

Motif Enrichment **without** Gene Expression Filtering

Description of results presented here.

In this document we show ALL enriched TFs based on our model, independent of its expression status.

We analyzed 3 possible motif enrichment scenarios. Each scenario composed of 2 genomic regions (up vs down OR hyper vs hypo) giving a total of 6 regions tested*:

- *Regulatory Region of Up- vs Downregulated genes*: This scenario is composed of the regulatory regions associated to upregulated and downregulated genes according to the definitions of such by Danilo (filter1 = p-value; filter2 = fold change). This input was generated solely with regard to the gene expression data. The length of each region is the exact length of the regulatory regions**.

- *Regulatory Region of Genes Associated with Hyper vs Hypo Meth.*: This scenario (not performed in the previous analysis) includes the regulatory regions** associated with the genes associated with hypermethylated and hypomethylated CpGs. The CpGs were retrieved from Danilo's table using his definitions. The length of each region is the exact length of the regulatory regions.

- *Hyper vs Hypo Methylated CpGs*: This scenario is composed of the hypermethylated vs hypomethylated CpG sites. These sites were retrieved from Danilo's table using his definitions. All CpGs were extended by 20bp both upstream and downstream (all regions were 40bp in length).

* All six regions were tested against a random background model for the final enrichment test.

** Regulatory regions = promoter (RefSeq annotation) and strong enhancers (annotated with *in situ* Hi-C).

*** All regulatory regions were further filtered using DNase-seq (wild-type MSCs).

Each scenario has two versions (in green at the end of each title):

- **V1**: All plots with "V1" correspond to the TFs where at least one of the two conditions has a corrected enrichment p-value < 0.05.

- **V2**: All plots with "V2" correspond to further filtering the TFs in V1 by keeping only the TFs in which in at least one of the conditions there was a TF with p-value > 0.05. V2 is always a subset of V1.

For each combination scenario x version, there are two slides with the results named **A-D** in red:

- **A**: Heatmap containing the enriched TFs for the scenario and version. P-values are given in $-\log_{10}$ scale (red colorkey). The gray zone in the colorkey corresponds to an enrichment p-value > 0.05 (not significant). The rows (TFs) of this heatmap were clustered using the hierarchical clustering with euclidean distance and Ward's method. This heatmap corresponds to the manuscript's Figure 4E.

- **B**: Differential heatmap. In this heatmap we subtract the $-\log_{10}$ pvalues from the two conditions [$(-\log_{10}(\text{p-value of condition1})) - (-\log_{10}(\text{p-value of condition2}))$]. In the color scale, darker blue corresponds to TFs associated to condition 1 (left column of heatmap in **A**); darker red corresponds to TFs associated to condition 2 (right column of heatmap in **A**). Gray values correspond to the TFs in which there was no more than one order of magnitude of difference between p-values. This "differential heatmap" is sorted.

- **C**: Barplot showing the frequency (x-axis) of foreground (input regions with motif; red) vs background (random regions with motif; blue) for each TF (y-axis) of **A**. The TF with lowest p-value were selected between the two regions. This graph is the same as manuscript's Figure 4F (without the motif logos).

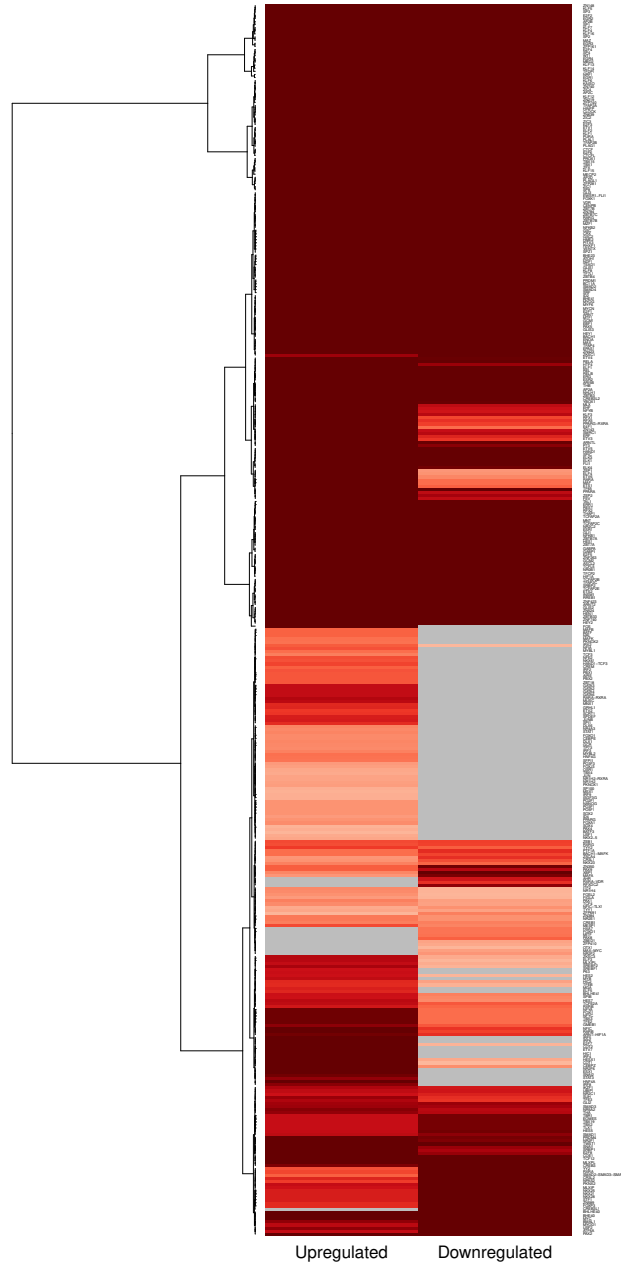
- **D**: Barplot showing the expression fold-change (x-axis) of TFs (y-axis) in **A**. The colorkey corresponds to the expression p-value intensities in $-\log_{10}$ scale. Vertical dashed lines were drawn at the cutoff fold-changes of 0.5 (blue) and 2 (green).

Regulatory Region of Up- vs Downregulated genes (V1)

A



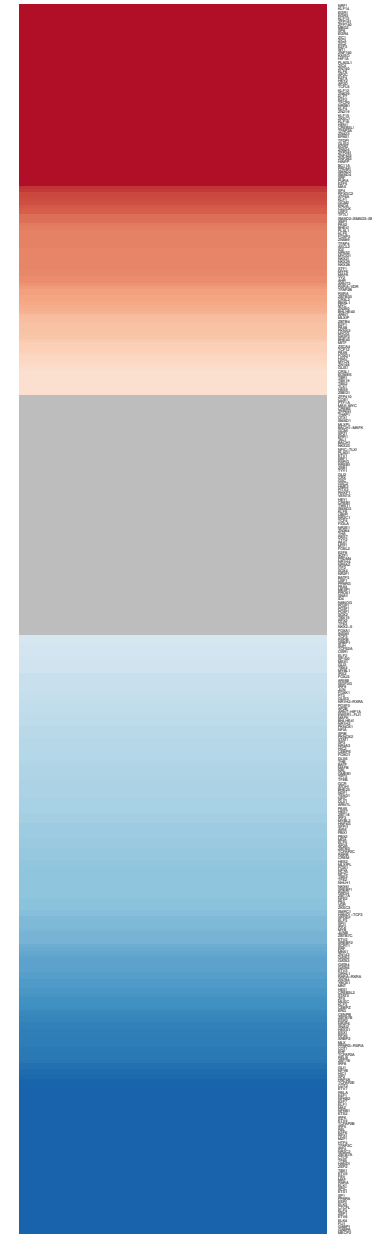
Up- and Down-regulated Genes (V1)
 $-\log_{10}(\text{p-value})$



B

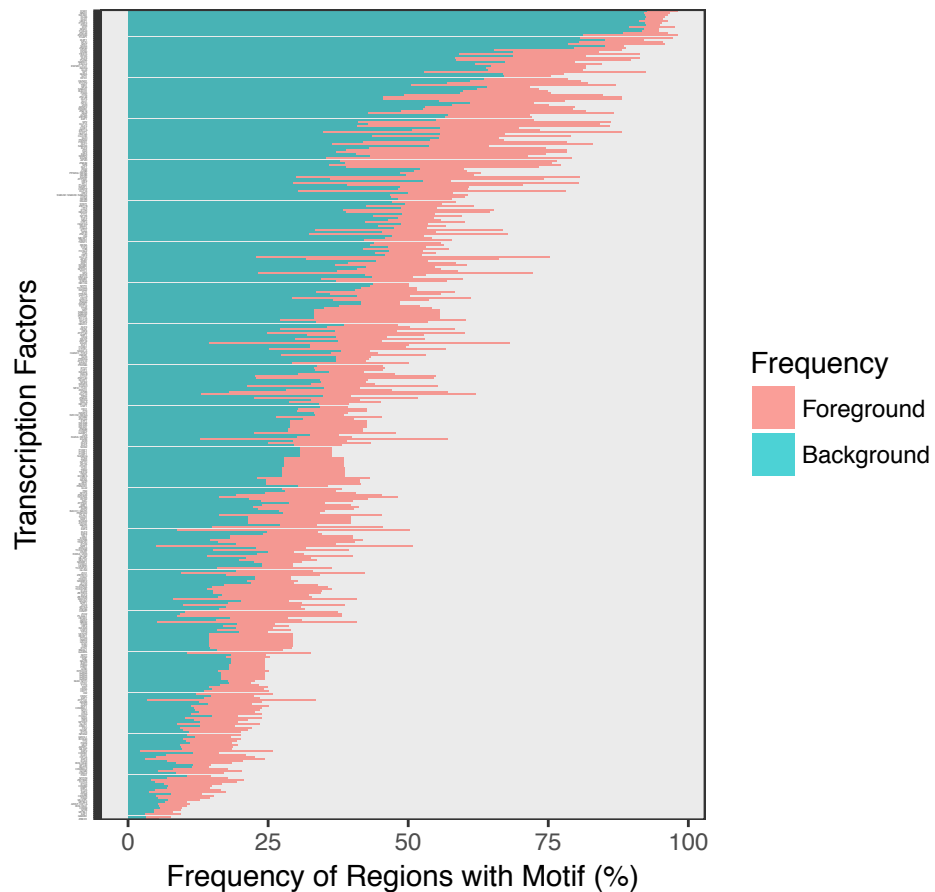


Up- and Down-regulated Genes (V1)
 $(-\log_{10}(\text{p-value up})) - (-\log_{10}(\text{p-value down}))$

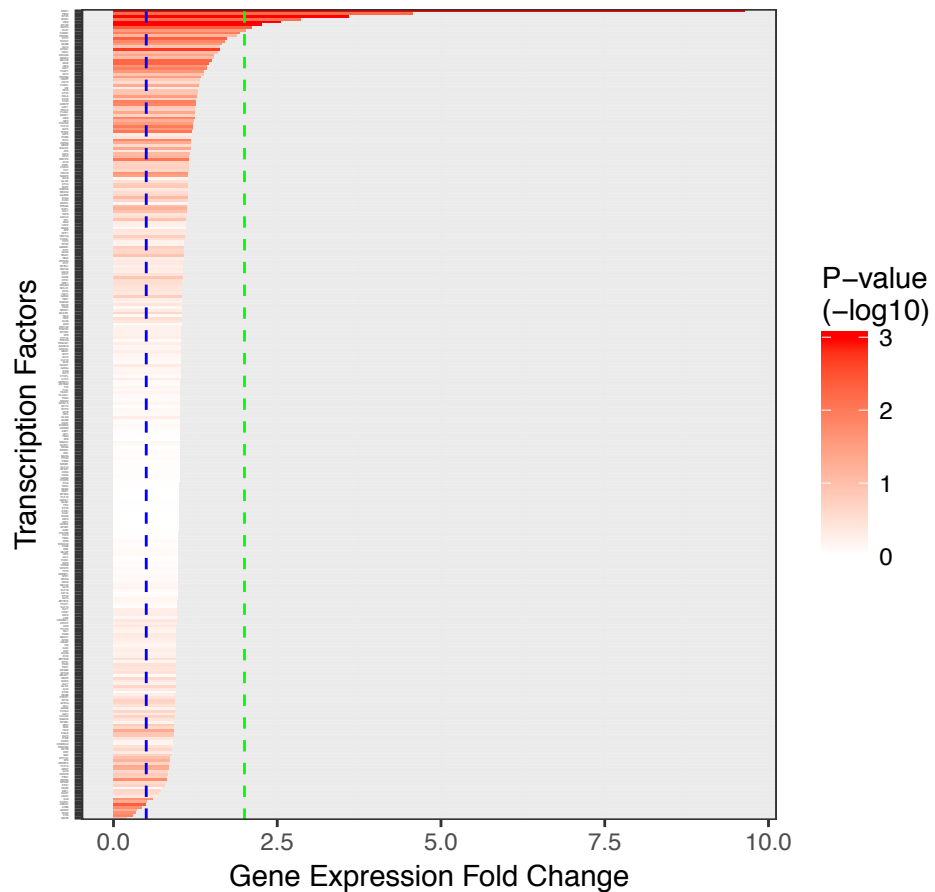


Regulatory Region of Up- vs Downregulated genes (V1)

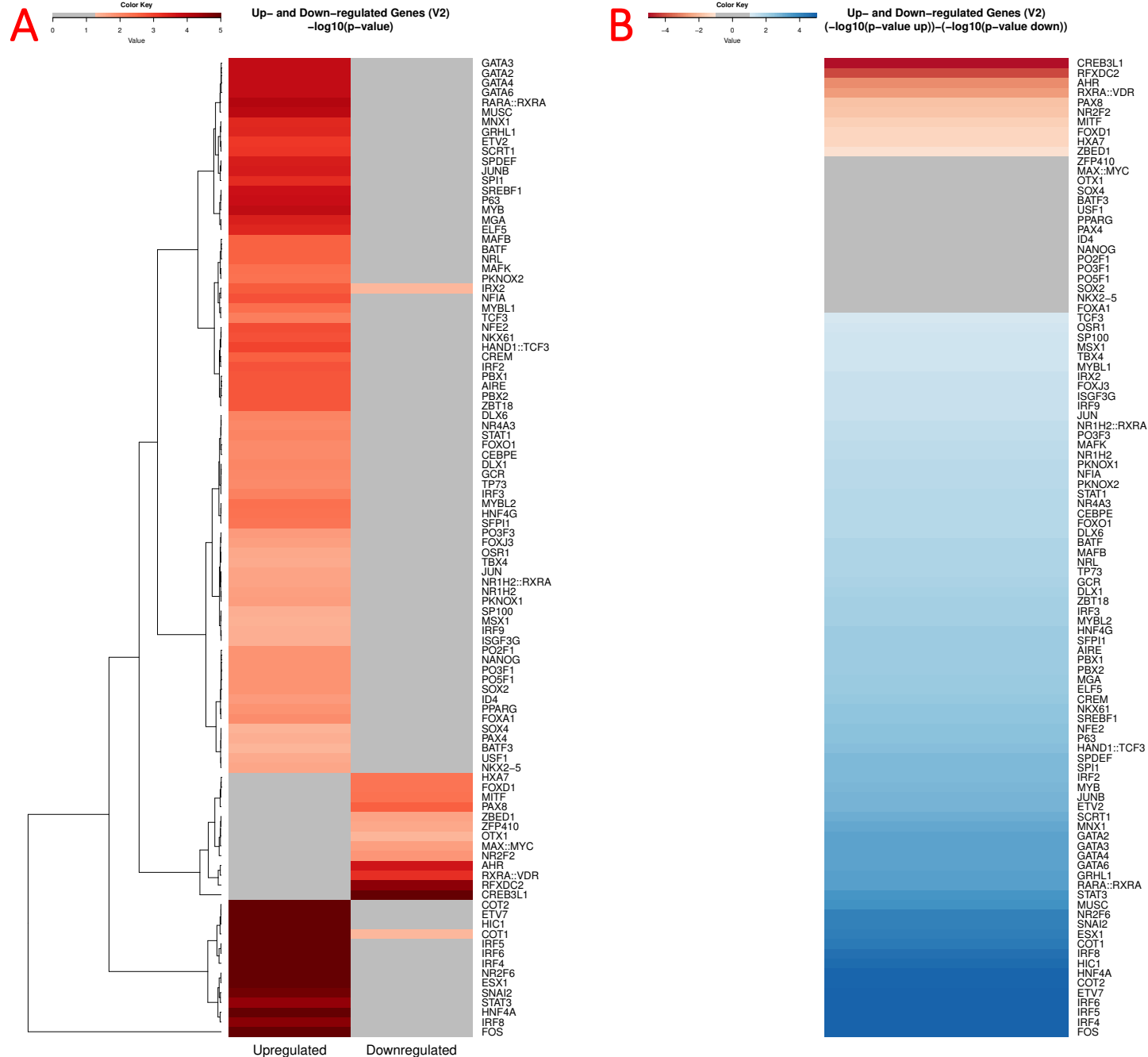
C Enriched Transcription Factors' Regions' Frequency



D Enriched Transcription Factors' Gene Expression

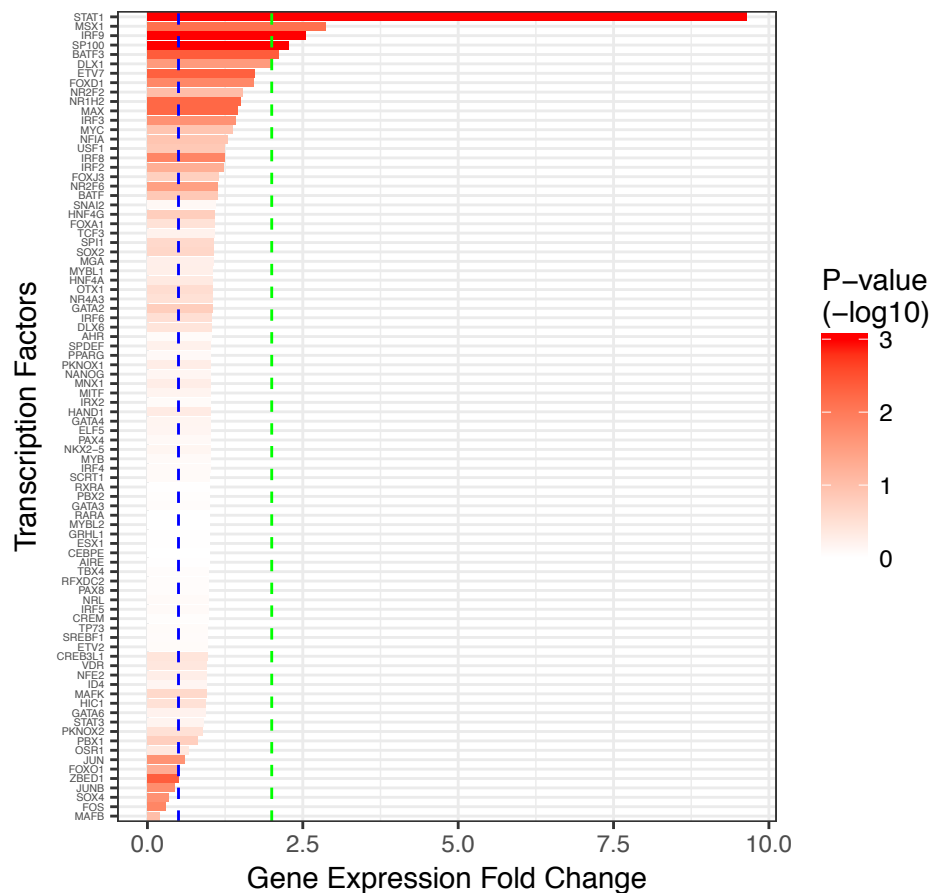


Regulatory Region of Up- vs Downregulated genes (V2)

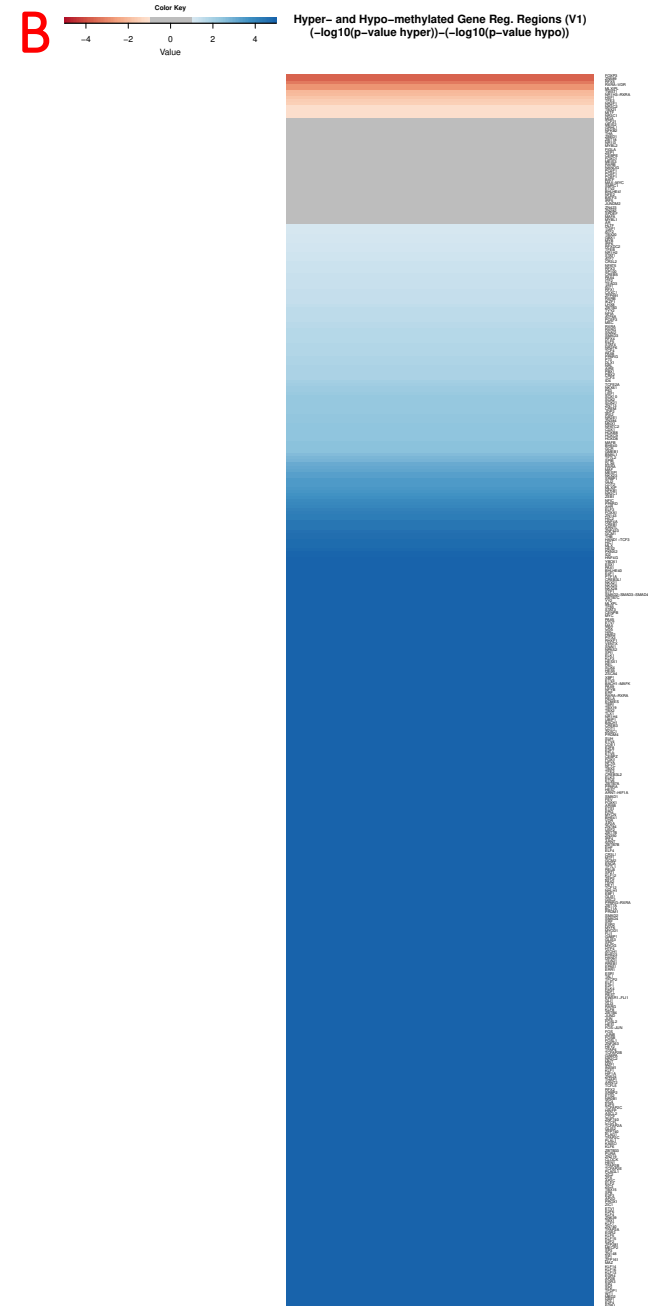
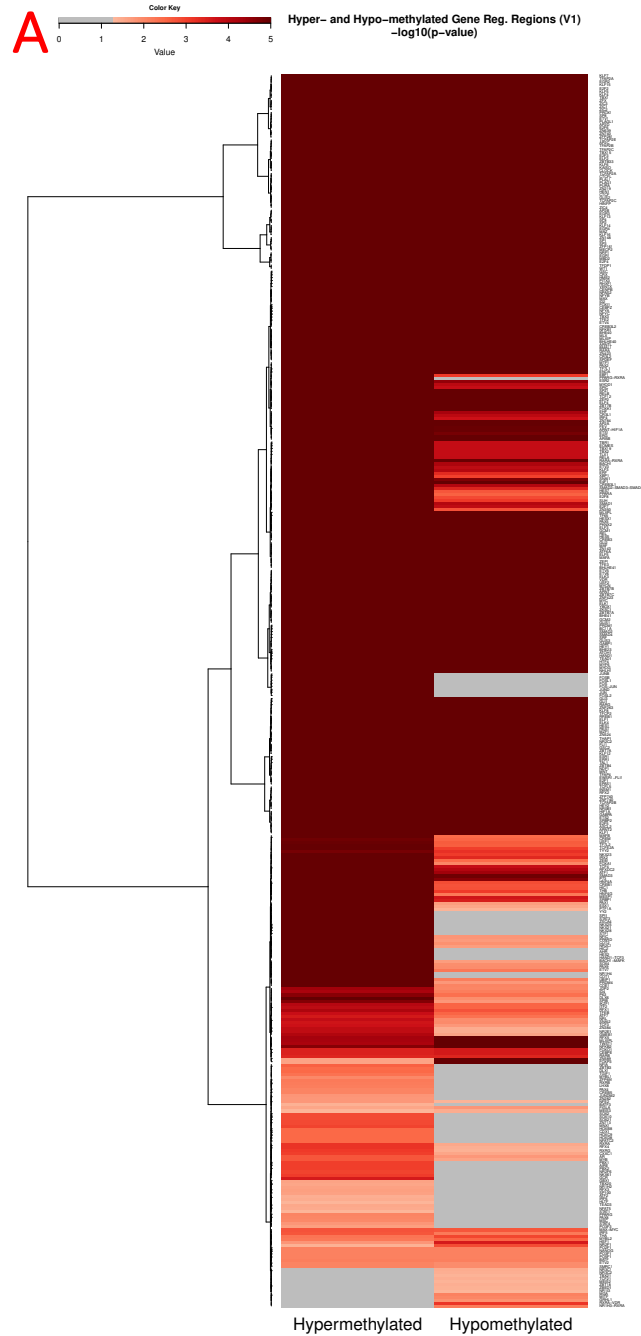


D

Enriched Transcription Factors' Gene Expression

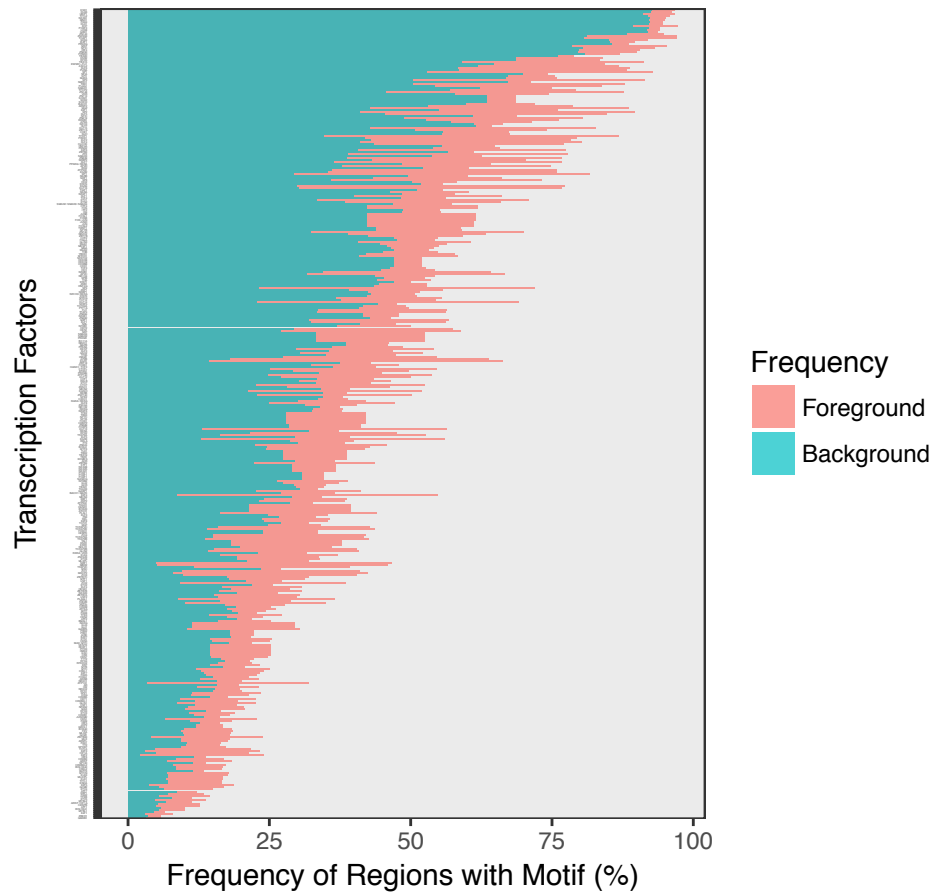


Regulatory Region of Genes Associated with Hyper vs Hypo Meth. (V1)

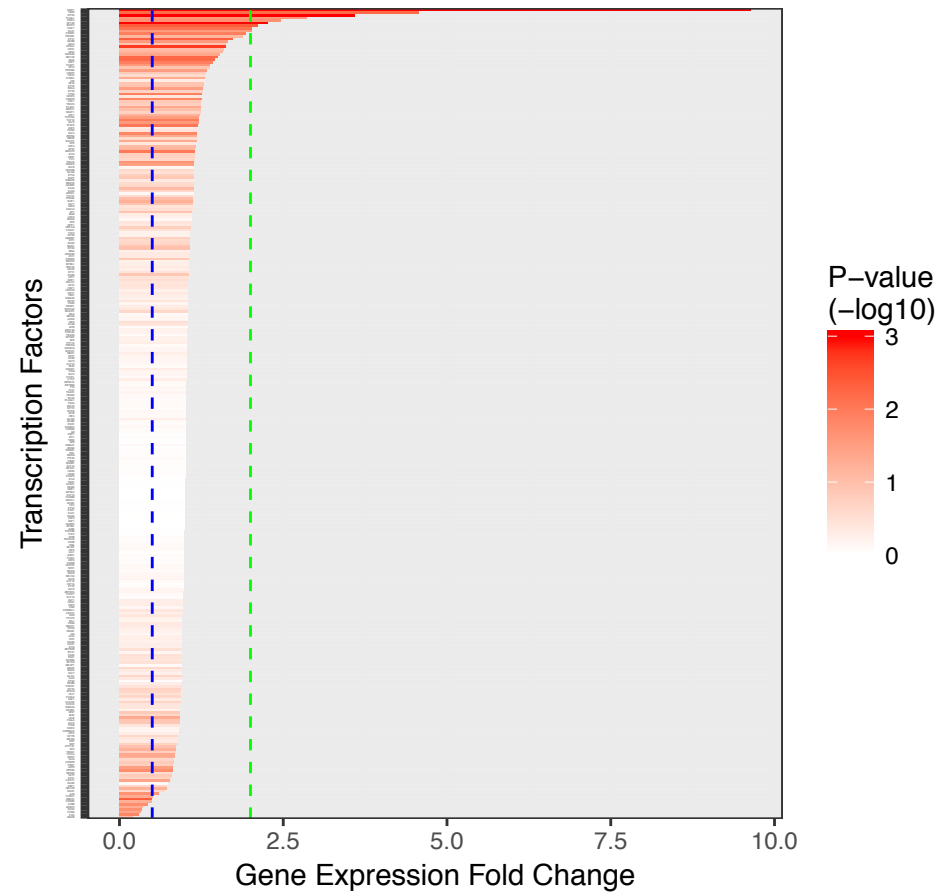


Regulatory Region of Genes Associated with Hyper vs Hypo Meth. (V1)

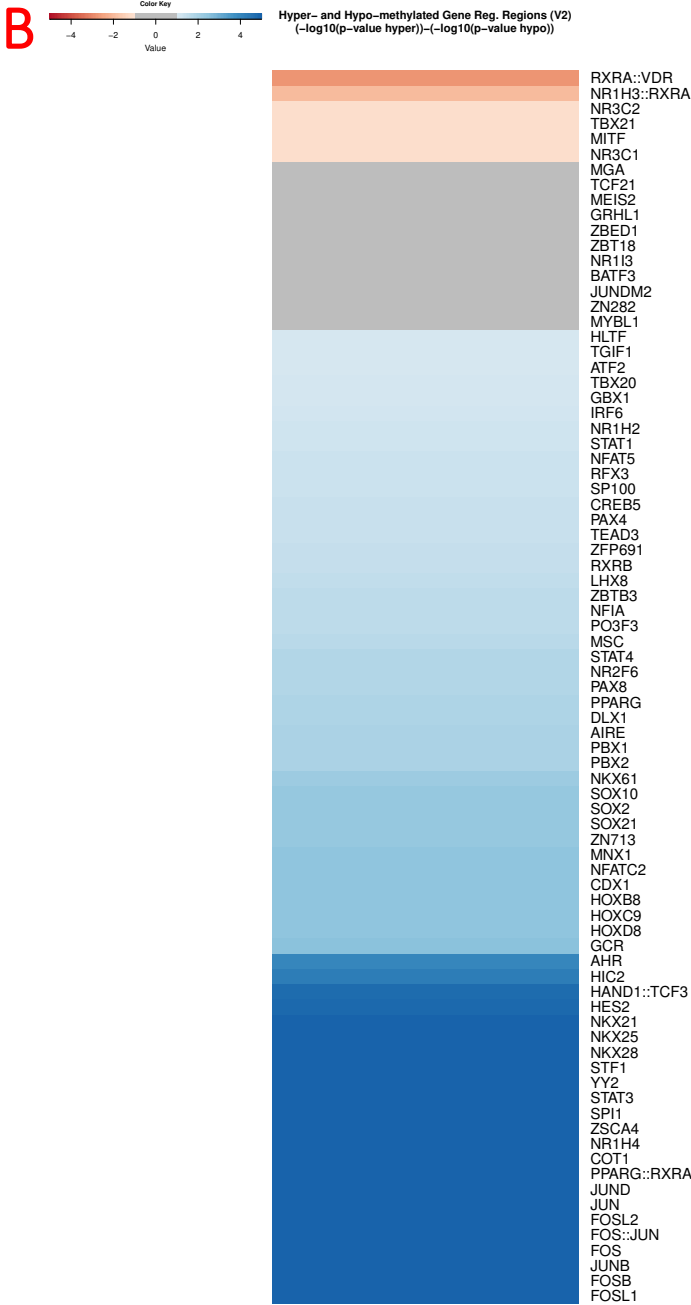
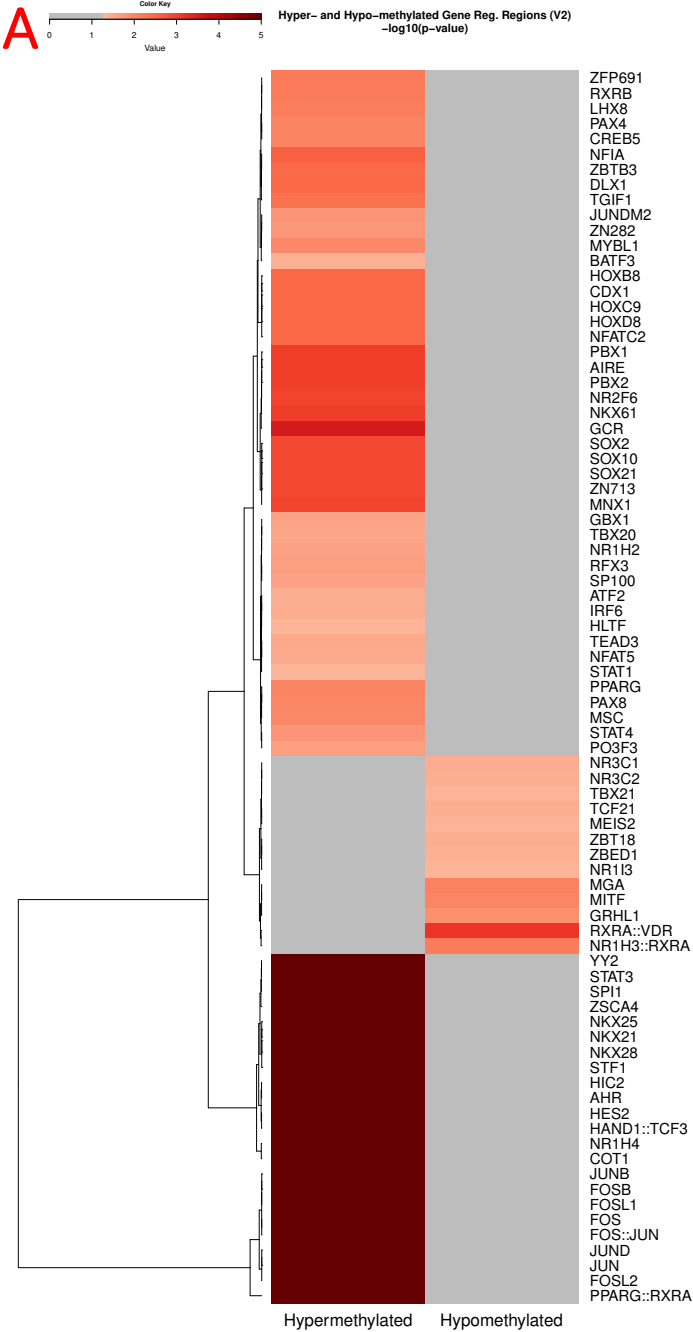
C Enriched Transcription Factors' Regions' Frequency



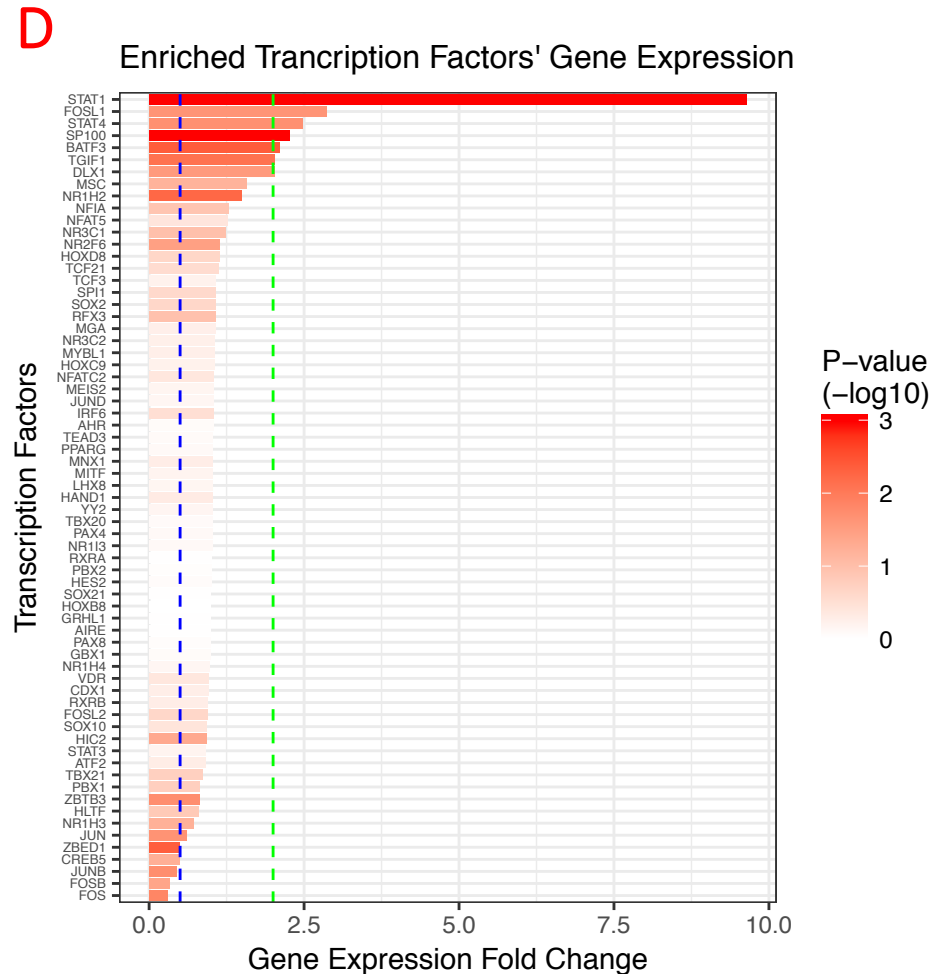
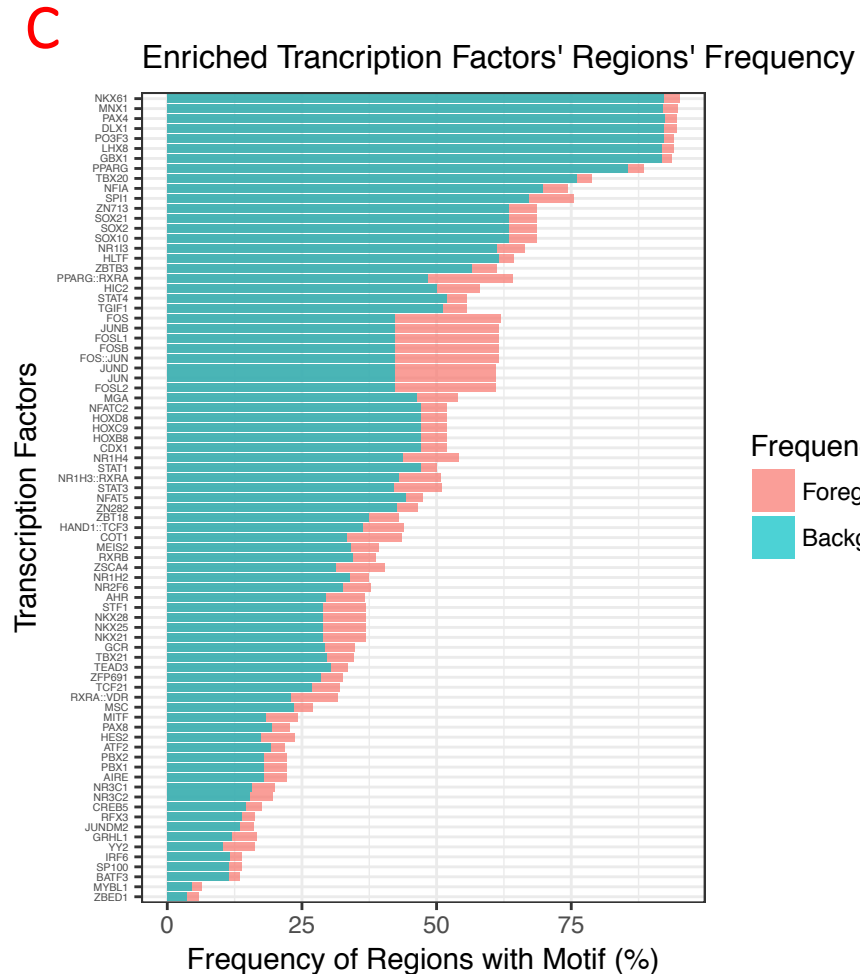
D Enriched Transcription Factors' Gene Expression



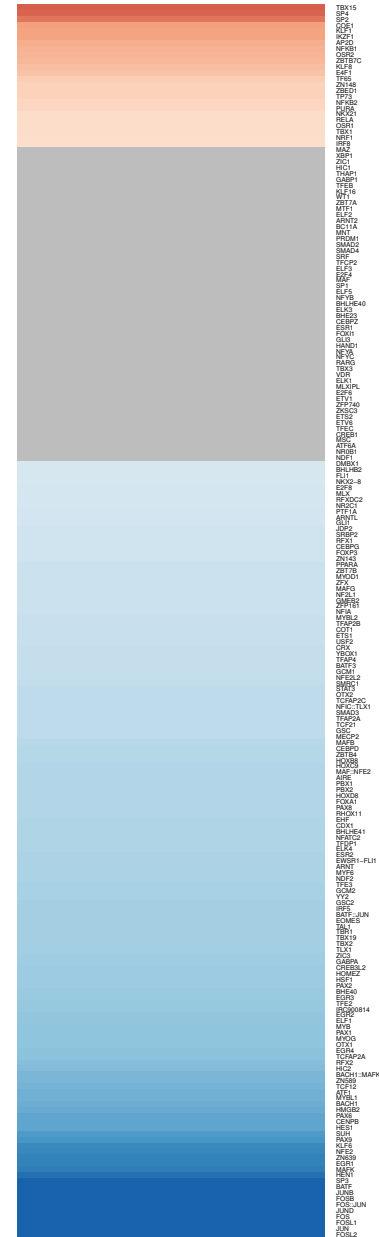
Regulatory Region of Genes Associated with Hyper vs Hypo Meth. (V2)



Regulatory Region of Genes Associated with Hyper vs Hypo Meth. (V2)

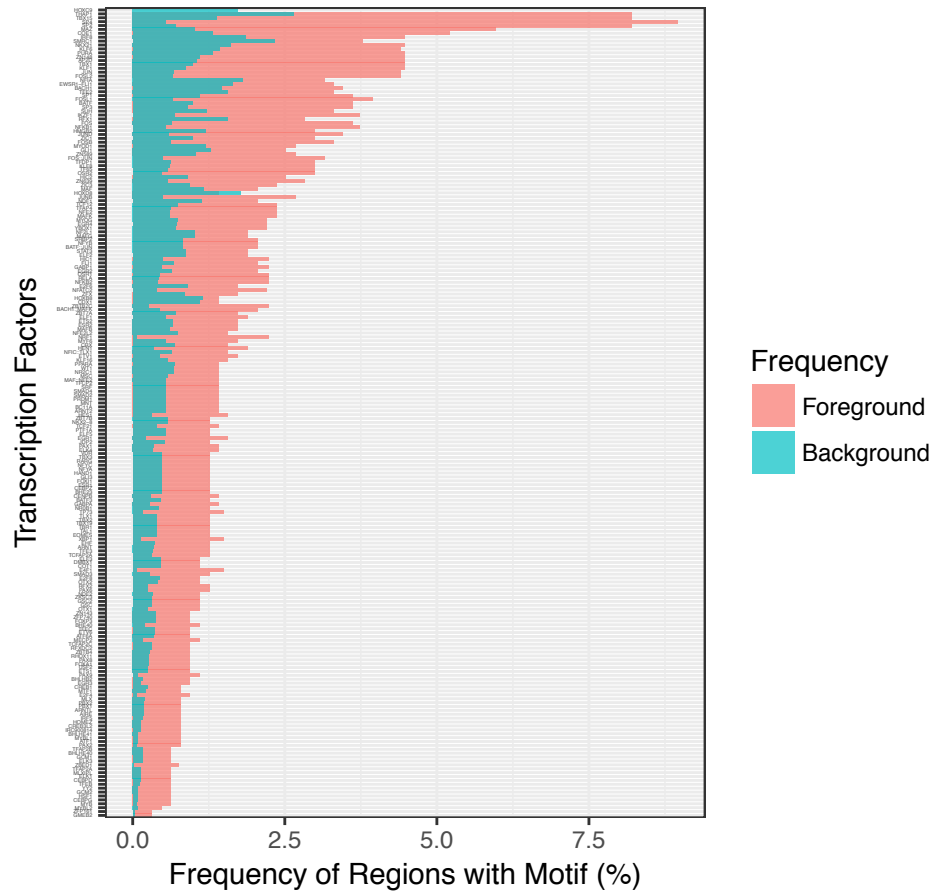


(V1)

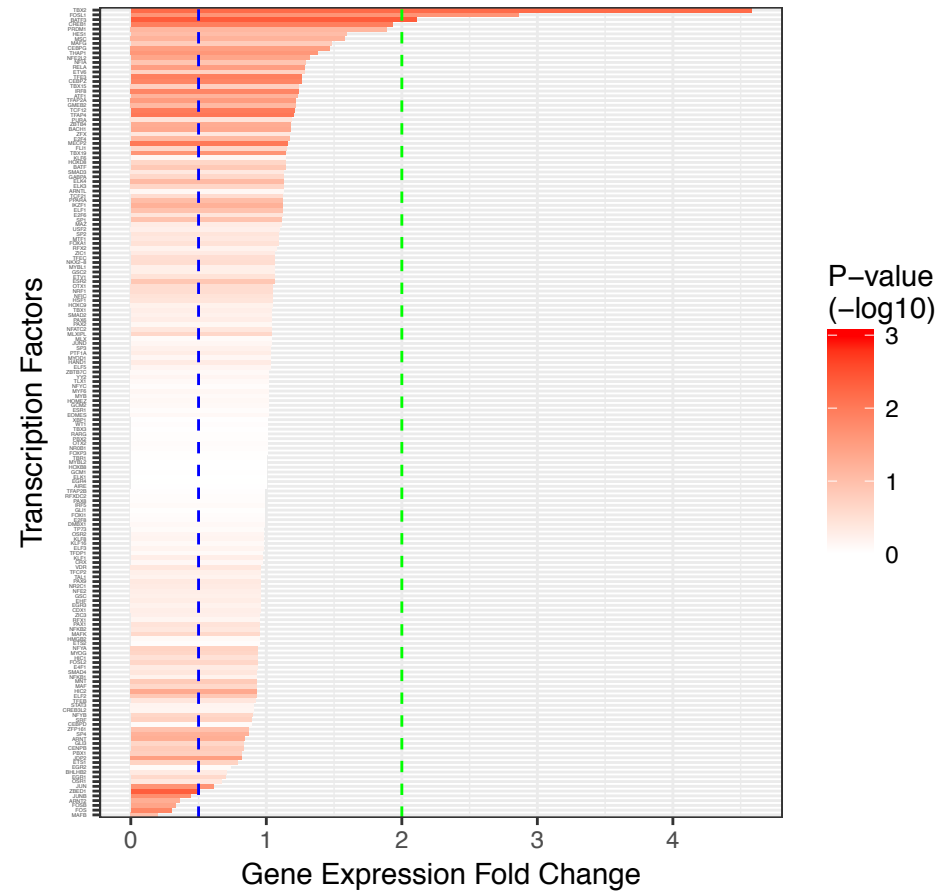


Hyper vs Hypo Methylated CpGs (V1)

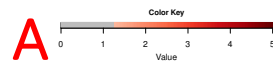
C Enriched Transcription Factors' Regions' Frequency



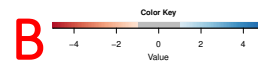
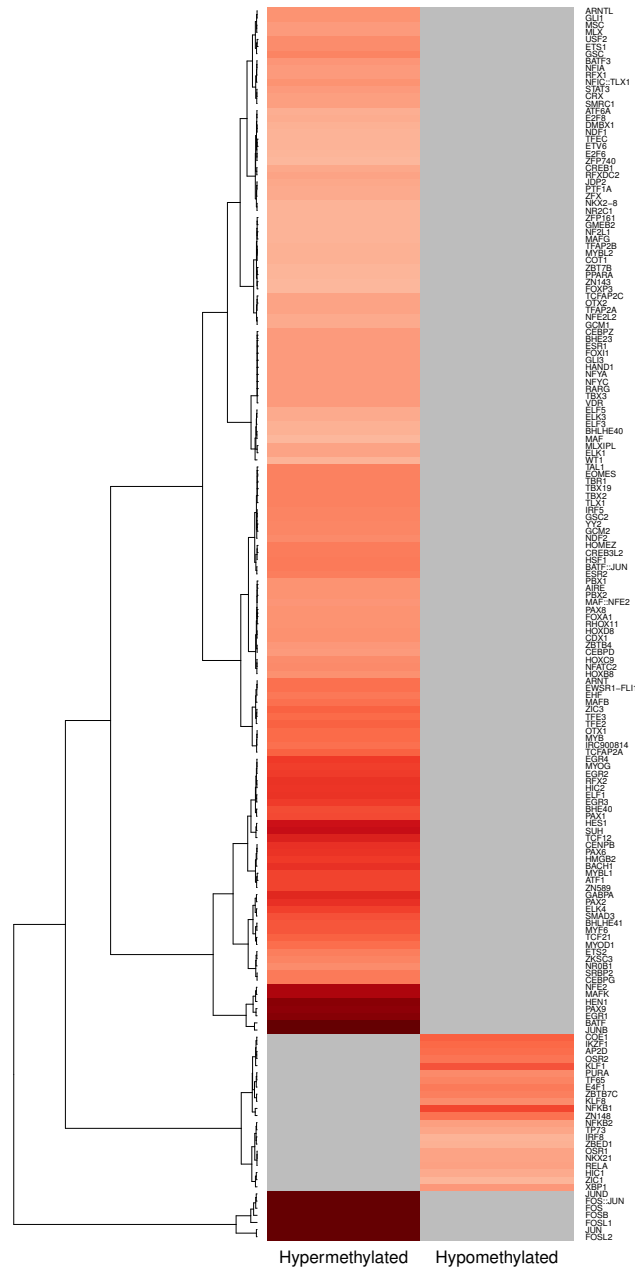
D Enriched Transcription Factors' Gene Expression



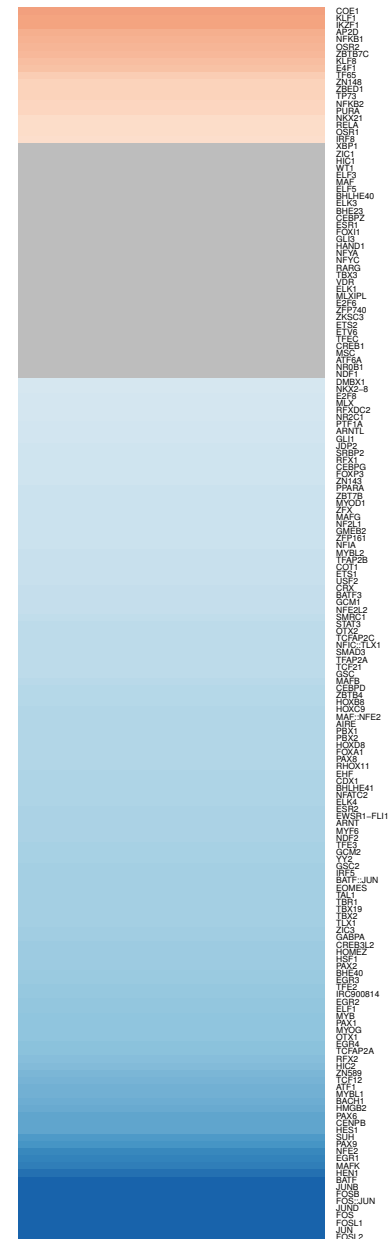
(V2)



Hyper- and Hypo-methylated CpGs (V2)
-log₁₀(p-value)

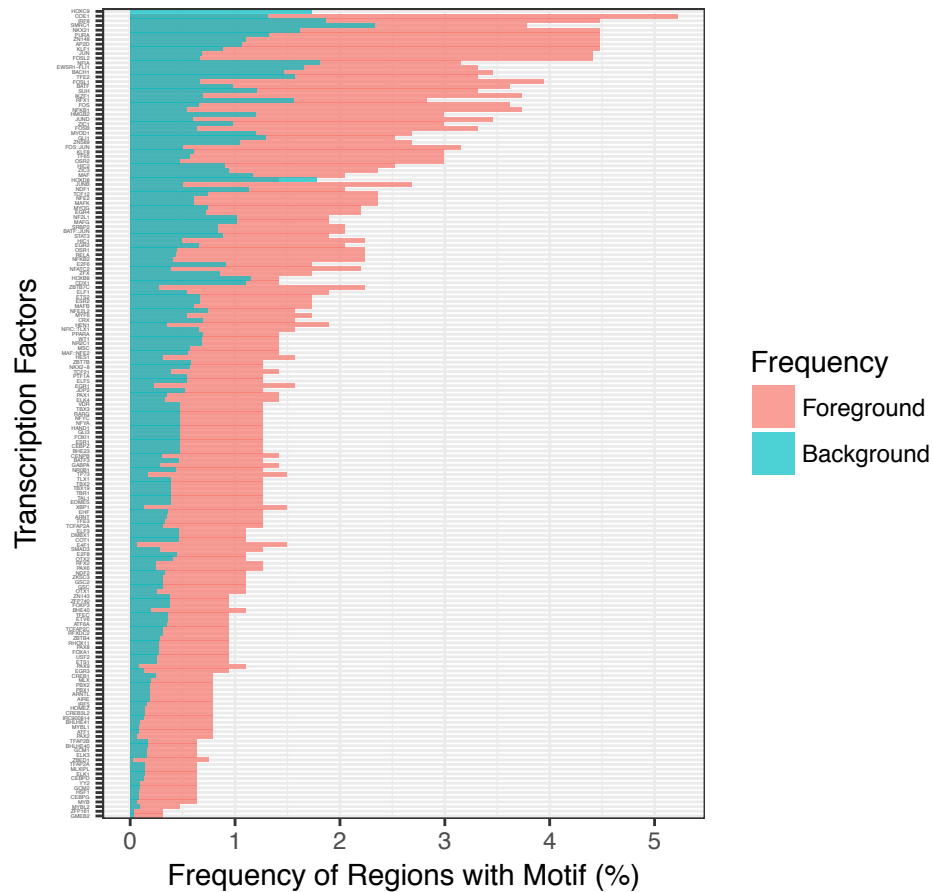


Hyper- and Hypo-methylated CpGs (V2)
 $(-\log_{10}(\text{p-value hyper})) - (-\log_{10}(\text{p-value hypo}))$



Hyper vs Hypo Methylated CpGs (V2)

C Enriched Transcription Factors' Regions' Frequency



D Enriched Transcription Factors' Gene Expression

