Core Promoter Analysis

Code

..

Andrey

14 August, 2018

1 General

Promoter data used is from ENCODE:

- K562 cell line (cell): two replicates

CAGE-seq was extracted from the R package **ENCODEprojectCAGE**.

Code

2 Extracting core set of promoters

Data was extracted from R package **ENCODEprojectCAGE**.

2.1 CAGEset

Make a CAGEobject with normalised CTSS and TCs (20bp distance metric) with interquantile widths determined (0.1 - 0.9).

Since the sample sizes may differ, it is important to normalise them. The normalised data can then be exported as bed graphs to be viewed in the UCSC Genome Browser.

Raw CAGE data are clustered into groups depending on how faw from each other the calls are. In this case, 20 nt is used as a threshold for one cluster. 0.5 cpm is a minimum amount of (normalised) signal for a cluster to be included in subsequent analysis.

After the clusters are obtained, promoter width can be determined. Cumulative distribution of TSS is calculated and promoter width is defined as a distance between bottom 10% and top 90% of the distribution.

Code

2.2 Quality control

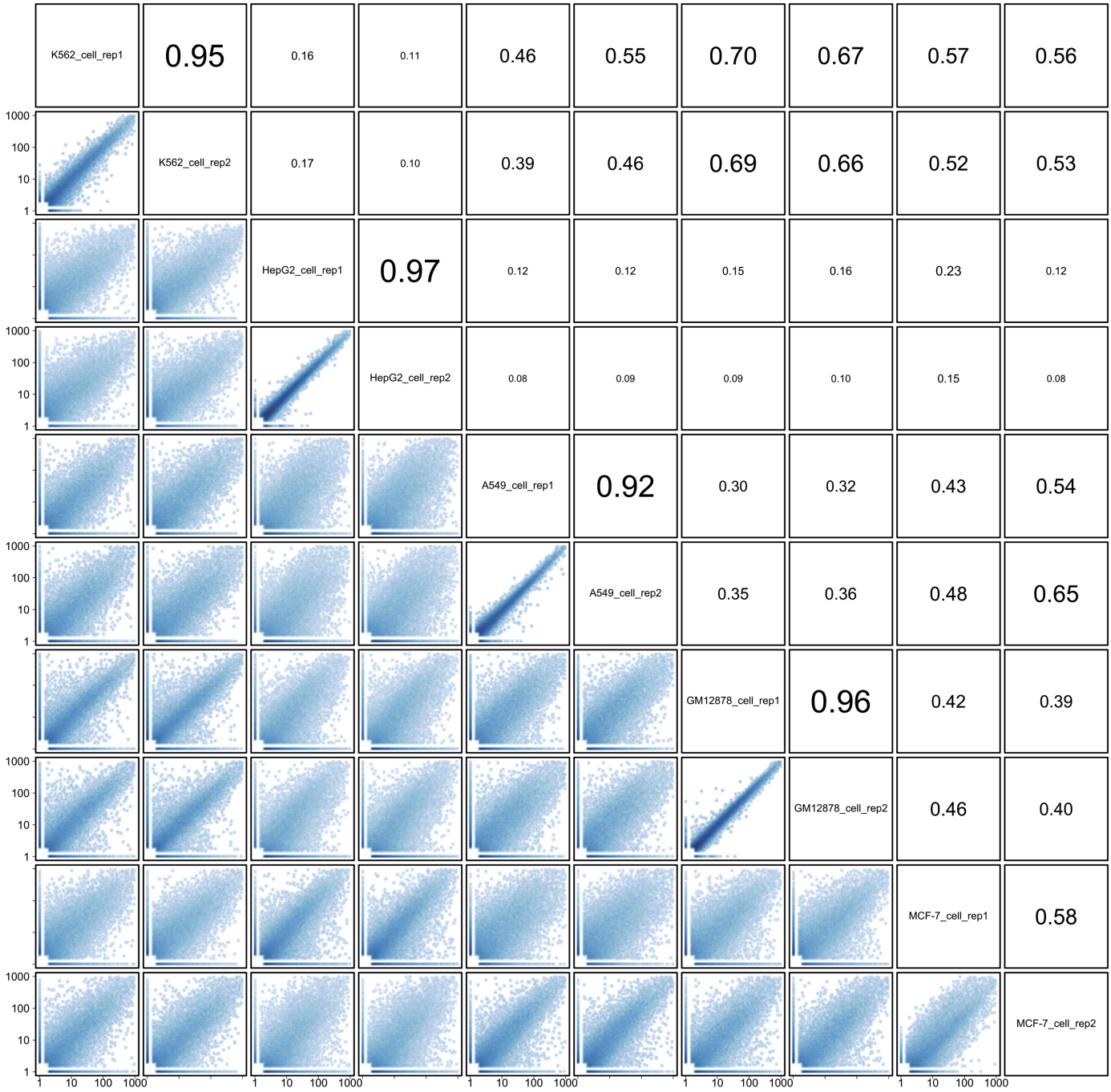
Correlation between the samples (here with CAGEr function), library sizes, number of TCs, and distribution of TC widths.

Code

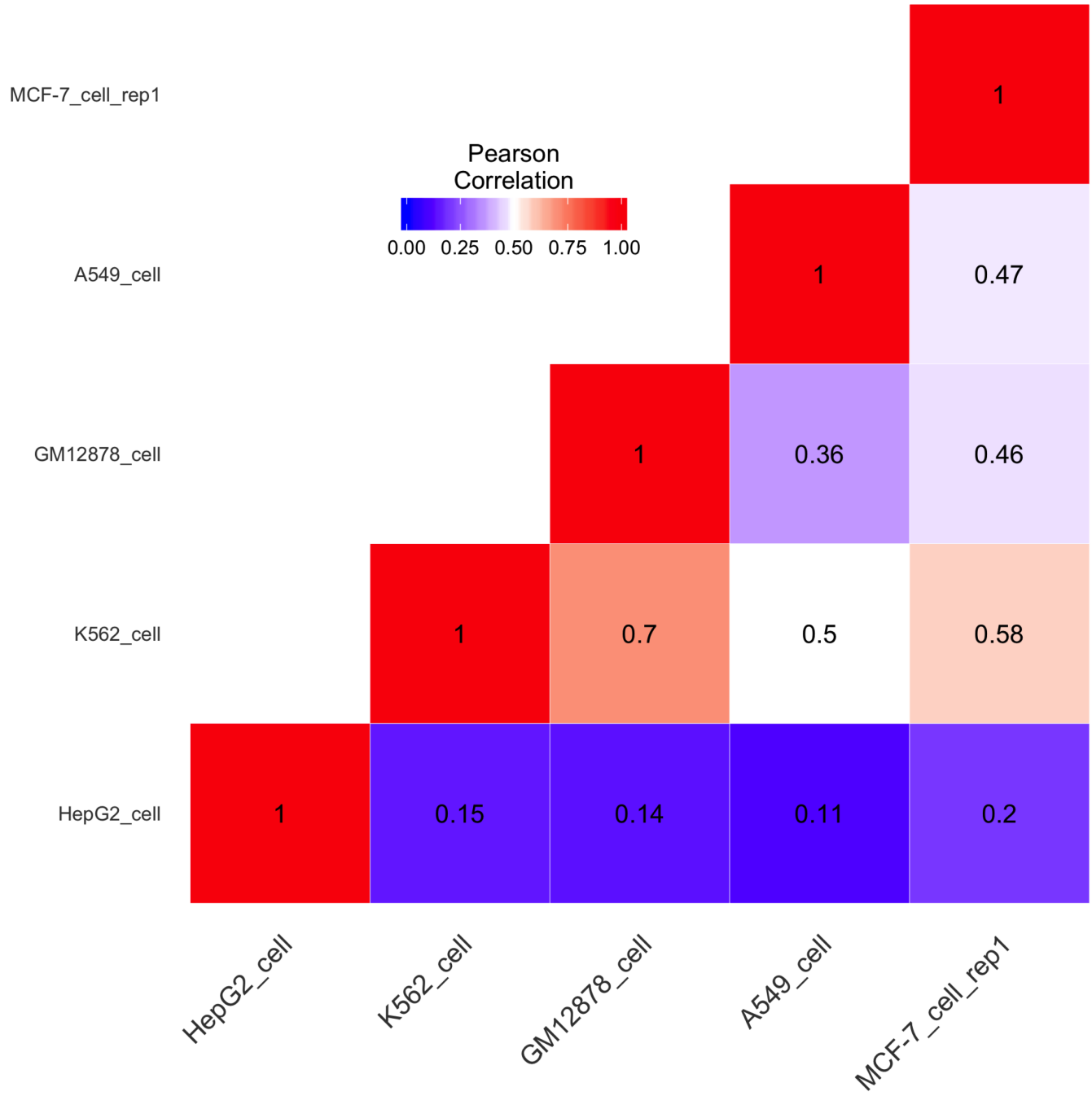
2.3 Produce a heatmap of correlations between all cell types

Code

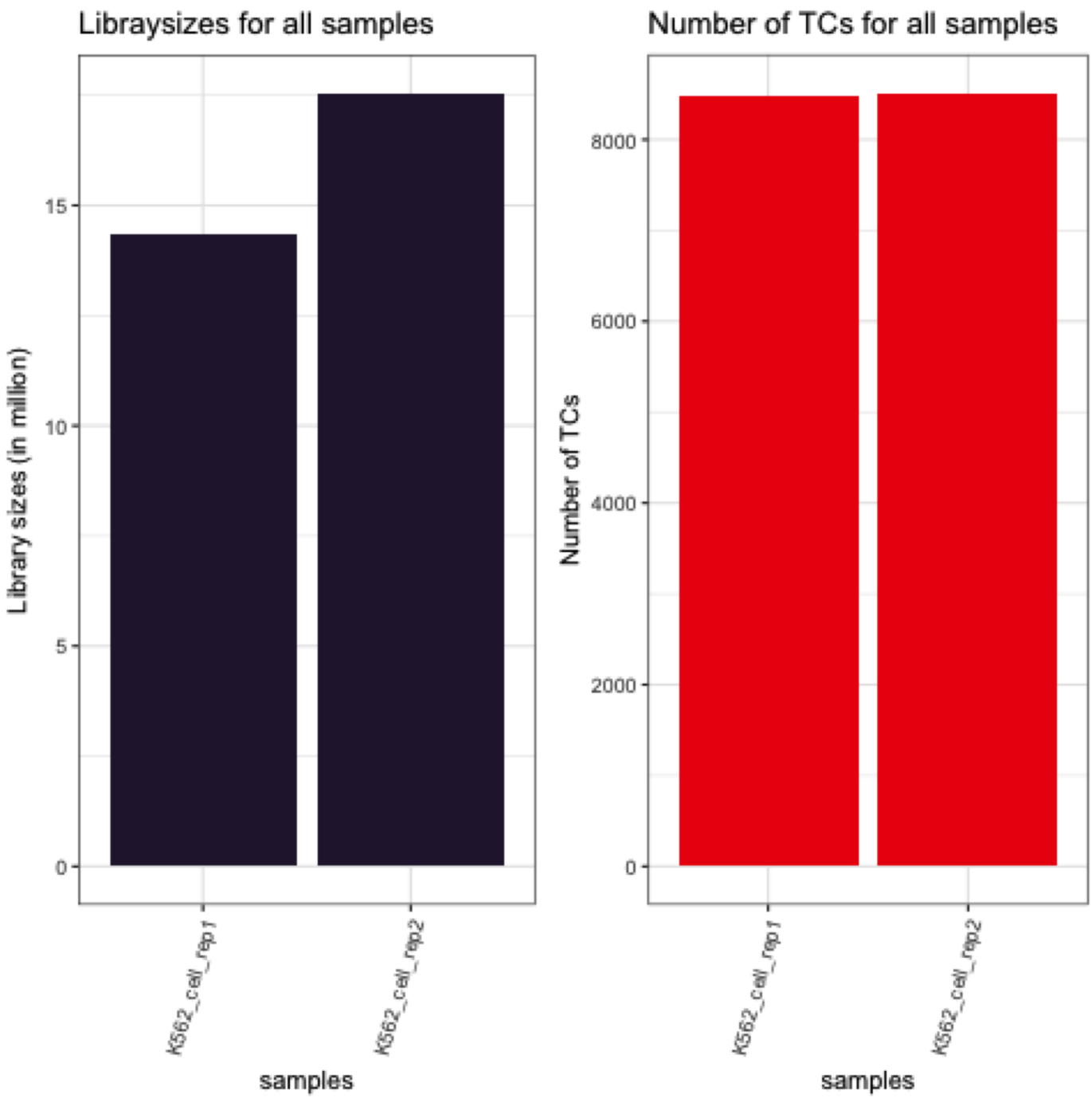
Correlation matrix. Since the samples are technical replicates, the correlation coefficient is close to 1, as expected. The MCF-7 line failed to demonstrate sufficient correlation level, thus its replicates were not merged and 2nd replicate was removed from subsequent analysis.



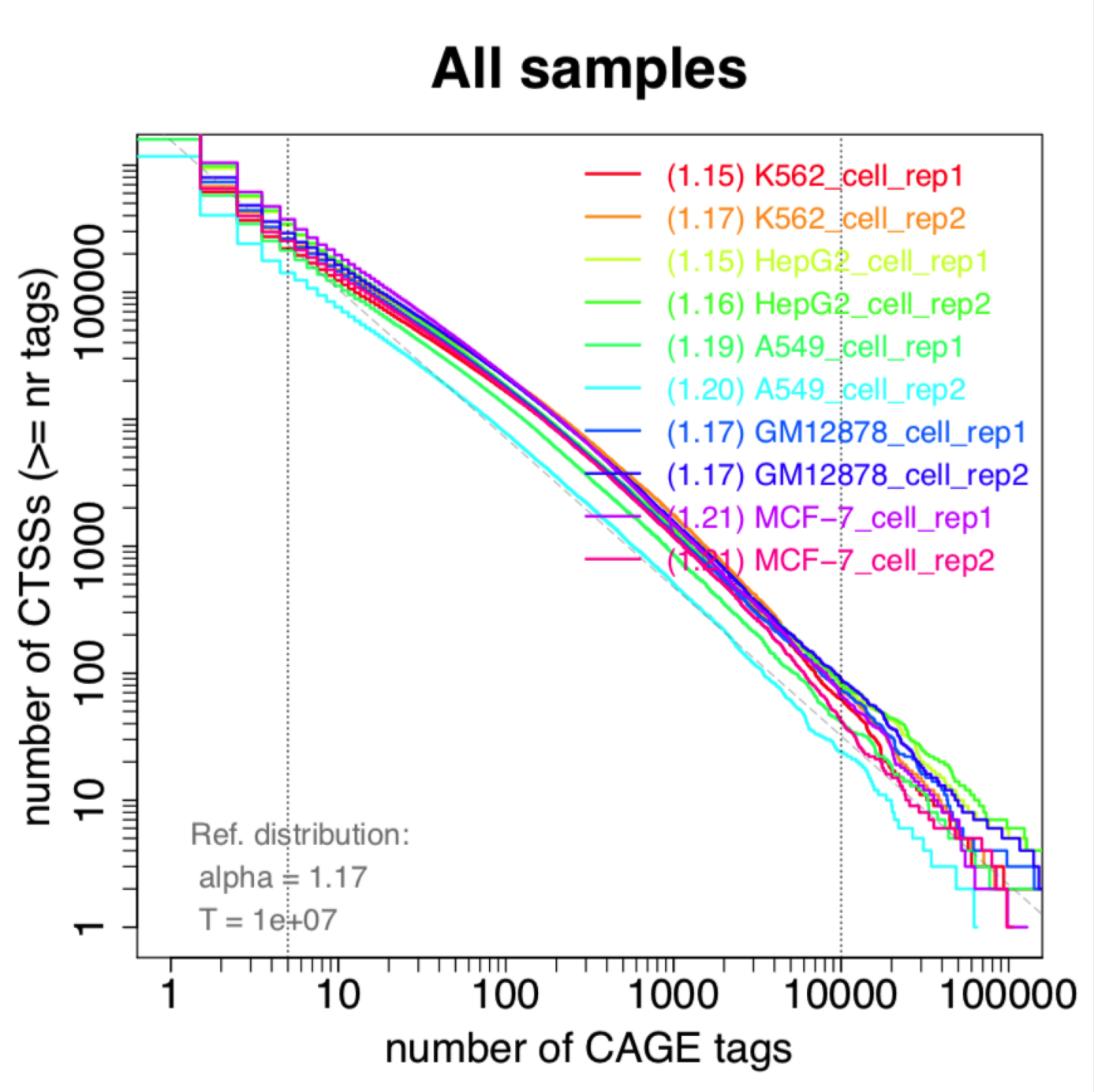
Correlation heatmap among consensus clusters. Whereas the majority of samples are relatively correlated, the HepG2 cell line shows nearly no correlation to other cell lines.



Library sizes



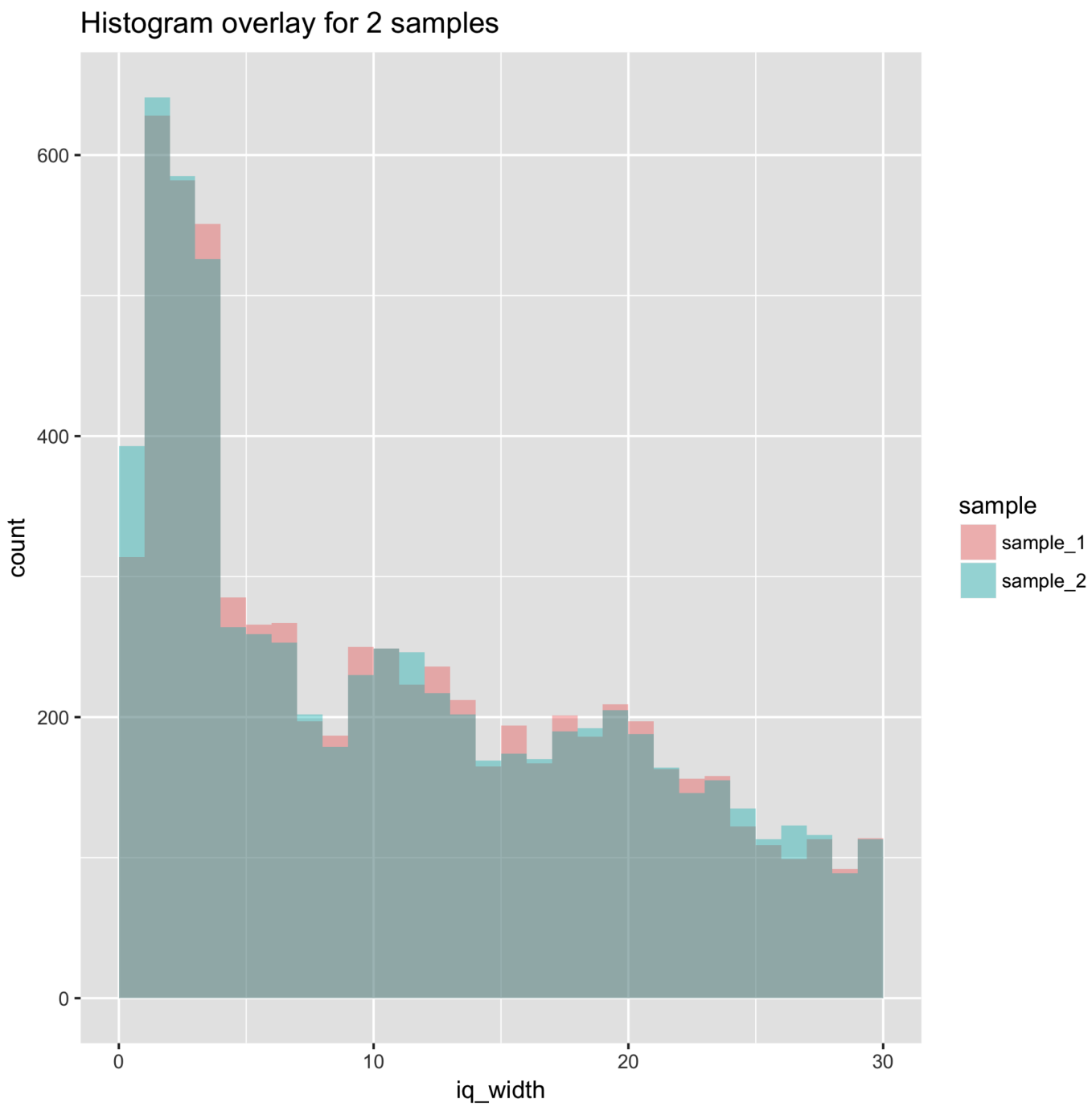
Cumulative distribution number of CTSS with respect to amount of CAGE signal



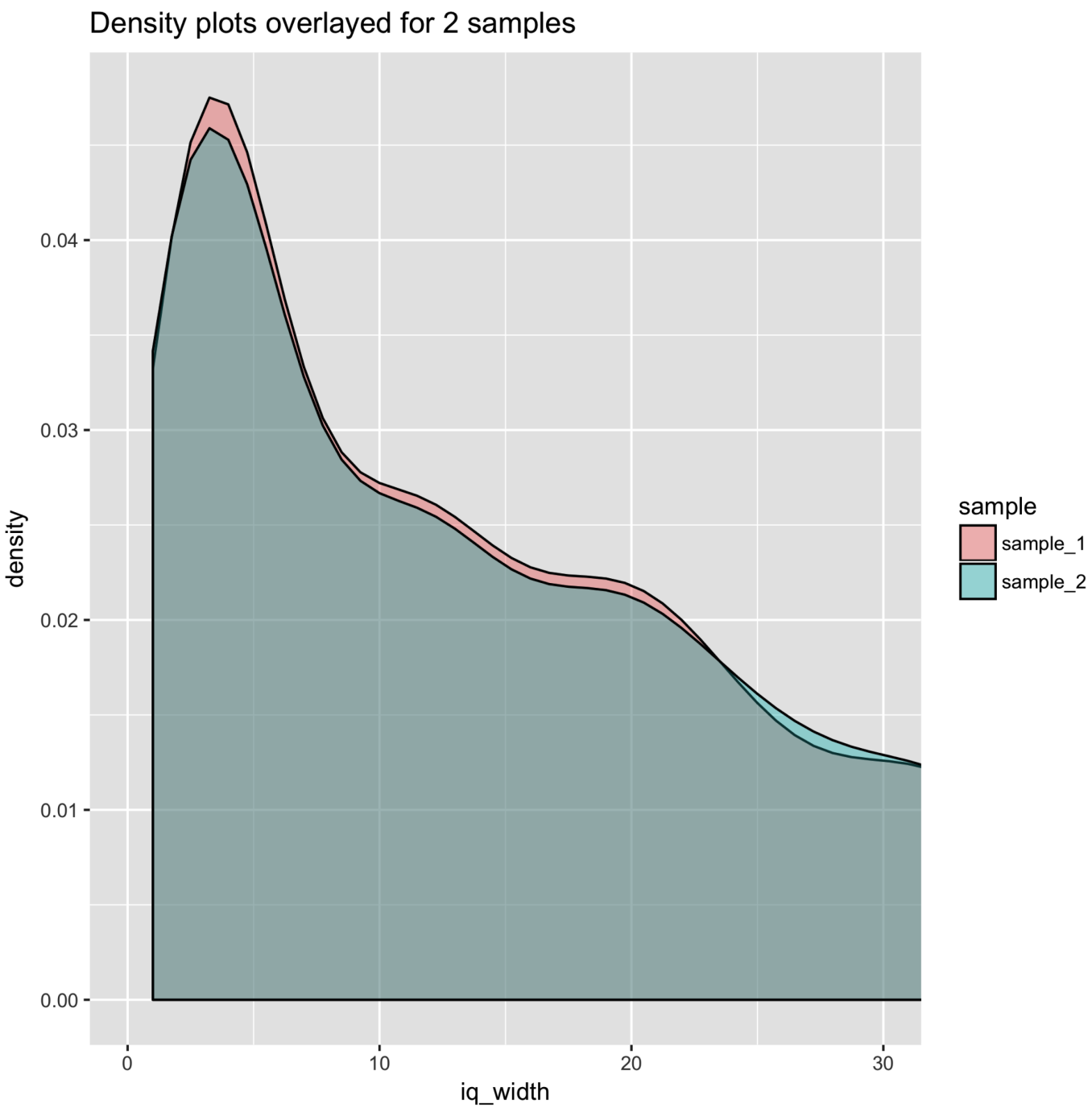
IQ width histograms

Code

Overlayed histograms of 2 samples



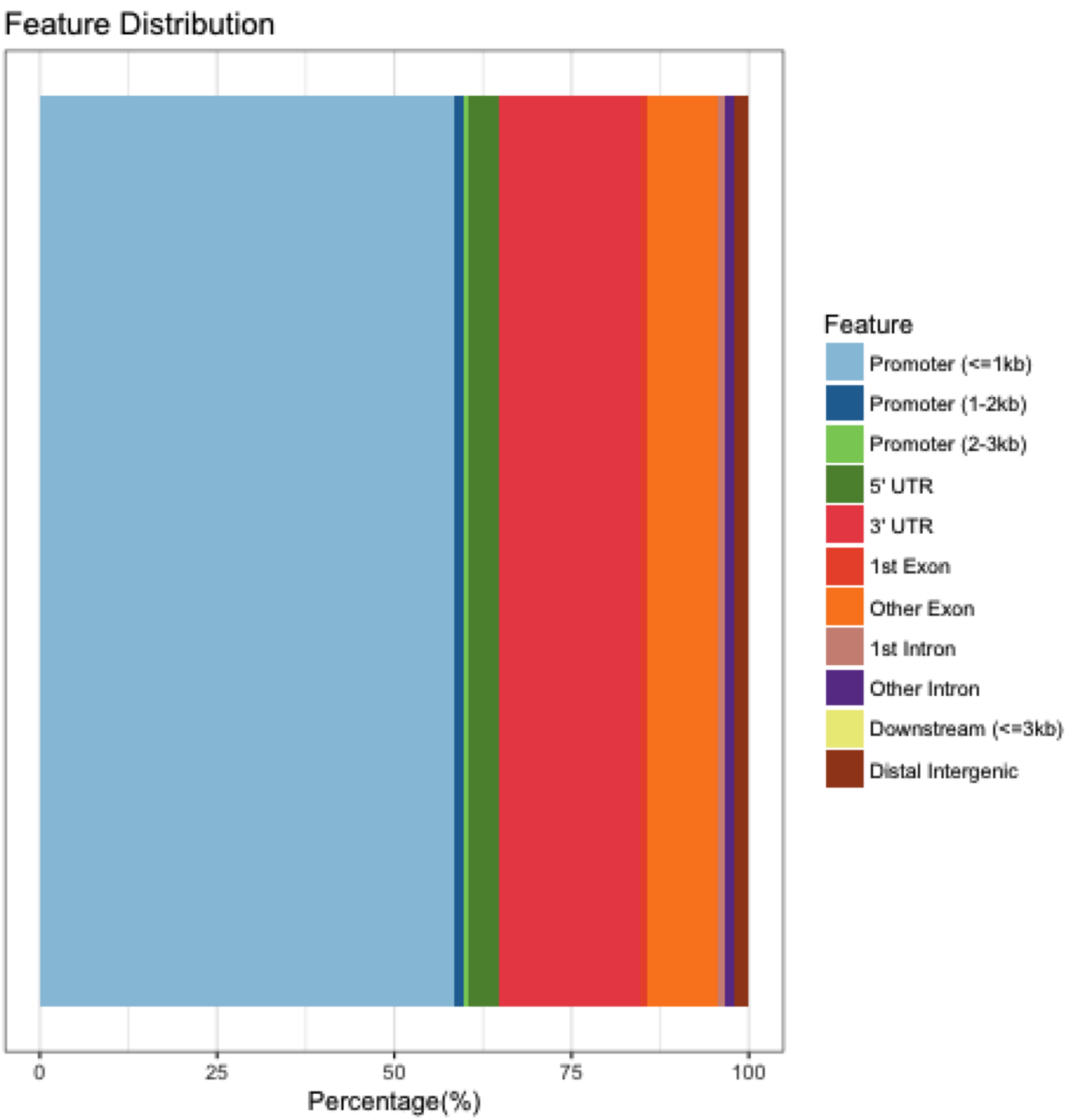
Overlayed Density Plots



2.4 Genomic features mapping

Code

Genomic feature distribution for the K562 cell line



2.5 Tracks

Tracks are found here: `../data/tracks/`

Code

3 Data analysis

3.1 GRanges

The tag clusters are extracted from the CAGEset object for downstream analysis. In order to match tag clusters to actual genomic sequences, the hg19 assembly of human genome is used. Dominant TSS is defined for each cluster (position with the highest amount of tags) and stored as GRanges object. Then, flanking seqeunces are retrived (250 bp upstream/downstream) with respect to the dominant TSS. It is important to trim the sequences which fall out of the range of chromosome lengths.

Code

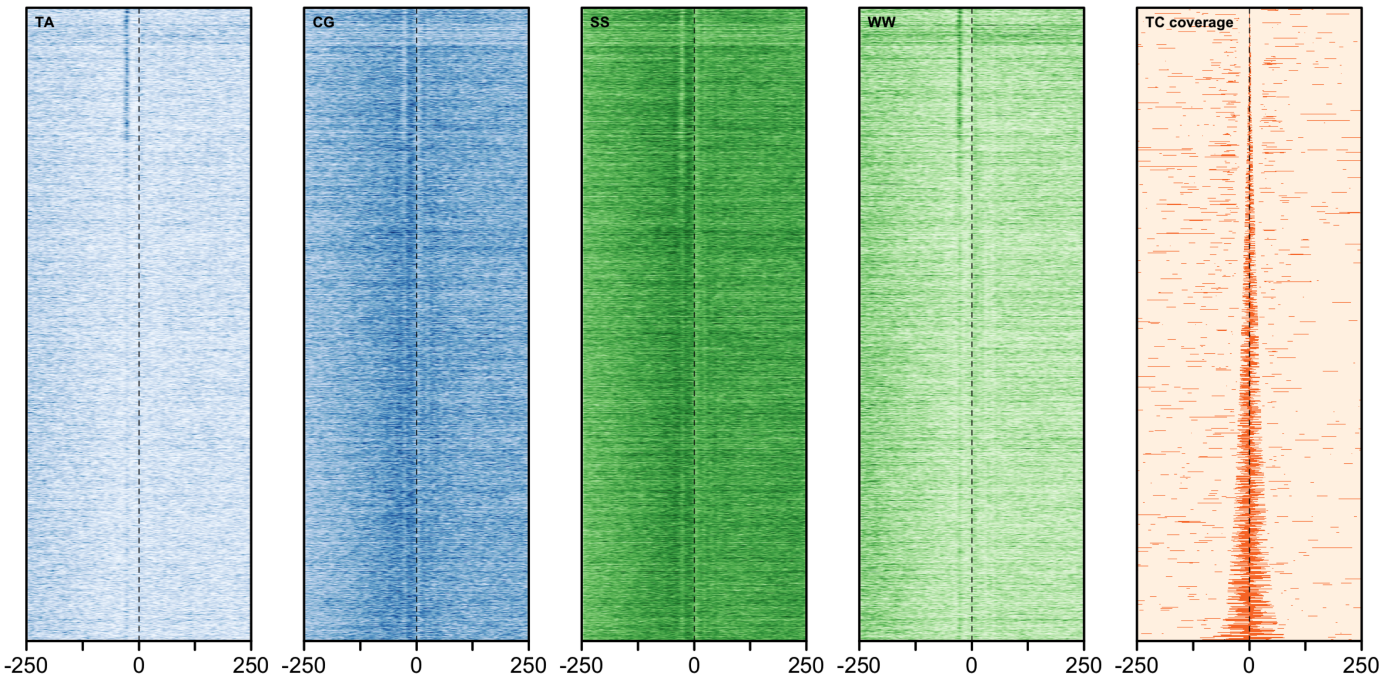
3.2 Dinucleotide heatmaps & pattern occurance

ChIPSeeker package allows to determine the locations of TF binding sites or specific histone modifications genome-wide and produce annotations for the TSS dataset (i.e. to which genomic location each TSS corresponds). Annotated peaks can further be filtered (e.g. Promoters <= 1kb from the TSS). Filtering step considerably decreases the number of clusters (~43%).

The Heatmaps package is used for plotting data. Heatmaps are produced for different nucleotide patterns and smoothed using Gaussian blur. Coverage heatmaps are produced to observe the distribution of iq_width.

Code

Heatmaps for clusters assigned to promoter regions (<=1kb) + TC coverage (K562 cell line)

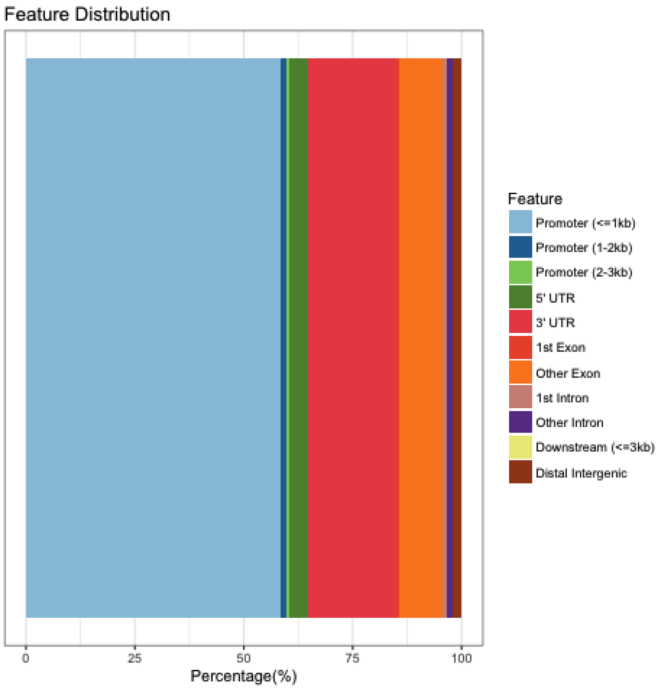


3.3 POLII collection of TF motif heatmaps

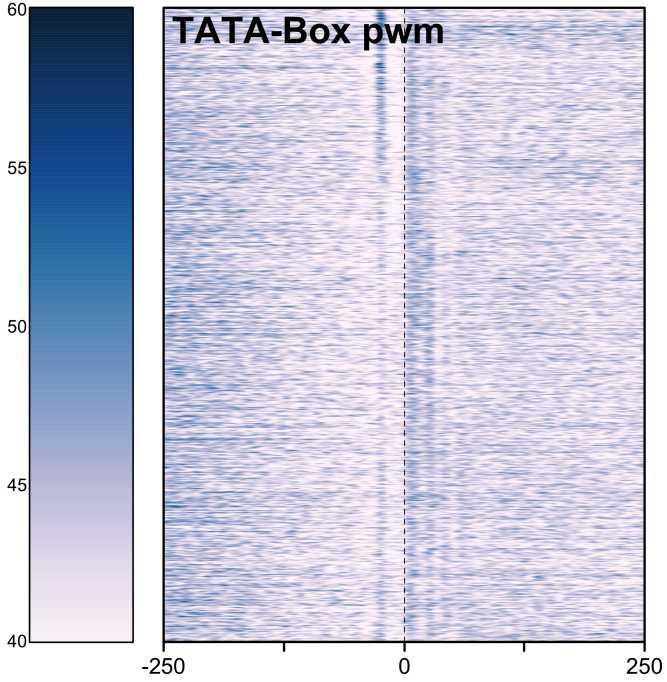
Transcription factor binding sites are extracted from the JASPAR database and converted to position-weight matrices.

Code

Annotation proportions from JASPAR. The majority of calls represent promoter sequence



Heatmap represneting the occurance of TATA-box motif within proximal promoter regions (<= 1kb).



3.3.1 kmeans

Apply k-means clustering to the heatmaps produced above (partition into 2 groups). This leads to reduction of noise and better partition of promoter sequences. Each heatmap was clustered based on the positions where the respective motif is found.

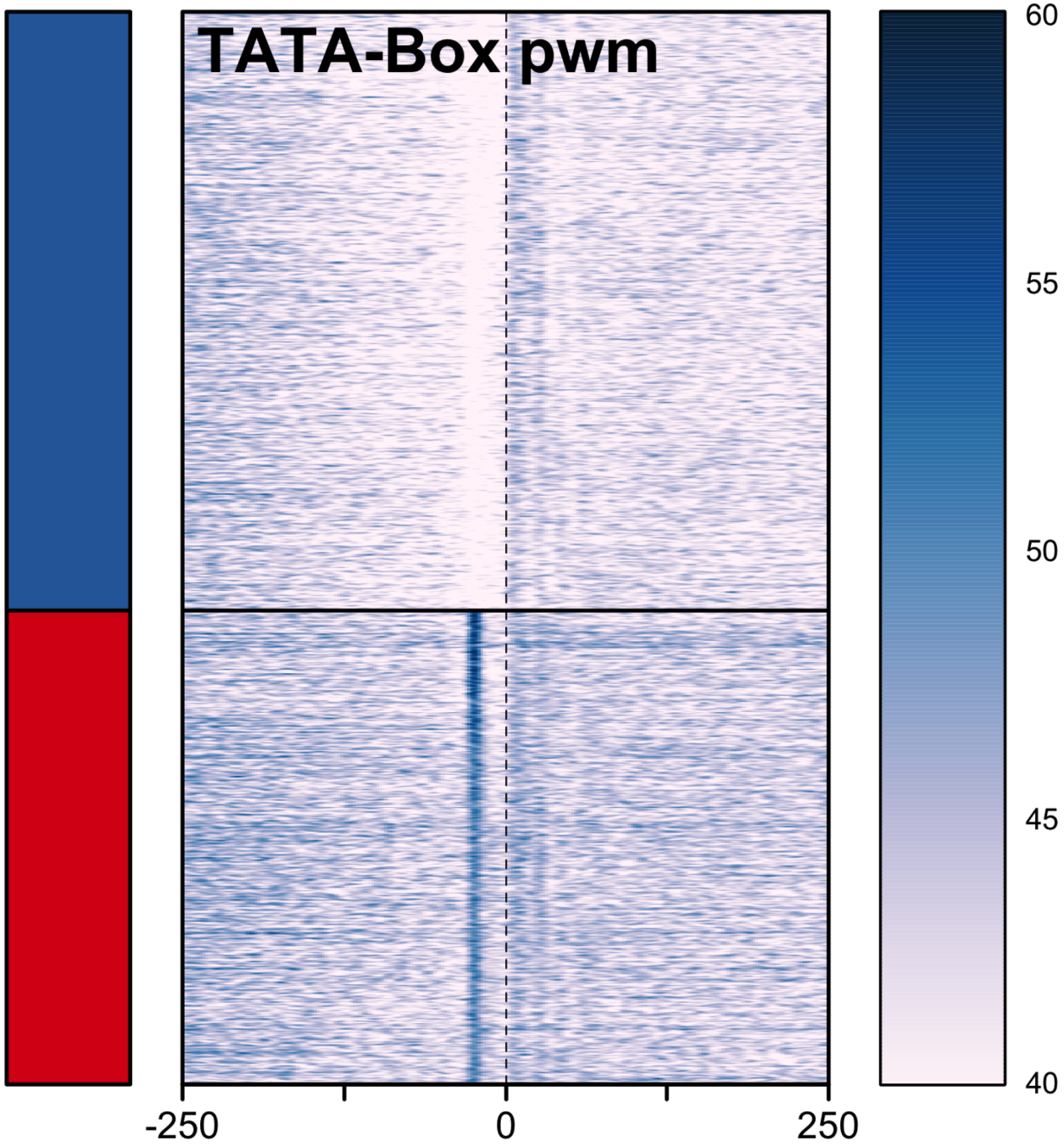
Code

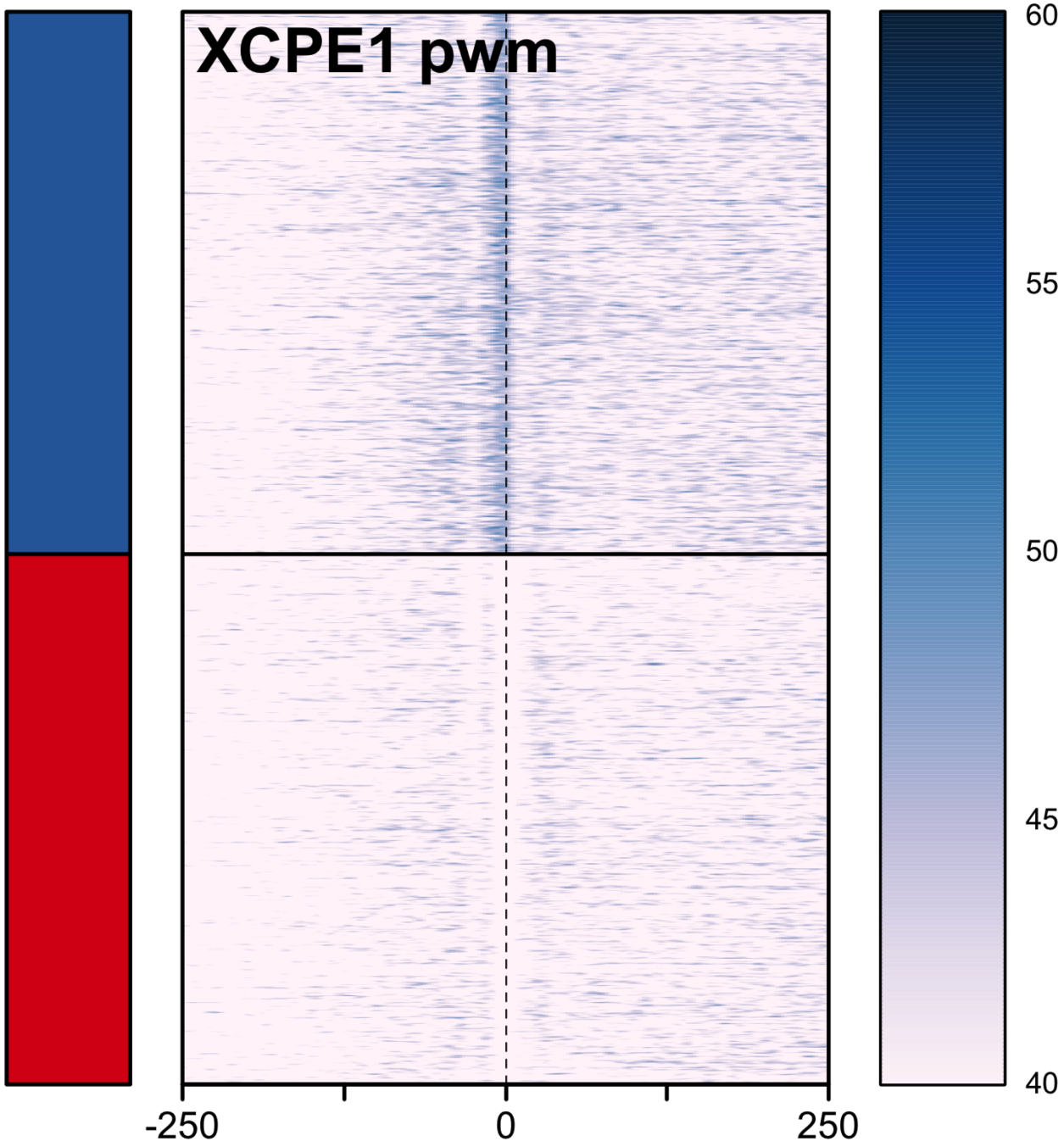
Percentage of sequences that contain a given motif (at 85% match) across cell types

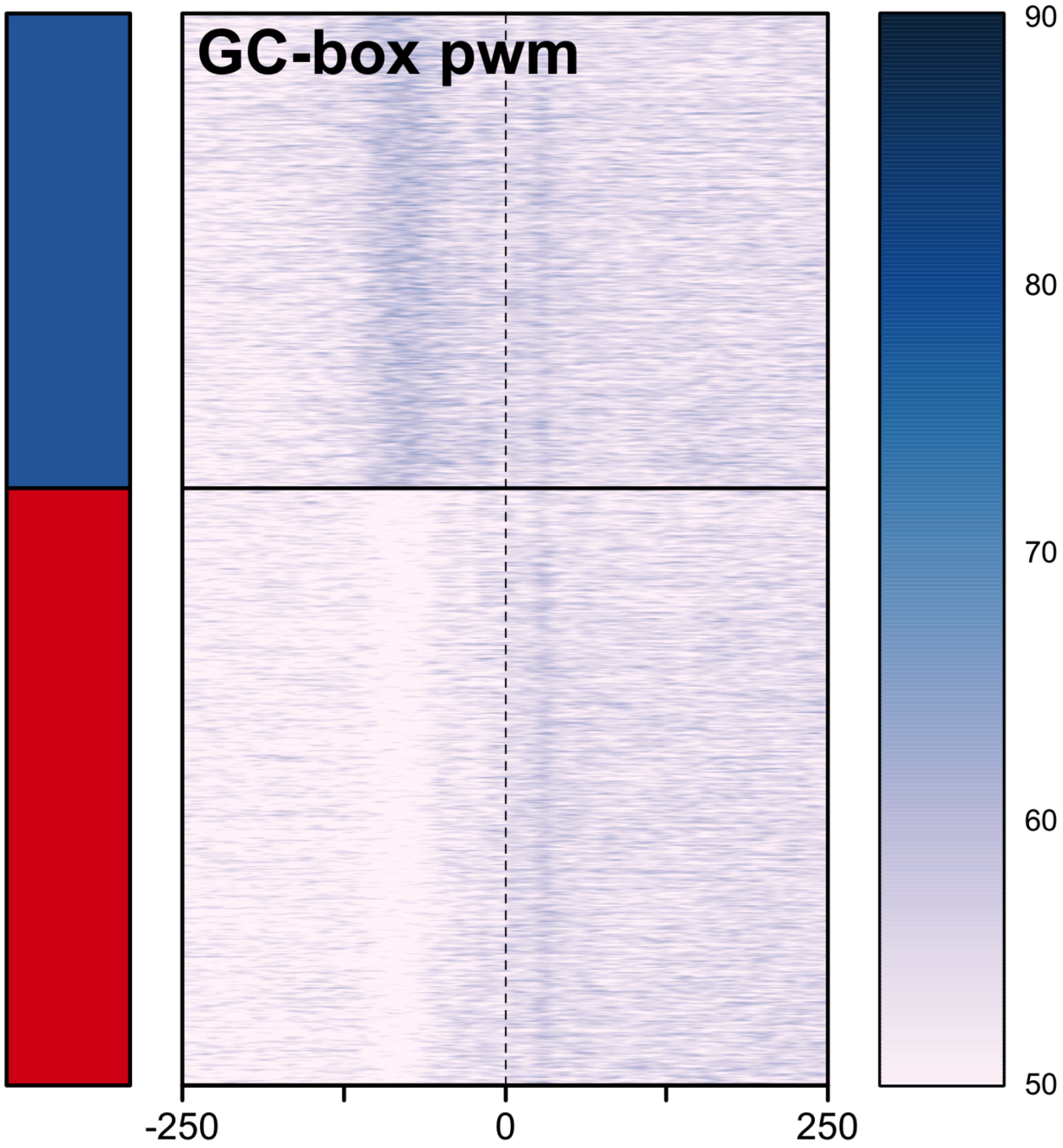
Code

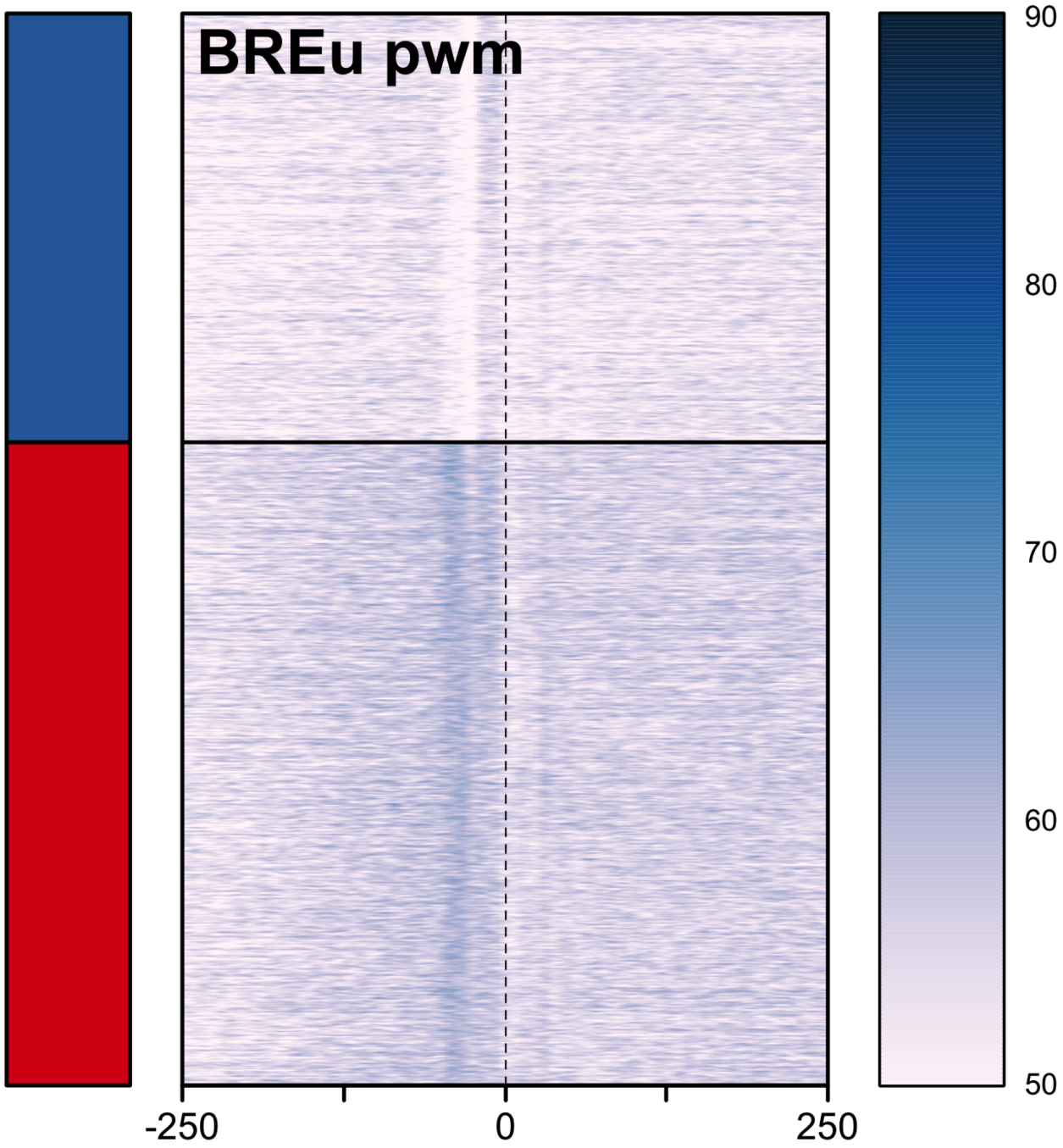
```
##            K562_cell HepG2_cell A549_cell GM12878_cell MCF-7_cell_rep1
## MTE                0          0         0            0               0
## INR               27         28        28           27              28
## GC-box            44         45        45           44              45
## CCAAT-box         10         10        10           10              10
## DPE               17         16        18           18              16
## BREu              42         40        41           40              41
## BREd              26         26        25           27              25
## DCE_S_I            7          7         8            7               8
## DCE_S_II           8          8         8            8               8
## DCE_S_III         16         15        15           15              16
## XCPE1              2          2         2            2               2
## TATA-Box           4          4         4            4               4
## MED-1             17         16        17           16              17
```

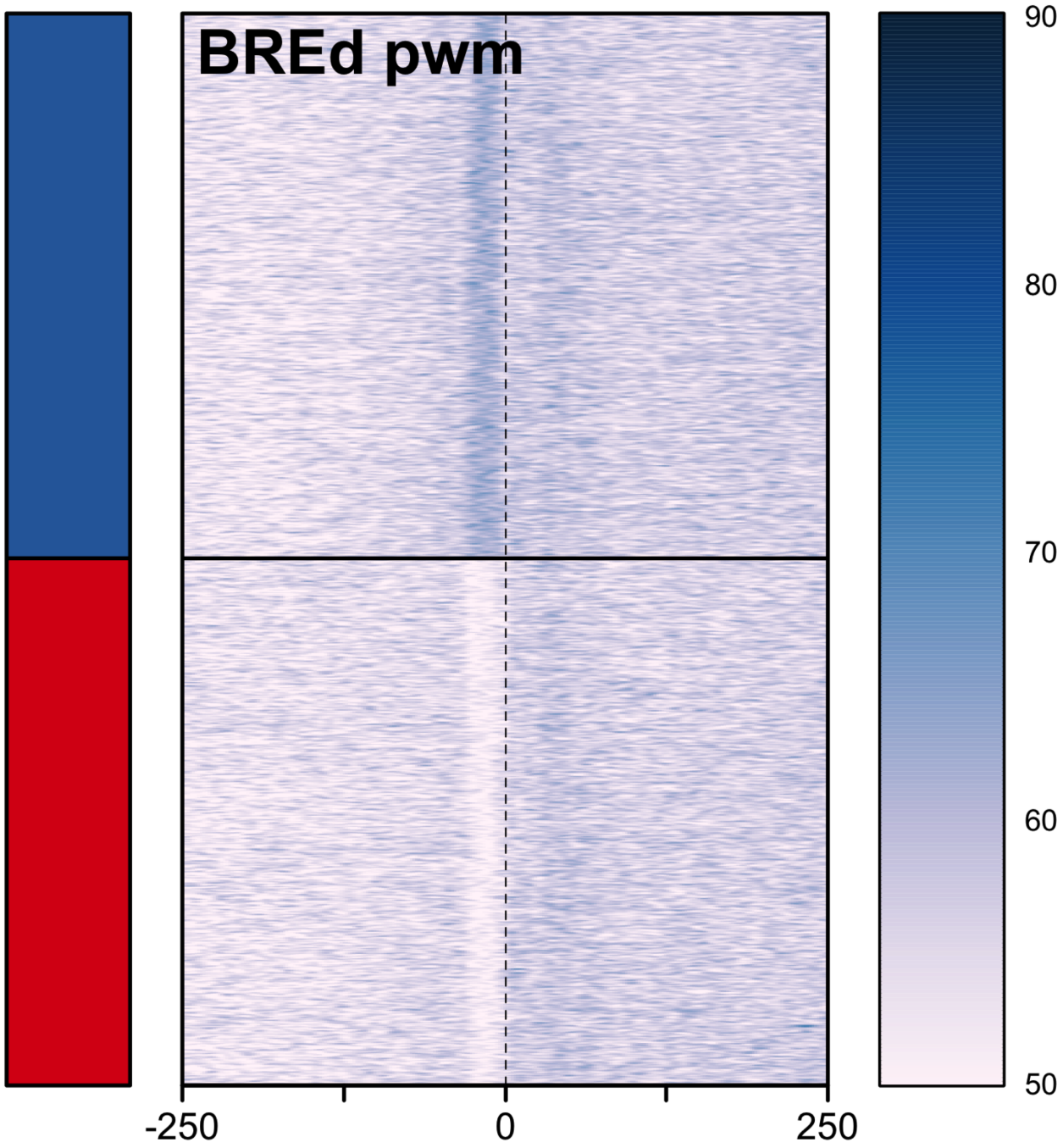
Clustered heatmaps


TATA-Box pwm
60
55
50
45
40
-250
0
250


XCPE1 pwm
60
55
50
45
40
-250
0
250


GC-box pwm
90
80
70
60
50
-250
0
250


BREu pwm
90
80
70
60
50
-250
0
250



3.3.2 add motif data to GRanges object

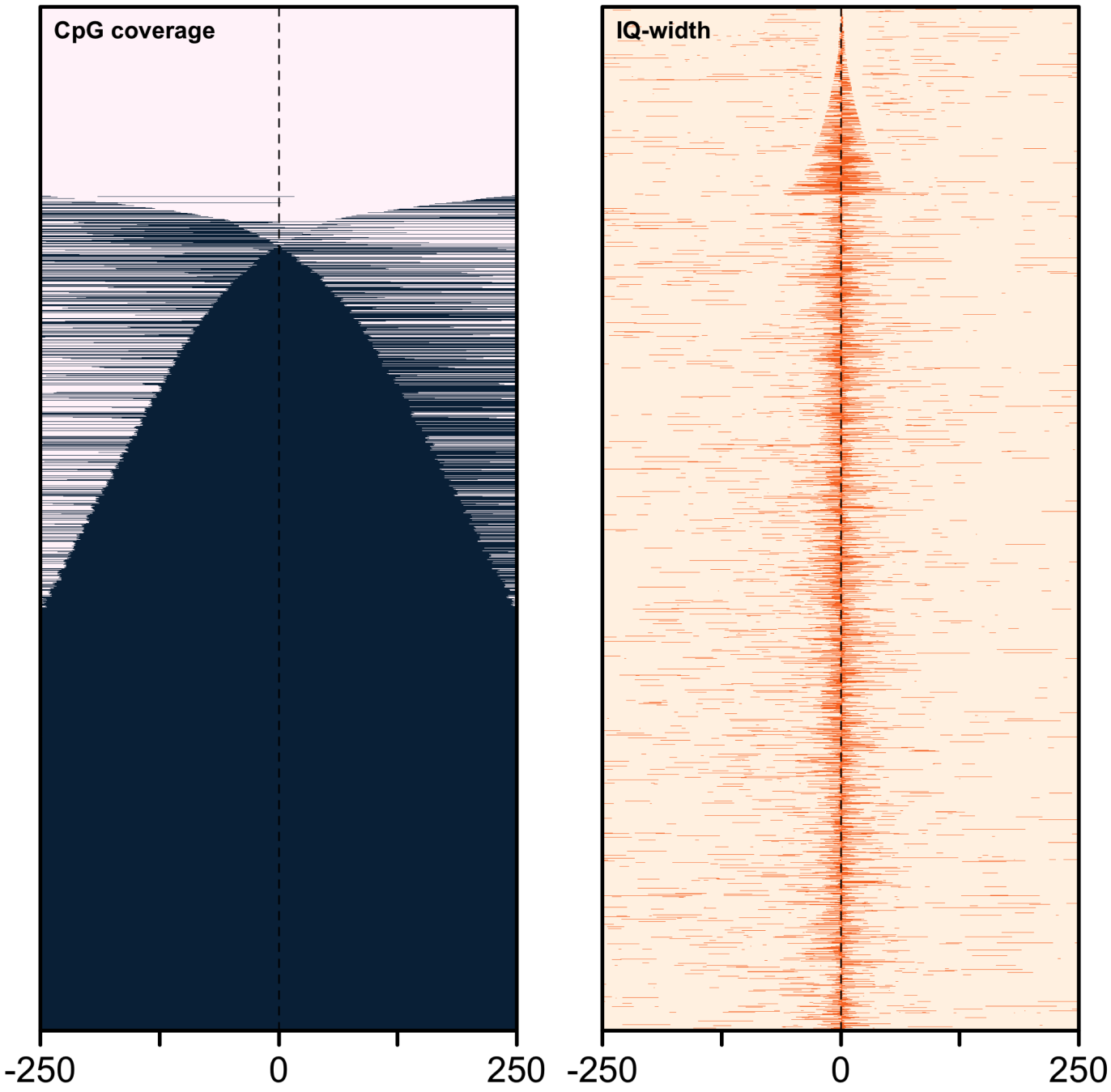
Data are stored as a matrix of seqs x motifs with boolean values indicating whether a given motif is found in a sequence with >85% match.

Code

3.4 CpG island analysis

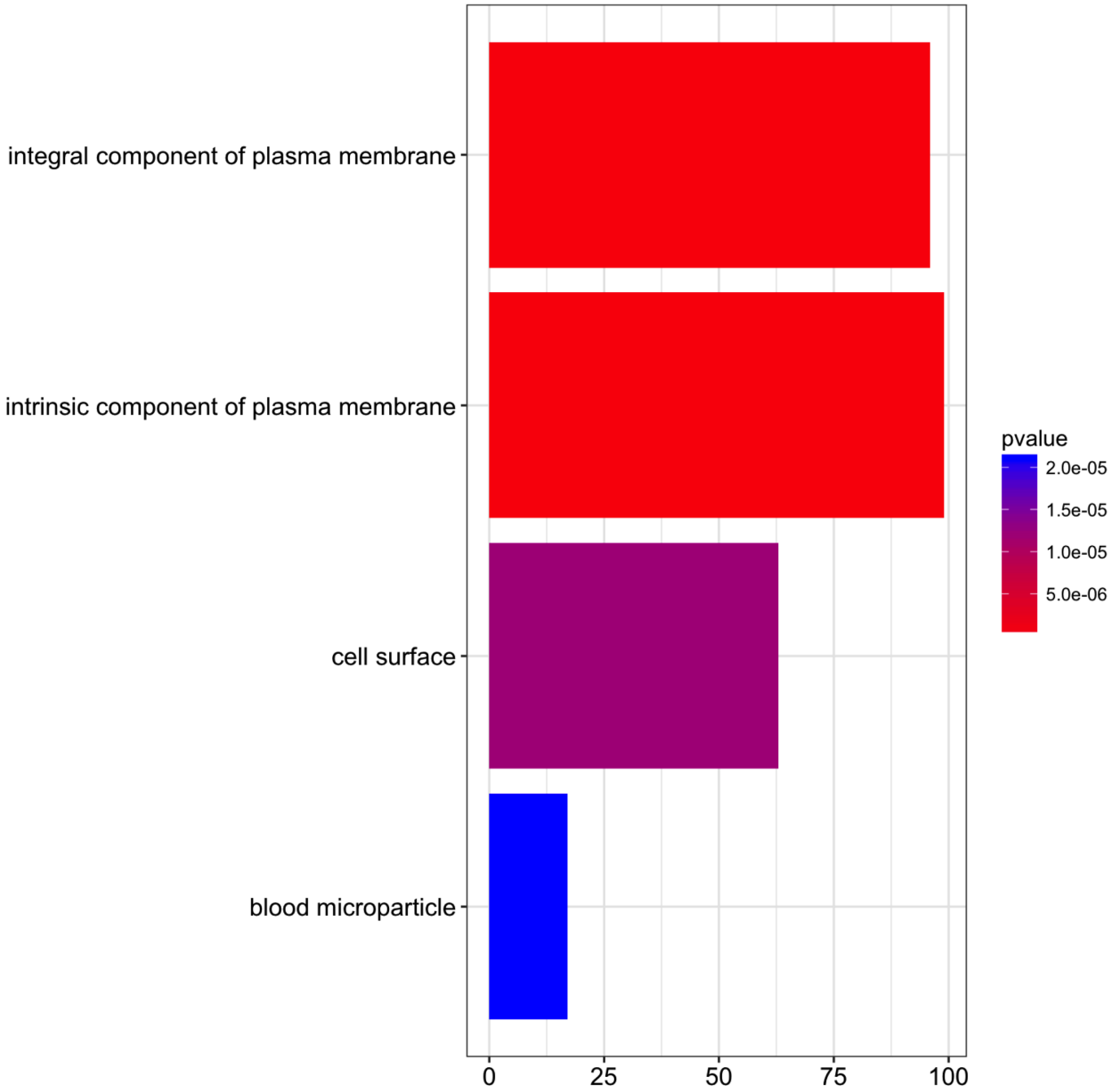
Code

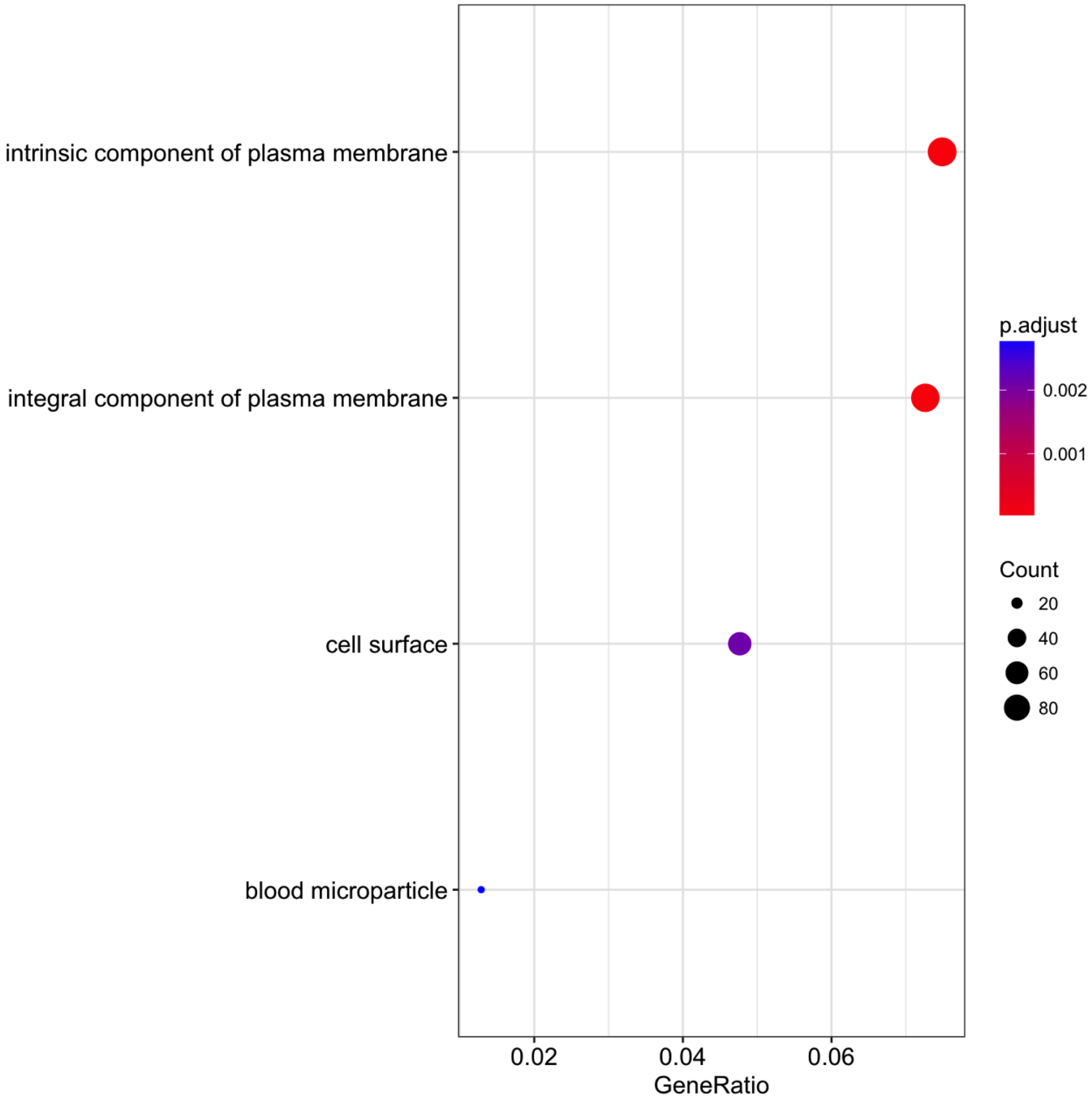
Coverage heatmap for CpG islands and IQ-width for K562 cells (sorted by degree of overlap) Notably, around 85% of promoters appear to have CpG islands and the heatmap indicates that they are centered at TSS.



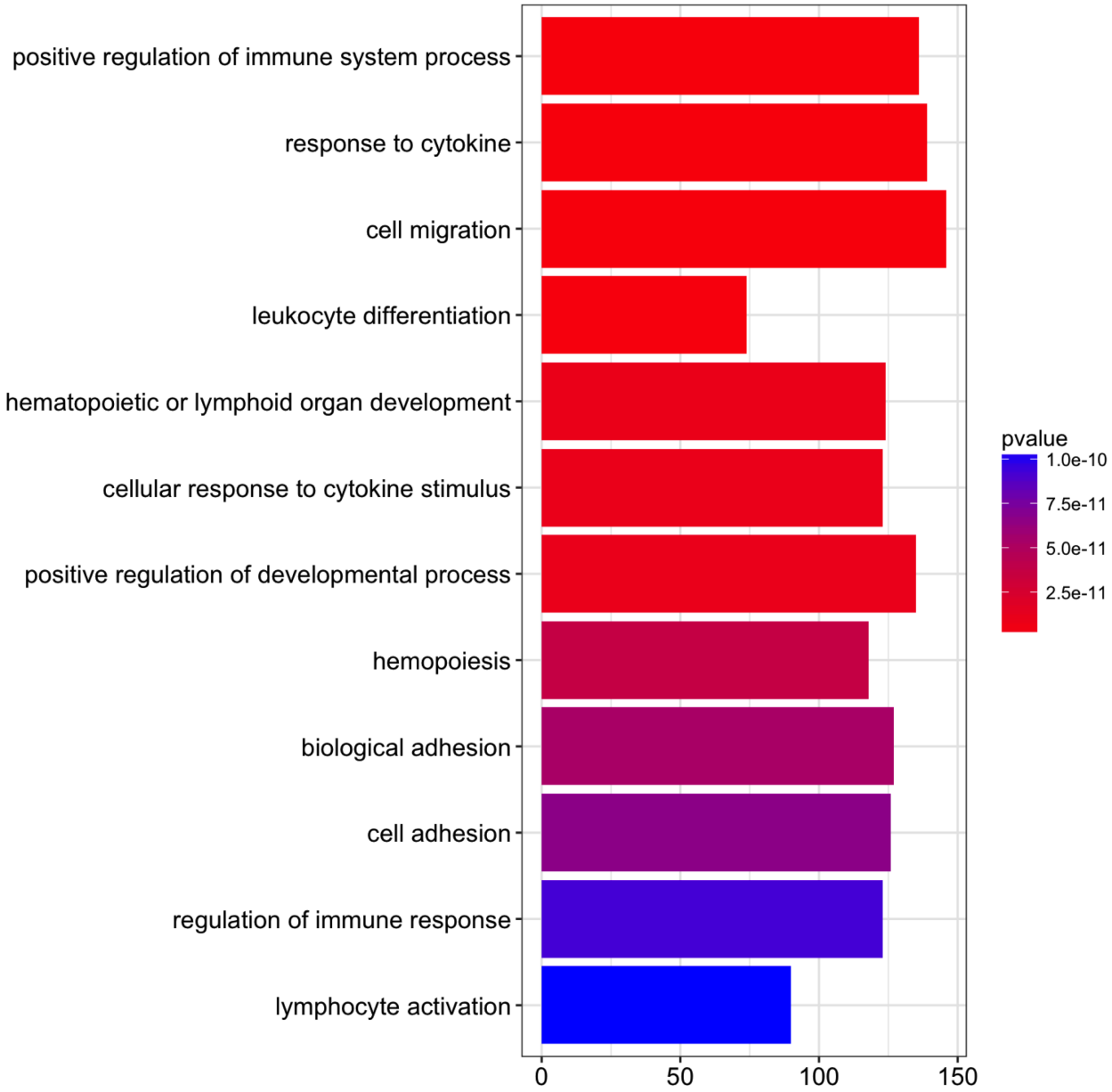
Gene onthology analysis on the subset of promoters without CGI

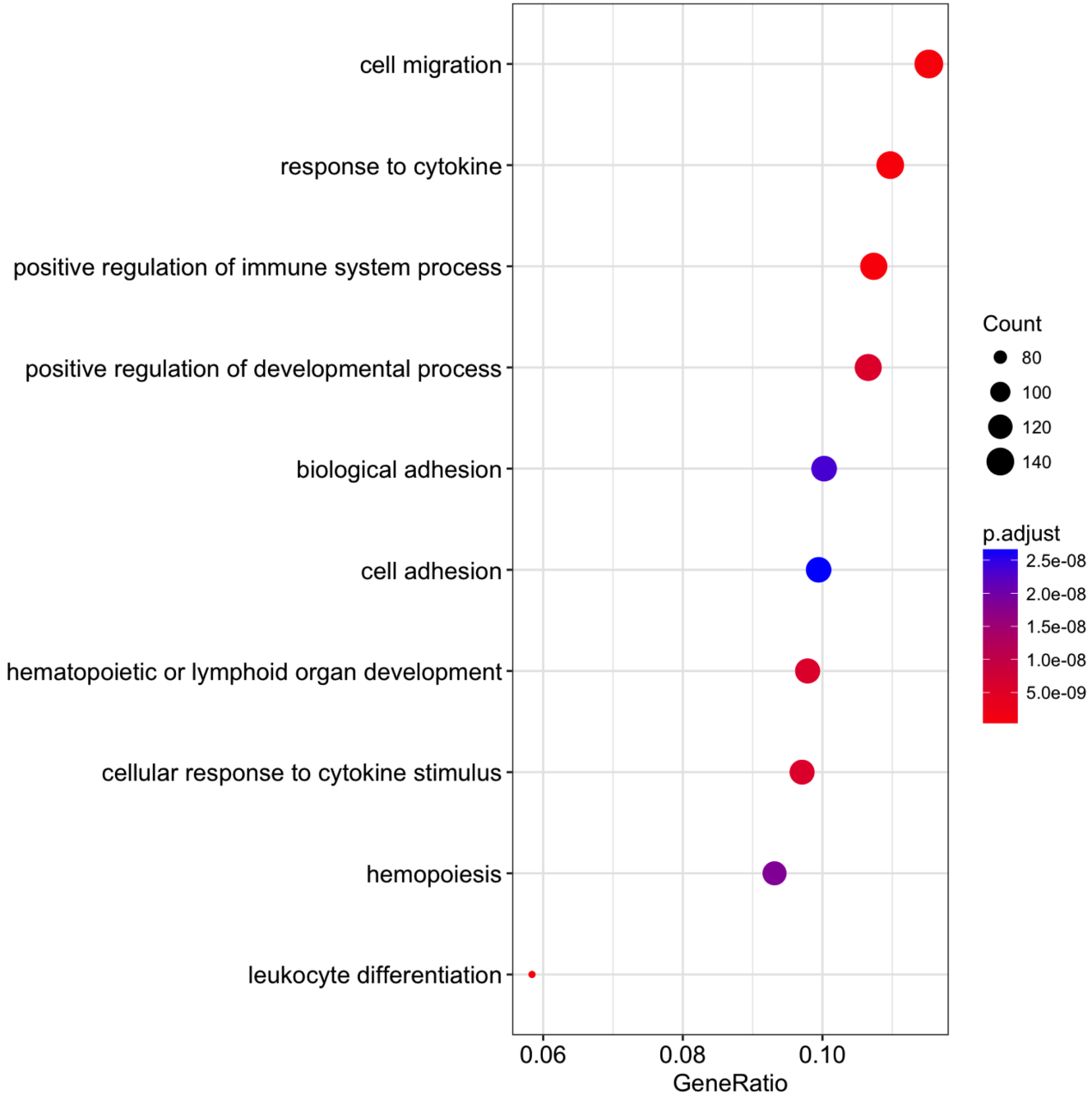
- Cellular Compartment


integral component of plasma membrane
intrinsic component of plasma membrane
cell surface
blood microparticle
0
25
50
75
100
pvalue
2.0e-05
1.5e-05
1.0e-05
5.0e-06



- Biological process


positive regulation of immune system process
response to cytokine
cell migration
leukocyte differentiation
hematopoietic or lymphoid organ development
cellular response to cytokine stimulus
positive regulation of developmental process
hemopoiesis
biological adhesion
cell adhesion
regulation of immune response
lymphocyte activation
0
50
100
150
pvalue
1.0e-10
7.5e-11
5.0e-11
2.5e-11



3.5 All human promoter motifs

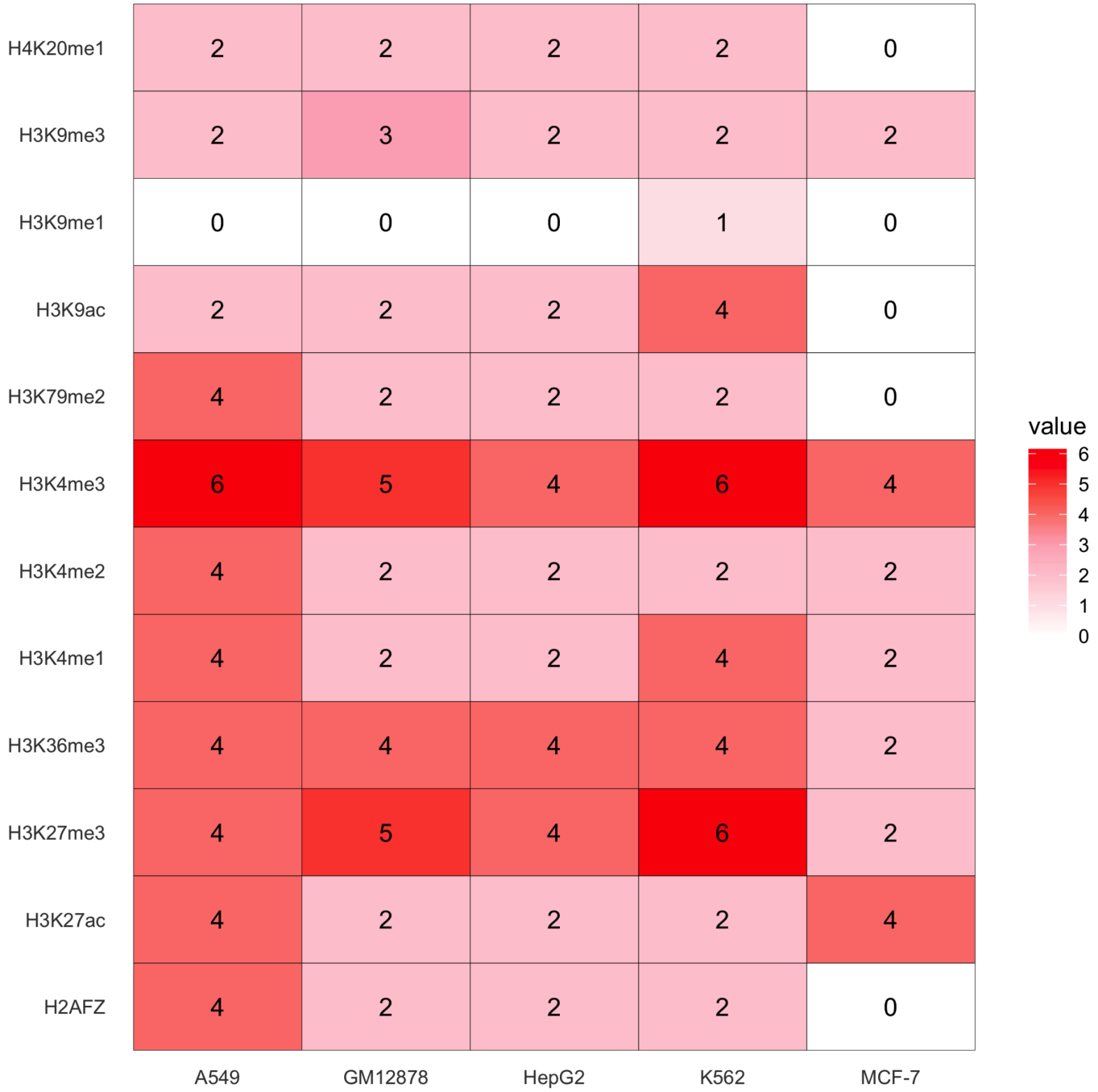
Code

3.6 Chromatin modifications

Extract all available chromatin marks (.bam alignment files) for given cell lines from the ENCODE website. Process data into format compatible with the STAN package input. Promoter sequences are extended to 1 kb upstream/downstream TSS and split into 100-nt bins. For each bin, number of overpals with ChIP-Seq reads (GAlignment objects; 36-nt window) is calculated. For each cell line/mark combination, a matrix is produced (seqs x bins). The matrices are then combined into 3-dimensional array and sliced into n matrices (bins x marks), where n is the number of promoters in a given cell line. The following script is computationally intense and should be run on a cluster.

Code

Heatmap of available histone marks v cell lines (number of files in ENCODE database)



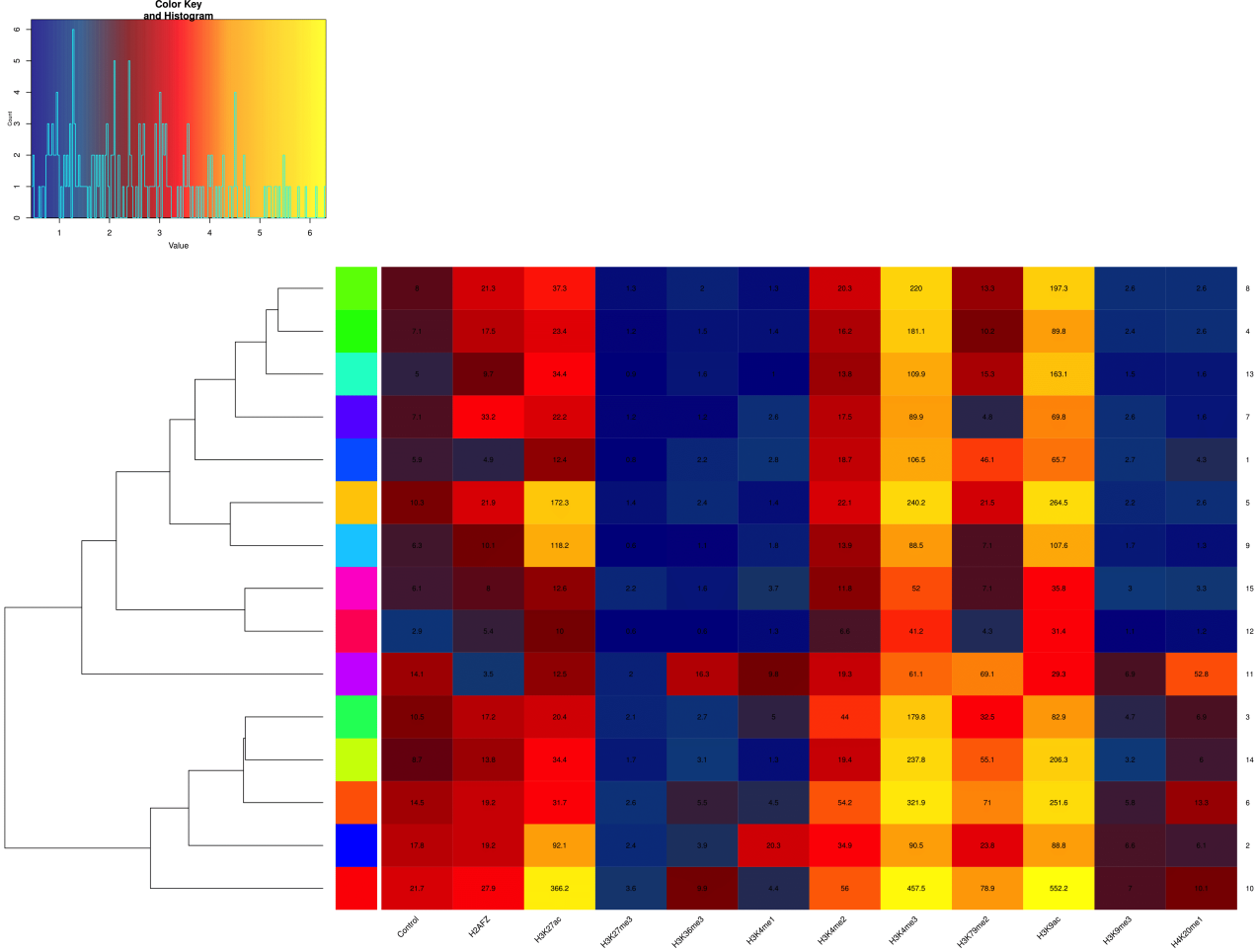
3.7 HMM via STAN package

The obtained matrices were used for unsupervised training of a Hidden Markov Model (HMM) which splits bins into a pre-defined number of states (i.e. a state is assigned to each bin). For each cell line, a separate set of models was produced. The STAN package allows to model crhomatin marks via different distributions, including Bernoulli, Poisson-LogNormal, and NegativeBinomial. The Bernoulli model was used for subsequent analysis, since it is quicker to train and signal gets binarised, reducing the amount of noise in the data. Models with varying number of states were produced, and optimal number of states was determined via assessing log-likelihood values.

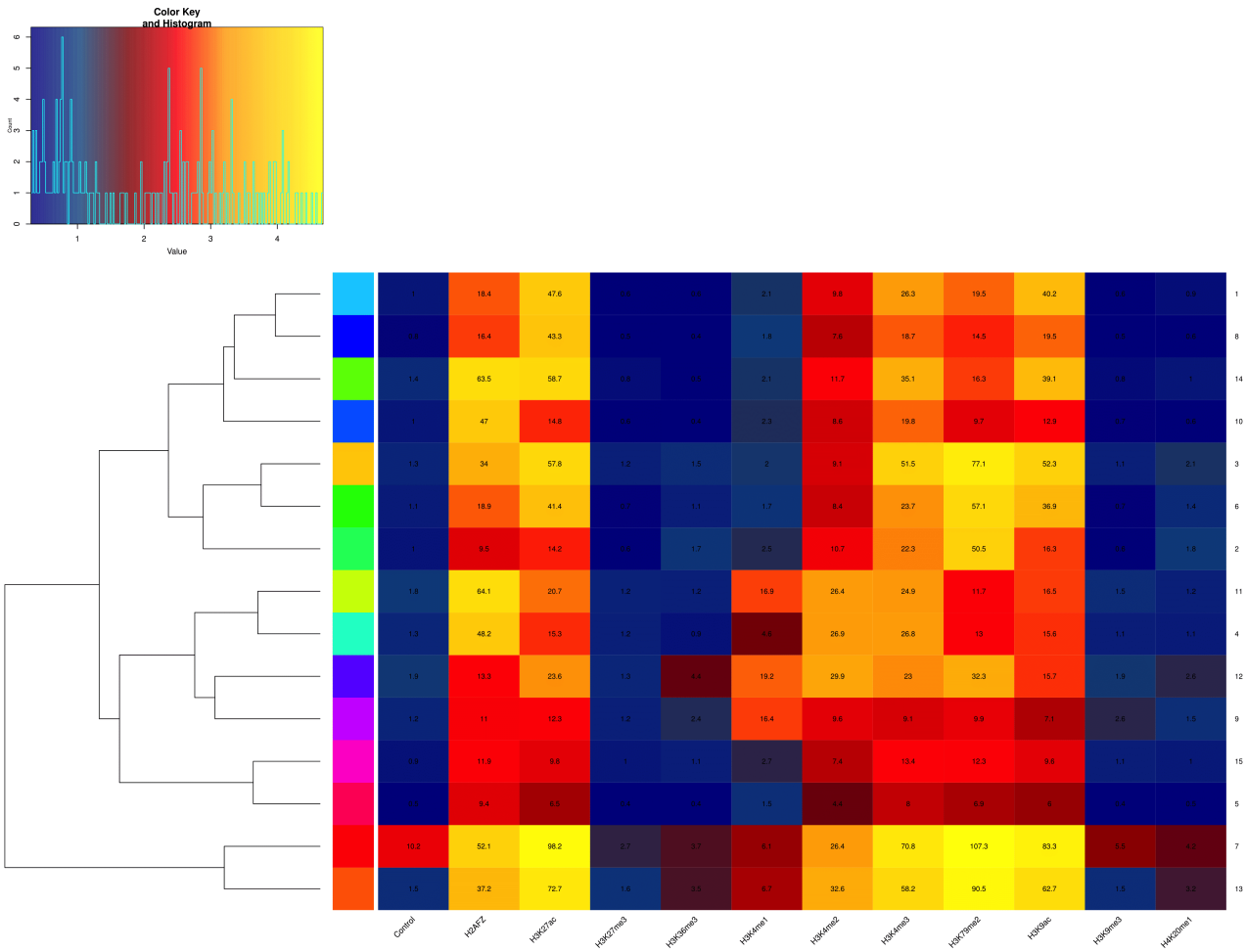
After model fitting, the decoding step produced vectors of most likely states for each promoter (50 values per sequence) via the Viterbi algorithm. The state paths were also converted to GRanges objects for downstream analysis. In order to assign biologically meaningful functions to the states obtained from the model, k-means clustering algorithm was applied to the state vectors (i.e. promoters were grouped based on similarity of their Viterbi paths). Optimal number of promoter groups was determined via the elbow method. After clustring, genes from each promoter group were extracted and analysed through GO annotation.

Code

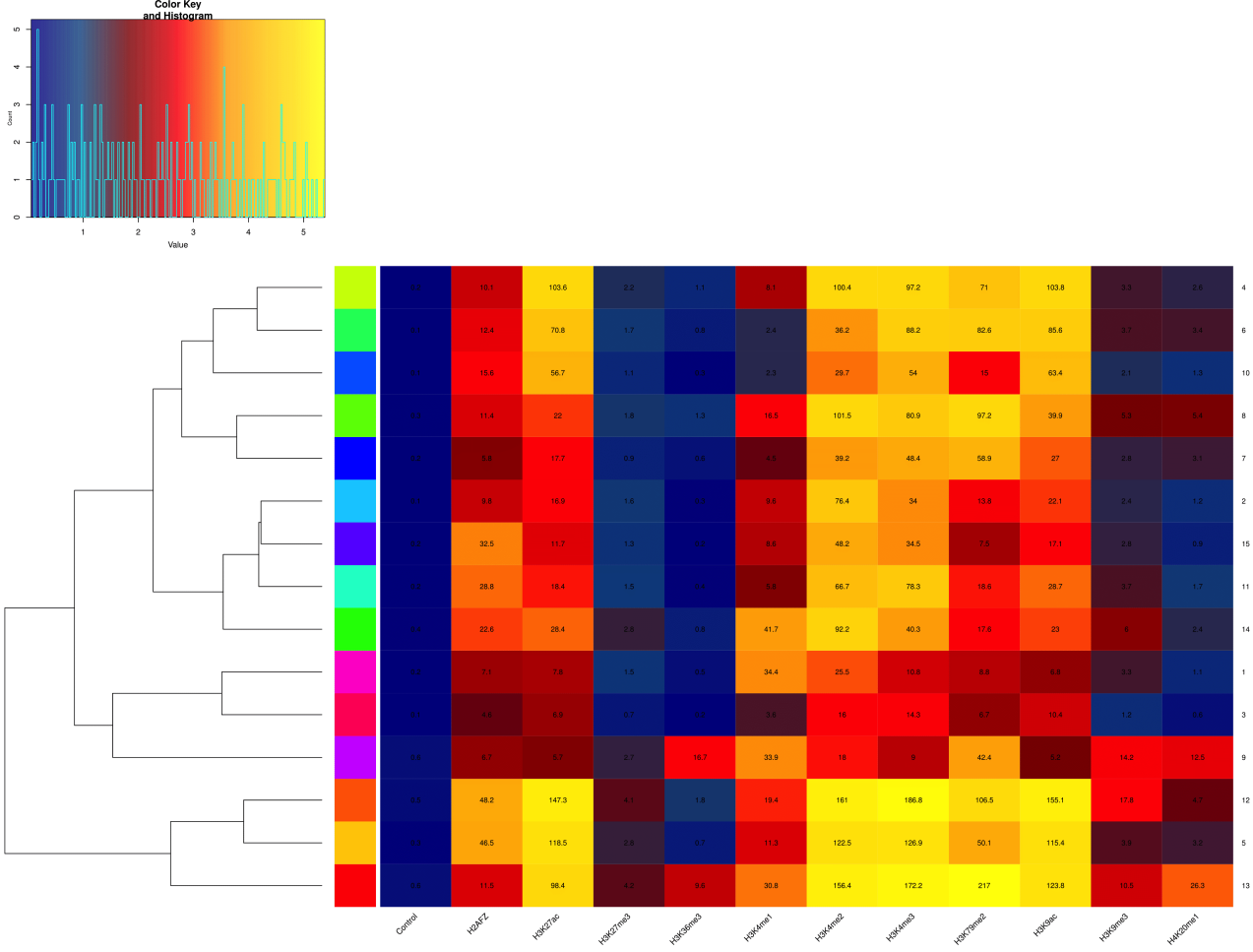
A549 heatmap (15 states):



GM12878:



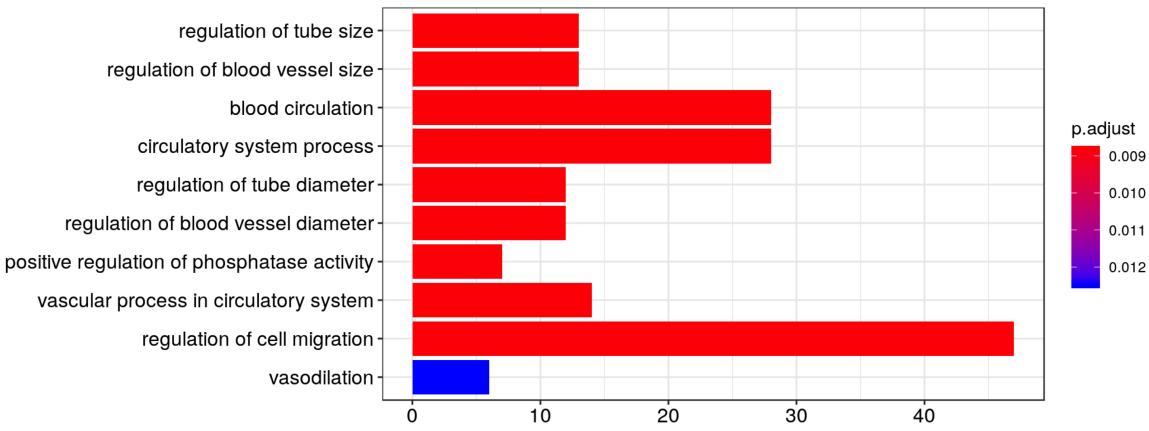
HepG2:



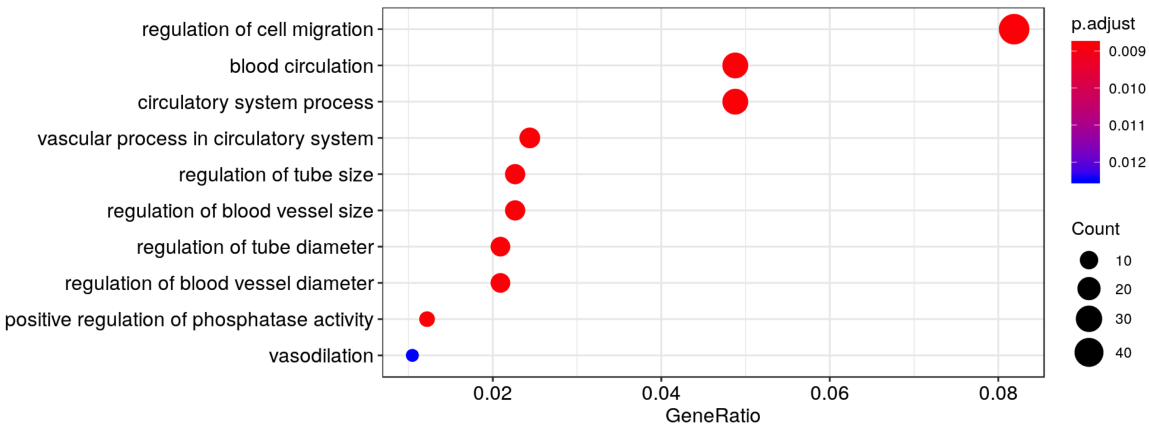
Enriched GO terms:

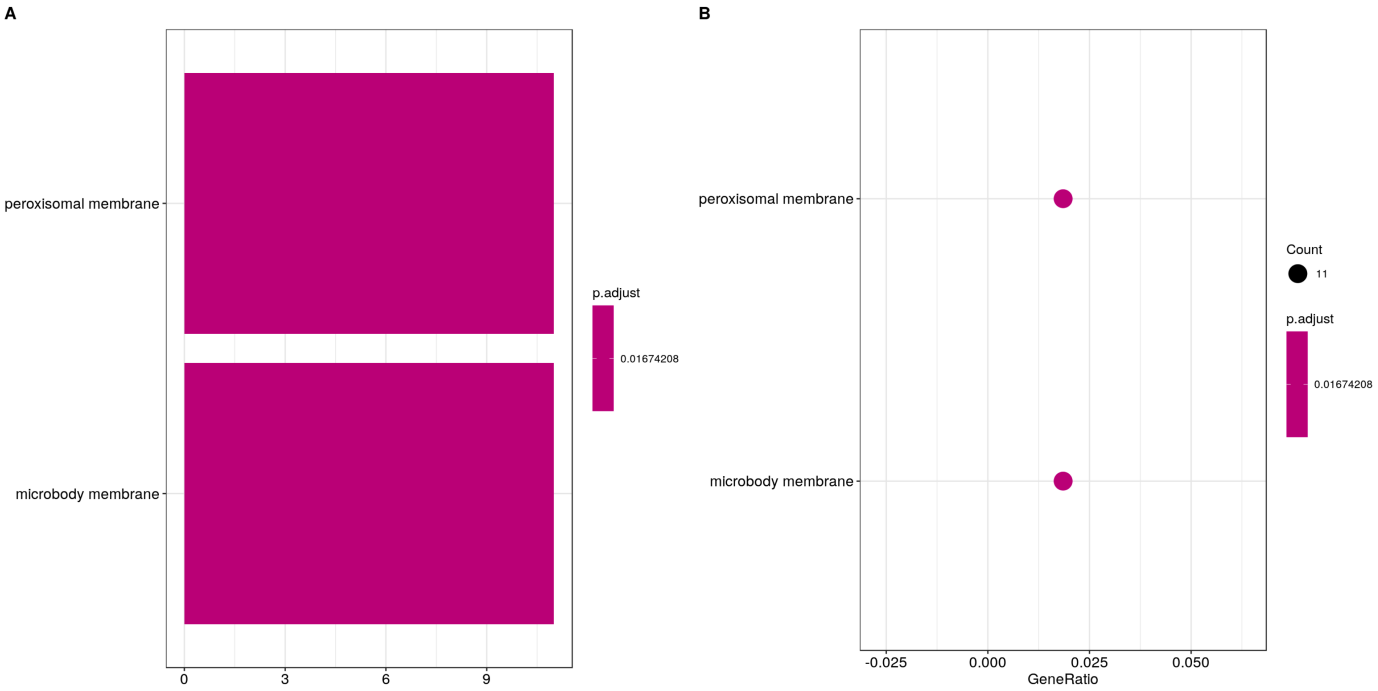
- A549 cell line, group 3

A

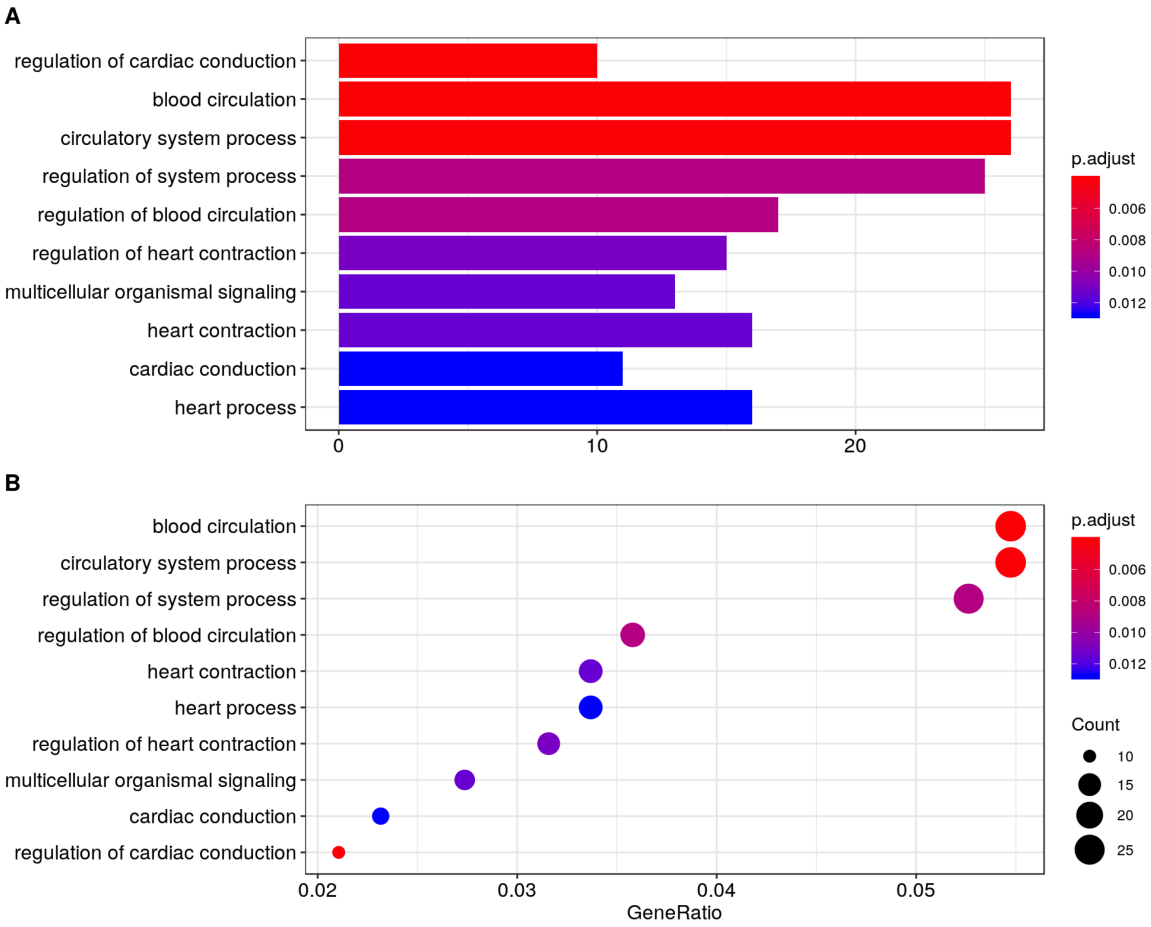


B

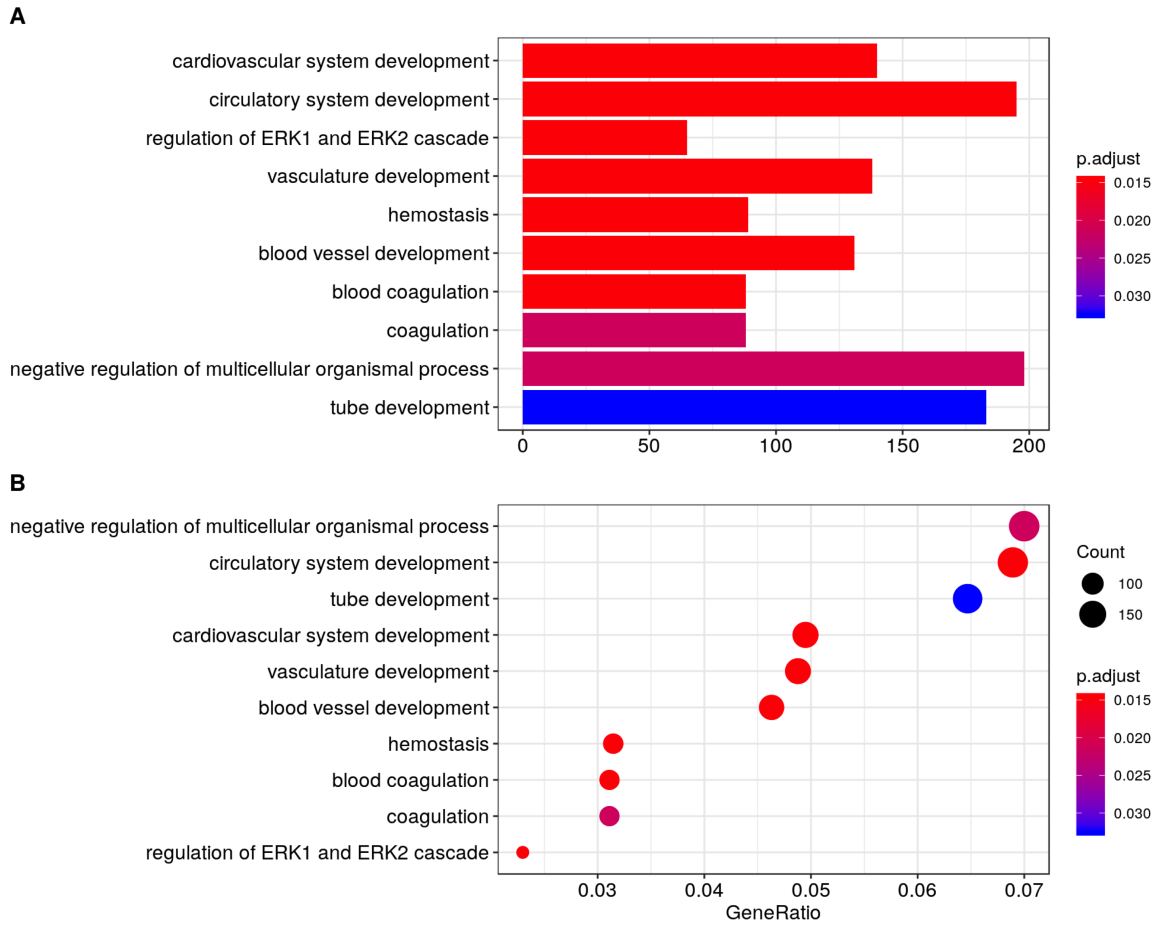


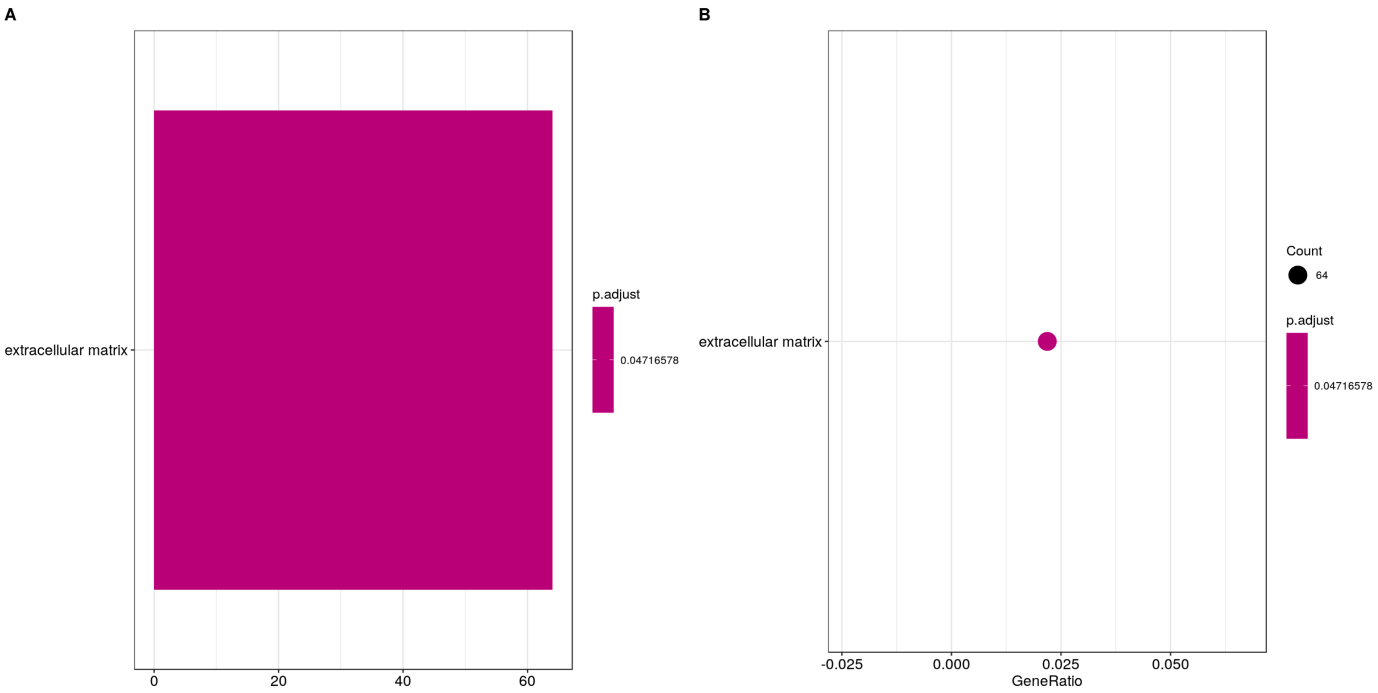


- A549, group 10

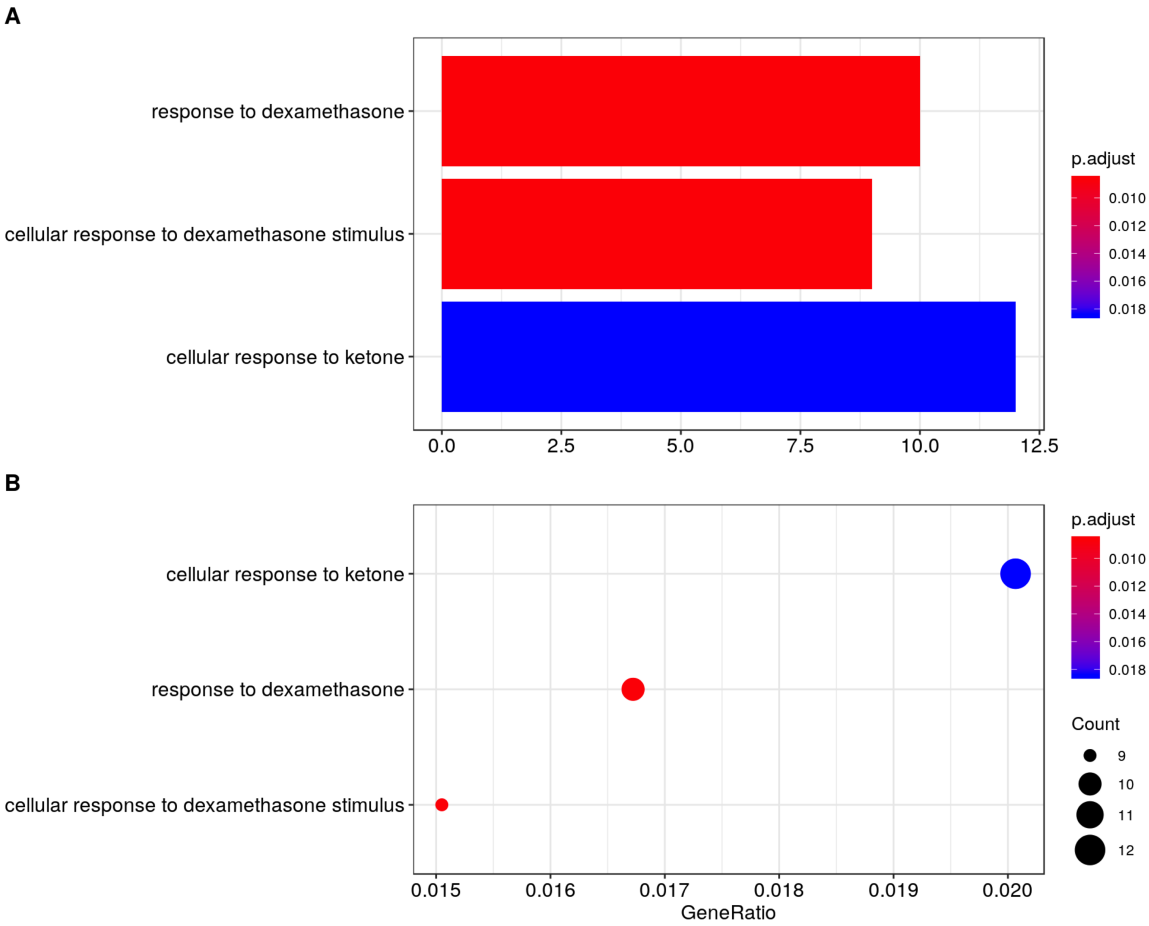


- HepG2, group 8


A
cardiovascular system development
circulatory system development
regulation of ERK1 and ERK2 cascade
vasculature development
hemostasis
blood vessel development
blood coagulation
coagulation
negative regulation of multicellular organismal process
tube development
0
50
100
150
200
p.adjust
0.015
0.020
0.025
0.030
B
negative regulation of multicellular organismal process
circulatory system development
tube development
cardiovascular system development
vasculature development
blood vessel development
hemostasis
blood coagulation
coagulation
regulation of ERK1 and ERK2 cascade
0.03
0.04
0.05
0.06
0.07
GeneRatio
Count
100
150
p.adjust
0.015
0.020
0.025
0.030


A
extracellular matrix
0
20
40
60
p.adjust
0.04716578
B
extracellular matrix
-0.025
0.000
0.025
0.050
GeneRatio
Count
64
p.adjust
0.04716578

- K562, group 8


A
response to dexamethasone
cellular response to dexamethasone stimulus
cellular response to ketone
0.0
2.5
5.0
7.5
10.0
12.5
p.adjust
0.010
0.012
0.014
0.016
0.018
B
cellular response to ketone
response to dexamethasone
cellular response to dexamethasone stimulus
0.015
0.016
0.017
0.018
0.019
0.020
GeneRatio
p.adjust
0.010
0.012
0.014
0.016
0.018
Count
9
10
11
12