A Quantitative Analysis of DNA Methylation and Gene Expression Correlation in Asthma

Ananya Chavadhal[1], Ashwajit Singh[2], and Nakul Agarwal[3]

[1]Department of Aerospace Engineering , 22b0045@iitb.ac.in
[2]Department of Electrical Engineering , ashwajit.singh@iitb.ac.in
[3]Department of Civil Engineering , 22b0755@iitb.ac.in

Our project addresses how environmental factors influence gene expression through DNA methylation, an epigenetic modification known to regulate gene activity. While numerous studies have established that methylation can alter gene function, the quantitative relationship between methylation levels and corresponding changes in gene expression remains incompletely understood. We aim to determine a quantitative correlation between them and explore novel approaches for grouping genes to identify potential biomarker signatures for asthma. We begin by attempting to replicate results in existing papers in order to understand what approach to use when developing our own methods.

Introduction

To develop a robust model linking methylation and gene expression, we are replicating the computational methods of two key asthma studies: Nicodemus-Johnson et al. (1) and Yang et al. (2). While both papers analyze methylation and expression data to find disease signatures, their approaches differ. By reproducing their work, we aim to validate their findings and test the feasibility of their methods for our own goals.

This effort builds on established epigenetic research, such as the work by Fakhar et al. (3), which has demonstrated that environmental factors alter gene expression via DNA methylation, contributing to diseases like asthma and highlighting their potential as biomarkers. The insights from these replications will guide the development of our own analytical model, by helping determine the methods to include in order to identify a comprehensive set of biomarkers for asthma.

Results

Replication of Nicodemus-Johnson et al.

We successfully replicated key results from the Nicodemus-Johnson et al. paper (1). Our analysis confirmed that methylation at the specific CpG site `cg11303839` is significantly lower in asthmatic subjects compared to controls (Figure 1A). We then performed a genome-wide differential methylation analysis. The results are summarized in a volcano plot, which shows the magnitude and statistical significance of methylation differences for all CpG sites (Figure 1B). A Manhattan plot was also generated to show the chromosomal locations of these significant changes (Figure 1C). Finally, we conducted a differential gene expression analysis. To validate the paper's conclusion that gene expression was a less informative marker than methylation, we tested if the differentially expressed (DE) genes were enriched for genes controlled by genetic variants (eQTLs). Our Fisher's exact test confirmed the paper's finding, supporting the conclusion that there is no significant enrichment of eQTLs among DE genes in this dataset.

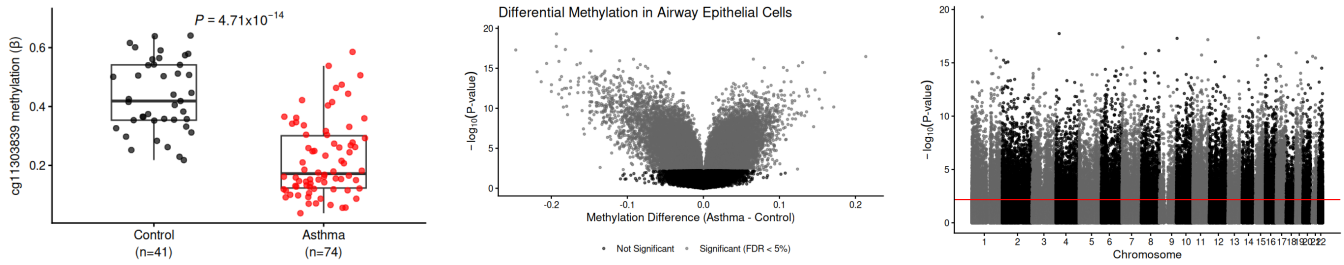
Replication of Yang et al.

Our replication of Yang et al. (2) was partially unsuccessful due to a lack of exact methods specified in the paper. Our analysis yielded 80 differentially expressed genes and 2128 differentially methylated probes (DMPs), the latter being a significant discrepancy from the 186 DMPs reported by the authors. We attribute this to our implementation lacking the paper's unspecified methods for batch correction and global p-value inflation adjustment (a few methods were tried but they did not give close results). We shall try fixing this using PEER (4).

Despite this, our data plots strongly support the paper's central hypothesis. The methylation volcano plot showed a clear trend of hypomethylation (skewed left). The expression plots (Figure 2A and 2B) confirmed a dominant trend of gene up-regulation (skewed right/top), with the scatter plot (Figure 2C) showing significant genes located above the y=x line. This provides evidence for the paper's main conclusion: a decrease in methylation is directly linked to an increase in gene expression in asthma.

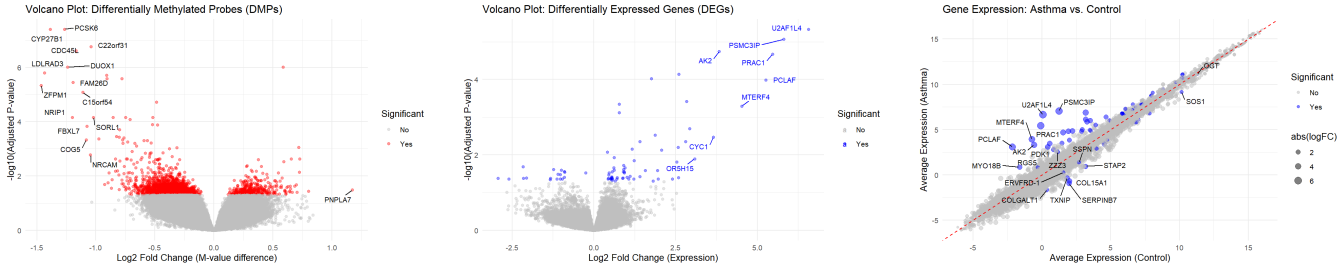
Replication of Fakhar et al.

Quality-control assessment showed that all samples had a low proportion of failed probes, confirming reliable methylation measurements across the dataset. The β-value density plot displayed the expected bimodal distribution, indicating appropriate separation between methylated and unmethylated CpG sites (Figure 3A). Differential methylation testing identified a number of CpG sites with statistically significant methylation differences, as visualized in the volcano plot where several loci displayed large effect sizes and low p-values (Figure 3B). These results demonstrate measurable, sample-wide methylation variation associated with asthma status.



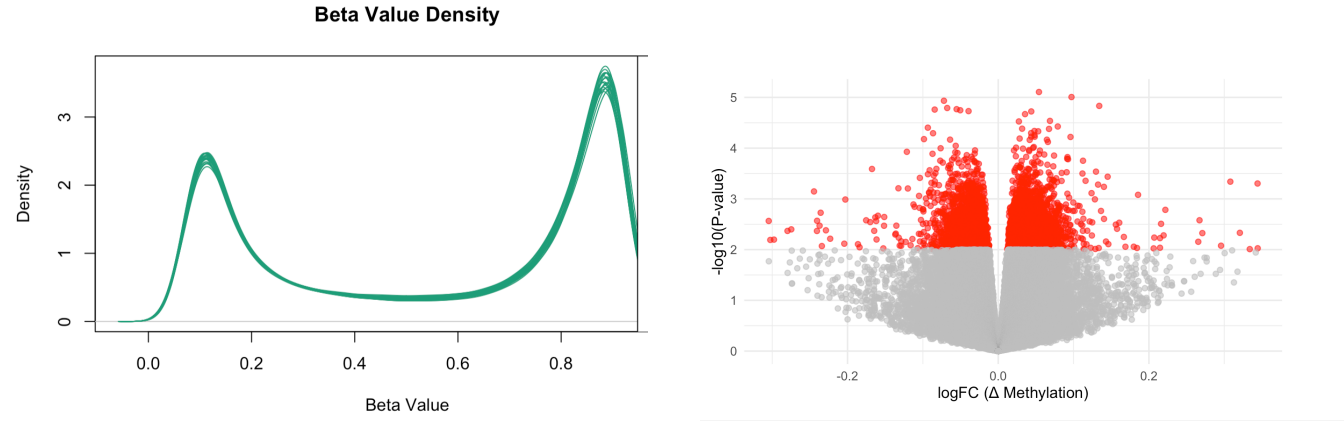
(a) Box Plot of `cg11303839` **(b)** Volcano Plot **(c)** Manhattan Plot

Figure 1. Replication of Differential Methylation Analyses (Nicodemus-Johnson et al. (1)). A) Box plot showing lower methylation at CpG site `cg11303839` in asthmatic subjects ($n = 74$) versus controls ($n = 41$). **B)** Volcano plot of genome-wide differential methylation, showing methylation difference on the x-axis and significance on the y-axis. **C)** Manhattan plot showing significance of methylation changes across all chromosomes.



(a) Volcano Plot for probe methylation **(b)** Volcano Plot for gene expression **(c)** Average expression of genes in asthma vs control

Figure 2. Replication of Differential Methylation Analyses (Yang et al. (2)). A) Volcano plot showing differential methylation in gene probes, showing methylation difference on the x-axis and significance on the y-axis. **B)** Volcano plot showing differential gene expression, showing expression difference on the x-axis and significance on the y-axis. **C)** Scatter plot showing Average expression of genes in asthma (n=36) vs control (n=36).



(a) Density Plot of β-values **(b)** Volcano Plot of Differential Methylation

Figure 3. Replication of Differential Methylation Analyses (Fakhar et al. (3)). A) Density plot showing the bimodal distribution of methylation β-values across all samples, confirming data quality. **B)** Volcano plot highlighting CpG sites with significant differential methylation between asthmatic and control groups.

Methods

Methods for Nicodemus-Johnson et al. (1) Replication

All analyses were performed in R. Data was sourced from the NCBI Gene Expression Omnibus (GEO) under accession number GSE85568.

Differential Methylation Analysis. We used the `limma` package to find differentially methylated CpG sites from the GSE85568 dataset. A linear model was fitted for each CpG site to compare asthma and control groups, while adjusting for age, gender, and ethnicity as covariates. Statistical significance was assessed using the `eBayes` function, and results were extracted with `topTable`.

Differential Gene Expression Analysis. We used the `edgeR` and `RUVSeq` packages on the raw gene counts from GSE85567. After filtering out genes with low expression, we used the `RUVg` function from the RUVSeq package to estimate and remove two factors of unwanted technical variation, using the 10% least variable genes as negative controls. We then fitted a negative binomial model using `glmQLFit` to find differentially expressed genes. The model included the disease group and the two unwanted variation factors as covariates.

eQTL Enrichment Analysis. We performed a one-sided Fisher's exact test to check if DE genes were enriched for eQTLs. DE genes were defined as those with a False Discovery Rate (FDR) < 0.05. The list of all known eQTL genes was taken from the paper's Supplemental Table 3.

Methods for Yang et al. (2) Replication

All analyses were performed in R (version 4.5.1). Data was sourced from the NCBI Gene Expression Omnibus (GEO) under SuperSeries accession GSE65205, specifically SubSeries GSE65163 (Methylation) and GSE65204 (Gene Expression).

Differential Methylation Analysis. Raw methylation data (IDAT files) from GSE65163 were loaded using the `minfi` package and normalized using the `preprocessSWAN` method. Probes associated with known SNPs were removed using `dropLociWithSnps`. Differential methylation analysis was then performed as done in Nicodemus-Johnson et al., using the `limma` package to fit a linear model to the resulting M-values. The model compared asthma and control groups while adjusting for age, gender, and African American race as covariates. P value < 0.05 was taken for significant differential methylation.

Differential Gene Expression Analysis. Raw gene expression data (Agilent text files) from GSE65204 were loaded using the `read.maimages` function from the `limma` package. The data was background corrected using the `"normexp"` method and subsequently quantile normalized using `normalizeBetweenArrays`. Differential expression analysis was then performed using `limma`, fitting a linear model to the normalized expression values. As in the methylation analysis, the model compared asthma and control groups while adjusting for age, gender, and African American race as covariates. P value < 0.05 was taken for significant differential expression.

Methylation-Expression Linking. To integrate the methylation and expression datasets, both statistical results tables were annotated with official gene symbols. For methylation, probe IDs were mapped to gene symbols using the `UCSC_RefGene_Name` column from the `IlluminaHumanMethylation450kanno.ilmn12.hg19` annotation; for probes mapping to multiple loci, only the first gene was retained. For expression, identifiers from the `SystematicName` column were translated to gene symbols using the `org.Hs.eg.db` package by mapping both RefSeq (`NM_`) and Ensembl (`ENST_`) identifiers. The two annotated results tables were then merged by gene symbol. Genes significantly associated with both methylation and expression changes were identified by filtering this merged table for an adjusted P-value < 0.05 in both analyses.

Methods for Fakhar et al. (3) Replication

Publicly available DNA methylation data in the form of raw .idat files were obtained for both asthmatic and non-asthmatic individuals. Because the IDAT files did not contain sample phenotype information, the corresponding .SOFT metadata file was used to extract clinical details and create a samplesheet.csv linking each sample to its asthma status. The IDAT files and metadata were then imported into the analysis environment, and normalized β-values were computed to represent methylation levels at individual CpG sites. Quality control procedures were applied to exclude low-quality probes and verify data consistency across samples. Principal component analysis was performed to assess overall clustering, and differential methylation analysis was conducted to identify CpG sites showing significant methylation differences between asthmatic and non-asthmatic groups. Statistical significance was evaluated using multiple testing corrections, and the results were visualized through density and volcano plots to illustrate both global methylation patterns and site-specific differences.

Next Steps

For the next phase, we will implement advanced batch normalization using PEER (4) to better correct for confounding factors. A primary task will be to quantify the relationship between methylation, expression, and asthma, likely via mediation analysis. To gain deeper biological insight, we will also identify and characterize differentially methylated regions (DMRs) to determine if significant CpG sites cluster within key gene promoters or regulatory elements. If feasible, pathway enrichment analyses may also be used to explore how these methylation changes could affect processes like immune signaling and airway inflammation. Finally, we will validate our model's robustness by testing if biomarkers derived from the Nicodemus-Johnson et al. (1) data are generalizable to the Yang et al. (2) dataset.

References

1. J. Nicodemus-Johnson, R. Myers, N. Sakabe, D. Sobreira, D. Hogarth, E. Naureckas, A. Sperling, J. Solway, S. White, M. Nobrega, *et al.*, "Dna methylation in lung cells is associated with asthma endotypes and genetic risk," *JCI insight*, vol. 1, no. 20, p. e90151, 2016.
2. I. V. Yang, B. S. Pedersen, A. H. Liu, G. T. O'Connor, D. Pillai, M. Kattan, R. T. Misiak, R. Gruchalla, S. J. Szefler, G. K. Khurana Hershey, C. Kercsmar, A. Richards, A. D. Stevens, C. A. Kolakowski, M. Makhija, C. A. Sorkness, R. Z. Krouse, C. Visness, E. J. Davidson, C. E. Hennessy, R. J. Martin, A. Togias, W. W. Busse, and D. A. Schwartz, "The nasal methylome and childhood atopic asthma," *Journal of Allergy and Clinical Immunology*, vol. 139, no. 5, pp. 1478–1488, 2017.
3. M. Fakhar, M. Gul, and W. Li, "Structural and functional studies on key epigenetic regulators in asthma," *Biomolecules*, vol. 15, no. 9, p. 1255, 2025.
4. O. Stegle, L. Parts, R. Durbin, and J. Winn, "A bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eqtl studies," *PLoS computational biology*, vol. 6, no. 5, p. e1000770, 2010.