MethReg: estimating the regulatory potential of DNA methylation in gene transcription

Tiago C. Silva[1], Juan I. Young[2,3], Eden R. Martin[2,3], Xi Chen[1,4], Lily Wang[1,2,3,4*]

[1] Department of Public Health Sciences, University of Miami, Miller School of Medicine, Miami, FL 33136, USA

[2] Dr. John T Macdonald Foundation Department of Human Genetics, University of Miami, Miller School of Medicine, Miami, FL 33136, USA

[3] John P. Hussman Institute for Human Genomics, University of Miami Miller School of Medicine, Miami, FL 33136, USA

[4] Sylvester Comprehensive Cancer Center, University of Miami, Miller School of Medicine, Miami, FL 33136, USA

* To whom correspondence should be addressed.

Corresponding Address:

Lily Wang
Soffer Clinical Research Ctr
University of Miami School of Medicine
1120 NW 14th St
Miami, Florida 33136-2107
lily.wang@miami.edu
Tel: 305-243-2927; Fax: 305-243-5544;
Email: lily.wangg@gmail.com

Abstract

Epigenome-wide association studies (EWAS) often detect a large number of differentially methylated sites or regions, many are located in distal regulatory regions. To further prioritize these significant sites, there is a critical need to better understand the functional impact of CpG methylation. Recent studies demonstrated CpG methylation-dependent transcriptional regulation is a widespread phenomenon. Here we present MethReg, an R/Bioconductor package that analyzes matched DNA-methylation and gene-expression data, along with external transcription factor (TF) binding information, to evaluate, prioritize, and annotate CpG sites with high regulatory potential. By simultaneous modeling three key elements that contribute to gene transcription (CpG methylation, target gene expression and TF activity), MethReg identifies TF-target gene associations that are present only in a subset of samples with high (or low) methylation levels at the CpG that influences TF activities, which can be missed in analyses that use all samples. Using real colorectal cancer and Alzheimer's disease datasets, we show MethReg significantly enhances our understanding of the regulatory roles of DNA methylation in complex diseases.

Background

DNA methylation is a widely studied epigenetic modification. Recent epigenome-wide association studies (EWAS) have identified numerous alterations in DNA methylation (DNAm) levels that are involved in many diseases such as various cancers [1-5] and neurodegenerative diseases [6-8]. Compared to genome-wide association studies (GWAS) of genetic variants, EWAS often detect a larger number of significant differences, often thousands of differentially methylated CpG sites (DMS), which are significantly associated with a disease or phenotype. Many of these DMS are located far from genes, complicating the interpretation of their functionality [9, 10]. Therefore, there is a critical need to better understand the functional impact of these CpG methylations and to further prioritize the significant methylation changes.

Transcription factors (TFs) are proteins that bind DNA and facilitate the transcription of DNA into RNA. A number of recent studies have observed that the binding of TFs onto DNA can be affected by DNA methylation, and in turn, DNA methylation can also be altered by proteins associated with TFs [11-15]. Using methylation-sensitive SELEX (systematic evolution of ligands by exponential enrichment), Yin et al. (2017) [16] classified 519 TFs into several categories: TFs whose binding strength increased, decreased, or was not affected by DNA methylation, as well as those not containing CpGs in their binding motifs.

Although a number of integrative analyses strategies [17-23] have been proposed to help assess the functional role of the DNA methylation changes in gene regulation, these methods typically integrate DNA methylation data with either gene expression [17-20] or TF binding data [21-23], but rarely both. For example, MethylMix [17] identifies CpGs that are predictive of transcription and then classifies the CpGs into different methylation states. Similarly, COHCAP [18] identifies a subset of CpGs within CpG islands that are most likely to regulate downstream gene expression. These methods which test the association between DNA methylation and target gene expression can be further improved by additionally incorporating information on TF activities.

To determine whether TF regulatory activity is enhanced or reduced by significant CpGs in EWAS, the gold standard would be to perform ChIP-Seq experiments for all candidate TFs with binding sites close to the CpGs in parallel with bisulfite sequencing of DNA methylation changes and transcriptome assessment.

However, performing ChIP-Seq experiments on primary tissues is technically challenging because of the limited number of cells in some samples, the large number of TF to be tested, and the lack of availability for specific antibodies. Therefore, in practice, computational approaches are often used to prioritize disease relevant TFs in DNA methylation studies. For example, LOLA [21] (Locus Overlap Analysis) performs enrichment analysis to identify regulatory elements such as TFs with binding sites enriched in candidate genomic regions (e.g., DMRs). Alternatively, Goldmine [22] annotates individual CpGs by TFs with binding sites that overlap with the CpG. However, because the binding motifs for TFs are often non-specific for different members of a TF family, there are often many TFs with binding sites that overlap with a given CpG. Moreover, TF binding can also occur without affecting the transcription of any gene [24, 25]. Therefore, methods which analyze DNA methylation and TF binding sites (TFBS) data would also be greatly enhanced by additionally modeling target gene expression.

To fill this critical gap in analytical methods and software for annotating and prioritizing DNA methylation changes identified in EWAS, here we present MethReg, an R/Bioconductor package that performs integrative modeling of three key components (DNA methylation, gene expression levels, TF) in gene regulation, to provide a more comprehensive functional assessment of CpG methylation in gene regulation. In particular, MethReg leverages information from external databases on TFBS, ChIP-seq experiments and TF-target interactions, performs both promoter and distal (enhancer) analyses, implements rigorous robust regression models, and can fully adjust for potential confounding effects such as copy number, age, and sex that are important in DNA methylation analysis. MethReg can be used either to evaluate the regulatory potentials of candidate regions identified in EWAS (in supervised analysis mode) or to search for methylation dependent TF regulatory processes in the entire genome (in unsupervised analysis mode).

Using simulated datasets, we showed that by simultaneous modeling of three key elements (DNA methylation, target gene, and TF), MethReg significantly improves prioritization for true positive DNA methylation changes with regulatory roles in gene transcription when compared to models that include only two key elements. In addition, we also analyzed the TCGA colorectal datasets and the ROSMAP

Alzheimer's dataset to show that MethReg was able to recover known biology, as well as nominate novel biologically meaningful DNA methylation-TF-target associations that influence transcription.

Results

An overview of the MethReg software

To systematically search for CpG methylation with significant regulatory effects on gene expression by influencing TF activities, we developed MethReg. Figure 1 illustrates the workflow for MethReg. Given matched (methylation arrays) DNA methylation data and (RNA-seq) gene expression data, MethReg additionally incorporates TF binding information from ReMap2020 [26] or the JASPAR2020 [27] database, and optionally additional TF-target gene interaction databases (Supplementary Table 1), to perform both promoter and distal (enhancer) analysis. In the unsupervised mode, MethReg analyzes all CpGs on the Illumina arrays. In supervised mode, MethReg analyzes and prioritizes differentially methylated CpGs (DMS) identified in EWAS. There are three main steps: (1) create a dataset with triplets of CpGs, TFs which bind near the CpGs, and putative target genes; (2) for each CpG-TF-target gene triplet, apply integrative statistical models to DNA methylation, target gene expression, and TF activity values; and (3) visualize and interpret results from statistical models to estimate the impact of DNA methylation on TF activity (synergistic effect of CpG methylation and TF on target gene), as well as to annotate the roles of TF and CpG methylation in regulating target gene expression. The results from the statistical models will allow us to identify a list of CpGs that work synergistically with TFs to influence target gene expression. Next, we discuss each of these analytical steps in detail. We will describe the analysis of TFs, but the methods and software tool are, in principle, also applicable to other types of chromatin proteins that crosstalk with DNA methylation. MethReg is an open-source R/Bioconductor package, available at https://bioconductor.org/packages/MethReg/.

Creating CpG-TF-target gene triplet dataset

MethReg first links CpGs to TFs with binding sites within a window of user-specified distance (e.g., ± 250 bp) using information from ReMap2020 [26] or JASPER2020 [27] database. The JASPER2020 [27] database includes curated transcription factor binding models, among which 637 are associated with human TFs with known DNA-binding profiles [28]. Similarly, the human atlas of the ReMap2020 database [25] contains regulatory regions for 1135 transcriptional regulators obtained using genome-wide DNA-binding experiments such as ChIP-seq. Next, in *promoter analysis*, CpGs located in promoter regions, defined as ± 2 kb regions around the transcription start sites (TSS), are linked to target genes with promoters that overlap with the CpG. On the other hand, CpGs in the distal regions (i.e., > 2kb from TSS) are linked to a specific number of genes (e.g., 5 genes) upstream or downstream and within 1M bp of the CpG. The CpG-TF pairs are then combined with CpG-target pairs to create triplets of CpG-TF-target genes.

Alternatively, CpGs can also be linked to genes within 1M bp using *regulon-based analysis*. A TF regulon consists of all the transcriptional targets of the TF. MethReg obtains TF-target pairs from curated external regulon databases [29-31] (Supplementary Table 1). Combining the CpG-TF pairs with TF-target gene pairs, we then obtain a triplet dataset where each row contains identifiers for a CpG, a TF, and the target gene.

Estimating regulatory effects of CpG methylation and TFs on target gene expression

Given a CpG-TF-target gene triplet, we then query the matched DNA methylation and gene expression datasets to obtain DNA methylation, target gene, and TF gene expression (or activity) values and fit the following statistical model to data:

$$target \sim TF + DNAm + DNAm \times TF \qquad \text{(Model 1)}$$

where *target* is $\log_2$ transformed target gene expression values, *TF* is $\log_2$ transformed TF gene expression values or estimated TF activity scores (see details in section "Modeling TF protein activities" below), and *DNAm* is DNA methylation beta-values.

Note that Model 1 partitions the effects of DNA methylation and TF on target gene expression into three categories: the direct effect of TF (modeled by term *TF*), the direct effect of DNA methylation

(modeled by term *DNAm*), and the synergistic effects of TF and DNA methylation (i.e., how the effect of TF on target gene expression is modified by DNA methylation; modeled by *DNAm* × *TF* interaction term).

For accurate statistical modeling, MethReg implements Model 1 by fitting a robust linear model which gives outlier gene expression values reduced weights [32]. Note that a key feature of Model 1 is that it provides more comprehensive modeling of gene regulation by incorporating the three components (TF activity, DNA methylation, target gene expression) simultaneously. In addition to Model 1 described above, which included *DNAm* as a continuous variable, we also considered another model that modeled methylation values as a binary variable. We also propose:

$$target \sim TF + DNAm.group + DNAm.group \times TF \qquad \text{(Model 2)}$$

where *DNAm.group* is high or low. That is, for a given CpG, the samples with the highest DNA methylation levels (top 25%) have *DNAm.group* = "high", and samples with lowest DNAm levels (bottom 25%) have *DNAm.group* = "low". In this model, only samples with DNA methylation values in the first and last quartiles are considered. Note also that statistically, the $DNAm.group \times TF$ effect is estimated by comparing the magnitude of TF-target gene association in high methylation group versus the magnitude of the TF-target gene association in low methylation group.

Modeling TF protein activities

Given that TF gene expression might not accurately reflect TF protein activities, which involve additional complex processes (e.g. post-translational modifications, protein-protein/ligand interactions, and localization changes), MethReg implements an additional option to model TF activities via the VIPER (virtual Inference by enriched regulon analysis) [33] or GSVA [34] methods, so that the TF effects in Model 1 and Model 2 described above can also be computed by replacing TF gene expression levels with estimated TF activities. Briefly, given RNA-seq data, these methods estimate the activity of a TF by performing a rank-based gene set enrichment analysis of its target genes (i.e., its regulon). MethReg can work with different regulon databases (Supplementary Table 1), such as those described Garcia-Alonso et al. (2019) [29], which were collected from four resources: manually curated databases, ChIP-seq binding

experimental data, prediction of TF binding motifs based on gene promoter sequences, or computational regulatory network analysis. The Genotype-Tissue Expression (GTEx) tissue-consensus regulons included 1,077,121 TF–target candidate regulatory interactions between 1,402 TFs and 26,984 targets, and the pan-cancer regulons included 636,753 TF–target candidate regulatory interactions between 1,412 TFs and 26,939 targets. Garcia-Alonso et al. (2019) annotated each TF-target interaction with a five-level confidence score, with "A" indicating most reliable, supported by multiple lines of evidence, and "E" indicating least confidence, supported only by computational predictions. Benchmark experiments using three separate datasets showed the GTEx tissue-consensus regulons performed similarly to tissue-specific regulons computed from GTEx data of specific tissue type. Notably, MethReg also provides options for users to input alternative TF regulon databases (Supplementary Table 1) and TF activities computed using alternative software such as Lisa [35].

Stage-wise method for controlling false discovery rate

MethReg implements two alternative methods for controlling false discovery rates, using the conventional approach [36] or a stage-wise approach [37]. To help improve power in high-throughput experiments where multiple hypotheses are tested for each gene, van de Berge et al. (2017) [37] proposed a stage-wise approach in the context of gene splicing analysis. First, in the screening step, a global test is applied to each gene to test the null hypothesis that there is a differential change in any of the transcripts within the gene. Second, in the confirmation step, for the genes selected in the screening step, individual transcripts are then tested while controlling family-wise error rate (FWER). By aggregating effects from individual transcripts within a gene in the screening step, the stage-wise procedure was shown to have superior power compared with the conventional approach that tests all individual transcripts in one-step.

In Model 1 and Model 2 described above, the synergistic activity of DNA methylation and TF is estimated by the interaction term $DNAm \times TF$. Because the standard error of interaction effects is typically much larger than those for main effects, the conventional approach for controlling false discovery rate often results in low power for discovering interaction effects [37]. To this end, MethReg additionally implements

the stage-wise procedure for testing interactions by first aggregating all CpG-TF-target triplets associated with the same CpG as a group. In the screening step, MethReg tests the null hypothesis that any of the individual triplets mapped to a CpG has a significant *DNAm* × *TF* effect. In the confirmation step, MethReg tests all the triplets associated with significant CpGs selected in the screening step while controlling FWER as described in Van den Berge et al. (2017) [37].

Visualizing and annotating roles of CpG and TF in gene transcription

To visualize how DNA methylation and TFs work together to influence gene expression, MethReg generates a suite of figures. Figure 2 shows an example output figure of Model 2 applied to the TCGA colorectal cancer dataset (Online Methods). In the first row are figures for assessing direct pairwise TF-target and DNA methylation-target associations. In the second row are figures for assessing TF-target gene expression, stratified by high or low DNA methylation levels. In the third row, we stratify by TF expression (or activity score) instead and plot gene expression against DNA methylation values, stratified by high or low TF expression (or activity) levels.

Note in Figure 2, without stratifying by DNA methylation, the overall TF-target effect (robust linear model P-value = 0.590) is not significant. In contrast, TF-target association is highly significant in samples with high methylation levels (robust linear model P-value = 0.001). Therefore, methylation at cg00328227 plays an important role in regulating TF activity in this case. This example also demonstrates that by additionally modeling DNA methylation, we can nominate TF-target associations that might have been missed otherwise.

Figure 3 shows the different biological scenarios where methylation and TF influences target gene expression synergistically. A TF repressor decreases transcription while a TF activator increases it, and the presence of methylation can either enhance or attenuate the binding affinity of the TF. For each triplet, MethReg annotates the role of TF on the target gene (repressor, activator, or dual) and how DNA methylation influences the TF (enhancing, attenuating, or invert).

Simulation analysis

We conducted a simulation study to compare the performance of 7 different approaches, including Model 1 & Model 2 as well as the conventional approach of directly correlating TF and target gene expressions, to identify TF-target associations where TF activities are modulated by CpG methylation (Supplementary Figure 1). We considered the scenario where CpG methylation affects TF binding affinity, so that TF only affects target gene expression when methylation level is low. To assess the statistical properties of these different methods, we estimated and compared type I error rate, power, and area under receiver operating characteristics curves (AUC) for each method.

More specifically, we simulated datasets for which target gene expression levels were dependent on TF expressions only in samples with low DNA methylation levels. We used 38 samples of TCGA COAD matched RNA-seq and DNA methylation data on chromosome 21 included in the MethReg R package as our input dataset, from which we randomly sampled TF gene expression and methylation expression levels. For each simulated triplet dataset, we randomly sampled one gene from RNA-seq dataset and one CpG site from methylation dataset to be our TF expression and DNA methylation expression. Next, target gene expression levels were simulated from negative binomial distributions as follows: the estimated median mean and variance over all genes in our input dataset were $\mu_0 = 10.59$ and $\sigma^2 = 16.90$, respectively. Therefore, for target gene expressions, we assumed a negative binomial distribution with parameters $\mu_0 = 10.59$ and $k_0 = {\mu_0}^2/(\sigma^2 - \mu_0) = 17.78$ for all samples except those with lowest DNA methylation levels in the first quartile. For the samples with low methylation levels in the first quartile, we generated target gene expression levels from negative binomial distribution ($\mu = \mu_0 + \beta \times$ TF gene expression, $k = k_0$), where $\beta = (0, 1, 2, \dots, 9)$ indicates different strengths of associations between TF and target gene expressions, corresponding to 10 different simulation scenarios. Therefore, by design of this simulation experiment, target gene expressions were associated with TF only when methylation levels were low. For each simulation scenario, we repeated this process 1,000 times to generate 1,000 triplets of TF, DNA methylation, and target gene expression levels.

Note that when $\beta = 0$, target gene expressions were generated randomly from negative binomial distribution and did not depend on TF gene expression in the samples, so the 1,000 triplet datasets simulated under this simulation scenario (null triplets) allowed us to estimate type I error rates of different models. We compared sensitivity and specificities for identifying TF-target associations based on p-values from 7 different approaches (Supplementary Figure 1):

a) lm.cont: p-value for *DNAm* × *TF* term in linear model implementation of Model 1.

b) lm.binary: p-value for *DNAm.group* × *TF* term in linear model implementation of Model 2.

c) rlm.cont: p-value for *DNAm* × *TF* term in robust linear model implementation of Model 1.

d) rlm.binary: p-value for *DNAm.group* × *TF* term in robust linear model implementation of Model 2.

e) rlm.binary.en: p-value for *DNAm.group* × *TF* term in robust linear model implementation of Model 2, estimated from empirical null distribution [38] (see details in Methods).

f) lm.main.tf: p-value for *TF* term in main effect model *target gene expression ~ TF*.

g) spearman.corr.tf: p-value for spearman correlation between target gene expression and TF.

Note that the last two methods lm.main.tf and spearman.corr.tf evaluate the conventional approach of directly correlating target gene expression with TF expression, without considering DNA methylation (Supplementary Figure 1). Also, in method e) rlm.binary.en, instead of the conventional approach which computes P-values by comparing test statistic for *DNAm.group* × *TF* to *t*-distribution, this method estimates P-values for *DNAm.group* × *TF* effect using empirical null distribution [38], which is a normal distribution with empirically estimated mean $\hat{\delta}$ and standard deviation $\hat{\sigma}$. Efron (2010) [38] showed that in large-scale simultaneous testing situations (e.g., when many triplets are tested in an analysis), serious defects in the theoretical null distribution may become obvious, while empirical Bayes methods can provide much more realistic null distributions.

The results showed that all methods had type I error rates close to 5%. In particular, the methods lm.main.tf, rlm.binary.en, and spearman.corr.tf had type I error rates below 5% (Supplementary Figure 2).

Among the methods, rlm.binary and rlm.binary.en had similar and the highest power across all simulation scenarios, while the main effect models that test association between target gene expression and TF expression using linear model (lm.main.tf) or Spearman correlation (spearman.corr.tf) lacked power (Supplementary Figure 3).

Given the known status of CpG methylation's role in association with target gene expression (i.e. true negative when $\beta = 0$ and true positive when $\beta > 0$), we next computed the area under the ROC curve (AUC) for each method. The receiver operating characteristic (ROC) curves show a trade-off between sensitivity and specificity as the significance cutoff is varied. AUC assesses the overall discriminative ability of the methods to determine whether a given methylation CpG is driving target gene expression over all possible cutoffs. The best performing methods with highest AUCs are methods rlm.binary (0.894), rlm.binary.en (0.882), and lm.binary (0.838), followed by p-values from models with continuous methylation levels, lm.cont (0.821) and rlm.cont (0.815) (Figure 4). Again, the main effect models lm.main.tf and spearman.corr.tf lacked sensitivity (i.e., power) therefore had the lowest AUCs at 0.631 and 0.582, respectively.

These simulation study results showed that the conventional approach of directly correlating gene expression with TF expression (lm.main.tf and spearman.corr.tf) lacked power for detecting TF activities on target gene expression when they are also modulated by DNA methylation. This is probably because without stratifying on DNA methylation levels, the main effect models detect TF-target gene associations averaged over all samples with various methylation levels. In contrast, the interaction terms in Model 1 & 2 compare the magnitude of TF-target gene association in low methylation samples versus the magnitude of the association in high methylation samples.

The results also indicated that the models that use a binary variable (low, high) to model methylation levels (rlm.binary, rlm.binary.en and lm.binary methods) performed best, probably because these models can reduce noise in data and thus can improve power. Among binary models, the robust linear model

rlm.binary and rlm.binary.en performed similarly and better than regular linear model lm.binary. Thus, we selected the rlm.binary model for our subsequent analyses.

Case Study: analyses of TCGA COAD-READ datasets

An unsupervised MethReg analysis

Colorectal cancer (CRC) is the third most commonly diagnosed cancer and the second leading cause of cancer death in the United States [39]. CRC, like many other cancers, is characterized by global hypo-methylation leading to oncogene activation, chromosomal instability, and locus-specific hyper-methylation which leads to silencing of tumor suppressor genes [40, 41]. In parallel, TFs also play instrumental roles in tumor development and metastasis [42-44]. Given the strong epigenetic basis of CRC, we next applied MethReg to the TCGA COAD-READ datasets, including 367 samples with matched DNA methylation, gene expression, and copy number alterations (CNAs). To account for potential confounding effects, we adjusted target gene expression values by CNA and tumor purity estimates [45] first, extracted the residuals, and then fitted the rlm.binary model to the residuals (see the "Methods" section for details).

We performed an unsupervised MethReg analysis, without selecting any CpGs *a priori*. First, we divided the CpGs into those in the promoter regions (within ± 2 kb regions around the transcription start sites (TSS)) or the distal regions (> 2kb from TSS). Next, we linked CpG sites in the promoter regions to genes with promoters that overlapped with them. On the other hand, CpG sites in the distal regions were linked to five genes upstream and five genes downstream within 1M bp. Alternatively, in the regulon-based approach, we also linked CpGs in either promoter or distal regions to genes regulated by TFs with binding sites close to the CpG (Figure 1). To more accurately model TF effects, we computed TF activity scores using the VIPER algorithm [33]. Stage-wise analysis using the rlm.binary model *residual target gene expression ~ TF.activity + DNAm.group + DNAm.group × TF.activity* was then applied to the triplet datasets to identify triplets with significant *DNAm.group × TF.activity* effect, in which CpGs had significant effects on target gene expression by influencing TF activities.

Stage-wise analysis results showed the number of triplets with significant *DNAm* × *TF.activity* terms in the promoter, distal, and regulon-based analysis were 31, 52, and 47, respectively (Table 1, Supplementary Table S2 – S4). There was no overlap between the significant triplets obtained in these three analyses. Our results agreed well with the previous study in Wang et al. (2020) [30], which also observed only a small number of transcriptional regulations were mediated by DNA methylation. To additionally validate our findings, we compared the TFs in our significant triplets with the MeDReaders database [46], which manually curated information from human and mouse studies for 731 TFs which could bind to methylated DNA sequences. Our comparison showed that among the 18, 44, and 11 unique TFs identified in the promoter, distal, and regulon-based analyses, respectively, the majority of them were previously shown to interact with methylated DNA. More specifically, 12 (68%), 32 (73%), and 8 (73%) of the TFs from the promoter, distal and regulon-based analyses were previously shown to exhibit CpG methylation-dependent DNA-binding activities using functional protein arrays [14] or systematic evolution of ligands by exponential enrichment (SELEX) [16] (Supplementary Tables S2-S4).

Moreover, the TFs and target genes in the significant triplets identified by MethReg have also been previously associated with CRC. Figure 2 shows the most significant triplet among all the triplets considered in promoter and distal analysis. In this example, the transcription factor NFATC2 represses the target gene *HENMT1* in samples with high DNA methylation at cg00328227 but is relatively independent of target gene expression in samples with high DNA methylation. Therefore, DNA methylation enhances TF activities in this case. NFATC2 belongs to the NFAT family of transcription factors, which are known to regulate T cell activation and differentiation as immune cells invades the malignant tissues [47]. In particular, NFATC2 is a critical regulator in intestinal inflammation that promotes development of colorectal cancer, and expression levels of NFATC2 were found to be elevated in colorectal cancer patients [48, 49]. The MethReg prediction that NFATC2 represses *HENMT1* is consistent with the recent study that demonstrated higher expression of *HENMT1* is a favorable prognostic biomarker for colorectal cancer [50], and that NFATC2 is associated with tumor initiation and progression. The *HENMT1* gene encodes a methyltransferase that adds a 2'-O-methyl group at the 3'-end of piRNAs and was previously shown to be

dysregulated in many cancers [51]. Our literature review showed the role of *HENMT1* is not well understood in colorectal cancer. Therefore, this example also shows MethReg can nominate plausible TF functions and targets that can be further studied experimentally. In contrast, without stratifying on DNA methylation levels, the direct TF-target association was only modest (Spearman correlation = - 0.033, P-value = 0.527).

Supplementary Figure 4 shows an example in which DNA methylation attenuates TF activity. In samples with low methylation level at cg02816729, the transcription factor TEAD3 down regulates target gene *SMOC2*. On the other hand, TEAD3 activity is relatively independent of target gene expression in samples with high methylation level at the CpG. TEAD3 belongs to the family of TEAD transcription factors, which also play critical roles in tumor initiation and progression in multiple types of malignancies including gastric, colorectal, breast, and prostate cancers [52]. Higher expression of TEADs, as well as association with poor patient survival, have been observed in many cancers including CRC [52, 53]. On the other hand, the target gene *SMOC2* was recently shown to be a favorable prognostic biomarker for better clinical outcomes in a large cohort of CRC patients [54]. Therefore, the MethReg prediction that transcription factor TEAD3 suppresses target gene *SMOC2* is consistent with the oncogenic role of TEAD3 and the favorable prognostic potential of *SMOC2*. Again, without stratifying on methylation levels, the TF-target association in all samples is only modest (Spearman correlation = 0.07, P-value = 0.183).

Interestingly, some of the TFs exhibited different modes of regulation depending on DNA methylation levels. In Supplementary Figure 5, the expression levels of target gene *BAHCC1* is increased by transcription factor POU3F4 in low methylation samples (P-value = 6.42×10^{-4}) but is decreased by POU3F4 in high methylation samples (P-value = 1.31×10^{-4}). Therefore, DNA methylation effectively increases the diversity of TF functions in this case. Without stratifying on DNA methylation, the TF-target association is not significant (Spearman correlation = 0.083, P-value = 0.112). POU3F4 belongs to the family of POU domain transcription factors, which are involved in different developmental processes and also recently found to be associated with the malignancy processes in different human tissues [55-59]. In addition,

expression levels of the target gene *BAHCC1* were shown to be associated with survival times in different types of cancers such as melanoma [60], liver cancer [61] and pancreatic cancer [62].

Among other transcription factors in the top 10 triplets (Table 2), SNAI1 is a zinc-finger protein involved with inducing the epithelial-mesenchymal transition (EMT) process during which tumor cells become invasive with increased apoptotic resistance [63-65]. The MEIS1 homeodomain protein is a tumor suppressor and was observed to be down-regulated in colorectal adenomas [66]. Similarly, ISL2 is also a tumor suppressor and ISL2 loss increased cell proliferation and enhanced tumor growth in pancreatic ductal adenocarcinoma cells [67]. ETV4 belongs to a subfamily of the ETS family with well-known oncogenic properties [68]. In particular, ETV4 is over-expressed in colorectal tumor samples and higher ETV4 expression is associated with shorter patient survival time [69-75]. ZNF384 is a zinc finger protein that is found to be over-expressed in several cancers, including leukemia [76, 77], liver cancer [78], and colorectal cancer [79]. Finally, FOXL1 is an important regulator of the Wnt/APC/β-catenin pathway, which frequently activates event in gastrointestinal carcinogenesis [80-82]. Similarly, the target genes in the top 10 triplets have also been implicated previously in CRC and other cancers (Table 2). Moreover, among the significant triplets identified by MethReg, the majority of the CpGs, TFs, and target genes (78%, 79% and 75%, respectively) were also differentially methylated or differentially expressed between tumor and normal samples (Supplementary Tables 2-4). Taken together, these results demonstrated that MethReg is able to identify biologically meaningful signals from real multi-omics datasets.

Comparative analysis of MethReg (in unsupervised mode) with alternative approaches

Encouraged by the consistency of MethReg results with recent colorectal cancer literature, we next performed a comparative analysis of MethReg with several alternative approaches. As evidence from large-scale analyses that a substantial number of TFs can interact with methylated DNA have only become available in recent years, few studies have analyzed DNA methylation, TF binding sites, and gene expression data simultaneously using large multi-omics datasets until the past year.

We considered a total of three alternative approaches: the recent studies by Wang et al. (2020) [30] and Liu et al. (2019) [83], as well as the conventional approach of directly correlating DNA methylation with gene expression. In Wang et al. (2020), a rewiring score constructed based on TF-target gene correlations in high and low methylation samples was proposed to identify methylation-sensitive TFs in different cancers using TCGA data. Similarly, using a conditional mutual information-based approach, Liu et al. (2019) identified CpG-TF-target triplets in which the TF-target regulation circuit is dependent on CpG methylation levels in different cancers using TCGA data. In both of these studies, sample permutations were used to estimate P-values for test statistics based on rewiring scores or conditional mutual information.

Conceptually, compared to these previous studies, MethReg makes several distinct contributions: (1) MethReg analysis is comprehensive. While both previous works [30, 83] analyzed only promoter CpG methylation, MethReg analyzes methylation CpGs at both promoter and distal (enhancer) regions. The identification of distal regulatory elements is crucial as they are often not well defined. (2) MethReg analysis is flexible. In addition to the example databases we have described here, MethReg is capable of incorporating any user-specified ChIP-seq or regulons database (Supplementary Table 1), including those that are tissue- or disease-specific. The power and potential of MethReg grows as more knowledge on TF regulation becomes available. (3) MethReg analysis is rigorous. Robust linear models are implemented to carefully model outlier samples in RNA-seq data. Further, with the permutation approaches used in the previous studies, it can be challenging to adjust for covariate information that is important for analyzing human datasets, for which confounding variables such as tumor purity are often significant contributors to both DNA methylation and gene expression. In contrast, MethReg's regression-based approach makes it easy to adjust for potential confounding variables. (4) Most importantly, while previous studies provided analysis results for TCGA cancer datasets, until now no software has been made available to the research community. This study provides an open-source software, MethReg, which implements careful bioinformatics and statistical analysis, to make it possible for researchers to analyze datasets beyond TCGA. In addition to providing estimated individual and joint regulatory effects of CpG methylation and TFs,

MethReg also annotates the potential regulatory roles of TFs and CpG methylation, as well as providing a rich suite of figures for visualizing analytical results.

Next, we performed a systematic comparison of MethReg analysis results and results from other approaches empirically using the TCGA COAD dataset. Because both previous studies analyzed TF gene expressions, we performed additional MethReg promoter analysis that model the TF effect based on TF gene expression, instead of TF activities. More specifically, to compare with results from Wang et al. (2020), which identified 3,244 TF-target gene pairs, we applied the robust linear model described above to the corresponding triplets (average promoter DNA methylation, TF gene expression, and target gene expression). Our results showed that for a majority (81.23%, n = 2613) of the significant TF-target pairs from Wang et al. (2020), the corresponding triplets also had significant DNAm × TF effect in the rlm.binary model from MethReg. Moreover, classification for promoter methylation effects on TFs in Wang et al. (2020) and MethReg agreed very well (kappa statistic = 0.975, P-value = 0) (Supplementary Table 5).

Similarly, we also fitted our rlm.binary model described above to the 47,029 triplets identified in Liu et al. (2020) for the TCGA COAD dataset. However, we observed less agreement between our significant results and those from Liu et al. (2019), as only 5,321 (11.3%) of the 47,029 triplets had significant DNAm x TF p-values by MethReg. The overlap between Liu et al. (2019) with Wang et al. (2020) was also very low, with only 8 TF-target associations identified by both studies. This discrepancy might be due to the differences in methodologies. The conditional mutual information approach used by Liu et al. (2019) detects any general associations which can be non-monotonic, while the robust linear model MethReg used and the rewiring score Wang et al. (2020) used mainly detects linear and monotonic associations in TF and target genes that are dependent on CpG methylation.

Finally, we compared MethReg promoter analysis results with the conventional approach of correlating methylation-target gene directly. More specifically, Spearman correlations were computed for each promoter CpG and target gene in the COAD-READ dataset. The correlation between rankings of P-values based on methylation-target gene correlation and $DNAm \times TF$ interaction effects in the MethReg model is

a significant but modest association (Spearman correlation = 0.0533, P-value = 2.2×10^{-16}), indicating these two approaches are identifying many different CpGs. This is not surprising, however, because the MethReg model identifies CpG methylation that influences the target gene by regulating TF activities (Figure 5B), instead of influencing the target gene directly (Figure 5A).

While MethReg is not designed to identify direct methylation-target associations, it can be useful for identifying those methylation-target correlations likely driven by TF effects (Figure 5C). In this case, TF activities are associated with DNA methylation levels, and TF also regulates gene expression independently of DNA methylation (Fig 5C). In statistics, TF represents a confounding effect in the methylation-target gene association. For example, Supplementary Figure 6 shows an example from the analysis of triplets in Wang et al. (2020) described above, where the observed correlation between promoter DNA methylation and target gene expression is highly significant (Spearman correlation = 0.238, P-value = 4.79×10^{-5} , FDR = 2.81×10^{-4}). However, the result of fitting the MethReg model indicated that for this triplet, neither *DNAm* nor *DNAm* × *TF* terms were significant (P-values = 0.395 and 0.477, respectively), but *TF* was highly significant (P-value = 5.75×10^{-5}). Therefore, the target gene expression is likely driven mainly by the TF EBF1, a tumor suppressor with prognostic value for CRC [84], and not by DNA methylation, even though we observed a highly significant methylation-target gene expression association.

Case study: analysis of ROSMAP Alzheimer's disease dataset

A supervised MethReg analysis

In this section, we demonstrate supervised MethReg analysis using an Alzheimer's disease (AD) dataset collected by the ROSMAP study [85]. In contrast to unsupervised analysis which tests triplets involving all CpGs, in a supervised analysis, we only test triplets involving differentially methylated CpG sites (DMS), typically obtained from EWAS studies. To study AD associated DNA methylation changes in the brain, we recently performed a meta-analysis of over 1,000 prefrontal cortex brain samples from four large brain studies [6, 86-88], and identified 3,751 significant CpGs at 5% FDR [89]. To help understand the regulatory roles of these DMS, we applied MethReg to analyze matched DNA methylation and gene expression

profiles measured on prefrontal cortex brain samples from 529 independent subjects in the ROSMAP dataset.

To illustrate the versatility of the MethReg analysis pipeline, we used alternative databases to analyze the ROSMAP dataset, compared to the analysis pipeline for the TCGA COAD-READ samples. More specifically, to map TF binding sites, instead of the JASPAR2020 database [30] here we used the ReMap database [26], which contains a large collection of regulatory regions obtained using genome-wide DNA-binding experiments such as ChIP-seq. In particular, the ReMap human atlas included binding regions for 1,135 transcriptional regulators. Also, to analyze these brain samples, instead of the pan-cancer regulons, we used the brain-specific TF regulons included in the ChEA3 [90] software website, along with TF activity scores estimated by Gene Set Variation Analysis (GSVA) [34].

More specifically, we first adjusted methylation and gene expression values separately by potential confounding effects, including age at death, sex, batch effects, and markers of different cell types. Next, we computed TF activities using GSVA method [34], which is an alternative method to VIPER [33] for computing enrichment scores of each TF, by comparing enrichment in target gene expressions for a TF (its regulons) with expression levels of background genes.

At 5% false discovery rate, MethReg identified 1, 20, and 103 triplets which included 1, 16, and 53 unique TFs that interact with DMS to influence target gene expressions in promoter, distal, and regulon-based analyses, respectively. A comparison with the MeDReaders database [46] shows more than half (58.6%, 41 out of 70) of these TFs were previously shown to interact with methylated DNA sequences (Supplementary Tables 6 and 7). In Table 3, many of the TFs and target genes in the top 10 triplets were previously implicated in AD pathology. For example, in the most significant triplet (Table 3, Supplementary Figure 7), the transcription factor SPI1 (PU.1) is a master regulator in the AD gene network [91, 92]. SPI1 is critical for regulating the viability and function of microglia [93], which are resident immune cells of the brain. Microglia functions as primary mediators of neuroinflammation and phagocytoses amyloid-beta peptides accumulated in AD brains [94]. Using transgenic mouse models for AD, and comparing with gene-level variations in recent human AD GWAS meta-analysis [95], Salih et al. (2019) [96] recently showed

the target gene *LAPTM5* (lysosome-associated protein, transmembrane 5) belongs to an amyloid-responsive microglial gene network, and predicted *LAPTM5* to be one of four new risk genes for AD. Moreover, *LAPTM5* was also shown to be a member of human microglia network in AD in multiple gene expression studies [97, 98]. Intriguingly, SPI1 and *LAPTM5* belonged to the same transcription co-expression network [96], and in mouse microglial-like BV-2 cells, results from ChIP-seq experiment showed SPI1 binds to the regulatory region of *LAPTM5* [99], consistent with the MethReg prediction that *LAPTM5* is regulated by SPI1. Interestingly, SPI1 appears to have dual roles in this case, depending on DNA methylation levels at cg17418085, which is located in the gene body of *LAPTM5*. More specifically, SPI1 upregulates the target gene *LAPTM5* in samples with low methylation at cg17418085 but down-regulates the target gene when DNA methylation is high, suggesting DNA methylation and the TF might have compensatory mechanisms that control gene expression at this locus (Supplementary Figure 7).

The triplet cg08824847- NR3C1- PDHX is an example in which methylation at cg08824847 appears to enhance activation of the target gene *PDHX* by NR3C1 (Table 3, Supplementary Figure 8). NR3C1 is the glucocorticoid receptor (GR) which can act as a transcription factor that binds glucocorticoid responsive genes to activate their transcriptions, or as a regulator for other TFs. NR3C1 regulates downstream processes such as glycolysis [100] and was observed to be dysregulated in AD [101-103]. The target gene *PDHX* encodes the component X of the pyruvate dehydrogenase complex (PDH), which is involved in the regulation of mitochondrial activity and glucose metabolism in the brain that is critical for neuron survival [104, 105]. Low levels of cerebral glucose metabolism often proceed with the onset of Alzheimer's Disease and have been proposed as a biomarker of AD risk [106-109]. Previously it was also shown that along with other regulators, glucocorticoids can regulate the efficacy of PDH [100, 110]. The MethReg prediction that methylation at cg08824847 enhances activation of *PDHX* by NR3C1 is consistent with these previous findings that demonstrated lower levels of *PDHX* in AD samples, and with results from our previous large meta-analysis of DNA methylation changes in AD, which discovered cg08824847 to be hypo-methylated in AD samples across all four analyzed brain samples cohorts and has a significant negative association with AD Braak stage even after multiple comparison correction [89].

Among the other TFs in the top 10 triplets (Table 3), ESR1 is estrogen receptor alpha, one of two subtypes of estrogen receptor. Genetic polymorphisms of ESR1 have been associated with risk of developing cognitive impairment in older women [111-115], as well as faster cognitive decline in women AD patients [116]. Induced by chronic inflammation in AD, CEBPD is associated with microglia activation and migration [117, 118]. TCF12 is a member of the basic helix-loop-helix (bHLH) E-protein family and plays important roles in developmental processes such as neurogenesis, mesoderm formation, and cranial vault development. Recently, TCF12 was predicted to be affected by SNP rs10498633 [119], a top AD-associated SNP identified in the IGAP AD meta-analysis study [120]. Moreover, TCF12 also belongs to a herpesvirus perturbed transcription factor regulatory network that is implicated in AD [121]. SRF is serum response factor, responsible for regulating the smooth muscle cells and blood flows in the brain, which are important for a blood vessel's ability to remove amyloid beta peptides accumulated in AD. Compared with healthy individuals, SRF was found to be four times higher in AD patients [122, 123]. GABPa belongs to the ETS family of DNA-binding factors and is a master regulator of multiple important processes including cell cycle control, apoptosis, and differentiations. Using evolutionary analysis and ChIP-seq experiments, Perdomo-Sabogal et al. (2016) [124] linked GABPa to several brain disorders including AD, autism, and Parkinson's disease. Finally, NFE2L2/NRF2 is another master regulator and regulates genes involved in response to oxidative stress and inflammation. Motivated by the encouraging therapeutic effect of NRF2 on AD pathology in animal models and cultured human cells [125-127], modulation of NRF2 pathway has recently been proposed as a strategy for AD drug development [125]. Similarly, Table 3 shows the target genes in these top 10 triplets identified by MethReg are also highly relevant to Alzheimer's disease. Taken together, these results demonstrated MethReg is also capable of identifying biologically meaningful regulatory effects of DNA methylation in other complex diseases, such as the AD brain DNA methylation data we used here, where association signals are expected to be much weaker than those observed in cancers.

Comparative analysis of MethReg (in supervised mode) with alternative approaches

To compare the performance of MethReg in supervised mode with currently available alternative tools, we next analyzed the ROSMAP dataset using the ReMapEnrich R package [26], which identifies regulators with binding sites enriched in user supplied regions. Several other tools such as LOLA [21] and ChIP-Enrich [128] perform similar analyses as ReMapEnrichR, but here we chose ReMapEnrichR because it uses the same ReMap database as MethReg for the ROSMAP dataset analysis. More specifically, for the ReMapEnrichR analysis, we also used the locations of the 3,751 AD-associated CpGs (these are the DMS) from previous AD meta-analysis as the input. The results showed that ReMapEnrich identified 143 TFs with binding sites enriched in the DMS, among which a substantial number (n = 28, 20%) were also identified by MethReg (Supplementary Table 8). These 28 TFs included many well-known regulators for AD such as TCF12, SPI1, NR3C1, CEBPB, GABPA and others. On the other hand, 115 and 32 TFs were uniquely identified by ReMapEnrich or MethReg, respectively.

Although many of these significant TFs have been previously implicated in AD pathology, their specific roles in transcription regulation and the identification of their targets in AD remain to be investigated. Notably, currently available tools such as ReMapEnrich only identify the TFs but do not consider CpGs or provide detailed information on the relevant target genes. In contrast, MethReg fills this critical gap by nominating plausible TF-target associations that are mediated by DNA methylation. Therefore, MethReg analysis which leverages additional gene expression data and provides more comprehensive information on transcription regulation for the TFs complements existing approaches.

Discussion

To evaluate the role of DNA methylation in gene regulation, we developed the MethReg R package. MethReg provides a systematic approach to dissect the variations in gene transcription into three different modes of regulation, which are direct effects by methylation and TF individually, and the synergistic effect from both DNA methylation and TF. By additionally modeling DNA methylation variations, MethReg complements existing approaches that analyze TF and target gene expression alone. In doing so, MethReg uncovers TF-target gene relations that are present only in samples with high (or low) methylation levels at

the CpG that mediates TF activities. On the other hand, compared to approaches that analyze DNA methylation and TFBS data alone, MethReg analysis also reduces the noisiness in TF binding site predictions by additionally modeling target gene expression data. Compared to the conventional approach of directly correlating DNA methylation with gene expression, MethReg can be useful for identifying methylation-gene expression associations which are likely driven by TFs instead of methylation due to confounding effects by TFs. Computationally, MethReg is efficient. The unsupervised analysis of TCGA COAD-READ datasets which considered all CpGs on the Illumina array took 5, 37, and 14 minutes for the promoter, distal, and regulon-based analyses, respectively, using a single Linux machine with 64 GB of RAM memory and Intel Xeon W-2175 (2.50 GHz) CPUs with 4 cores for parallel computing (Supplementary Table 9).

Because of current limitations in technology, directly measuring DNA methylation, TF binding and target gene expression in high throughput is still a difficult task, especially for a large cohort of primary tissue samples. Therefore, computational approaches are needed to prioritize regulatory elements in gene transcription. To this end, a main computational challenge is the accurate assessment of TF activities. Many integrative studies have used TF gene expression data, which are often widely available, as surrogate measurements for TF activities. However, the abundance of TF expression does not necessarily correspond to more TF binding events, which needs to be confirmed by cell type specific ChIP-Seq experiments. On the other hand, TF binding events are sometimes non-functional and might not lead to changes in gene expression [24, 25]. To this end, we implemented the option to model TF effects based on VIPER [33] or GSVA [34] estimated TF protein activities in MethReg. Both of these methods assume that collectively, the target genes of a TF represent an optimal reporter of its activity and estimate TF activity based on enrichment of its target genes expressions compared to background genes. Both VIPER and GSVA approaches have been widely used for modeling protein activities using gene expression datasets.

While the motivation of MethReg is to rank significant and functional DNA methylation changes identified in EWAS, a useful by-product of this analysis is the identification of enhancers, which are often located several hundreds of kilobases (kb) away from the target gene, where TFs bind and interact with

DNA methylation to activate gene expression by looping DNA segments [129]. Growing evidence indicates that in addition to promoter methylation, DNA methylation at enhancers also plays an equally or more important role in activating gene expression [130]. Active cancer-specific enhancers are typically hypomethylated at CpG sites [10, 131, 132] in open chromatin regions free of nucleosomes [133, 134]. In many cases, hypermethylation of CpG sites can interfere with TF binding and lead to decreased enhancer activities in various cancers [9, 135]. It has been observed that TF activities often correlate with levels of demethylation at enhancer regions and subsequent target gene expressions [135-137].

Although recently many cis-regulatory regions have been identified using genomic and epigenomic data [136, 138], assigning these candidate enhancers to target genes on a genome-wide scale remains challenging and is currently an active area of research. A recent study [139] compared several published computational approaches for enhancer-gene linking using a collection of experimentally derived genomic interactions. It was shown that the best-performing method, TargetFinder [140], is only modestly better than the baseline distance-based approach, and the authors suggested further improvement in current computational methods in this area is needed. Although many of these computational methods leverage various information from histone marks, chromatin accessibility and interaction, TF binding models and gene expression levels, few if any also model DNA methylation at candidate enhancer regions. As many recent studies suggested substantial crosstalk between DNA methylation and other regulatory elements such as transcription factors [14, 16], we developed MethReg to specifically estimate and evaluate the regulatory potential of DNA methylation for interacting with candidate TFs at both promoter and distal regions. In particular, MethReg links target genes with CpGs in distal regions using two alternative approaches: linking to a fixed number of nearby genes or by using annotations in regulon databases (Figure 1). Indeed, among significant triplets identified in MethReg analysis of ROSMAP dataset, a substantial number of the CpGs (6 and 22 in distal and regulon-based analyses, respectively) were located in brain-specific enhancer regions annotated in the EnhancerAtlas 2.0 database [141] (Supplementary Tables 6-7). Notably, the power and potential of MethReg will also grow as more knowledge on TF regulatory activities are accumulated and new ChIP-seq and TF regulon databases become available.

The aim of MethReg is to prioritize functional elements and to generate useful testable hypothesis for subsequent mechanistic studies. For the DNA methylation changes identified by MethReg which couple with TF activities, additional experimental studies are needed to determine whether the changes in methylation are causing or are caused by TF activities. Nevertheless, even if the DNA methylation changes are passive markers that accumulated as result of TF binding (or lack of binding), they can still be useful in the clinics as biomarkers. Currently, many of the large DNA methylation datasets are measured using methylation arrays because of their lower cost and the simplicity in benchwork and analysis. MethReg has been tested successfully on microarrays and can also be easily extended to analyze large cohorts of samples measured using high throughput sequencing such as WGBS or RRBS. In addition to transcription factors, MethReg can also be applied to analyze other types of chromatin proteins including histones which are known to crosstalk with DNA methylation [142-144]. Finally, MethReg can further be extended to incorporate TF-target associations based on spatial enhancer-gene linking when 3D chromatin and genetic interaction data on primary cells become available.

Conclusions

We have presented an integrative analysis and annotation software, MethReg, which has several critical roles. First, given the large number of DMS identified from EWAS, supervised MethReg analysis can be used to analyze CpG-TF-target gene triplets involving the DMS, to identify and prioritize important CpG methylations that influence target gene expressions by interacting with TFs that bind in proximity. Second, MethReg annotates CpG methylation and the TFs that bind in close proximity with their regulatory roles (i.e., activator or repressor TFs, DNA methylation that attenuate or enhance TF effects). Third, because some TFs affect target gene expression only in samples with high methylation levels (or only in samples with low methylation levels), MethReg can help uncover TF-target gene associations that are not obvious in an analysis that uses all samples. Finally, for a particular target gene, MethReg partitions the variances in gene regulation into direct impact by methylation, direct impact by TF, or joint impact of both methylation and TF, which allows MethReg to identify methylation-target gene associations that are likely

driven by TF instead of DNA methylation. Using two case studies in colorectal cancer and Alzheimer's disease, we have shown the power of MethReg to uncover biologically relevant transcriptional regulation in both diseases which have vastly different biology. Open source software scripts, along with extensive documentation and example data for MethReg are freely available from the Bioconductor repository. We hope MethReg will empower researchers to gain a better understanding of the important regulatory roles of CpG methylation in many complex diseases.

Methods

Unsupervised MethReg analysis of TCGA COAD-READ samples

Datasets

Level 3 TCGA COAD and READ datasets, including genomic profiles in copy number alterations (CNAs; Gene Level Copy Number Scores), gene expressions (HTSeq - FPKM-UQ values from RNA-seq), and DNA methylation levels (beta values from 450k Illumina arrays) were downloaded from the NCI's Genomic Data Commons (GDC) using TCGAbiolinks R package (version 2.17.4) with functions GDCquery, GDCdownload, and GDCprepare [145]. These level 3 datasets were previously pre-processed and normalized as described in https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines. We included only primary tumors samples in our analysis. Tumor purity estimates were downloaded from Supplementary Data 1 of Aran D et al. (2015) [45]. We included 367 samples with DNA methylation, gene expression, and copy number alterations (CNAs) profiles. To compute TF activity scores, we used the pan-cancer regulon database from Garcia-Alonso et al. (2019) [29], downloaded from https://github.com/saezlab/dorothea/tree/deprecated/data/TFregulons/consensus/Robjects_VIPERformat/pancancer

Pre-processing data

First, we removed genes and CpGs with little expression variations across samples. More specifically, from a total of 19,456 genes with matched DNA methylation, gene expression, and CNA values, we removed

2,315 genes with zero counts in more than 50% of the samples in RNA-seq data. For each CpG, we computed inter-quartile range (IQR) of the beta values and removed CpGs with IQR less than 0.2. In addition, to avoid confounding by TF effect (Figure 5C), we focused the analysis on those CpG-TF-target gene triplets where CpG was associated with the target gene (Wilcoxon test p-value less than 0.05), but was not associated with TF (Wilcoxon test p-value greater than 0.05).

Next, for each gene, the get_residuals function in MethReg was used to remove potential confounding effects in gene expression data by fitting a linear model which included $\log_2$ transformed gene expression as the outcome variable, and CNA and tumor purity estimates as predictor variables. The residuals obtained from this model were then used for subsequent analyses. To estimate TF activity scores, we used the VIPER algorithm [33], which is implemented in function run_viper in R package dorothea, along with the pan-cancer regulons described above.

Creating triplet datasets

In the promoter analysis, the function create_triplet_distance_based with parameter target.method = "genes.promoter.overlap" were used to link CpGs in the promoter regions (within ± 2 kb regions around the transcription start sites (TSS)) to the corresponding target genes. Similarly, in distal analysis, the function create_triplet_distance_based with parameters target.method = "nearby.genes" and target.num.flanking.genes = 5 were used to link CpGs in the distal region (beyond ± 2 kb regions around the TSS) to five genes upstream and downstream of the CpG. Alternatively, in regulon-based analysis, the function create_triplet_regulon_based was used to link CpGs to the TF-target gene pairs in the pan-cancer regulons. In all three analyses, we obtained TF binding sites information from the JASPAR2020 [27] database and linked CpGs to TFs with binding sites within 250 bp distance using *motifmatchr* R package (version 1.10.0). The CpG-TF pairs were then combined with CpG-target gene pairs (promoter and distal analyses) or TF-target pairs (regulon-based analysis) to obtain CpG-TF-target triplets. In all analysis, we specified the parameter max.distance.region.target = 10^6 to obtain triplets for which the CpG is within

1M bp distance from the target gene. These analyses were performed using hg38 genome assembly coordinates.

Statistical analysis

Given CpG-TF-target gene triplets from "Creating triplet datasets" section, and DNA methylation, estimated TF activities, and residual gene expression data from "Pre-processing data" section, the function interaction_model was then used to fit robust linear statistical model *residual.target.gene.expression* ~ *TF* + *DNAm.group* + *DNAm.group* × *TF*. Within the interaction_model function, the robust linear model analysis is implemented by rlm function from the MASS R package, with parameter psi = psi.bisquare (bisquare weighting) [146], which gives outlier gene expression values reduced weight. After p-values for the interaction term $DNAm.group \times TF$ was computed for each triplet, we next performed stage-wise analysis by setting stage.wise.analysis = TRUE in interaction_model function. The stage-wise analysis in *MethReg* is implemented by calling stageRTx function of stageR R package [37]. For each CpG, we grouped all triplets associated with it and used the smallest p-value for *DNAm.group* × *TF* to represent the CpG. These smallest p-values for each CpG were then used to define values for pScreen in function stageRTx. The parameter pConfirmation was defined as $DNAm.group \times TF$ p-values for each triplet. In the screening step, the null hypothesis that any of the individual triplets mapped to a CpG had a significant *DNAm.group* × *TF* effect was tested. In the confirmation step, all the triplets associated with significant CpGs selected in the screening step were tested while controlling FWER [37].

Annotating roles of TFs and effects of CpG methylations

For triplets with significant *DNAm.group* × *TF*, we next annotated roles of TF and CpG methylations in gene regulation (Figure 3). To this end, the function stratified_model fits the robust linear model $residual.target.gene.expression \sim TF$ in samples with high or low methylation levels at the relevant CpG separately, to obtain p-values for TF effect (pval.tf.high and pval.tf.low). Let the most significant p-

value among the two p-values be pval.most.sig. A TF is annotated only if pval.most.sig < 0.001 (that is pval.tf.high < 0.001 or pval.tf.low < 0.001). If the estimated TF effect (i.e., slope) corresponding to the most significant p-value (pval.most.sig) is positive, the TF is annotated as an activator. If the estimated TF effect corresponding to pval.most.sig is negative, the TF is annotated as a repressor. To annotate CpG methylation effect, if pval.most.sig is from the model fitted to high methylation samples, the CpG is labeled as enhancing; otherwise the CpG is labeled as attenuating. If pval.tf.high < 0.001 and pval.tf.low < 0.001, and the TF effects (i.e., slopes) in high and low CpG methylation samples have opposite signs, then the TF is annotated to have dual roles and the CpG is labeled to have invert effect.

Tumor vs. normal samples analysis

For each significant triplet from the stageR analysis described above, we additionally assessed differential methylation of the CpGs and differential expressions of the TFs and target genes in tumor vs. normal samples. To this end, 21 adjacent normal samples with matched gene expression (HTSeq - FPKM-UQ values from RNA-seq) and DNA methylation levels (beta values from 450k Illumina arrays) were downloaded from NCI's Genomic Data Commons (GDC) using TCGAbiolinks R package (version 2.17.4) using functions GDCquery, GDCdownload *and* GDCprepare [145]. To assess the statistical significance of differential methylation levels at CpGs (or differential gene expression levels at TF and target genes) in tumor vs. normal samples, we used two-tailed Wilcoxon test.

Simulation study

The *rlm.binary.en* method estimates P-values for each triplet using empirical null distribution [38, 147], which is a normal distribution with empirically estimated mean $\hat{\delta}$ and standard deviation $\hat{\sigma}$. There are five steps:

1) For each triplet, fit rlm.binary model to obtain *t*-statistic and degrees of freedom (*df*) corresponding to *DNAm.group* × *TF* effect.

2) Convert the t-statistic to a z-score. For example, let t_i be the *t*-statistic for triplet *i*, the corresponding z-score can be obtained by $z_i = \Phi^{-1}(F_d(t_i, df))$ where Φ and F_d are distribution functions for standard normal distribution and *t* distribution with *df* degrees of freedom.
3) Pool the z-scores for all triplets tested in an analysis and calculate their median value (m). Subtract m from the z-scores so that they have median of 0.
4) Given the median-centered z-scores, use the locfdr package, estimate location ($\hat{\delta}$) and scale ($\hat{\sigma}$) parameters of the empirical null distribution.
5) Calculate standardized z-scores, $s_i = (z_i - m - \hat{\delta}) / \hat{\sigma}$, and compute the *P*-value for each triplet as $p_i = 1 - \Phi(s_i)$.

In triplets simulated under the scenario β = 0, target gene expressions did not depend on TF expression levels, so the genes simulated under this simulation scenario were the null triplets. Type I error rate for each model was computed as the number of triplets declared to be significant by the model when β = 0, divided by the total number of null triplets. P-values less than 0.05 for each method were considered to be significant. On the other hand, power was computed using triplets simulated under the scenarios where β = 1, .., 9. For each of the 7 compared methods (Supplementary Figure 1), power was estimated as the number of triplets with P-values less than 0.05 for each compared method, divided by the total number of simulated triplets (i.e. 1,000). The pROC R package was used to compute area under ROC curves.

Comparative analysis with alternative approaches

For the comparison with Wang et al. (2020) [30], we downloaded data from the DMTDB database (http://bio-bigdata.hrbmu.edu.cn/DMTDB; accessed on June 5, 2020). The file "COAD_DMTD" contained 3,244 COAD DNA methylation dependent TF-target gene pairs, where target gene expressions were regulated by the TF and average DNA methylation at promoter region (± 2k bp) of the target gene. For the comparison with promoter analysis results from Liu et al. (2019) [83], we downloaded promoter CpG

methylation-dependent TF-target associations from file "COAD.txt" in Supplementary Data S1 of Liu et al. (2019). We included 42,590 triplets with non-masked CpGs [148] in our analysis.

Analysis of ROSMAP Alzheimer's disease dataset

Datasets

The ROSMAP dataset, which included samples from the Religious Order Study (ROS) and the Memory and Aging Project (MAP) [85], included 529 samples with matched DNA methylation and gene expression data [6]. More specifically, normalized FPKM (Fragments Per Kilobase of transcript per Million mapped reads) gene expression values and Illumina HumanMethylation 450K beadchip files (.idat format) for the ROSMAP study were downloaded from the AMP-AD Knowledge Portal (Synapse ID: syn3388564). Brain-specific regulons (file "brain.TFs.gmt") were downloaded from ChEA3 software website https://maayanlab.cloud/chea3/ (accessed Oct 10, 2020).

Pre-processing datasets

DNA methylation pre-processing included the removal of CpG probes with detection P-value > 0.01 across all the samples in the cohort, and CpG probes in which a single nucleotide polymorphism (SNP) with minor allele frequency (MAF) ≥ 0.01 was present in the last 5 base pairs of the probe. The quality controlled methylation datasets were next subjected to the QN.BMIQ normalization procedure [149]. More specifically, we first applied quantile normalization as implemented in the lumi R package to remove systematic effects between samples. Next, we applied the β-mixture quantile normalization (BMIQ) procedure [150] as implemented in the wateRmelon R package [151] to normalize beta values of type 1 and type 2 design probes within the Illumina arrays.

Next, we removed confounding effects in DNA methylation data by fitting the model *median methylation M value ~ neuron.proportions + batch + sample.plate array + ageAtDeath + sex* and extracting residuals from this model, which are the methylation residuals. Similarly, we also removed potential confounding effects in RNA-seq data by fitting model log_2 *(normalized FPKM values + 1) ~*

ageAtDeath + sex + markers for cell types. The last term *markers for cell types* included multiple covariate variables to adjust for the multiple types of cells in the brain samples. More specifically, we estimated expression levels of genes that are specific for the main five cell types present in the CNS: ENO2 for neurons, GFAP for astrocytes, CD68 for microglia, OLIG2 for oligodendrocytes, and CD34 for endothelial cells, and included these as variables in the above linear regression model, as was done in a previous large study of AD samples [6]. The residuals extracted from this model are the gene expression residuals.

TF activities were calculated using the collection of brain-specific TF Regulons (downloaded from ChEA3 software website at https://maayanlab.cloud/chea3/assets/tflibs/brain.TFs.gmt) and gene expression residuals, using the R package GSVA (Gene Set Variation Analysis) with the following parameters: method = "gsva", kcdf = "Gaussian", abs.ranking = TRUE, min.sz = 5, max.sz = Inf, mx.diff = TRUE, ssgsea.norm = TRUE.

Creating triplet datasets

For supervised MethReg analysis, we considered the 3,751 differently methylated CpGs associated with Braak stage identified in Zhang et al. (2020) [89]. First, these CpGs were filtered to remove probes with beta-values IQR < 0.03, resulting in 2,761 CpGs. For CpGs in the promoter regions, the function create_triplet_distance_based with parameters target.method = "genes.promoter.overlap" was used to link each of the 1002 CpGs in promoter regions (within ± 2 kb regions around the TSS) to the target gene with promoter region that overlaps with the CpG. In distal analysis, the function create_triplet_distance_based with parameters target.method = "nearby.genes" and target.num.flanking.genes = 5 were used to link each of the 1,759 CpGs in distal regions (beyond ± 2 kb regions around TSS of the target genes) to five genes upstream and downstream of the CpG. Alternatively, in regulon-based analysis, the function create_triplet_regulon_based was used to link CpGs to the TF-target gene pairs in the brain-specific regulons (downloaded from https://maayanlab.cloud/chea3/). In all three analyses, we obtained TF binding sites information from the REMAP2020 database and linked CpGs

to TFs with binding sites within 250 bp. The CpG-TF pairs were then combined with CpG-target gene pairs (promoter and distal analyses) or TF-target pairs (regulon-based analysis) to obtain CpG-TF-target triplets. In all analyses, we specified the parameter max.distance.region.target = 10^6 to obtain triplets for which the CpG is within 1M bp distance from the target gene. These analyses were performed using hg19 genome assembly coordinates.

Overlap with cis-regulatory elements in enhancer database

We assessed the overlap between regions in significant triplets (CpG location +/- 250bp) with enhancer regions in EnhancerAtlas 2.0 database [152]. The following brain/neuron specific tissues or cell lines for Homo sapiens (hg19) were considered: Astrocyte, hNCC, KELLY, BE2C, ESC_NPC, NGP, SK.N.MC, Cerebellum, Gliobla, SH.SY5Y, SK.N.SH_RA, T98G, Fetal_brain, U87, H54, SK.N.SH, SK.N.MC, NH.A, Macrophage, ESC_neuron.

Acknowledgement

This research was supported by US National Institutes of Health grants R21AG060459 (L.W), R01AG061127 (L.W.), and R01AG062634 (E.R.M, L.W.). The ROSMAP study data were provided by the Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago. Data collection was supported through funding by NIA grants P30AG10161, R01AG15819, R01AG17917, R01AG30146, R01AG36836, U01AG32984, U01AG46152, the Illinois Department of Public Health, and the Translational Genomics Research Institute.

Availability of data and materials

The MethReg R package is available from the Bioconductor repository at https://bioconductor.org/packages/MethReg/.

Author Contributions

L.W., T.C.S., J.Y. designed the computational analysis. T.C.S., L.W. analyzed the data. L.W., J.Y., E.R.M, X.C. contributed to interpretation of the results. T.C.S, L.W. wrote the paper, and all authors participated in the review and revision of the manuscript. L.W. conceived the original idea and supervised the project.

Ethics approval and consent to participate

Not applicable.

Competing Interest

The authors declare that they have no conflict of interest.

References

1. Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, Cui H, Gabo K, Rongione M, Webster M, et al: **The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores.** *Nat Genet* 2009, **41:**178-186.
2. Cancer Genome Atlas N: **Comprehensive molecular characterization of human colon and rectal cancer.** *Nature* 2012, **487:**330-337.
3. McInnes T, Zou D, Rao DS, Munro FM, Phillips VL, McCall JL, Black MA, Reeve AE, Guilford PJ: **Genome-wide methylation analysis identifies a core set of hypermethylated genes in CIMP-H colorectal cancer.** *BMC Cancer* 2017, **17:**228.
4. Kirby MK, Ramaker RC, Roberts BS, Lasseigne BN, Gunther DS, Burwell TC, Davis NS, Gulzar ZG, Absher DM, Cooper SJ, et al: **Genome-wide DNA methylation measurements in prostate tissues uncovers novel prostate cancer diagnostic biomarkers and transcription factor binding patterns.** *BMC Cancer* 2017, **17:**273.
5. Kuang Y, Wang Y, Zhai W, Wang X, Zhang B, Xu M, Guo S, Ke M, Jia B, Liu H: **Genome-Wide Analysis of Methylation-Driven Genes and Identification of an Eight-Gene Panel for Prognosis Prediction in Breast Cancer.** *Front Genet* 2020, **11:**301.
6. De Jager PL, Srivastava G, Lunnon K, Burgess J, Schalkwyk LC, Yu L, Eaton ML, Keenan BT, Ernst J, McCabe C, et al: **Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci.** *Nat Neurosci* 2014, **17:**1156-1163.
7. Young JI, Sivasankaran SK, Wang L, Ali A, Mehta A, Davis DA, Dykxhoorn DM, Petito CK, Beecham GW, Martin ER, et al: **Genome-wide brain DNA methylation analysis suggests epigenetic reprogramming in Parkinson disease.** *Neurol Genet* 2019, **5:**e342.
8. Tarr IS, McCann EP, Benyamin B, Peters TJ, Twine NA, Zhang KY, Zhao Q, Zhang ZH, Rowe DB, Nicholson GA, et al: **Monozygotic twins and triplets discordant for amyotrophic lateral sclerosis display differential methylation and gene expression.** *Sci Rep* 2019, **9:**8254.
9. Aran D, Sabato S, Hellman A: **DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes.** *Genome Biol* 2013, **14:**R21.
10. Heyn H, Vidal E, Ferreira HJ, Vizoso M, Sayols S, Gomez A, Moran S, Boque-Sastre R, Guil S, Martinez-Cardus A, et al: **Epigenomic analysis detects aberrant super-enhancer DNA methylation in human cancer.** *Genome Biol* 2016, **17:**11.
11. Zhu H, Wang G, Qian J: **Transcription factors as readers and effectors of DNA methylation.** *Nat Rev Genet* 2016, **17:**551-565.
12. Bonder MJ, Luijk R, Zhernakova DV, Moed M, Deelen P, Vermaat M, van Iterson M, van Dijk F, van Galen M, Bot J, et al: **Disease variants alter transcription factor levels and methylation of their binding sites.** *Nat Genet* 2017, **49:**131-138.
13. Banovich NE, Lan X, McVicker G, van de Geijn B, Degner JF, Blischak JD, Roux J, Pritchard JK, Gilad Y: **Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels.** *PLoS Genet* 2014, **10:**e1004663.
14. Hu S, Wan J, Su Y, Song Q, Zeng Y, Nguyen HN, Shin J, Cox E, Rho HS, Woodard C, et al: **DNA methylation presents distinct binding sites for human transcription factors.** *Elife* 2013, **2:**e00726.
15. Lioznova AV, Khamis AM, Artemov AV, Besedina E, Ramensky V, Bajic VB, Kulakovskiy IV, Medvedeva YA: **CpG traffic lights are markers of regulatory regions in human genome.** *BMC Genomics* 2019, **20:**102.
16. Yin Y, Morgunova E, Jolma A, Kaasinen E, Sahu B, Khund-Sayeed S, Das PK, Kivioja T, Dave K, Zhong F, et al: **Impact of cytosine methylation on DNA binding specificities of human transcription factors.** *Science* 2017, **356**.

17. Cedoz PL, Prunello M, Brennan K, Gevaert O: **MethylMix 2.0: an R package for identifying DNA methylation genes.** *Bioinformatics* 2018, **34:**3044-3046.
18. Warden CD, Lee H, Tompkins JD, Li X, Wang C, Riggs AD, Yu H, Jove R, Yuan YC: **COHCAP: an integrative genomic pipeline for single-nucleotide resolution DNA methylation analysis.** *Nucleic Acids Res* 2019, **47:**8335-8336.
19. Saghafinia S, Mina M, Riggi N, Hanahan D, Ciriello G: **Pan-Cancer Landscape of Aberrant DNA Methylation across Human Tumors.** *Cell Rep* 2018, **25:**1066-1080 e1068.
20. Klett H, Balavarca Y, Toth R, Gigic B, Habermann N, Scherer D, Schrotz-King P, Ulrich A, Schirmacher P, Herpel E, et al: **Robust prediction of gene regulation in colorectal cancer tissues from DNA methylation profiles.** *Epigenetics* 2018, **13:**386-397.
21. Sheffield NC, Bock C: **LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor.** *Bioinformatics* 2016, **32:**587-589.
22. Bhasin JM, Ting AH: **Goldmine integrates information placing genomic ranges into meaningful biological contexts.** *Nucleic Acids Res* 2016, **44:**5550-5556.
23. Lawson JT, Tomazou EM, Bock C, Sheffield NC: **MIRA: an R package for DNA methylation-based inference of regulatory activity.** *Bioinformatics* 2018, **34:**2649-2650.
24. Li XY, MacArthur S, Bourgon R, Nix D, Pollard DA, Iyer VN, Hechmer A, Simirenko L, Stapleton M, Luengo Hendriks CL, et al: **Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm.** *PLoS Biol* 2008, **6:**e27.
25. Fisher WW, Li JJ, Hammonds AS, Brown JB, Pfeiffer BD, Weiszmann R, MacArthur S, Thomas S, Stamatoyannopoulos JA, Eisen MB, et al: **DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in Drosophila.** *Proc Natl Acad Sci U S A* 2012, **109:**21330-21335.
26. Cheneby J, Menetrier Z, Mestdagh M, Rosnet T, Douida A, Rhalloussi W, Bergon A, Lopez F, Ballester B: **ReMap 2020: a database of regulatory regions from an integrative analysis of Human and Arabidopsis DNA-binding sequencing experiments.** *Nucleic Acids Res* 2020, **48:**D180-D188.
27. Fornes O, Castro-Mondragon JA, Khan A, van der Lee R, Zhang X, Richmond PA, Modi BP, Correard S, Gheorghe M, Baranasic D, et al: **JASPAR 2020: update of the open-access database of transcription factor binding profiles.** *Nucleic Acids Res* 2020, **48:**D87-D92.
28. Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, Chen X, Taipale J, Hughes TR, Weirauch MT: **The Human Transcription Factors.** *Cell* 2018, **172:**650-665.
29. Garcia-Alonso L, Holland CH, Ibrahim MM, Turei D, Saez-Rodriguez J: **Benchmark and integration of resources for the estimation of human transcription factor activities.** *Genome Res* 2019, **29:**1363-1375.
30. Wang Z, Yin J, Zhou W, Bai J, Xie Y, Xu K, Zheng X, Xiao J, Zhou L, Qi X, et al: **Complex impact of DNA methylation on transcriptional dysregulation across 22 human cancer types.** *Nucleic Acids Res* 2020, **48:**2287-2302.
31. Mei S, Meyer CA, Zheng R, Qin Q, Wu Q, Jiang P, Li B, Shi X, Wang B, Fan J, et al: **Cistrome Cancer: A Web Resource for Integrative Gene Regulation Modeling in Cancer.** *Cancer Res* 2017, **77:**e19-e22.
32. Venables WN, Ripley BD: *Modern Applied Statistics with S.* Fourth edition edn: Springer, New York. ; 2002.
33. Alvarez MJ, Shen Y, Giorgi FM, Lachmann A, Ding BB, Ye BH, Califano A: **Functional characterization of somatic mutations in cancer using network-based inference of protein activity.** *Nat Genet* 2016, **48:**838-847.
34. Hanzelmann S, Castelo R, Guinney J: **GSVA: gene set variation analysis for microarray and RNA-seq data.** *BMC Bioinformatics* 2013, **14:**7.

35. Qin Q, Fan J, Zheng R, Wan C, Mei S, Wu Q, Sun H, Brown M, Zhang J, Meyer CA, Liu XS: **Lisa: inferring transcriptional regulators through integrative modeling of public chromatin accessibility and ChIP-seq data.** *Genome Biol* 2020, **21:**32.
36. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *Journal of the Royal Statistical Society, Series B* 1995, **57:**289-300.
37. Van den Berge K, Soneson C, Robinson MD, Clement L: **stageR: a general stage-wise method for controlling the gene-level false discovery rate in differential expression and differential transcript usage.** *Genome Biol* 2017, **18:**151.
38. Efron B: **Correlated z-values and the accuracy of large-scale statistical estimates.** *J Am Stat Assoc* 2010, **105:**1042-1055.
39. **Facts and stats** [https://fightcolorectalcancer.org/prevent/about-colorectal-cancer/facts-stats/]
40. Berman BP, Weisenberger DJ, Aman JF, Hinoue T, Ramjan Z, Liu Y, Noushmehr H, Lange CP, van Dijk CM, Tollenaar RA, et al: **Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains.** *Nat Genet* 2011, **44:**40-46.
41. Ehrlich M: **DNA hypomethylation in cancer cells.** *Epigenomics* 2009, **1:**239-259.
42. Lee TI, Young RA: **Transcriptional regulation and its misregulation in disease.** *Cell* 2013, **152:**1237-1251.
43. Ell B, Kang Y: **Transcriptional control of cancer metastasis.** *Trends Cell Biol* 2013, **23:**603-611.
44. Puisieux A, Brabletz T, Caramel J: **Oncogenic roles of EMT-inducing transcription factors.** *Nat Cell Biol* 2014, **16:**488-494.
45. Aran D, Sirota M, Butte AJ: **Systematic pan-cancer analysis of tumour purity.** *Nat Commun* 2015, **6:**8971.
46. Wang G, Luo X, Wang J, Wan J, Xia S, Zhu H, Qian J, Wang Y: **MeDReaders: a database for transcription factors that bind to methylated DNA.** *Nucleic Acids Res* 2018, **46:**D146-D151.
47. Daniel C, Gerlach K, Vath M, Neurath MF, Weigmann B: **Nuclear factor of activated T cells - a transcription factor family as critical regulator in lung and colon cancer.** *Int J Cancer* 2014, **134:**1767-1775.
48. Gerlach K, Daniel C, Lehr HA, Nikolaev A, Gerlach T, Atreya R, Rose-John S, Neurath MF, Weigmann B: **Transcription factor NFATc2 controls the emergence of colon cancer associated with IL-6-dependent colitis.** *Cancer Res* 2012, **72:**4340-4350.
49. Lang T, Ding X, Kong L, Zhou X, Zhang Z, Ju H, Ding S: **NFATC2 is a novel therapeutic target for colorectal cancer stem cells.** *Onco Targets Ther* 2018, **11:**6911-6924.
50. **HENMT1** [https://www.proteinatlas.org/ENSG00000162639-HENMT1/pathology]
51. Begik O, Lucas MC, Liu H, Ramirez JM, Mattick JS, Novoa EM: **Integrative analyses of the RNA modification machinery reveal tissue- and cancer-specific signatures.** *Genome Biol* 2020, **21:**97.
52. Zhou Y, Huang T, Cheng AS, Yu J, Kang W, To KF: **The TEAD Family and Its Oncogenic Role in Promoting Tumorigenesis.** *Int J Mol Sci* 2016, **17**.
53. Liu Y, Wang G, Yang Y, Mei Z, Liang Z, Cui A, Wu T, Liu CY, Cui L: **Increased TEAD4 expression and nuclear localization in colorectal cancer promote epithelial-mesenchymal transition and metastasis in a YAP-independent manner.** *Oncogene* 2016, **35:**2789-2800.
54. Jang BG, Kim HS, Bae JM, Kim WH, Kim HU, Kang GH: **SMOC2, an intestinal stem cell marker, is an independent prognostic marker associated with better survival in colorectal cancers.** *Sci Rep* 2020, **10:**14591.
55. Purkayastha BP, Roy JK: **Cancer cell metabolism and developmental homeodomain/POU domain transcription factors: a connecting link.** *Cancer Lett* 2015, **356:**315-319.

56. Dunne J, Gascoyne DM, Lister TA, Brady HJ, Heidenreich O, Young BD: **AML1/ETO proteins control POU4F1/BRN3A expression and function in t(8;21) acute myeloid leukemia.** *Cancer Res* 2010, **70:**3985-3995.
57. Diss JK, Faulkes DJ, Walker MM, Patel A, Foster CS, Budhram-Mahadeo V, Djamgoz MB, Latchman DS: **Brn-3a neuronal transcription factor functional expression in human prostate cancer.** *Prostate Cancer Prostatic Dis* 2006, **9:**83-91.
58. Leblond-Francillard M, Picon A, Bertagna X, de Keyzer Y: **High expression of the POU factor Brn3a in aggressive neuroendocrine tumors.** *J Clin Endocrinol Metab* 1997, **82:**89-94.
59. Jin T, Branch DR, Zhang X, Qi S, Youngson B, Goss PE: **Examination of POU homeobox gene expression in human breast cancer cells.** *Int J Cancer* 1999, **81:**104-112.
60. Gao Y, Li Y, Niu X, Wu Y, Guan X, Hong Y, Chen H, Song B: **Identification and Validation of Prognostically Relevant Gene Signature in Melanoma.** *Biomed Res Int* 2020, **2020:**5323614.
61. Nalesnik MA, Tseng G, Ding Y, Xiang GS, Zheng ZL, Yu Y, Marsh JW, Michalopoulos GK, Luo JH: **Gene deletions and amplifications in human hepatocellular carcinomas: correlation with hepatocyte growth regulation.** *Am J Pathol* 2012, **180:**1495-1508.
62. **BAHCC1** [https://www.proteinatlas.org/ENSG00000266074-BAHCC1/pathology]
63. Brzozowa M, Michalski M, Wyrobiec G, Piecuch A, Dittfeld A, Harabin-Slowinska M, Boron D, Wojnicz R: **The role of Snail1 transcription factor in colorectal cancer progression and metastasis.** *Contemp Oncol (Pozn)* 2015, **19:**265-270.
64. Zhou BP, Deng J, Xia W, Xu J, Li YM, Gunduz M, Hung MC: **Dual regulation of Snail by GSK-3beta-mediated phosphorylation in control of epithelial-mesenchymal transition.** *Nat Cell Biol* 2004, **6:**931-940.
65. Huang X, Xiang L, Li Y, Zhao Y, Zhu H, Xiao Y, Liu M, Wu X, Wang Z, Jiang P, et al: **Snail/FOXK1/Cyr61 Signaling Axis Regulates the Epithelial-Mesenchymal Transition and Metastasis in Colorectal Cancer.** *Cell Physiol Biochem* 2018, **47:**590-603.
66. Crist RC, Roth JJ, Waldman SA, Buchberg AM: **A conserved tissue-specific homeodomain-less isoform of MEIS1 is downregulated in colorectal cancer.** *PLoS One* 2011, **6:**e23665.
67. Tufan T, Yang J, Tummala KS, Cingoz H, Kuscu C, Adair SJ, Comertpay G, Nagdas S, Goudreau BJ, Luleyap HU, et al: **ISL2 is an epigenetically silenced tumor suppressor and regulator of metabolism in pancreatic cancer.** *bioRXiv https://doiorg/101101/20200523112839* 2020.
68. Oh S, Shin S, Janknecht R: **ETV1, 4 and 5: an oncogenic subfamily of ETS transcription factors.** *Biochim Biophys Acta* 2012, **1826:**1-12.
69. Horiuchi S, Yamamoto H, Min Y, Adachi Y, Itoh F, Imai K: **Association of ets-related transcriptional factor E1AF expression with tumour progression and overexpression of MMP-1 and matrilysin in human colorectal cancer.** *J Pathol* 2003, **200:**568-576.
70. Boedefeld WM, 2nd, Soong R, Weiss H, Diasio RB, Urist MM, Bland KI, Heslin MJ: **E1A-F is overexpressed early in human colorectal neoplasia and associated with cyclooxygenase-2 and matrix metalloproteinase-7.** *Mol Carcinog* 2005, **43:**13-17.
71. Nosho K, Yoshida M, Yamamoto H, Taniguchi H, Adachi Y, Mikami M, Hinoda Y, Imai K: **Association of Ets-related transcriptional factor E1AF expression with overexpression of matrix metalloproteinases, COX-2 and iNOS in the early stage of colorectal carcinogenesis.** *Carcinogenesis* 2005, **26:**892-899.
72. Moss AC, Lawlor G, Murray D, Tighe D, Madden SF, Mulligan AM, Keane CO, Brady HR, Doran PP, MacMathuna P: **ETV4 and Myeov knockdown impairs colon cancer cell line proliferation and invasion.** *Biochem Biophys Res Commun* 2006, **345:**216-221.
73. Liu HY, Zhou B, Wang L, Li Y, Zhou ZG, Sun XF, Xu B, Zeng YJ, Song JM, Luo HZ, Yang L: **Association of E1AF mRNA expression with tumor progression and matrilysin in human rectal cancer.** *Oncology* 2007, **73:**384-388.

74. Jung Y, Lee S, Choi HS, Kim SN, Lee E, Shin Y, Seo J, Kim B, Jung Y, Kim WK, et al: **Clinical validation of colorectal cancer biomarkers identified from bioinformatics analysis of public expression data.** *Clin Cancer Res* 2011, **17:**700-709.
75. Deves C, Renck D, Garicochea B, da Silva VD, Giulianni Lopes T, Fillman H, Fillman L, Lunardini S, Basso LA, Santos DS, Batista EL, Jr.: **Analysis of select members of the E26 (ETS) transcription factors family in colorectal cancer.** *Virchows Arch* 2011, **458:**421-430.
76. McClure BJ, Heatley SL, Kok CH, Sadras T, An J, Hughes TP, Lock RB, Yeung D, Sutton R, White DL: **Pre-B acute lymphoblastic leukaemia recurrent fusion, EP300-ZNF384, is associated with a distinct gene expression.** *Br J Cancer* 2018, **118:**1000-1004.
77. Yao L, Cen J, Pan J, Liu D, Wang Y, Chen Z, Ruan C, Chen S: **TAF15-ZNF384 fusion gene in childhood mixed phenotype acute leukemia.** *Cancer Genet* 2017, **211:**1-4.
78. He L, Fan X, Li Y, Chen M, Cui B, Chen G, Dai Y, Zhou D, Hu X, Lin H: **Overexpression of zinc finger protein 384 (ZNF 384), a poor prognostic predictor, promotes cell growth by upregulating the expression of Cyclin D1 in Hepatocellular carcinoma.** *Cell Death Dis* 2019, **10:**444.
79. Lo Sasso G, Bovenga F, Murzilli S, Salvatore L, Di Tullio G, Martelli N, D'Orazio A, Rainaldi S, Vacca M, Mangia A, et al: **Liver X receptors inhibit proliferation of human colorectal cancer cells and growth of intestinal tumors in mice.** *Gastroenterology* 2013, **144:**1497-1507, 1507 e1491-1413.
80. Perreault N, Katz JP, Sackett SD, Kaestner KH: **Foxl1 controls the Wnt/beta-catenin pathway by modulating the expression of proteoglycans in the gut.** *J Biol Chem* 2001, **276:**43328-43333.
81. Perreault N, Sackett SD, Katz JP, Furth EE, Kaestner KH: **Foxl1 is a mesenchymal Modifier of Min in carcinogenesis of stomach and colon.** *Genes Dev* 2005, **19:**311-315.
82. Kaestner KH: **The Intestinal Stem Cell Niche: A Central Role for Foxl1-Expressing Subepithelial Telocytes.** *Cell Mol Gastroenterol Hepatol* 2019, **8:**111-117.
83. Liu Y, Liu Y, Huang R, Song W, Wang J, Xiao Z, Dong S, Yang Y, Yang X: **Dependency of the Cancer-Specific Transcriptional Regulation Circuitry on the Promoter DNA Methylome.** *Cell Rep* 2019, **26:**3461-3474 e3465.
84. Shen Z, Chen Y, Li L, Liu L, Peng M, Chen X, Wu X, Sferra TJ, Wu M, Lin X, et al: **Transcription Factor EBF1 Over-Expression Suppresses Tumor Growth in vivo and in vitro via Modulation of the PNO1/p53 Pathway in Colorectal Cancer.** *Front Oncol* 2020, **10:**1035.
85. Bennett DA, Buchman AS, Boyle PA, Barnes LL, Wilson RS, Schneider JA: **Religious Orders Study and Rush Memory and Aging Project.** *J Alzheimers Dis* 2018, **64:**S161-S189.
86. Gasparoni G, Bultmann S, Lutsik P, Kraus TFJ, Sordon S, Vlcek J, Dietinger V, Steinmaurer M, Haider M, Mulholland CB, et al: **DNA methylation analysis on purified neurons and glia dissects age and Alzheimer's disease-specific changes in the human cortex.** *Epigenetics Chromatin* 2018, **11:**41.
87. Lunnon K, Smith R, Hannon E, De Jager PL, Srivastava G, Volta M, Troakes C, Al-Sarraj S, Burrage J, Macdonald R, et al: **Methylomic profiling implicates cortical deregulation of ANK1 in Alzheimer's disease.** *Nat Neurosci* 2014, **17:**1164-1170.
88. Smith RG, Hannon E, De Jager PL, Chibnik L, Lott SJ, Condliffe D, Smith AR, Haroutunian V, Troakes C, Al-Sarraj S, et al: **Elevated DNA methylation across a 48-kb region spanning the HOXA gene cluster is associated with Alzheimer's disease neuropathology.** *Alzheimers Dement* 2018, **14:**1580-1588.
89. Zhang L, Silva TC, Young JI, Gomez L, Schmidt MA, Hamilton-Nelson KL, Kunkle B, Chen X, Martin ER, Wang L: **Epigenome-wide meta-analysis of DNA methylation differences in prefrontal cortex highlights the immune processes in Alzheimer's disease.** *Nature Communications* 2020:In Press.
90. Keenan AB, Torre D, Lachmann A, Leong AK, Wojciechowicz ML, Utti V, Jagodnik KM, Kropiwnicki E, Wang Z, Ma'ayan A: **ChEA3: transcription factor enrichment analysis by orthogonal omics integration.** *Nucleic Acids Res* 2019, **47:**W212-W224.

91. Huang KL, Marcora E, Pimenova AA, Di Narzo AF, Kapoor M, Jin SC, Harari O, Bertelsen S, Fairfax BP, Czajkowski J, et al: **A common haplotype lowers PU.1 expression in myeloid cells and delays onset of Alzheimer's disease.** *Nat Neurosci* 2017, **20:**1052-1061.
92. Rustenhoven J, Smith AM, Smyth LC, Jansson D, Scotter EL, Swanson MEV, Aalderink M, Coppieters N, Narayan P, Handley R, et al: **PU.1 regulates Alzheimer's disease-associated genes in primary human microglia.** *Mol Neurodegener* 2018, **13:**44.
93. Smith AM, Gibbons HM, Oldfield RL, Bergin PM, Mee EW, Faull RL, Dragunow M: **The transcription factor PU.1 is critical for viability and function of human brain microglia.** *Glia* 2013, **61:**929-942.
94. Cunningham C: **Microglia and neurodegeneration: the role of systemic inflammation.** *Glia* 2013, **61:**71-90.
95. Kunkle BW, Grenier-Boley B, Sims R, Bis JC, Damotte V, Naj AC, Boland A, Vronskaya M, van der Lee SJ, Amlie-Wolf A, et al: **Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Abeta, tau, immunity and lipid processing.** *Nat Genet* 2019, **51:**414-430.
96. Salih DA, Bayram S, Guelfi S, Reynolds RH, Shoai M, Ryten M, Brenton JW, Zhang D, Matarin M, Botia JA, et al: **Genetic variability in response to amyloid beta deposition influences Alzheimer's disease risk.** *Brain Commun* 2019, **1:**fcz022.
97. Zhang B, Gaiteri C, Bodea LG, Wang Z, McElwee J, Podtelezhnikov AA, Zhang C, Xie T, Tran L, Dobrin R, et al: **Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease.** *Cell* 2013, **153:**707-720.
98. Forabosco P, Ramasamy A, Trabzuni D, Walker R, Smith C, Bras J, Levine AP, Hardy J, Pocock JM, Guerreiro R, et al: **Insights into TREM2 biology by network analysis of human brain gene expression data.** *Neurobiol Aging* 2013, **34:**2699-2714.
99. Satoh J, Asahina N, Kitano S, Kino Y: **A Comprehensive Profile of ChIP-Seq-Based PU.1/Spi1 Target Genes in Microglia.** *Gene Regul Syst Bio* 2014, **8:**127-139.
100. Landfield PW, Blalock EM, Chen KC, Porter NM: **A new glucocorticoid hypothesis of brain aging: implications for Alzheimer's disease.** *Curr Alzheimer Res* 2007, **4:**205-212.
101. Canet G, Chevallier N, Zussy C, Desrumaux C, Givalois L: **Central Role of Glucocorticoid Receptors in Alzheimer's Disease and Depression.** *Front Neurosci* 2018, **12:**739.
102. Dharshini SAP, Taguchi YH, Gromiha MM: **Investigating the energy crisis in Alzheimer disease using transcriptome study.** *Sci Rep* 2019, **9:**18509.
103. Wetzel DM, Bohn MC, Kazee AM, Hamill RW: **Glucocorticoid receptor mRNA in Alzheimer's diseased hippocampus.** *Brain Res* 1995, **679:**72-81.
104. Jha MK, Jeon S, Suk K: **Pyruvate Dehydrogenase Kinases in the Nervous System: Their Principal Functions in Neuronal-glial Metabolic Interaction and Neuro-metabolic Disorders.** *Curr Neuropharmacol* 2012, **10:**393-403.
105. Vaughn AE, Deshmukh M: **Glucose metabolism inhibits apoptosis in neurons and cancer cells by redox inactivation of cytochrome c.** *Nat Cell Biol* 2008, **10:**1477-1483.
106. Piquet J, Toussay X, Hepp R, Lerchundi R, Le Douce J, Faivre E, Guiot E, Bonvento G, Cauli B: **Supragranular Pyramidal Cells Exhibit Early Metabolic Alterations in the 3xTg-AD Mouse Model of Alzheimer's Disease.** *Front Cell Neurosci* 2018, **12:**216.
107. Mosconi L, Mistur R, Switalski R, Brys M, Glodzik L, Rich K, Pirraglia E, Tsui W, De Santi S, de Leon MJ: **Declining brain glucose metabolism in normal individuals with a maternal history of Alzheimer disease.** *Neurology* 2009, **72:**513-520.
108. Mosconi L, Mistur R, Switalski R, Tsui WH, Glodzik L, Li Y, Pirraglia E, De Santi S, Reisberg B, Wisniewski T, de Leon MJ: **FDG-PET changes in brain glucose metabolism from normal cognition to pathologically verified Alzheimer's disease.** *Eur J Nucl Med Mol Imaging* 2009, **36:**811-822.

109. Reiman EM, Chen K, Alexander GE, Caselli RJ, Bandy D, Osborne D, Saunders AM, Hardy J: **Functional brain abnormalities in young adults at genetic risk for late-onset Alzheimer's dementia.** *Proc Natl Acad Sci U S A* 2004, **101:**284-289.
110. Huang B, Wu P, Bowker-Kinley MM, Harris RA: **Regulation of pyruvate dehydrogenase kinase expression by peroxisome proliferator-activated receptor-alpha ligands, glucocorticoids, and insulin.** *Diabetes* 2002, **51:**276-283.
111. Olsen L, Rasmussen HB, Hansen T, Bagger YZ, Tanko LB, Qin G, Christiansen C, Werge T: **Estrogen receptor alpha and risk for cognitive impairment in postmenopausal women.** *Psychiatr Genet* 2006, **16:**85-88.
112. Yaffe K, Lui LY, Grady D, Stone K, Morin P: **Estrogen receptor 1 polymorphisms and risk of cognitive impairment in older women.** *Biol Psychiatry* 2002, **51:**677-682.
113. Boada M, Antunez C, Lopez-Arrieta J, Caruz A, Moreno-Rey C, Ramirez-Lorca R, Moron FJ, Hernandez I, Mauleon A, Rosende-Roca M, et al: **Estrogen receptor alpha gene variants are associated with Alzheimer's disease.** *Neurobiol Aging* 2012, **33:**198 e115-124.
114. Janicki SC, Schupf N: **Hormonal influences on cognition and risk for Alzheimer's disease.** *Curr Neurol Neurosci Rep* 2010, **10:**359-366.
115. Yaffe K, Lindquist K, Sen S, Cauley J, Ferrell R, Penninx B, Harris T, Li R, Cummings SR: **Estrogen receptor genotype and risk of cognitive impairment in elders: findings from the Health ABC study.** *Neurobiol Aging* 2009, **30:**607-614.
116. Corbo RM, Gambina G, Ruggeri M, Scacchi R: **Association of estrogen receptor alpha (ESR1) PvuII and XbaI polymorphisms with sporadic Alzheimer's disease and their effect on apolipoprotein E concentrations.** *Dement Geriatr Cogn Disord* 2006, **22:**67-72.
117. Ko CY, Wang WL, Wang SM, Chu YY, Chang WC, Wang JM: **Glycogen synthase kinase-3beta-mediated CCAAT/enhancer-binding protein delta phosphorylation in astrocytes promotes migration and activation of microglia/macrophages.** *Neurobiol Aging* 2014, **35:**24-34.
118. Ko CY, Chang WC, Wang JM: **Biological roles of CCAAT/Enhancer-binding protein delta during inflammation.** *J Biomed Sci* 2015, **22:**6.
119. Karch CM, Ezerskiy LA, Bertelsen S, Alzheimer's Disease Genetics C, Goate AM: **Alzheimer's Disease Risk Polymorphisms Regulate Gene Expression in the ZCWPW1 and the CELF1 Loci.** *PLoS One* 2016, **11:**e0148717.
120. Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, DeStafano AL, Bis JC, Beecham GW, Grenier-Boley B, et al: **Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease.** *Nat Genet* 2013, **45:**1452-1458.
121. Readhead B, Haure-Mirande JV, Funk CC, Richards MA, Shannon P, Haroutunian V, Sano M, Liang WS, Beckmann ND, Price ND, et al: **Multiscale Analysis of Independent Alzheimer's Cohorts Finds Disruption of Molecular, Genetic, and Clinical Networks by Human Herpesvirus.** *Neuron* 2018, **99:**64-82 e67.
122. Bell RD, Deane R, Chow N, Long X, Sagare A, Singh I, Streb JW, Guo H, Rubio A, Van Nostrand W, et al: **SRF and myocardin regulate LRP-mediated amyloid-beta clearance in brain vascular cells.** *Nat Cell Biol* 2009, **11:**143-153.
123. Dotti CG, De Strooper B: **Alzheimer's dementia by circulation disorders: when trees hide the forest.** *Nat Cell Biol* 2009, **11:**114-116.
124. Perdomo-Sabogal A, Nowick K, Piccini I, Sudbrak R, Lehrach H, Yaspo ML, Warnatz HJ, Querfurth R: **Human Lineage-Specific Transcriptional Regulation through GA-Binding Protein Transcription Factor Alpha (GABPa).** *Mol Biol Evol* 2016, **33:**1231-1244.
125. Bahn G, Park JS, Yun UJ, Lee YJ, Choi Y, Park JS, Baek SH, Choi BY, Cho YS, Kim HK, et al: **NRF2/ARE pathway negatively regulates BACE1 expression and ameliorates cognitive deficits in mouse Alzheimer's models.** *Proc Natl Acad Sci U S A* 2019, **116:**12516-12523.

126. Ren P, Chen J, Li B, Zhang M, Yang B, Guo X, Chen Z, Cheng H, Wang P, Wang S, et al: **Nrf2 Ablation Promotes Alzheimer's Disease-Like Pathology in APP/PS1 Transgenic Mice: The Role of Neuroinflammation and Oxidative Stress.** *Oxid Med Cell Longev* 2020, **2020:**3050971.
127. Pajares M, Jimenez-Moreno N, Garcia-Yague AJ, Escoll M, de Ceballos ML, Van Leuven F, Rabano A, Yamamoto M, Rojo AI, Cuadrado A: **Transcription factor NFE2L2/NRF2 is a regulator of macroautophagy genes.** *Autophagy* 2016, **12:**1902-1916.
128. Welch RP, Lee C, Imbriano PM, Patil S, Weymouth TE, Smith RA, Scott LJ, Sartor MA: **ChIP-Enrich: gene set enrichment testing for ChIP-seq data.** *Nucleic Acids Res* 2014, **42:**e105.
129. Mukherjee S, Erickson H, Bastia D: **Enhancer-origin interaction in plasmid R6K involves a DNA loop mediated by initiator protein.** *Cell* 1988, **52:**375-383.
130. Sur I, Taipale J: **The role of enhancers in cancer.** *Nat Rev Cancer* 2016, **16:**483-493.
131. Schubeler D: **Function and information content of DNA methylation.** *Nature* 2015, **517:**321-326.
132. Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Scholer A, van Nimwegen E, Wirbelauer C, Oakeley EJ, Gaidatzis D, et al: **DNA-binding factors shape the mouse methylome at distal regulatory regions.** *Nature* 2011, **480:**490-495.
133. Song L, Zhang Z, Grasfeder LL, Boyle AP, Giresi PG, Lee BK, Sheffield NC, Graf S, Huss M, Keefe D, et al: **Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity.** *Genome Res* 2011, **21:**1757-1767.
134. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, et al: **The accessible chromatin landscape of the human genome.** *Nature* 2012, **489:**75-82.
135. Blattler A, Farnham PJ: **Cross-talk between site-specific transcription factors and DNA methylation states.** *J Biol Chem* 2013, **288:**34287-34294.
136. Yao L, Shen H, Laird PW, Farnham PJ, Berman BP: **Inferring regulatory element landscapes and transcription factor networks from cancer methylomes.** *Genome Biol* 2015, **16:**105.
137. Shakya A, Callister C, Goren A, Yosef N, Garg N, Khoddami V, Nix D, Regev A, Tantin D: **Pluripotency transcription factor Oct4 mediates stepwise nucleosome demethylation and depletion.** *Mol Cell Biol* 2015, **35:**1014-1025.
138. Consortium EP, Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, Adrian J, Kawli T, Davis CA, Dobin A, et al: **Expanded encyclopaedias of DNA elements in the human and mouse genomes.** *Nature* 2020, **583:**699-710.
139. Moore JE, Pratt HE, Purcaro MJ, Weng Z: **A curated benchmark of enhancer-gene interactions for evaluating enhancer-target gene prediction methods.** *Genome Biol* 2020, **21:**17.
140. Whalen S, Truty RM, Pollard KS: **Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin.** *Nat Genet* 2016, **48:**488-496.
141. Gao T, He B, Liu S, Zhu H, Tan K, Qian J: **EnhancerAtlas: a resource for enhancer annotation and analysis in 105 human cell/tissue types.** *Bioinformatics* 2016, **32:**3543-3551.
142. Schmidl C, Klug M, Boeld TJ, Andreesen R, Hoffmann P, Edinger M, Rehli M: **Lineage-specific DNA methylation in T cells correlates with histone methylation and enhancer activity.** *Genome Res* 2009, **19:**1165-1174.
143. Reddington JP, Perricone SM, Nestor CE, Reichmann J, Youngson NA, Suzuki M, Reinhardt D, Dunican DS, Prendergast JG, Mjoseng H, et al: **Redistribution of H3K27me3 upon DNA hypomethylation results in de-repression of Polycomb target genes.** *Genome Biol* 2013, **14:**R25.
144. Brinkman AB, Gu H, Bartels SJ, Zhang Y, Matarese F, Simmer F, Marks H, Bock C, Gnirke A, Meissner A, Stunnenberg HG: **Sequential ChIP-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk.** *Genome Res* 2012, **22:**1128-1138.

145. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, Sabedot TS, Malta TM, Pagnotta SM, Castiglioni I, et al: **TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data.** *Nucleic Acids Res* 2016, **44:**e71.
146. Yu C, Yao W: **Robust linear regression: A review and comparison.** *Communications in Statistics - Simulation and Computation* 2017, **46:** 6261-6282.
147. Wang L, Jia P, Wolfinger RD, Chen X, Grayson BL, Aune TM, Zhao Z: **An efficient hierarchical generalized linear mixed model for pathway analysis of genome-wide association studies.** *Bioinformatics* 2011, **27:**686-692.
148. Zhou W, Triche TJ, Jr., Laird PW, Shen H: **SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions.** *Nucleic Acids Res* 2018, **46:**e123.
149. Wang T, Guan W, Lin J, Boutaoui N, Canino G, Luo J, Celedon JC, Chen W: **A systematic study of normalization methods for Infinium 450K methylation data using whole-genome bisulfite sequencing data.** *Epigenetics* 2015, **10:**662-669.
150. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, Beck S: **A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data.** *Bioinformatics* 2013, **29:**189-196.
151. Pidsley R, CC YW, Volta M, Lunnon K, Mill J, Schalkwyk LC: **A data-driven approach to preprocessing Illumina 450K methylation array data.** *BMC Genomics* 2013, **14:**293.
152. Gao T, Qian J: **EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species.** *Nucleic Acids Res* 2020, **48:**D58-D64.
153. Tripathi MK, Deane NG, Zhu J, An H, Mima S, Wang X, Padmanabhan S, Shi Z, Prodduturi N, Ciombor KK, et al: **Nuclear factor of activated T-cell activity is associated with metastatic capacity in colon cancer.** *Cancer Res* 2014, **74:**6947-6957.
154. Kroepil F, Fluegen G, Vallbohmer D, Baldus SE, Dizdar L, Raffel AM, Hafner D, Stoecklein NH, Knoefel WT: **Snail1 expression in colorectal cancer and its correlation with clinical and pathological parameters.** *BMC Cancer* 2013, **13:**145.
155. Betge J, Schneider NI, Harbaum L, Pollheimer MJ, Lindtner RA, Kornprat P, Ebert MP, Langner C: **MUC1, MUC2, MUC5AC, and MUC6 in colorectal cancer: expression profiles and clinical significance.** *Virchows Arch* 2016, **469:**255-265.
156. Shvab A, Haase G, Ben-Shmuel A, Gavert N, Brabletz T, Dedhar S, Ben-Ze'ev A: **Induction of the intestinal stem cell signature gene SMOC-2 is required for L1-mediated colon cancer progression.** *Oncogene* 2016, **35:**549-557.
157. Traicoff JL, De Marchis L, Ginsburg BL, Zamora RE, Khattar NH, Blanch VJ, Plummer S, Bargo SA, Templeton DJ, Casey G, Kaetzel CS: **Characterization of the human polymeric immunoglobulin receptor (PIGR) 3'UTR and differential expression of PIGR mRNA during colon tumorigenesis.** *J Biomed Sci* 2003, **10:**792-804.
158. Wang L, Ai M, Nie M, Zhao L, Deng G, Hu S, Han Y, Zeng W, Wang Y, Yang M, Wang S: **EHF promotes colorectal carcinoma progression by activating TGF-beta1 transcription and canonical TGF-beta signaling.** *Cancer Sci* 2020, **111:**2310-2324.
159. Gimeno-Valiente F, Riffo-Campos AL, Vallet-Sanchez A, Siscar-Lewin S, Gambardella V, Tarazona N, Cervantes A, Franco L, Castillo J, Lopez-Rodas G: **ZNF518B gene up-regulation promotes dissemination of tumour cells and is governed by epigenetic mechanisms in colorectal cancer.** *Sci Rep* 2019, **9:**9339.
160. Kim SH, Park YY, Cho SN, Margalit O, Wang D, DuBois RN: **Kruppel-Like Factor 12 Promotes Colorectal Cancer Growth through Early Growth Response Protein 1.** *PLoS One* 2016, **11:**e0159899.

161. Zhang X, Xu J, Zhang H, Sun J, Li N, Huang X: **MicroRNA-758 acts as a tumor inhibitor in colorectal cancer through targeting PAX6 and regulating PI3K/AKT pathway.** *Oncol Lett* 2020, **19:**3923-3930.
162. Janecki DM, Sajek M, Smialek MJ, Kotecki M, Ginter-Matuszewska B, Kuczynska B, Spik A, Kolanowski T, Kitazawa R, Kurpisz M, Jaruzelska J: **SPIN1 is a proto-oncogene and SPIN3 is a tumor suppressor in human seminoma.** *Oncotarget* 2018, **9:**32466-32477.
163. Cera I, Whitton L, Donohoe G, Morris DW, Dechant G, Apostolova G: **Genes encoding SATB2-interacting proteins in adult cerebral cortex contribute to human cognitive ability.** *PLoS Genet* 2019, **15:**e1007890.
164. Jaitner C, Reddy C, Abentung A, Whittle N, Rieder D, Delekate A, Korte M, Jain G, Fischer A, Sananbenesi F, et al: **Satb2 determines miRNA expression and long-term memory in the adult central nervous system.** *Elife* 2016, **5**.
165. Velasco-Estevez M, Mampay M, Boutin H, Chaney A, Warn P, Sharp A, Burgess E, Moeendarbary E, Dev KK, Sheridan GK: **Infection Augments Expression of Mechanosensing Piezo1 Channels in Amyloid Plaque-Reactive Astrocytes.** *Front Aging Neurosci* 2018, **10:**332.
166. Fehlbaum-Beurdeley P, Jarrige-Le Prado AC, Pallares D, Carriere J, Guihal C, Soucaille C, Rouet F, Drouin D, Sol O, Jordan H, et al: **Toward an Alzheimer's disease diagnosis via high-resolution blood gene expression.** *Alzheimers Dement* 2010, **6:**25-38.
167. Wirz KT, Bossers K, Stargardt A, Kamphuis W, Swaab DF, Hol EM, Verhaagen J: **Cortical beta amyloid protein triggers an immune response, but no synaptic changes in the APPswe/PS1dE9 Alzheimer's disease mouse model.** *Neurobiol Aging* 2013, **34:**1328-1342.
168. Pelucchi S, Stringhi R, Marcello E: **Dendritic Spines in Alzheimer's Disease: How the Actin Cytoskeleton Contributes to Synaptic Failure.** *Int J Mol Sci* 2020, **21**.
169. Arendt T, Bruckner MK: **Linking cell-cycle dysfunction in Alzheimer's disease to a failure of synaptic plasticity.** *Biochim Biophys Acta* 2007, **1772:**413-421.
170. Xu Z, Wu C, Pan W, Alzheimer's Disease Neuroimaging I: **Imaging-wide association study: Integrating imaging endophenotypes in GWAS.** *Neuroimage* 2017, **159:**159-169.
171. Bahn G, Jo DG: **Therapeutic Approaches to Alzheimer's Disease Through Modulation of NRF2.** *Neuromolecular Med* 2019, **21:**1-11.
172. Wang R, Reddy PH: **Role of Glutamate and NMDA Receptors in Alzheimer's Disease.** *J Alzheimers Dis* 2017, **57:**1041-1048.

Table 1 Stage-wise analysis results of TCGA COAD-READ datasets. The robust linear model target gene expression ~ TF activity + DNAm + DNAm x TF was used to analyze CpG-TF-target gene triplets. Shown are significant triplets at 5% overall FDR level, and the unique CpGs, TFs, and targets in these triplets.

analysis	unique triplets	unique CpGs	unique TFs	unique targets
Promoter analysis				
screening	165	49	56	42
confirmation	31	29	18	25
Distal analysis				
screening	2358	315	155	570
confirmation	52	43	44	44
Regulon-based analysis				
screening	78	23	19	19
confirmation	47	18	11	6

Table 2 Top 10 most significant triplets identified by unsupervised MethReg analysis of TCGA COAD-READ samples. CRC = colorectal cancer.

Triplet data				Annotations			DNAm x TF effect			TF-target		P-values for tumor vs. normal comparison			CRC literature	
ProbeID	position	TF	target gene	CpG-target distance	TF.role	DNAm.effect	Estimate	P-value	Adj. P-value	Correlation	P-value	CpG	TF	target gene	TF	target gene
promoter analysis																
cg00328227	chr1: 108661703	NFATC2	HENMT1	-176	Repressor	Enhancing	-0.663	0	0	-0.033	5.27E-01	3.19E-08	2.53E-01	8.23E-08	[47] [153] [49]	[51] [50]
cg25083481	chr11: 1034989	SNAI1	MUC6	1727	Activator	Attenuating	-0.413	1.77E-04	0	0.117	2.52E-02	8.98E-03	4.99E-10	1.97E-02	[154] [63, 65]	[155]
cg02816729	chr6: 168442419	TEAD3	SMOC2	1267	Repressor	Attenuating	0.323	7.77E-05	0	0.07	1.83E-01	2.96E-02	3.47E-02	7.32E-01	[52]	[54] [156]
cg12751565	chr1: 206946642	NFATC2	PIGR	-175	Activator	Enhancing	0.282	1.09E-09	0	0.315	8.67E-10	6.49E-03	2.53E-01	1.88E-11	[47] [153] [49]	[157]
cg05503887	chr11: 34620378	MEIS1	EHF	-713	Activator	Attenuating	-0.268	7.24E-05	0	0.347	1.10E-11	6.17E-01	1.84E-08	1.16E-02	[66]	[158]
cg13542964	chr4: 10457043	ETV4	ZNF518B	381	Activator	Enhancing	0.217	2.08E-04	0	0.023	6.59E-01	1.31E-06	1.32E-14	1.77E-02	[68, 72]	[159]
distal analysis																
cg14043104	chr13: 74075149	ISL2	KLF12	-80092	Activator	Attenuating	-0.222	1.03E-04	0	0.232	7.60E-06	3.89E-02	3.96E-02	1.20E-03	[67]	[160]
cg09217215	chr11: 31808034	ZNF384	PAX6	9925	Repressor	Enhancing	-0.398	6.90E-07	1.48E-04	-0.028	5.94E-01	7.64E-03	2.15E-01	1.68E-02	[76] [77]	[161]
cg11871337	chrX: 56991382	FOXL1	SPIN3	4443	Activator	Attenuating	-0.294	2.18E-07	2.68E-04	0.183	4.32E-04	7.26E-01	9.95E-01	8.21E-07	[80] [81]	[162]
cg04665204	chr17: 81453398	POU3F4	BAHCC1	57922	Dual	Invert	-0.37	6.80E-07	2.91E-04	0.083	1.12E-01	2.69E-05	1.47E-06	2.38E-02	[55] [59]	[62] [61]

Table 3 Top 10 most significant triplets identified by supervised MethReg analysis of ROSMAP Alzheimer's disease (AD) dataset.

Triplet data				Annotation			DNAm x TF effect			TF-target		AD literature	
ProbeID	position	TF	target gene	CpG-target distance	TF.role	DNAm.effect	Estimate	P-value	Adj. P-value	Correlation	P-value	TF	target gene
promoter analysis													
cg17418085	chr1: 31229122	SPI1	LAPTM5	1,543	Dual	Invert	-7.177	7.36E-12	3.06E-07	-0.011	8.05E-01	[92, 93]	[96]
distance-based analysis													
cg11556846	chr2: 200468728	ESR1	SATB2	-132,738	Dual	Invert	-7.266	3.03E-07	7.80E-03	-0.058	1.82E-01	[111, 113]	[163] [164]
cg00153919	chr16: 88859944	CEBPD	PIEZO1	-8,324	Repressor	Enhancing	-8.122	4.92E-07	9.14E-03	-0.044	3.10E-01	[117, 118]	[165]
cg08824847	chr11: 35052388	NR3C1	PDHX	115,011	Activator	Enhancing	4.383	6.13E-07	1.02E-02	0.06	1.66E-01	[101] [103]	[104, 105]
cg08760493	chr4: 109994039	TCF12	SEC24B	-360,887	Dual	Invert	8.120	1.26E-06	1.49E-02	0.033	4.44E-01	[119, 120]	[166]
cg05715492	chr7: 98991138	SRF	ARPC1B	19,265	Repressor	Attenuating	11.581	1.34E-06	1.49E-02	0.017	7.00E-01	[122]	[167]
cg09316954	chr16: 67687754	TCF12	CARMIL2	8,931	Repressor	Attenuating	8.376	2.28E-06	2.24E-02	-0.07	1.08E-01	[119, 120]	[168]
cg21155834	chr2: 149282209	GABPA	ORC4	-503,061	Activator	Enhancing	6.407	2.75E-06	2.41E-02	0.045	2.98E-01	[124]	[169]
cg13819552	chr9: 95799870	TCF12	ZNF484	-159,565	Activator	Enhancing	7.989	2.92E-06	2.41E-02	0.039	3.71E-01	[119, 120]	[170]
cg21535772	chr2: 171679906	NFE2L2	AC007405.4	52,282	Dual	Invert	7.065	3.10E-06	2.41E-02	-0.026	5.47E-01	[125] [171]	[172]

Figure 1 Analysis workflow of MethReg.

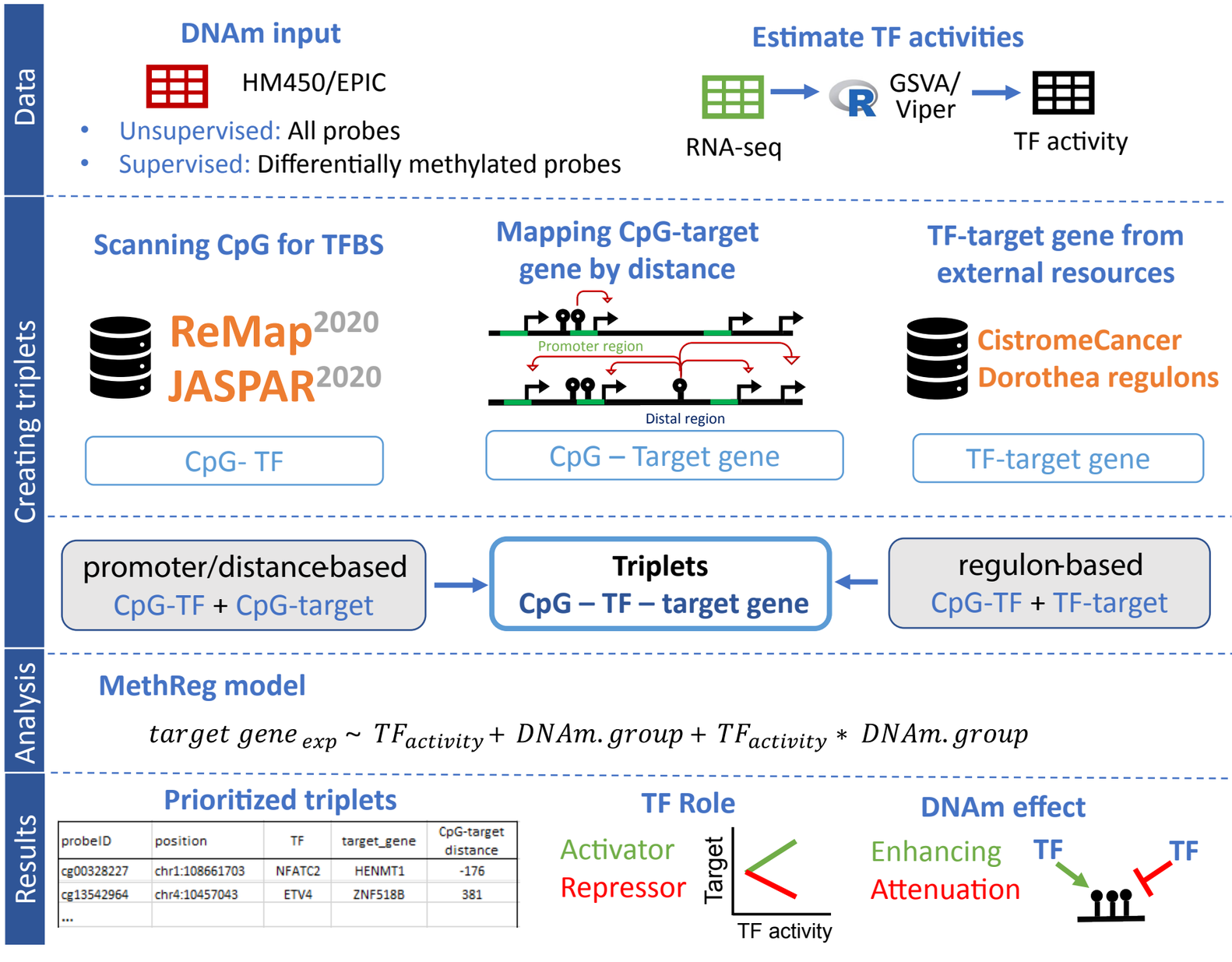
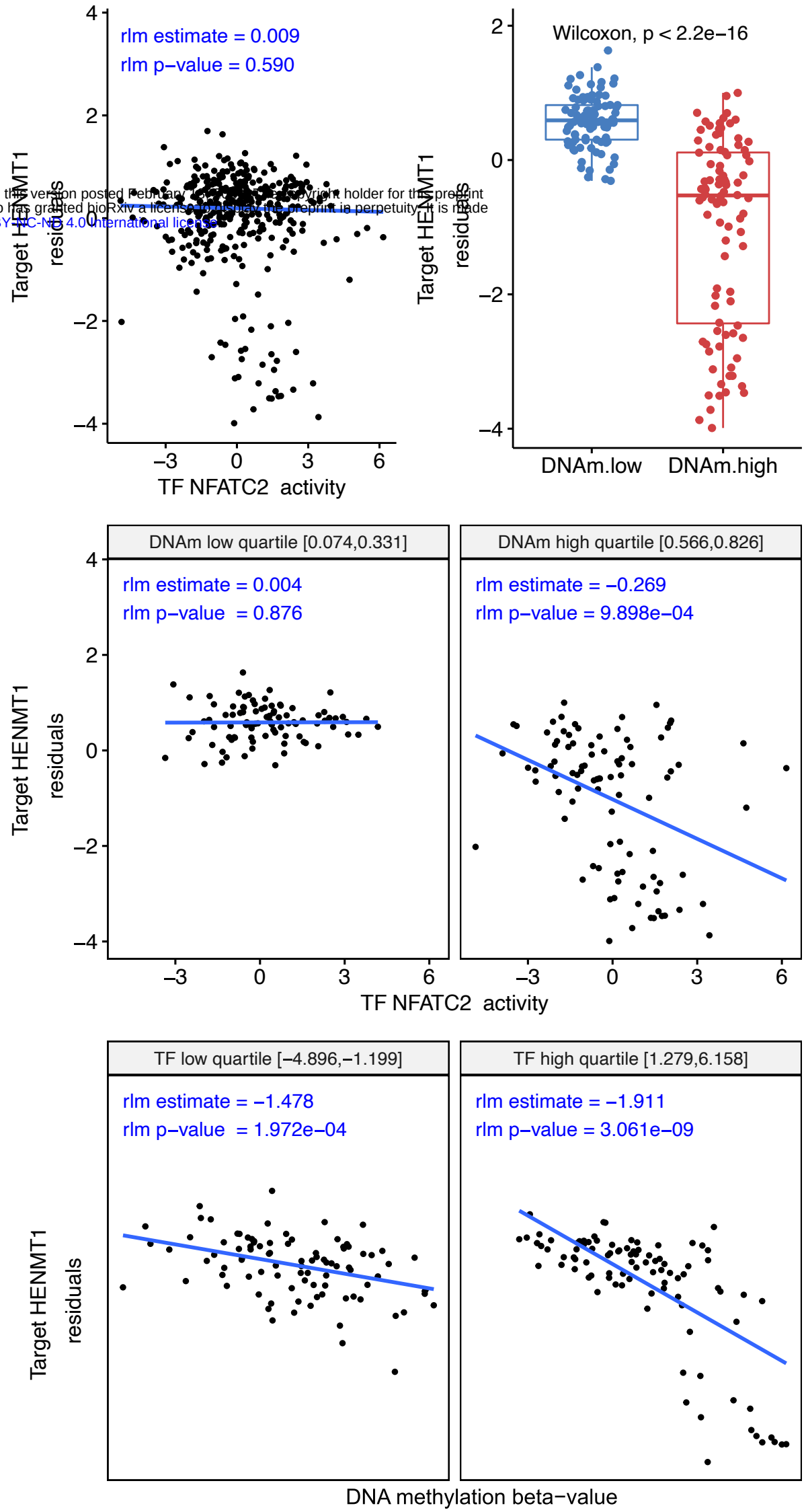


Figure 2. An example output from MethReg unsupervised analysis of TCGA COAD-READ samples, in which CpG methylation enhances TF activity.

Genome of reference	hg38
Region ID	chr1:108661703–108661704
Probe ID	cg00328227
Target gene ID	ENSG00000162639
Target gene Symbol	HENMT1
TF gene ID	ENSG00000101096
TF gene Symbol	NFATC2
TF role	Repressor
DNAm effect	Enhancing

Target ~ TF + DNAm Quant. Group + TF * DNAm Quant. Group	Estimate	P–Values
Direct effect of DNAm	–1.95	< 2e–16
Direct effect of TF	0.00523	0.91851
Synergistic effect of DNAm and TF	–0.663	< 2e–16

Legend
rlm: robust linear model



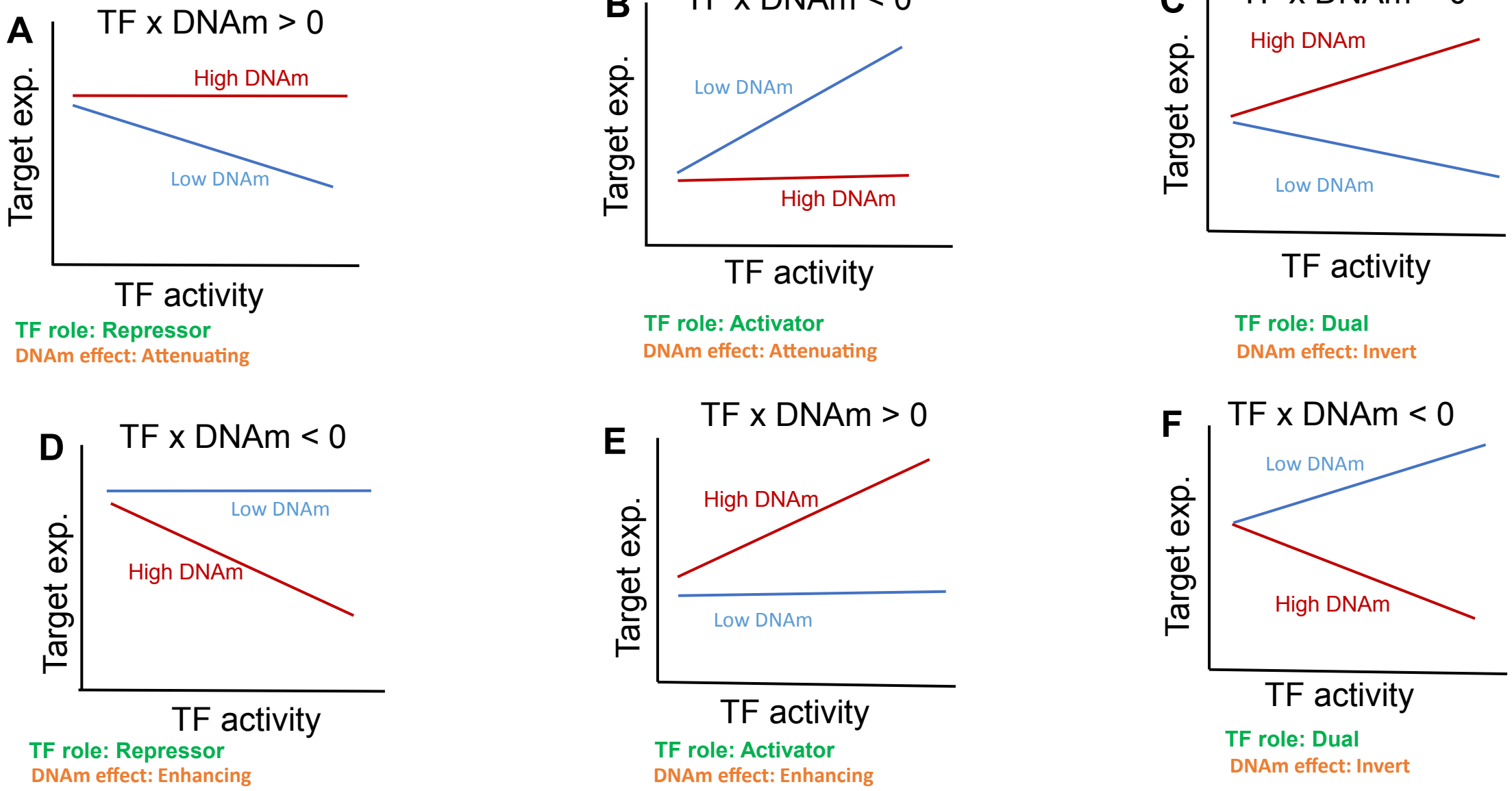
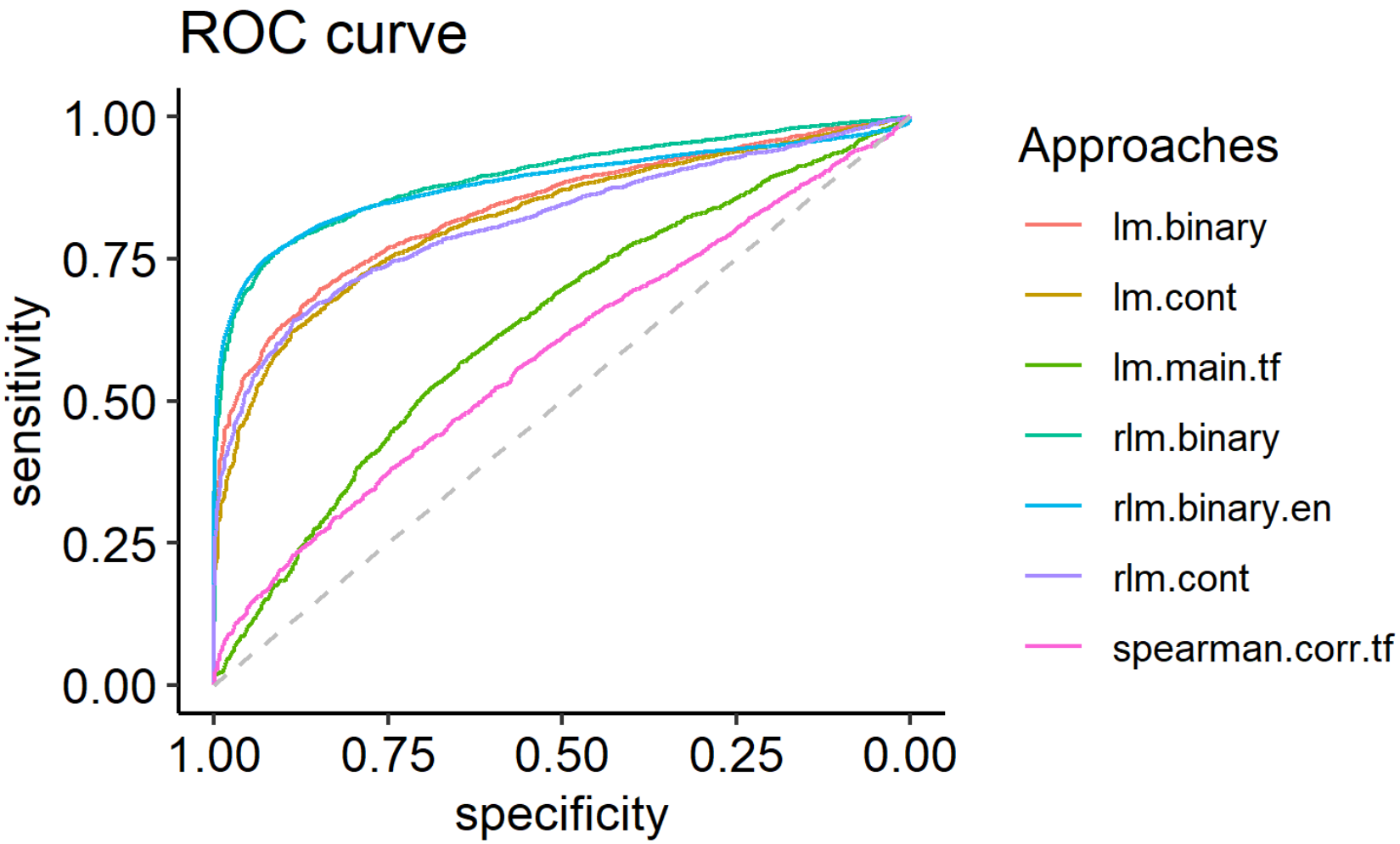


Figure 3 Scenarios modeled by MethReg.

Figure 4 Receiver Operating Characteristics curves for different methods compared in simulation study. lm.binary = linear model implementation of Model 2; lm.cont = linear model implementation of Model 1; lm.main.tf = linear model gene expression ~ TF; rlm.binary = robust linear model implementation of Model 2; rlm.binary.en = same as rlm.binary model, except P-values were estimated using empirical null distribution; rlm.cont = robust linear model implementation of Model 1; spearman.corr.tf = Spearman correlation between gene expression and TF expression; Model 1: gene expression ~ TF + DNAm + TF*DNAm; Model 2: gene expression ~ TF + DNAm.group + TF*DNAm.group



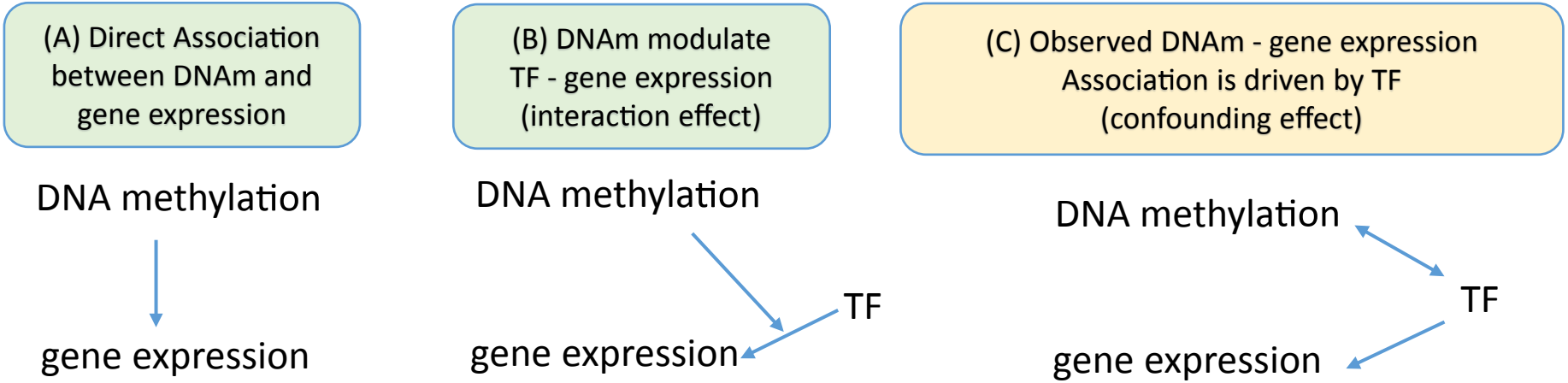
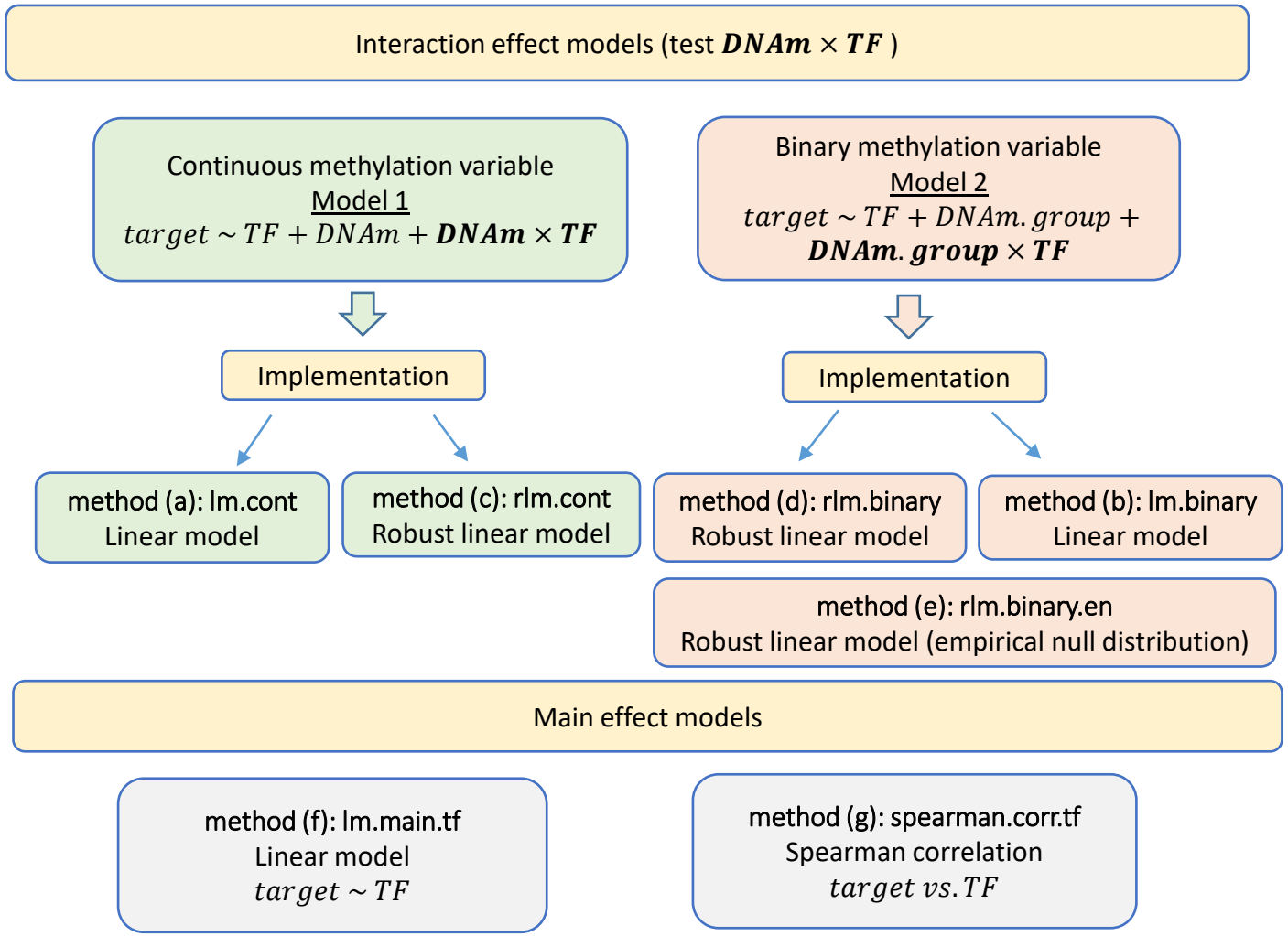
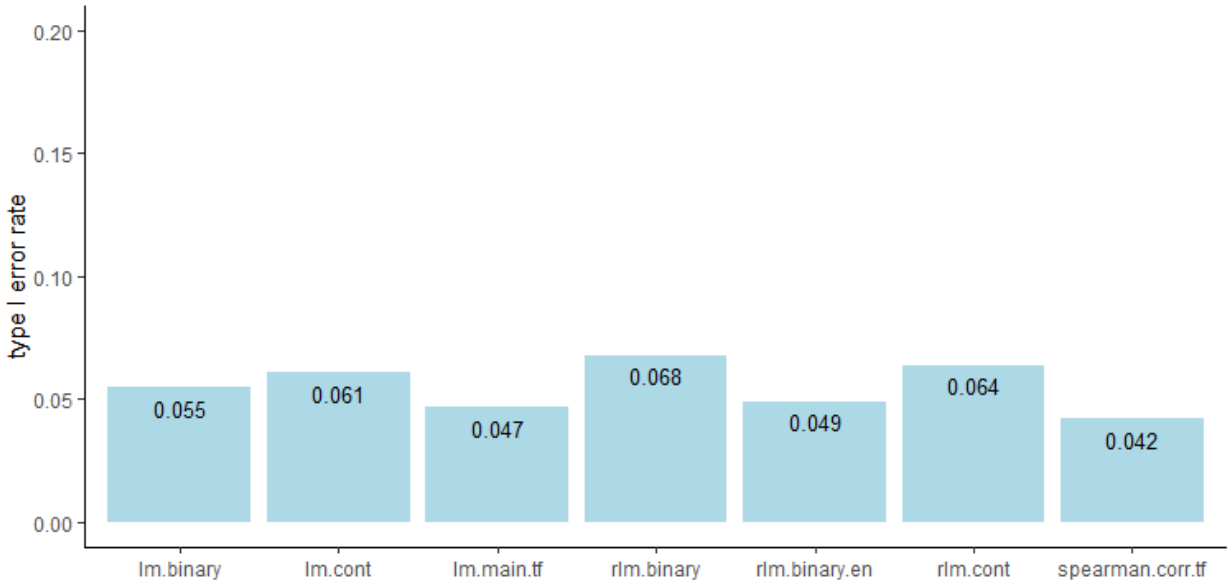


Figure 5 Direct, indirect, and confounding effects in gene regulation.

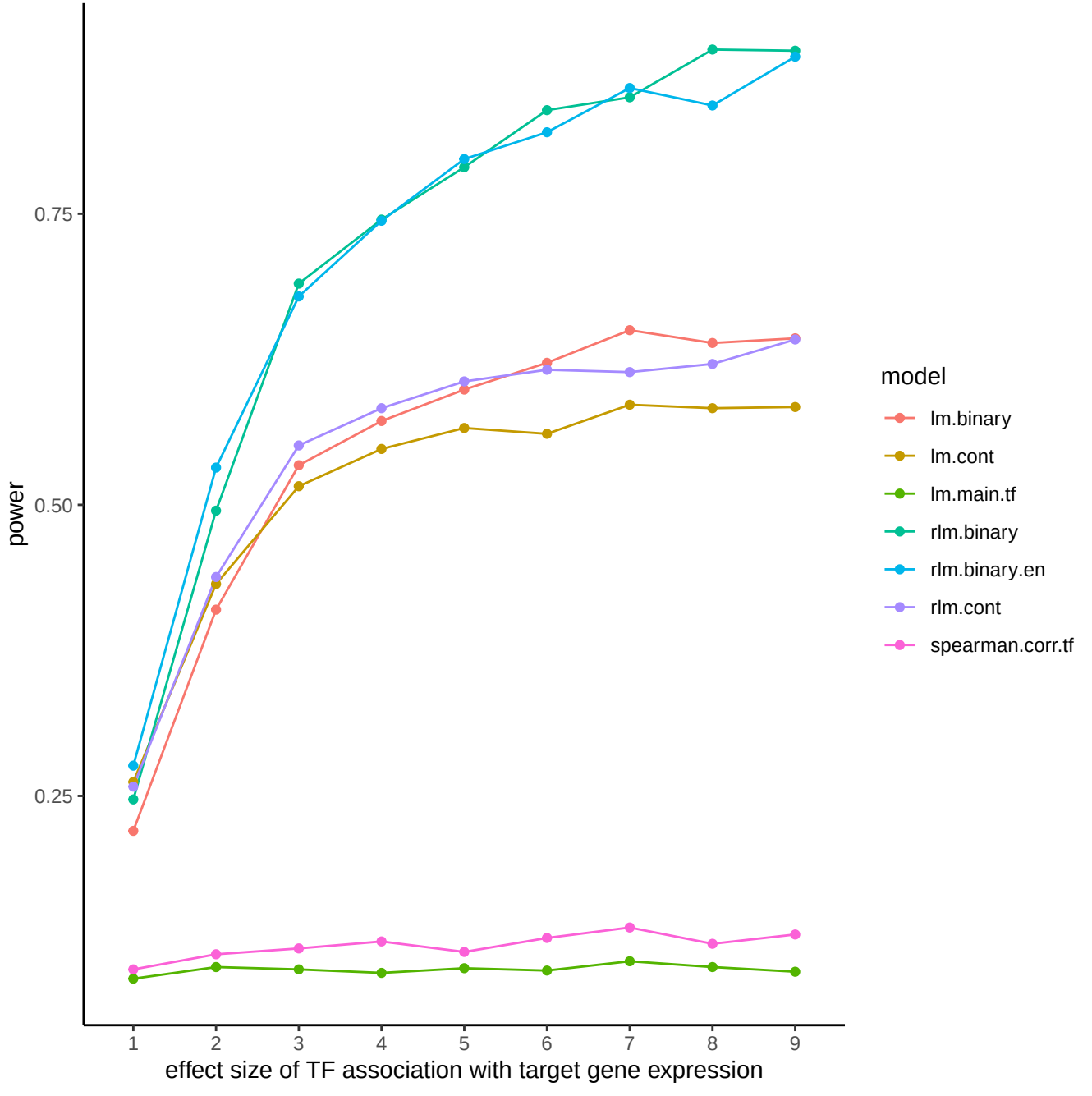


Supplementary Figure 1 Methods compared in simulation study.

Supplementary Figure 2 Type I error rates of different methods in simulated datasets. lm.binary = linear model implementation of Model 2; lm.cont = linear model implementation of Model 1; lm.main.tf = linear model *gene expression ~ TF*; rlm.binary = robust linear model implementation of Model 2; rlm.binary.en = same as rlm.binary model, except P-values were estimated using empirical null distribution; rlm.cont = robust linear model implementation of Model 1; spearman.corr.tf = Spearman correlation between gene expression and TF expression; Model 1: *gene expression ~ TF + DNAm + TF*DNAm*; Model 2: *gene expression ~ TF + DNAm.group + TF*DNAm.group*

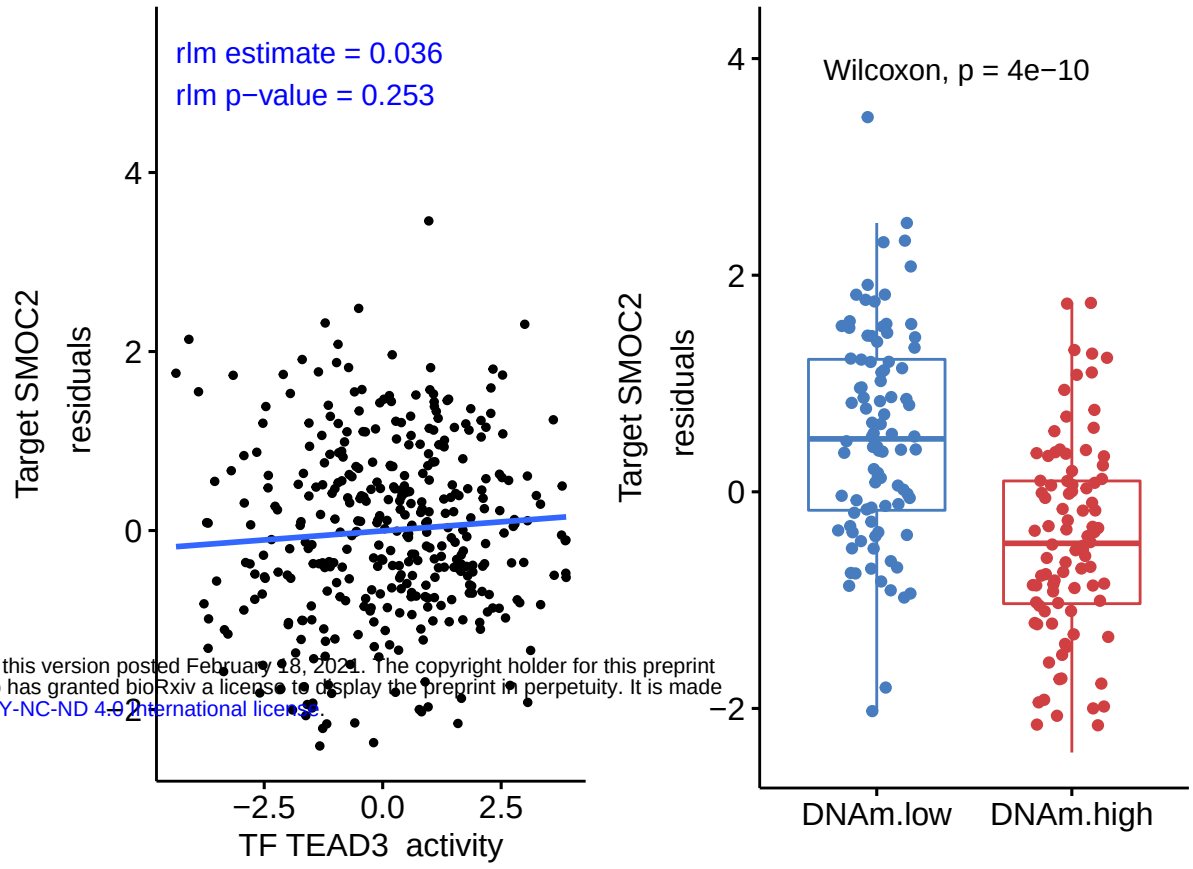


Supplementary Figure 3 Power of different methods in simulation study. lm.binary = linear model implementation of Model 2; lm.cont = linear model implementation of Model 1; lm.main.tf = linear model *gene expression ~ TF*; rlm.binary = robust linear model implementation of Model 2; rlm.binary.en = same as rlm.binary model, except P-values were estimated using empirical null distribution; rlm.cont = robust linear model implementation of Model 1; spearman.corr.tf = Spearman correlation between gene expression and TF expression; Model 1: *gene expression ~ TF + DNAm + TF*DNAm;* Model 2: *gene expression ~ TF + DNAm.group + TF*DNAm.group*

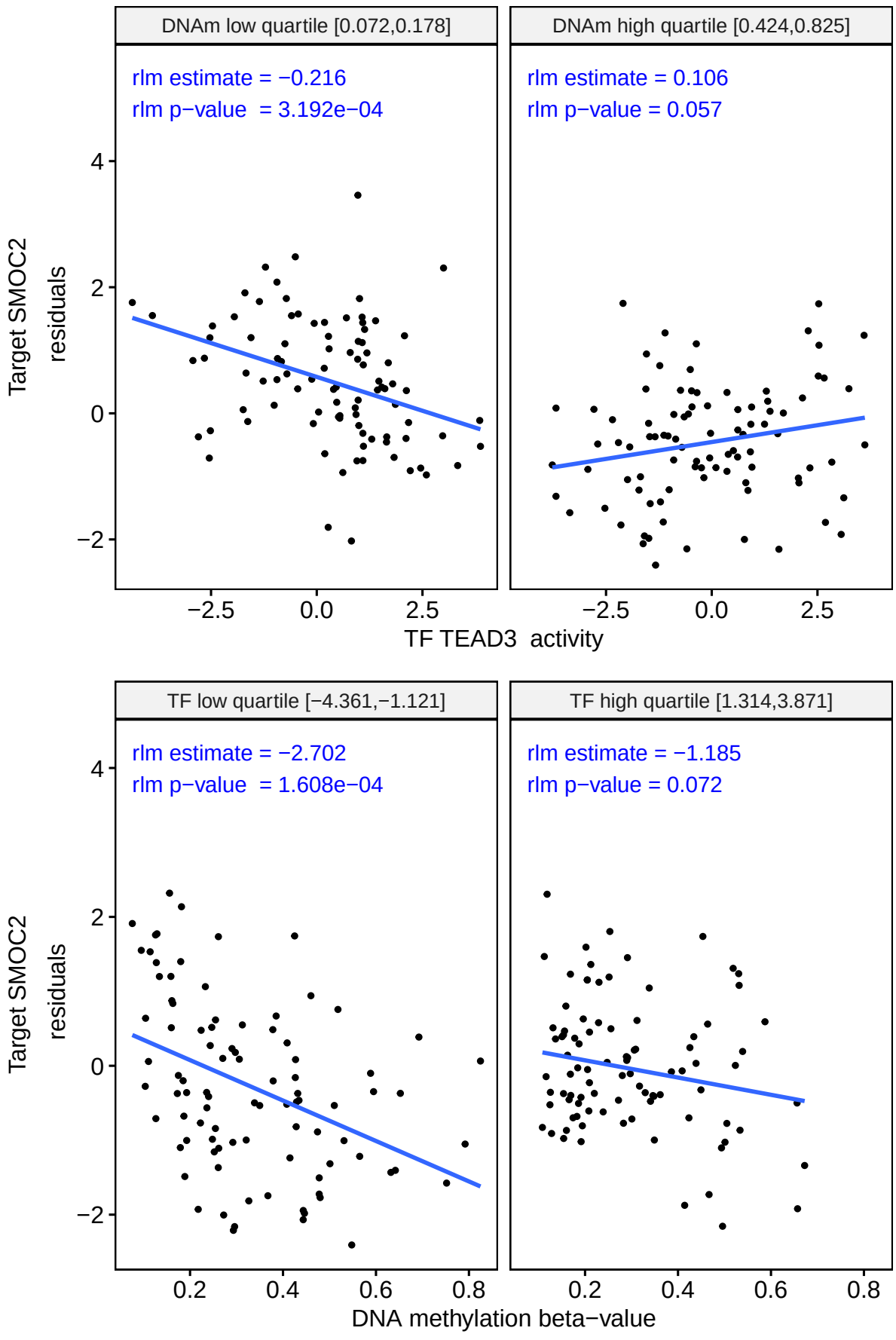


Supplementary Figure 4 An example of CpG methylation attenuating TF activities, from unsupervised MethReg analysis of TCGA COAD-READ samples.

Genome of reference	hg38
Region ID	chr6:168442419−168442420
Probe ID	cg02816729
Target gene ID	ENSG00000112562
Target gene Symbol	SMOC2
TF gene ID	ENSG00000007866
TF gene Symbol	TEAD3
TF role	Repressor
DNAm effect	Attenuating



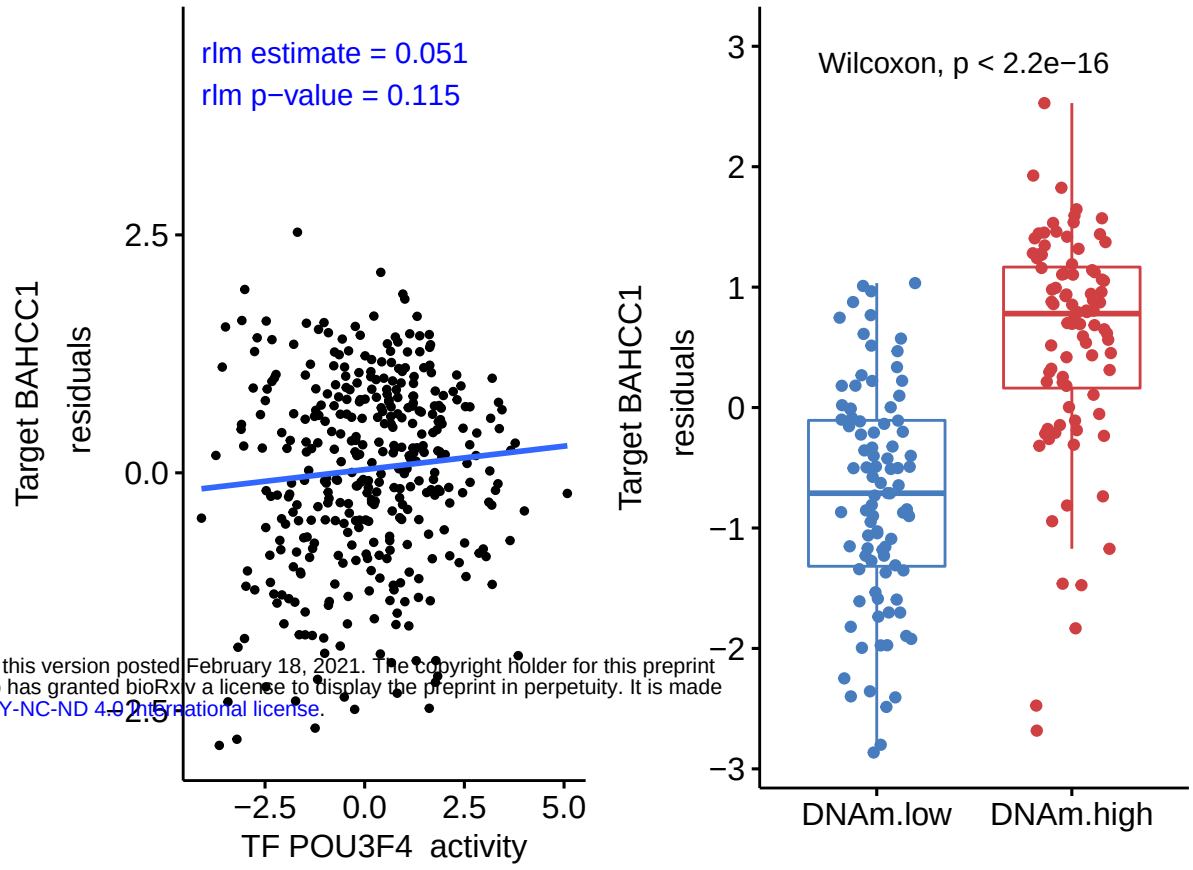
Target ~ TF + DNAm Quant. Group + TF * DNAm Quant. Group	Estimate	P−Values
Direct effect of DNAm	−1.02	5.28e−12
Direct effect of TF	−0.217	0.000287
Synergistic effect of DNAm and TF	0.323	7.77e−05



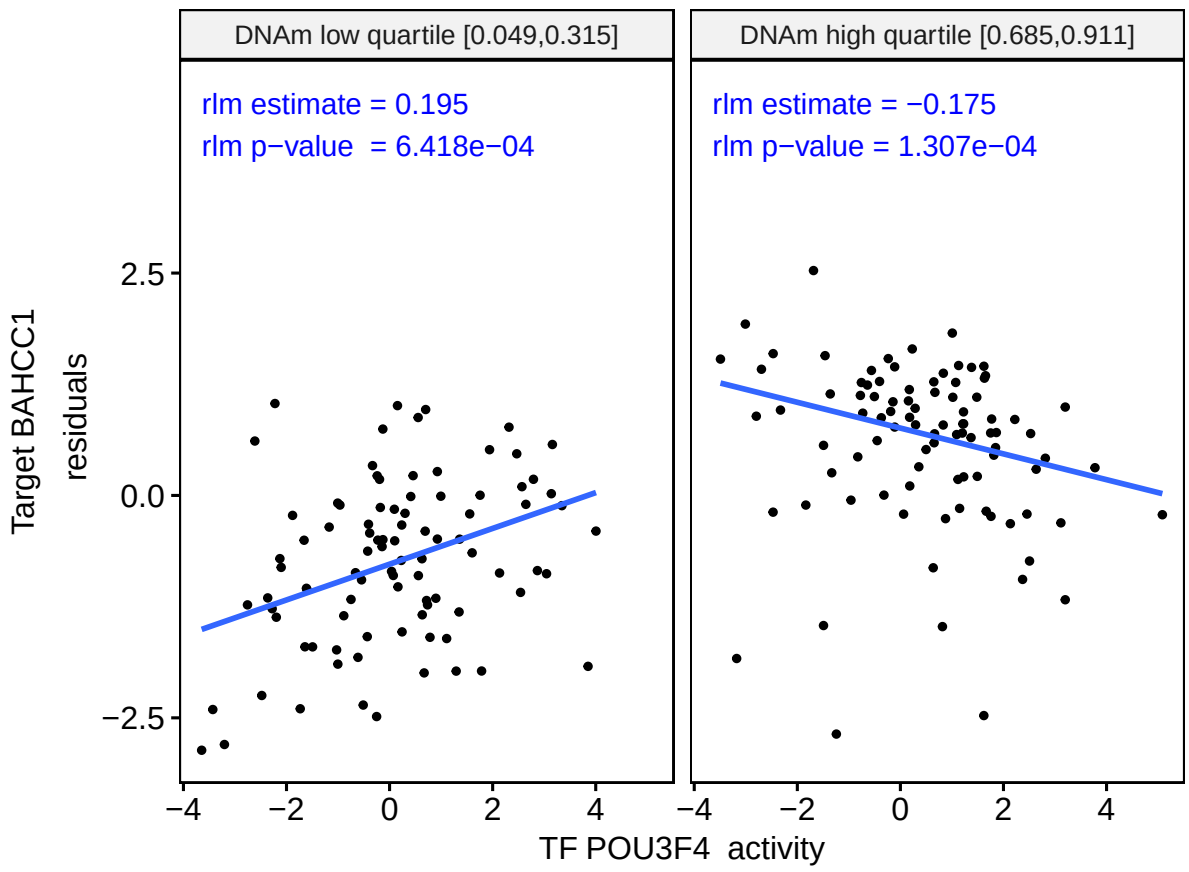
Legend
rlm: robust linear model

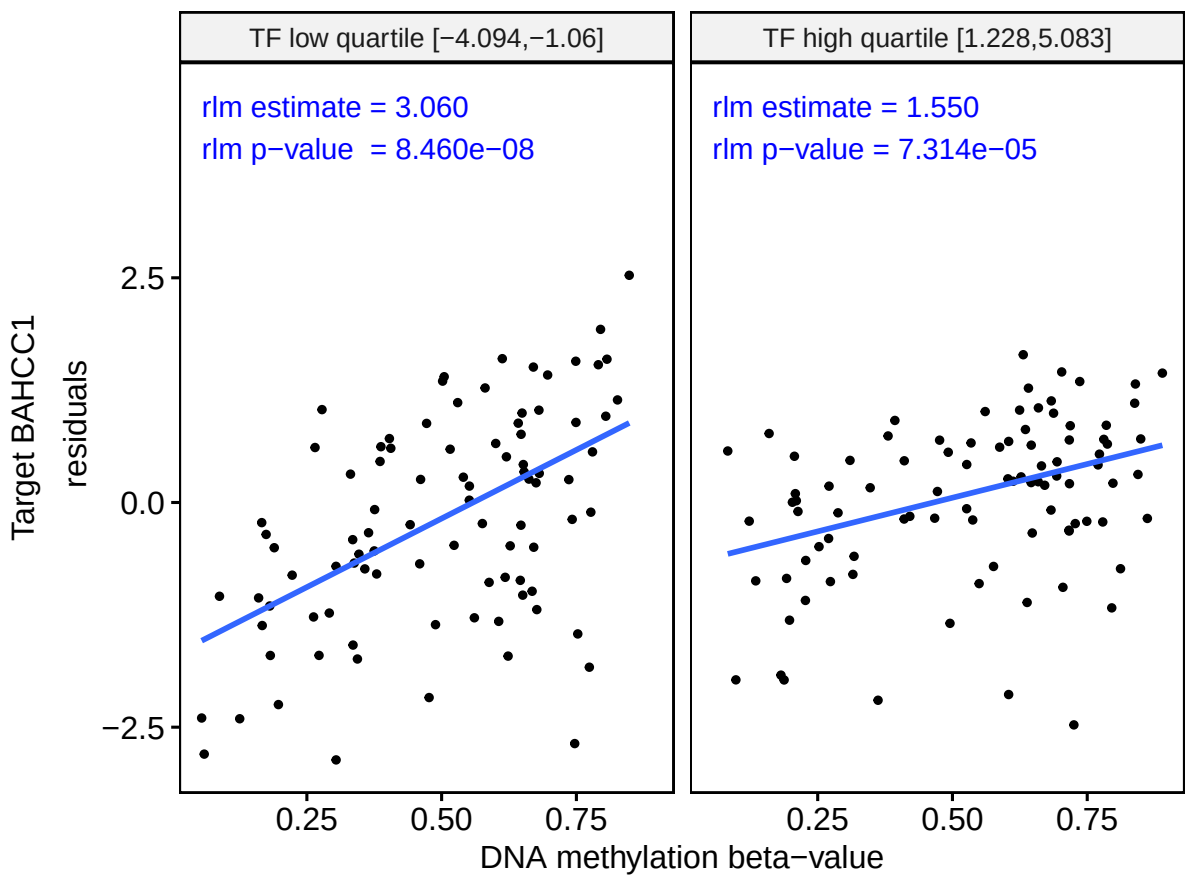
Supplementary Figure 5 An example of TF with dual modes of regulations depending on CpG methylation levels, from unsupervised MethReg analysis of TCGA COAD-READ datasets.

Genome of reference	hg38
Region ID	chr17:81453398−81453399
Probe ID	cg04665204
Target gene ID	ENSG00000266074
Target gene Symbol	BAHCC1
TF gene ID	ENSG00000196767
TF gene Symbol	POU3F4
TF role	Dual
DNAm effect	Invert



Target ~ TF + DNAm Quant. Group + TF * DNAm Quant. Group	Estimate	P−Values
Direct effect of DNAm	1.56	< 2e−16
Direct effect of TF	0.201	8.27e−05
Synergistic effect of DNAm and TF	−0.37	6.80e−07

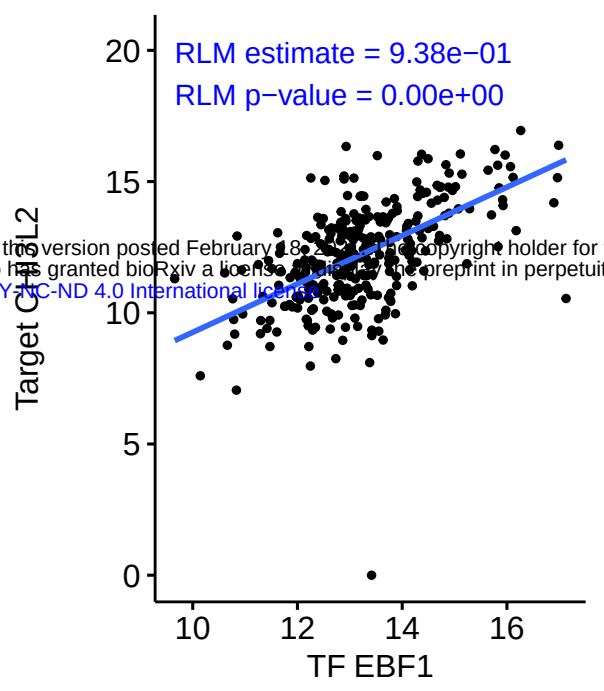


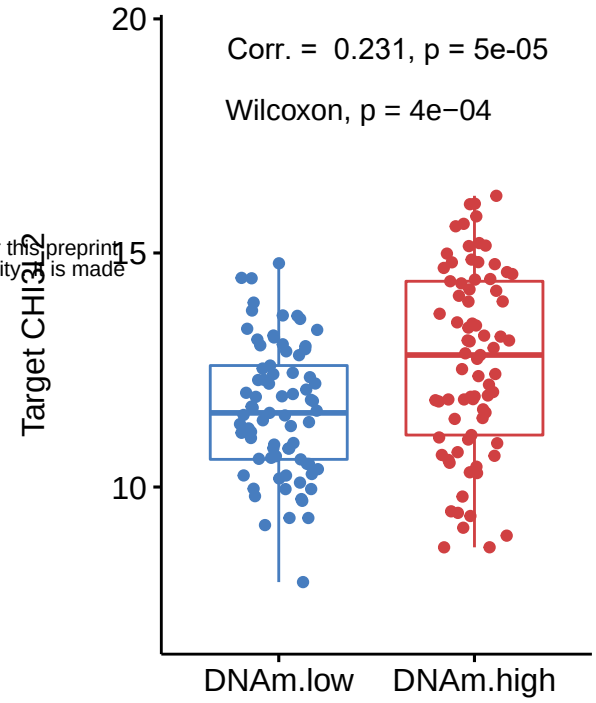


Legend
rlm: robust linear model

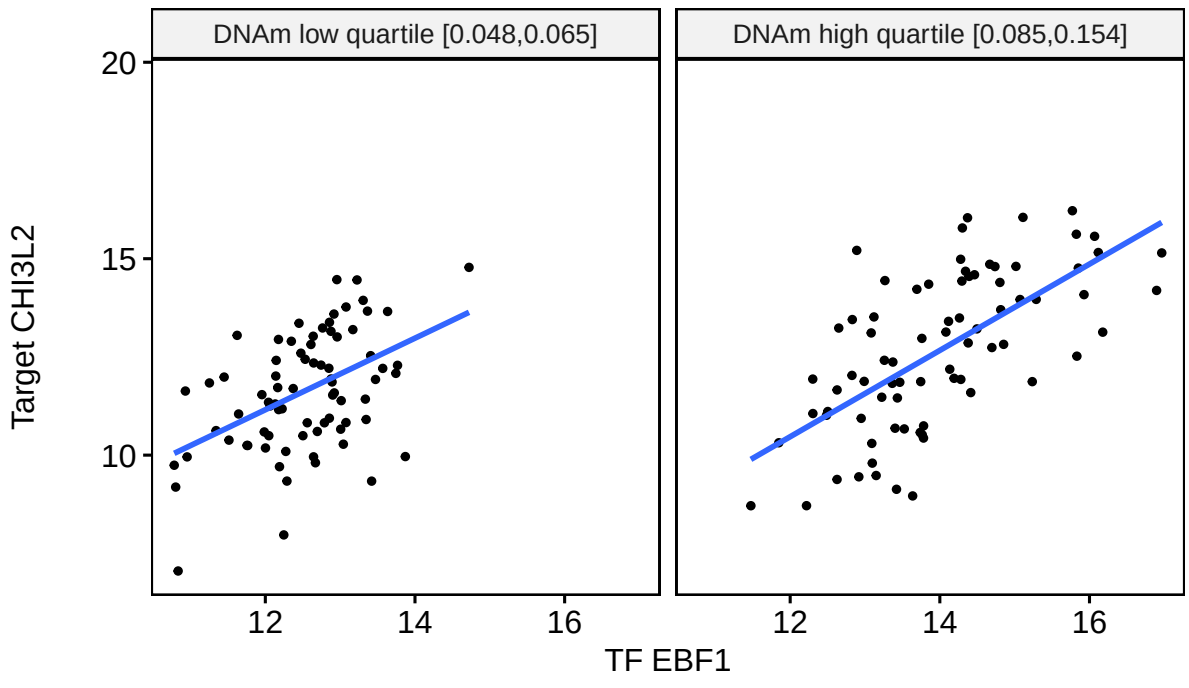
Supplementary Figure 6 Example of confounding TF effect. The target gene expression is mainly driven by the TF EBF1, and not by DNA methylation, even though a highly significant methylation-target gene association was observed.

Region ID	chr1:111202536−111206535
Target gene ID	ENSG00000064886
Target gene Symbol	CHI3L2
TF gene ID	ENSG00000164330
TF gene Symbol	EBF1
Diff. DNAm (Q4 − Q1)	0.02032327
TF role	Activator
DNAm effect	Enhancing



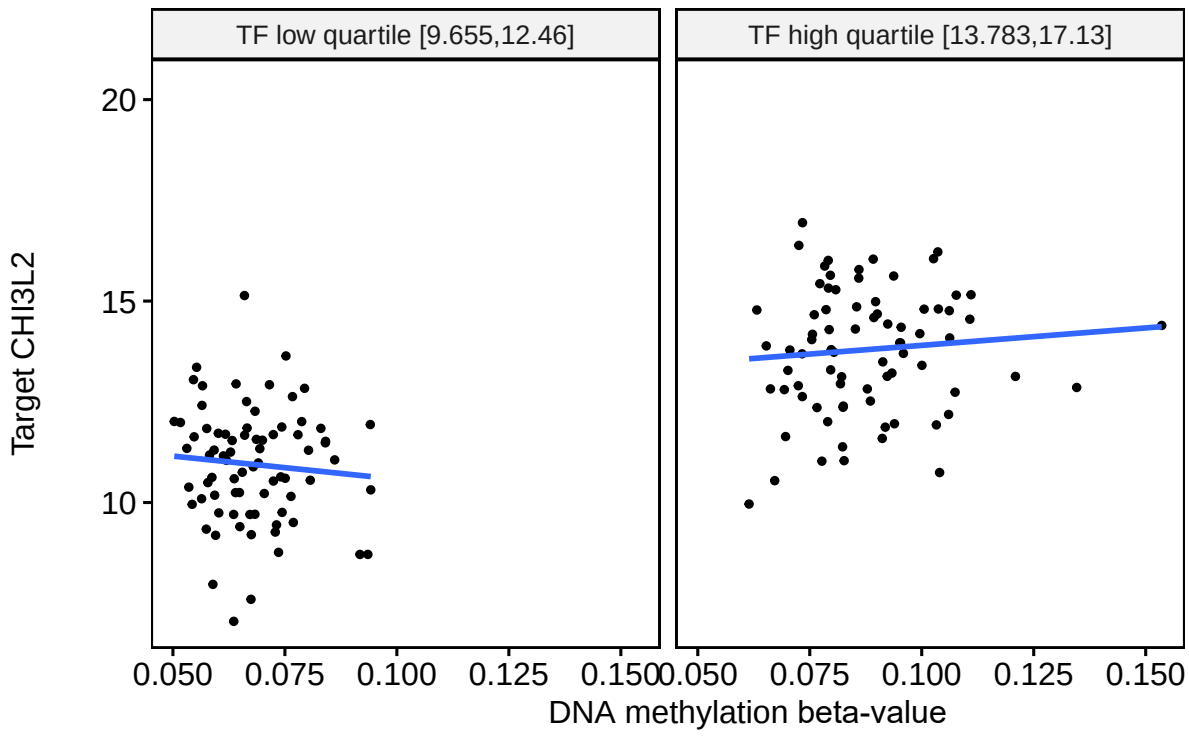


Target ~ TF + DNAm Quant. Group + TF * DNAm Quant. Group	Estimate	P−Values
Direct effect of DNAm	−2.94	0.395081
Direct effect of TF	0.914	5.75e−05
Synergistic effect of DNAm and TF	0.188	0.477157



Target ~ TF DNAm low samples	Estimate	P−Values
Direct effect of TF	0.914	2.23e−05

Target ~ TF DNAm high samples	Estimate	P−Values
Direct effect of TF	1.1	6.91e−10

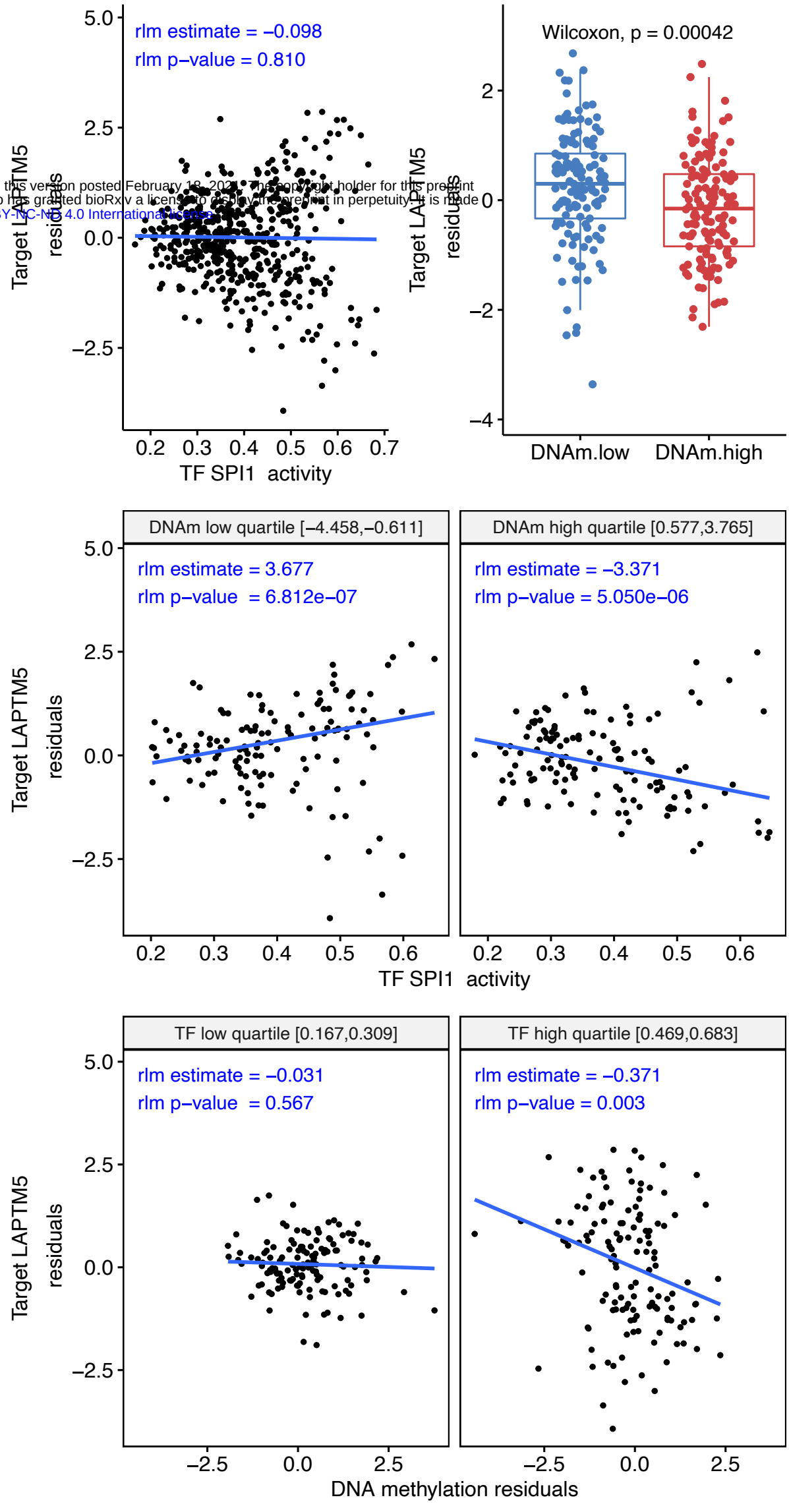


Supplementary Figure 7 An example of TF having dual regulatory roles depending on CpG methylation levels, from supervised MethReg analysis ROSMAP dataset.

Genome of reference	hg19
Region ID	chr1:31229122–31229123
Probe ID	cg17418085
Target gene ID	ENSG00000162511
Target gene Symbol	LAPTM5
TF gene ID	ENSG00000066336
TF gene Symbol	SPI1
TF role	Dual
DNAm effect	Invert

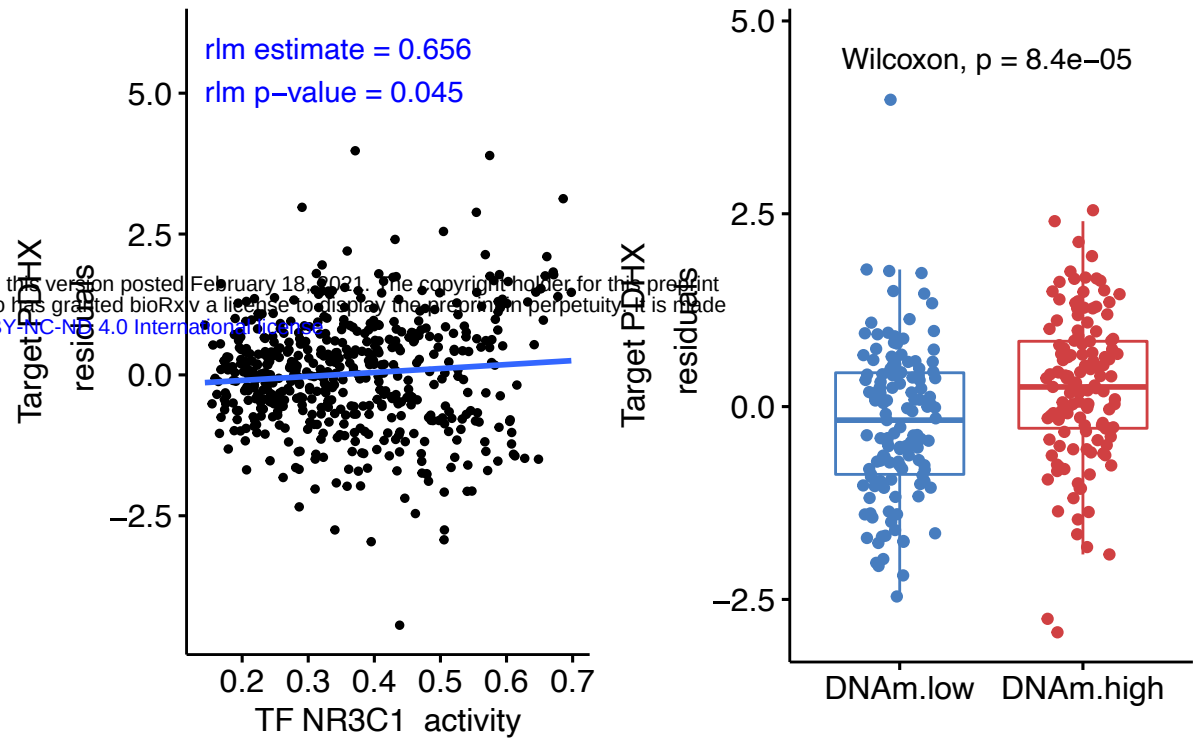
Target ~ TF + DNAm Quant. Group + TF * DNAm Quant. Group	Estimate	P–Values
Direct effect of DNAm	2.12	2.91e–07
Direct effect of TF	3.6	1.03e–06
Synergistic effect of DNAm and TF	–7.18	7.36e–12

Legend
rlm: robust linear model

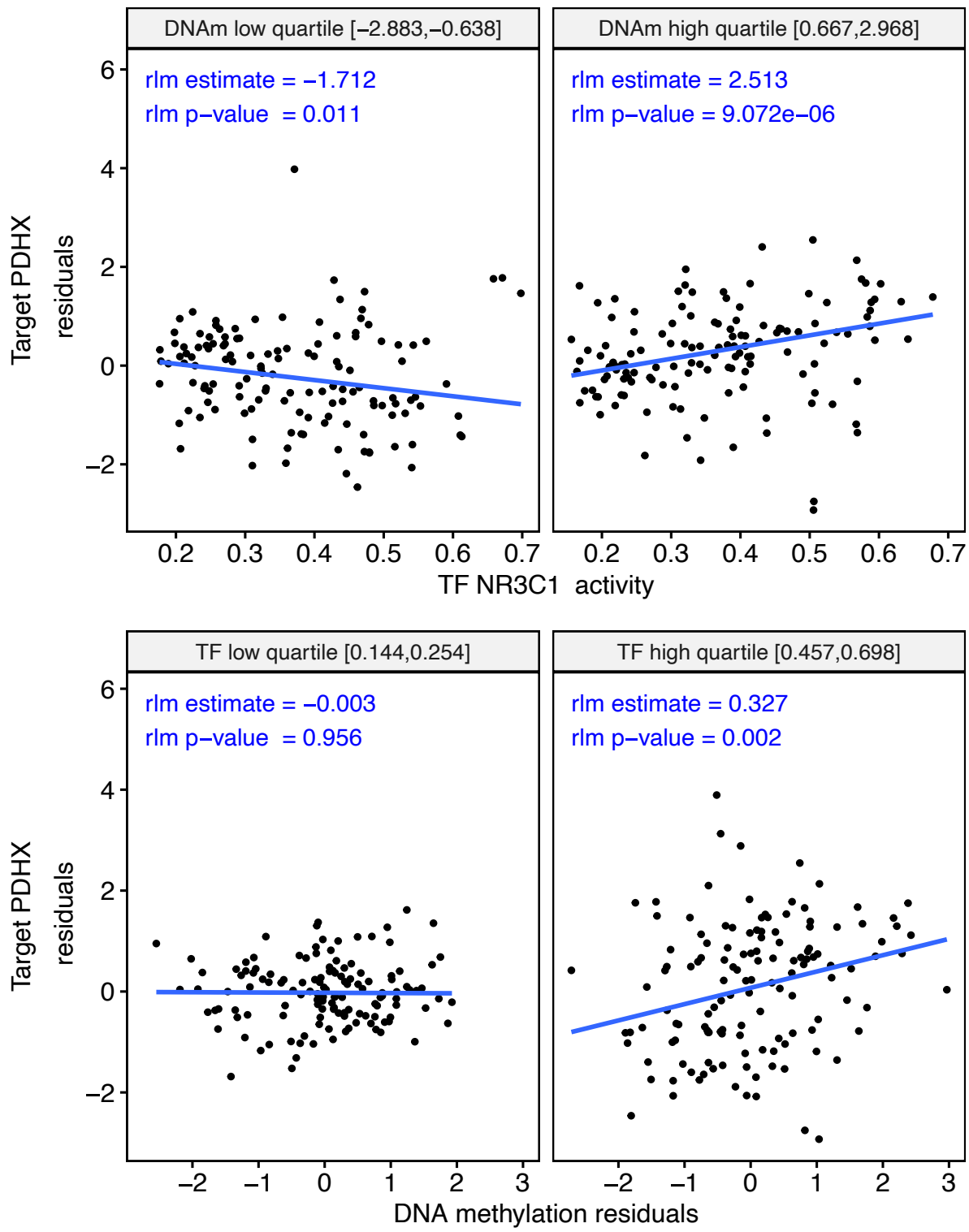


Supplementary Figure 8 An example of CpG methylation enhancing TF activities, from supervised MethReg analysis of ROSMAP dataset.

Genome of reference	hg19
Region ID	chr11:35052388–35052389
Probe ID	cg08824847
Target gene ID	ENSG00000110435
Target gene Symbol	PDHX
TF gene ID	ENSG00000113580
TF gene Symbol	NR3C1
TF role	Activator
DNAm effect	Enhancing



Target ~ TF + DNAm Quant. Group + TF * DNAm Quant. Group	Estimate	P–Values
Direct effect of DNAm	–1.03	0.002425
Direct effect of TF	–2.02	0.001407
Synergistic effect of DNAm and TF	4.38	6.13e–07



Legend
rlm: robust linear model