Studying the regulatory effect of DNA methylation on gene expression in human dorsolateral prefrontal cortex (DLPFC)

Tiam Heydari, [1] Sina Jafarzadeh, [2] Lisa Leung, [3] Zohreh Sharafian, [4] Hiwot Tafessu [3]

1. Department of Biophysics; 2. Department of Bioinformatics; 3. Department of Statistics 4. Department of Experimental Medicine

Introduction

DNA methylation is an epigenetic mechanism which plays a role in regulating tissue specific gene expression[1]. DNA methylation mostly occurs when a cytosine base is next to a guanine base, forming a CpG site. Early studies showed that methylation of CpG sites prevents the expression of genes. However, recent studies revealed that methylation can be linked with both decreasing and increasing of gene expressions[2]. In this study, we characterize the relationship between DNA methylation and gene expression in a quantitative manner, i.e. evaluating the statistical significance of the effect of methylated CpG sites on gene expressions colloquiality known as **expression Quantitative Trait Methylation(eQTM)** in human dorsolateral prefrontal cortex (DLPFC) [3].

Objective

1) We analyze the correlation of each gene and the probes located in the vicinity of that gene.
2) we extend our model to include the collective effect of multiple CpG sites on the expression of genes.
3) we quantitatively assess the roles of chromatin states[7] and CpG site distance from gene on the regulatory relationship between methylation probes and a gene expression on brain tissue. This provides us more biological intuition behind the regulatory processes in human brain.

Fig 1. Graphical overview of our data analysis pipeline. We first correct the data for confounders and outliers. Next, we apply eQTM analysis to assess significant correlations between the pairs of methylation probe and gene. We then extend the analysis to multivariate linear and non-linear regression models. We verify the biological plausibility of results using other data types including probe-gene distances and chromatin states.

Dataset

We use a combination of two publicly available datasets including the Religious Orders Study (ROS)[4] and The Memory and Aging Project (MAP)[5]. Both datasets are longitudinal cohort studies containing DNA methylation and gene expression values. The gene expression data is RNA-sequencing data extracted from DLPFC from 540 subjects using Illumina HiSeq with 101-bp paired-end reads. The DNA methylation data was generated using 450K Illumina array from 740 subjects. We use the samples of 468 individuals have both gene expression and methylation data. The data is quality-controlled and corrected for common batch-effect and confounders[4,5].

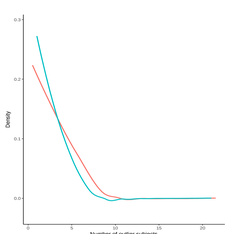


Fig2. The distribution of outlier frequency in gene expression and methylation data. the expected value of this empirical distribution are 1.31 and 2.91 for genes and methylation sites respectively. It shows that the data is corrected for outliers by the data provider.

Methods

Data Preprocessing

- Convert and organize Matlab and CSV data files into R format using **R base libraries and Matrix**
- Assess and correct for multicollinearity using **A-clust method** (figure 3)
- Analyze and remove outliers on gene expression and methylation data. We consider values outside the 3*standard deviation window around the mean as an outlier.
- Correct hidden confounders in gene expression and methylation data using **Principal Component Analysis (PCA)**[6] (figure 4&5)

Finding significant correlated pairs of probe-gene (eQTM)

- Consider methylation probes within a 1 MB window from the Transcription Start Sites (TSS) of genes
- Assess the correlations of probe-gene pairs
- Correct the p-values for multiple hypothesis testing using the **Family-Wise Error Rate (FWER)** and **False Discovery Rate (FDR) techniques**

Extend the model to analyze the collective effect of methylation probes on gene expression value

- Use linear regression model
- Feature selection: **Most significant probe, Forward probe selection, Backward probe selection (MASS), Lasso-(GLMnet), and Full probes** (figure 6&7)
- Non-linear models: Neural Network (NN) and Conditional Inference Tree (CIT)
- Use 80% of subjects as train and 20% as test. We train the model using 3-fold cross validation.

Biological Plausibility

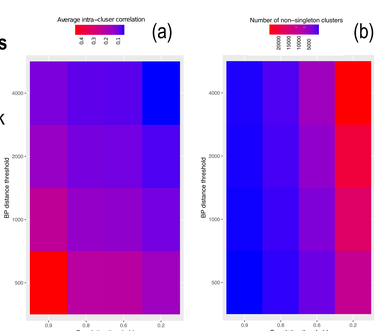
- Assess the spatial distribution of significant methylation probes according to their distance from genes
- Analyze the distribution of significant methylation probes in different chromatin states using **ChromHMM**[7](figure 8)
- Rank the genes based on the number of significant corresponding probes as an indicator for the multifunctionality of the gene
- Gene set enrichment analysis to find the relationship between the number of significant probes and the complexity of gene regulation using **erminR**[8] (Table1)

Results

Fig 3. The results of A-clust algorithm for different settings of parameters. A-clust splits methylation probes into highly correlated clusters and then pick one probe per cluster as a representative of that cluster to form a new set of probes corrected for multicollinearity.

(a) The average value of intra-cluster correlation for clusters with more than one probe.

(b) The number of clusters with more than one methylation probe.



Results (Cont'd)

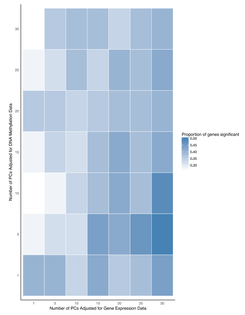


Fig4. Proportion of genes with at least one significant eQTM For different confounder correction settings. We remove variable number of first set of principle components for gene expression and methylation data. Significance levels are adjusted according to FDR values < 0.1. we select 30 PCs adjustment for genes and 10 PCs adjustment for methylation data as it provides the highest number of genes with at least one significant probe.

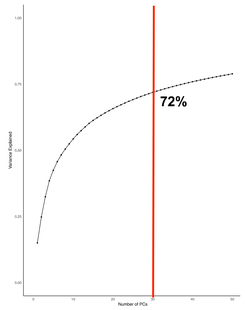


Fig5. Cumulative variance explained by PCs on gene expression data. The plot shows how much variance is explained when different numbers of PCs are included in the gene expression dataset. For 30 PCs, we see that 0.72 proportion of variance is explained.

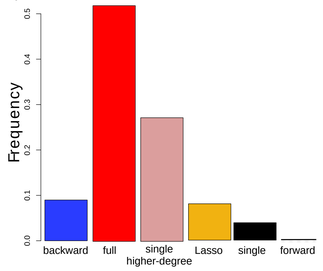


Fig6 . The percentage of genes best explained by models.
The plot shows that the full probes set is the best model for >50% of the genes. The second best uses the most significant probes transformed by a high degree function.

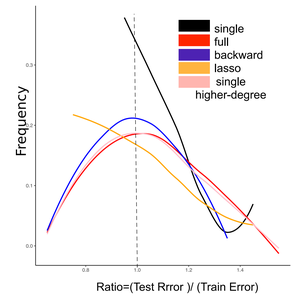
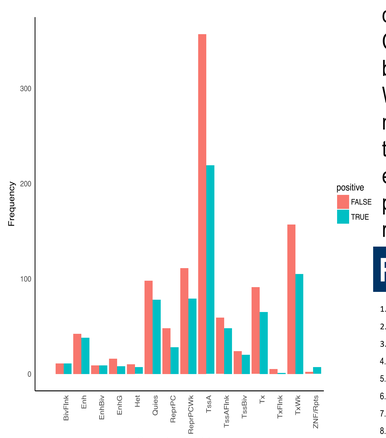


Fig7 . Ratio of test error to the training error for different models. The results indicate that the distribution of test error is very similar to the train in full, backward and non-linear models. Due to the distribution of test error, we conclude that our models generalize the data accurately.

Fig 8. The distribution of positive and negative correlations across different chromatin states. The red and green bars show the frequency of total number of positive and negative regulating CpG sites for different chromatin states. It shows that TSS enriches with the highest number of significant methylation probes compared to other chromatin states. Moreover, we see that the number of negative regulating methylation probes in TSS region is higher compared to the positive ones. Biological experiments also verify that most of the methylated CpG sites near the TSS and gene promotor repress the expression value of the gene resulting in a down-regulation effect.



Results (Cont'd)

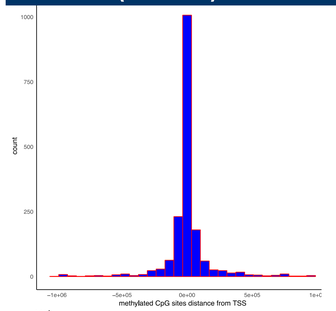


Fig9. The spatial distribution of significant CpG sites
a)The histogram shows that the majority of methylated CpG sites are concentrated near the transcription start sites (TSS) of genes.
b)The bar graph shows that the methylated CpG sites with significant P-values are located close to TSS. This result is consistent with previous data which showed that CpG sites are relatively enriched around TSS of genes.

Name	Number of genes	Corrected Pvalue	Corrected MFPvalue
oxoacid metabolic process	64	0.2675250	1
cellular lipid metabolic process	85	0.2952000	1
lipid metabolic process	107	0.2952000	1
oxidation-reduction process	83	0.2952000	1
carboxylic acid metabolic process	62	0.3057429	1

Table 1. **Gene set enrichment analysis.** The analysis shows that there is not any significant relationship between the numbers of significant methylated CpG sites corresponding to genes and the multifunctionality of those genes.

Discussion

Our study presents a large-scale analysis on the regulatory effect of methylation CpG sites on the expression value of genes in brain dorsolateral prefrontal cortex. We develop a set of regression models to assess the relationship between expression value of each gene and the methylation value of CpG sites. We investigate the distribution pattern of CpG sites considering the chromatin state and CpG-Gene distance. these findings are consistent with previous biological studies[2,7].

We observed that despite our initial expectation, the non-linear models (NN and CIT) didn't provide better estimations in comparison to linear regression. We suspect that using more complicated models, e.g. increasing the number of hidden layers in NN may capture the potential non-linearity between the CpG-gene relationship. A line of research that we will address in future.

References

1. Acharya, C. R., Owzar, K. & Allen, A. S. Mapping eQTL by leveraging multiple tissues and DNA methylation. *BMC Bioinformatics* **18,** 1–20 (2017).
2. Gutierrez-Arcelus, M. *et al.* Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *Elife* **2013,** 1–18 (2013).
3. Ng, B. *et al.* An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. *Nat. Neurosci.* **20,** 1418–1426 (2017).
4. Bennett, D. A. *et al.* Overview and findings from the rush Memory and Aging Project. *Curr. Alzheimer Res.* **9,** 646–63 (2012).
5. Bennett, D. A., Schneider, J. A., Arvanitakis, Z. & Wilson, R. S. overvie and findings from the religious orders study. *Curr. Alzheimer Res.* **29,** 628–645 (2012).
6. Ma, S. & Dai, Y. Principal component analysis based Methods in bioinformatics studies. *Brief. Bioinform.* **12,** 714–722 (2011).
7. Ernst, J. & Kellis, M. ChromHMM: Automating chromatin-state discovery and characterization. *Nat. Methods* **9,** 215–216 (2012).
8. Gillis, J., Mistry, M. & Pavlidis, P. Gene function analysis in complex data sets using erminej. *Nat. Protoc.* **5,** 1148–1159 (2010).