac and H3K4me1 have a high correlation. Once more ATAc-seq has a low correlation to the other experiments.

Figure 6 shows the fingerprints of all histone modifications and transcription factors in MEF and ESC. Concerning the transcription factors in both stages the differences between control and signal are less clear than in the plots with the histone modification. There is a higher difference between control and the most histone modifications, especially in H3K79me2 and H3K36me3 in the MEF and H3K4me3, H3K4me2 and H3K9ac in the ESC. Nevertheless, we can assume an acceptable quality from the plots.

4.3.2 Visualization

Heatmaps on Figure 9 show that most signals for MEF come from H3K4me3, H3K4me2, H3K27ac, H3K79me2 modifications, as the color of the those reads becomes blue. There is also reverse dependance between acetylation on 9^{th} and 27^{th} Lysin.

4.3.3 Prediction model

The model derived from the histone modifications in MEF is fairly well correlated to the gene expression with r = 0.65 and a p-value below 2.2e-16.

(Table 1) H3K27ac, H3K9ac, H3K4me3 seem to be the most important modifications since their absolute values are the highest. We also used this model to predict the gene expression of the test data set. The correlation coefficient between the predicted and true expression is 0.7988768.
We did the same process for histone modifications in ESC and derived a model with r = 0.6997, p-value below 2.2e-16 and coefficients that are listed in Table 3. In this model the histone modification H3K27me3 has one of the highest absolute values, unlike in the MEF model. The predicted and true expression have a correlation coefficient of 0.7376995, which is just a bit worse than the MEF model.

Table 1. β coefficients of the multivariate regression for MEF

modification	value	p-value
(Intercept)	-1.03583	6.36e-11
H3K4me3	1.09168	< 2e-16
H3K4me2	0.64727	< 2e-16
H3K4me1	0.18478	0.00665
H3K9ac	-1.26799	< 2e-16
H3K27ac	1.40176	< 2e-16
H3K27me3	-0.27458	4.80e-08
H3K79me2	0.47766	< 2e-16
H3K36me3	0.06039	0.20053
H3K9me3	-0.89854	< 2e-16

5 Discussion

Heatmaps showed an expected pattern (K.Barth and Imhof, 2010) of histone modification signals. At this point we were looking for the linear model coefficients to prove this trend.

The number of significant peaks of H3K4me3, H3K9ac and H3K9me3 rise drastically from MEF to ESC, although the number of significant peaks for all HMs decrease. H3K4me3 and H3K9ac both activate transcription and are important in the MEF model, whereas H3K9me3 is a repressive mark(M.Lawrence *et al.*, 2016). This suggests that H3K4me3 and H3K9ac mark many genes which are expressed in ESC and H3K9ac represses those genes, which were previously expressed in MEF but aren't in ESC.
The HMs H3K4me3, H3K9ac, and H3K27ac have high values in both the multivariate regression for MEF and ESC. In the MEF model all of these coefficients are positive, which would mean that they increase the level of gene expression. In the ESC model the β coefficient for H3K9ac is negative which contradicts the other model. However these HMs are all correlated with active transcription[M.Lawrence *et al.* (2016),**?**], so the ESC model must be faulty somehow.

6 Conclusion

1. this is item, use enumerate
2. this is item, use enumerate
3. this is item, use enumerate

Acknowledgements

We would like to thank Alena von Bömmel and Robert Schöpflin for their support and helpful advices.

References

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data.

B.Langmead, S. (2007). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**, 357–359.

Brazma, A. *et al.* (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, **21**(16), 3439–3440.

Chen, H. and Boutros, P. C. (2011). Venndiagram: a package for the generation of highly-customizable venn and euler diagrams in r. *BMC Bioinformatics*, **12**(1), 35.

Chronis, C. *et al.* (2017). Cooperative binding of transcription factors orchestrates reprogramming. *Cell*, **168**(3), 442 – 459.e20.

Dobin, A. *et al.* (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.

Durinck, S. *et al.* (2009). Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. *Nature Protocols*, **4**, 1184 EP –.

K.Barth, T. and Imhof, A. (2010). Fast signals and slow marks: the dynamics of histone modifications. *Trends in Biochemical sciences*, **35**, 618–626.

Lawrence, M. *et al.* (2013). Software for computing and annotating genomic ranges. *PLOS Computational Biology*, **9**(8), 1–10.

Love, M. I. *et al.* (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, **15**(12), 550.

Mammana, A. and Helmuth, J. (2016). bamsignals: Extract read count signals from bam files. *R package version*, **1**(0).

Menssen, A. and Hermeking, H. (2002). Characterization of the c-myc-regulated transcriptome by sage: Identification and analysis of c-myc target genes. *Proceedings of the National Academy of Sciences*, **99**(9), 6274–6279.

M.Lawrence *et al.* (2016). Lateral Thinking: How Histone Modifications Regulate Gene Expression. *Trends in Genetics*, **32**, 42–56.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**(6), 841–842.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ramírez *et al.* (2016). deepTools2: A next Generation Web Server for Deep-Sequencing Data Analysis. *Nucleic Acids Research*.

Reynolds, A. P. *et al.* (2012). BEDOPS: high-performance genomic feature operations. *Bioinformatics*, **28**(14), 1919–1920.

Robinson, J. T. *et al.* (2011). Integrative Genomics Viewer. *Nature Biotechnology*, **29**, 24–26.

Thorvaldsdóttir, H. *et al.* (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, **14**, 178–192.

Zhang, Y. *et al.* (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**(9), R137.

Appendices

A Graph Appendix

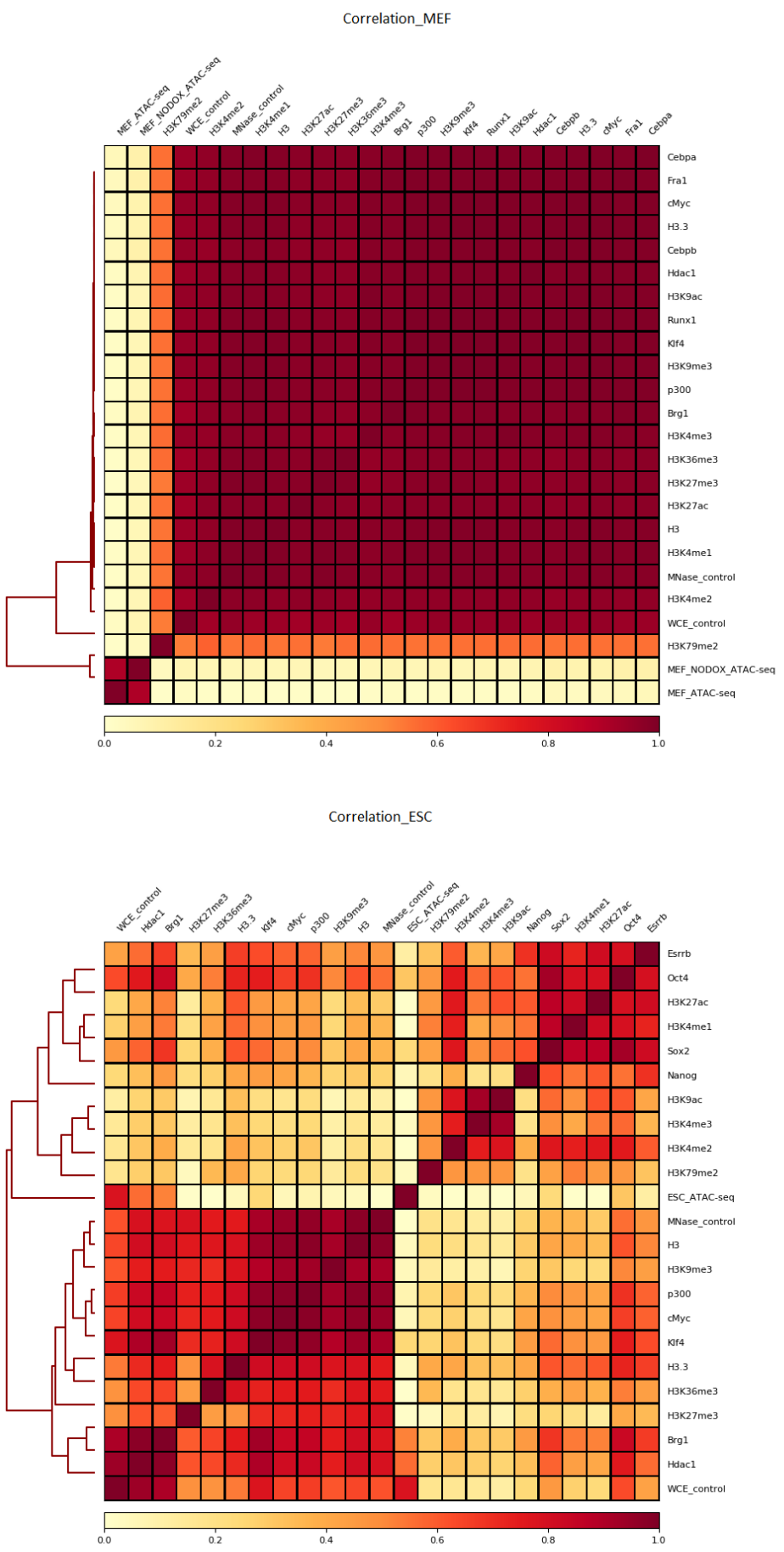

Fig. 5. Correlation of histone modifications and transcription factors in MEF and ESC

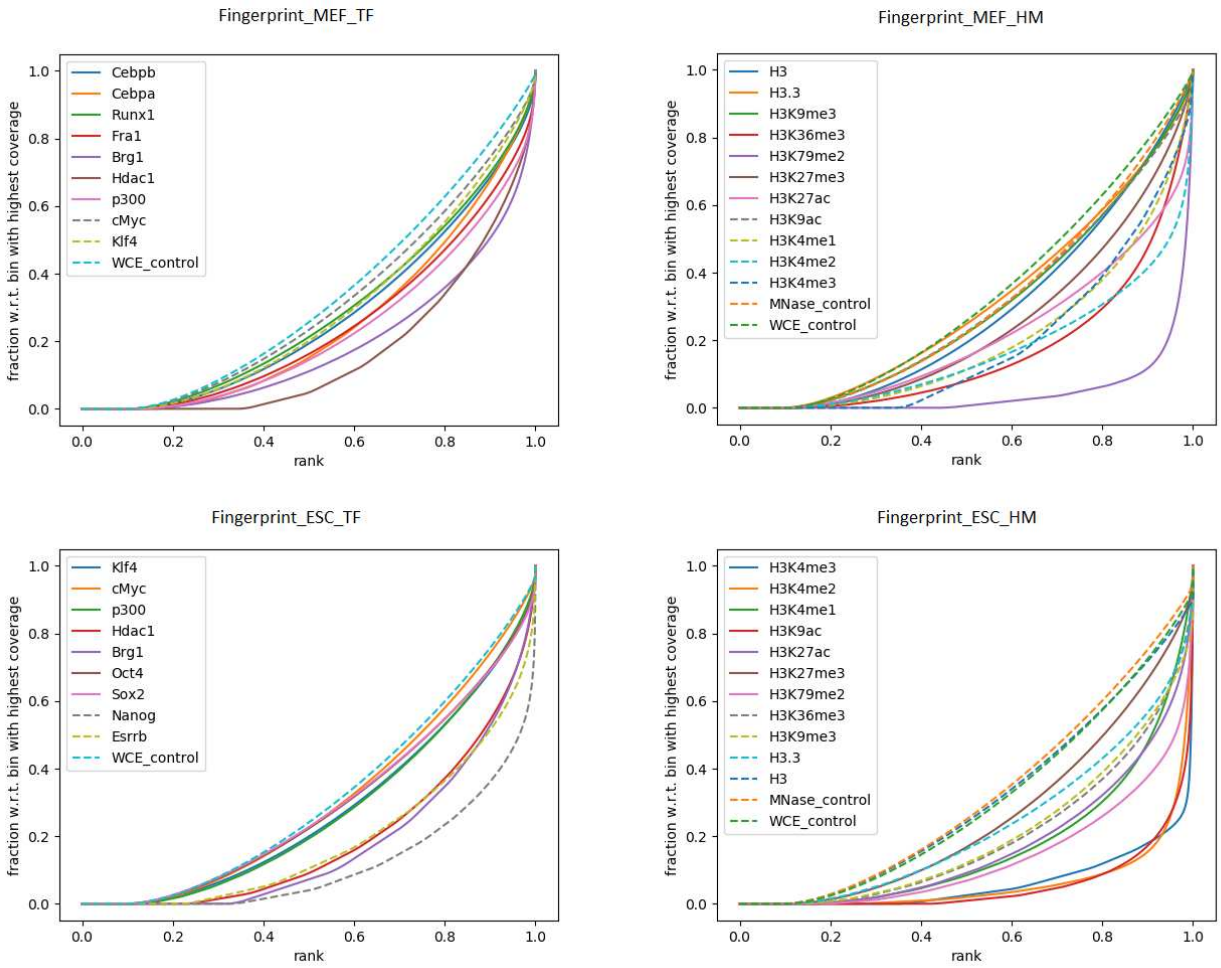

Fig. 6. fingerprints of histone modifications and transcription factors in MEF and ESC

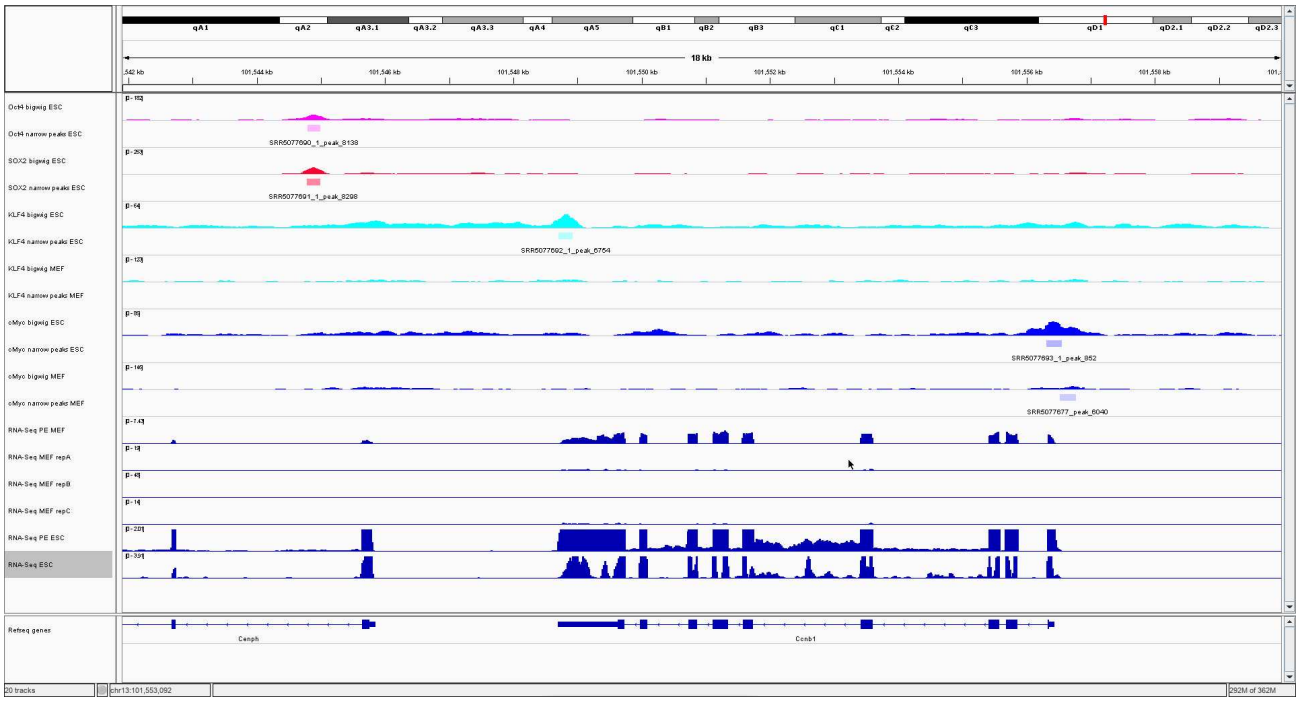

Fig. 7. IGV Image of signals in bigWig, Rna-seq and peak files

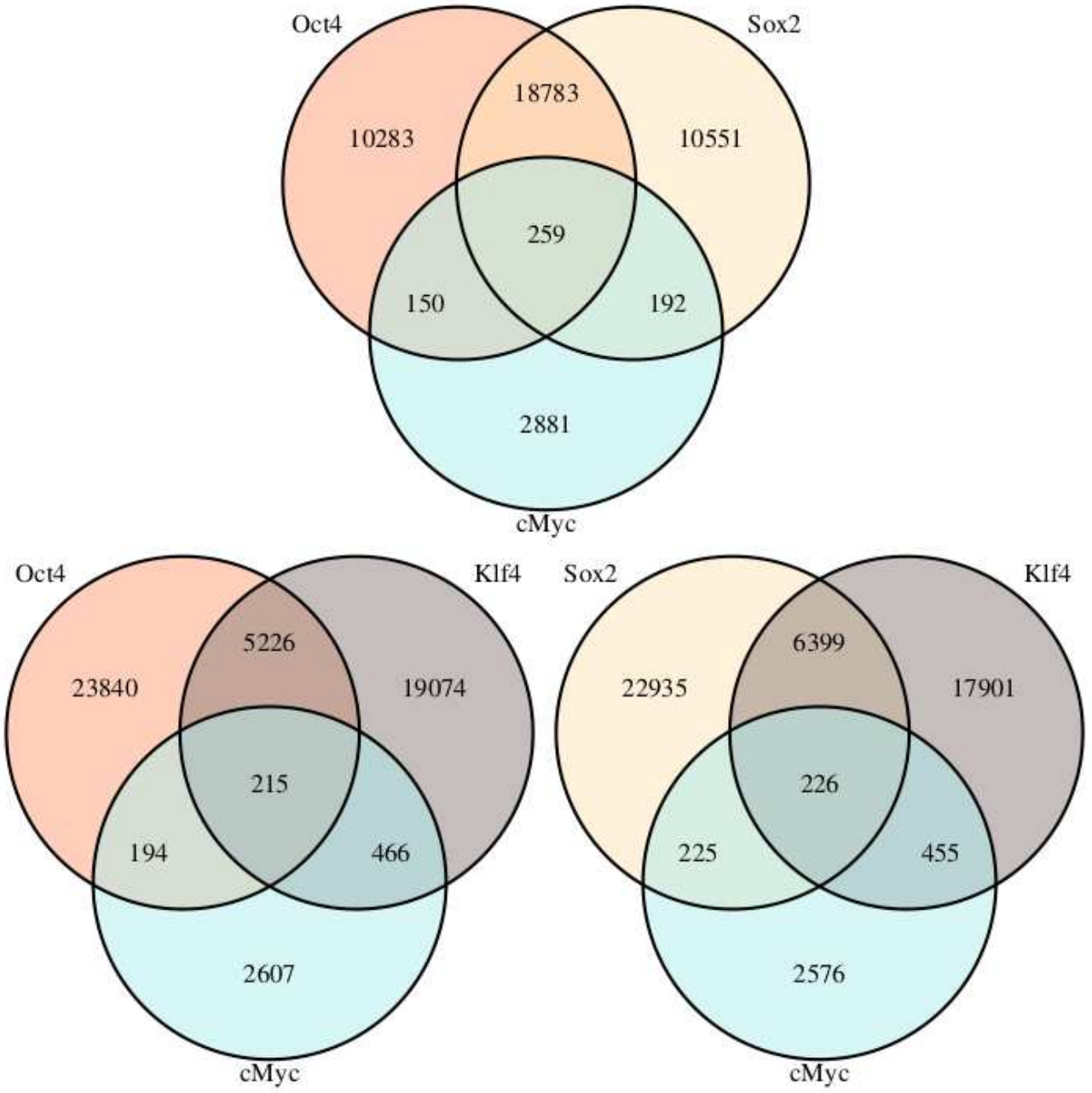

Fig. 8. Shared peaks of different transcription factors in embryonic stem cells

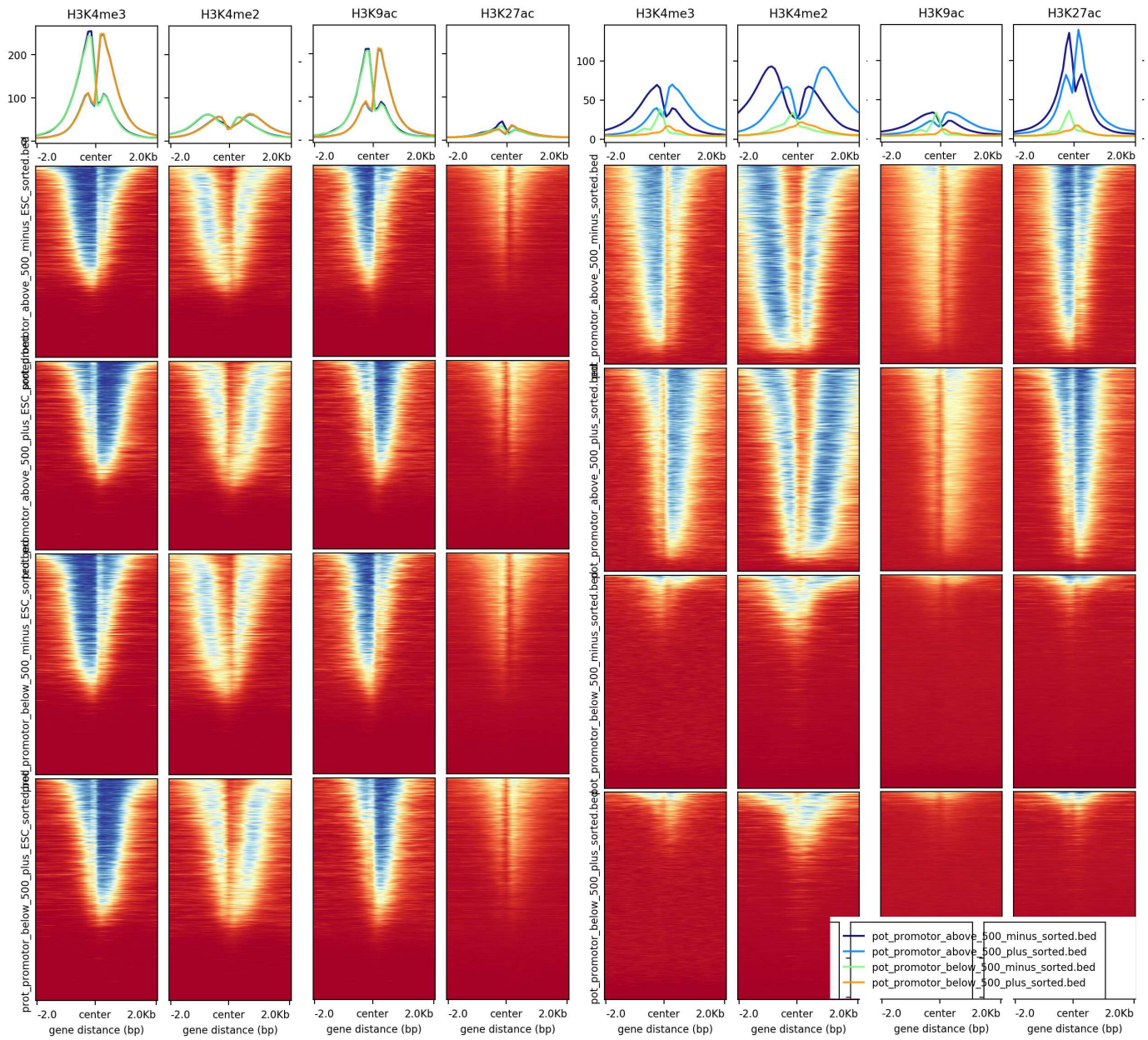

Fig. 9. Heatmap with most remarkable signals in promoter region according to the strands in ESC and MEF

Table 2. Number of significant peaks in all experiments

HM/TF	MEF	ESC
H3K4me3	23124	42355
H3K4me2	147290	171863
H3K4me1	261130	199966
H3K9ac	30337	122129
H3K27ac	105054	106031
H3K27me3	88768	24295
H3K79me2	91976	67879
H3K36me3	388342	98844
H3K9me3	7037	65494
H3.3	16583	61574
H3	132	0
p300	97458	13787
Hdac1	37438	48029
Klf4	20209	24981
cMyc	22039	3482
Brg1	139174	30365
Fra1	68117	
Runx1	33612	
Cebpa	24036	
Cebpb	48592	
Oct4		29475
Sox2		29785
Nanog		72111
Esrrb		59666
HM	1294669	1022246
TF	355779	249865
total	1650448	1272111

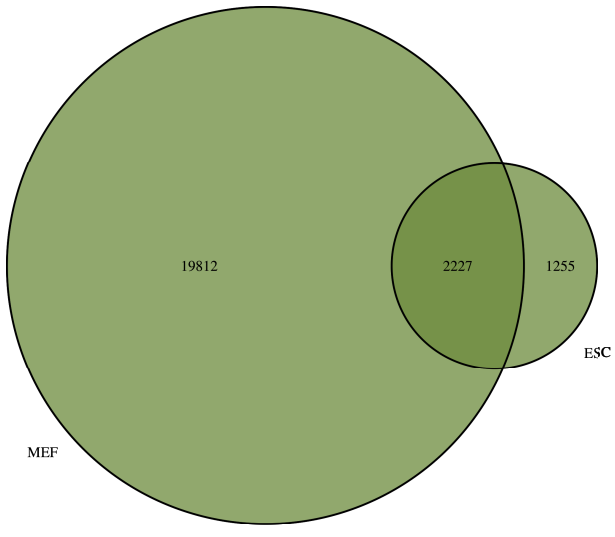

Fig. 10. Number of unique and shared of cMyc peaks in mouse embryonic fibroblasts and embryonic stem cells

Table 3. β coefficients of the multivariate regression for ESC

modification	value	p-value
(Intercept)	0.16989	0.0584
H3K4me3	0.57822	< 2e-16
H3K4me2	-0.26536	< 2e-16
H3K4me1	0.35192	0.< 2e-16
H3K9ac	0.42256	< 2e-16
H3K27ac	0.49311	< 2e-16
H3K27me3	-0.52931	< 2e-16
H3K79me2	0.03307	0.2481
H3K36me3	-0.01427	0.6122
H3K9me3	-0.27798	1.19e-13

B Code Appendix