This folder contains scripts to generate the transcription factors and K-means clustering as presented in the manuscript. TCGA CRC methylation and expression data is downloaded using the commands in the script itself. Additional data needed for running the script including CIMP labels, autosomal probes, ENCODE cCRE probes, and normalized expression data for the combined CRC datasets (TCGA + GSE39582) for clustering analysis are provided as supplementary data.

Below are two major analysis pipelines utilized in our study.

1. Transcription Factor (TF) Analysis using ELMER V2:
2. Refer to the directory “overall\_ana” for identifying TFs related to overall analysis of TCGA CRC vs normal samples. Refer to the directory “cimp\_ana” for identifying TFs related to CIMP specific analysis of TCGA CRC samples
3. You will need to sequentially run three scripts for the TF calculation. Change to either the “overall\_ana” directory or the “cimp\_ana” directory.
4. Run the script “get\_Data.R” in order to download and prepare the methylation and expression data
5. Determine differentially methylated probes using the script “limma-difMeth.R”
6. Identify TFs using the script “get\_TFs\_ELMERv2.R”
7. Filter the TFs based on the adjusted p-value using the script “filter\_TFs.R”

1. Carry out K-means clustering on the expression data:

Run the script “kmeans\_clust.R”. This will carry out the K-means clustering as well as generate a PCA plot of the gene expression, along with the K-means cluster, CIMP and CMS labels.