A Quantitative Comparison of Methylation Association with Age, BMI and Smoking Status

Aditya Garg
Data Science Institute
Columbia University
New York, NY
ag3741@columbia.edu

Chenchao Zang
Data Science Institute
Columbia University
New York, NY
chenchao.zang@columbia.edu

Jun Guo
Data Science Institute
Columbia University
New York, NY
jg3555@columbia.edu

Papiya Sen
Data Science Institute
Columbia University
New York, NY
ps2893@columbia.edu

***Abstract*—In this research, we study the interactions between gene methylation and age, BMI and smoking habits of individuals by analyzing the recently released large skin aging dataset and finding sites that are significantly associated with the 3 factors- age, BMI, and smoking status.**

***Keywords*—Gene, Methylation, Age, BMI, Smoking, Association.**

I. Introduction

DNA methylation is a widely studied epigenetic modification in organ development, aging, and different disease statuses. DNA methylation occurs when a methyl (CH3) group is added to a CpG site without changing the DNA sequence. Up to 80% of all CpG sites in human DNA are methylated, and the process is reversible and heritable. The extent of methylation can be affected by both the environmental and genetic factors. Global methylation is typical in aging cells, while a decrease in global methylation has been proposed as a molecular marker of cancer. A large-scale study focusing on BMI has identified associations with health, and DNA methylation occurred at a few CpG sites of blood cells and adipose tissue. A study using blood samples from five young and five old male adults has identified significant differential methylation of more than 10000 CpG sites. Another study using Blood DNA and self-reported smoking status of 645 individuals has recognized 66 differentially methylated CpG sites. The same research has also found that people quitting smoking and having high methylation levels recovered to the levels of methylation of people who never smoked. All three studies have identified changes in methylation at several CpG sites associated with one of following factors: age, BMI, and smoking status.

The MuTHER cohort from skin tissue was genotyped with Illumina Infinium HumanMethylation450 array, which assays about 485,000 CpG sites spanning 99% of gene. There are two metrics used to measure methylation: Beta-value and M-value. Beta-value is the ratio of the methylated probe intensity and the sum of methylated, unmethylated intensities and a predetermined constant. Therefore, beta-value ranges from 0 to 1 with zero meaning completely unmethylated and one an entirely methylated CpG site. The M-value, which can be calculated as the logit of beta value, is the log2 of ratio between methylated intensities and unmethylated intensities. A positive M- value means that more molecules are methylated than unmethylated and vice versa. M-value gives more insight of the distribution of methylation level across genomes.

II. Project Goal

Using the (MuTHER) methylation data from skin tissue, our minimum goal is to find the list of CpG sites that are significantly linked to age, BMI and smoking. The goals will be achieved through following sub-goals:

- Identify and remove extreme values caused by technical factors such as noise, manual error, etc. However, keep the CpG sites that are outliers due to biological reasons.
- Generate linear regression models for each CpG site with age, BMI and smoking as factors, and m-values as dependent variable. In addition, include interactions among factors as independent variables for regressions.
- Identify and extract the CpG sites based on their statistically significant factors e.g. group all the CpG sites for which BMI is statistically significant.
- Within a group, check for correlations between the CpG sites adn cluster them based on their methylation values. This would indicate groups of CpG sites which are correlated, which is significant for the genomic community.

III. Dataset

The dataset used here is the Multiple Tissue Human Expression Resource (MuTHER) methylation data, which is publicly available as part of the Gene Expression Omnibus (GEO) dataset on the National Center for Biotechnology Information (NCBI) portal.

The genetic methylation data contains DNA methylation analysis of skin samples from punch tissue biopsy of 322 healthy female individuals. The data is collected using genome wide genomic DNA profiling with the HumanMethylation450 BeadChip (450k) of skin samples. The skin tissue DNA was derived from a peri-umbilical punch biopsy (adipose tissue was removed from the biopsy before freezing) from 322 healthy female individuals of the TwinsUK cohort.

The dataset contains the beta-value associated with the degree of methylation for different CpG sites. Besides the beta-value, the dataset contains associated information about the samples gender, age, body mass index (BMI) and smoking status.

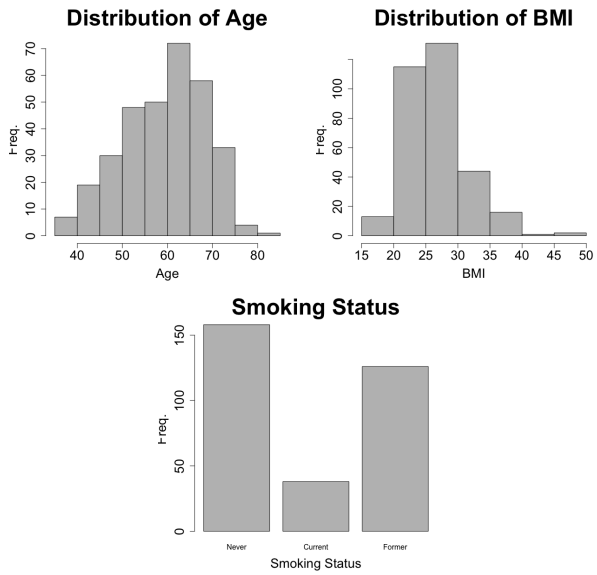


Fig. 1: Age, BMI and Smoking Status Distribution

IV. Normalization Method

A. Why Normalization?

The purpose of normalization is to remove the technical and systematic variability from the data so that measurements can be compared throughout the entire sample. The systematic variation occurs because the Infinium HumanMethylation450 BeadChip 450k platform assesses methylation levels using two types of probes, Type I and Type II. The two probes quantify methylation levels differently. As a result, the distributions of methylation levels for the two probes are different as well.

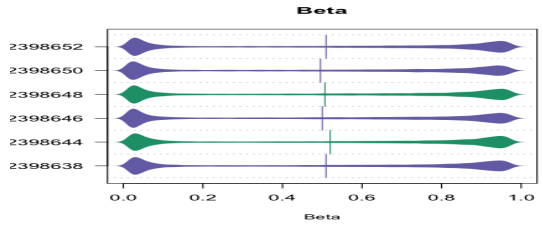


Fig. 2: Beta Value Distribution of 6 Random Sites

B. SWAN vs BMIQ

Two common normalization methods are BMIQ (the beta-mixture quantile normalization) and the SWAN (the subset-quantile within array normalization). Both methods perform normalization within sample array. SWAN select subsets of probes biologically similar based on CpG content and assume Type I and Type II intensity distribution should be the same. The normalization enforces the distribution to be similar. BMIQ approach transforms the Type II distribution to be similar to that of Type I. BMIQ sampling method is also different. For Type I and Type II probes on a sample, they are assigned to be methylated, hemi-methylated, and unmethylated. Type II probes classified as methylated or unmethylated are quantile normalized to have identical distribution as Type I probes of the corresponding class. Then the hemi-methylated Type II probe is centered and scaled to span between the methylated and unmethylated Type II probes.

After comparing BMIQ-normalized, SWAN-normalized and unnormalized data, Wu et al. found that both normalization methods do not significantly improve the reproducibility of data. BMIQ method is superior for probe bias reduction. For the association analysis, different normalizations do not make much of a difference for CpG sites that are highly statistically significant.

Importantly, since the samples are all collected by using Illumina HumanMethylation450 BeadChip from the same platform (GPL13534), there is no need to concern the variations because of the differences of platform. Finally, we successfully validated that the dataset should and has been performed on BMIQ to correct probe-type bias.

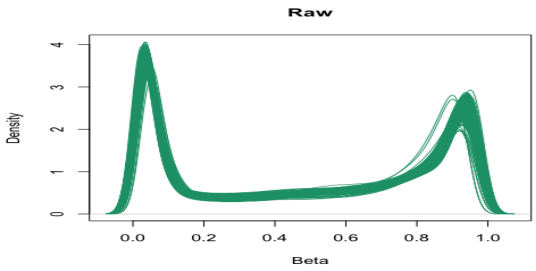


Fig. 3: Beta Value Distribution of All Sites

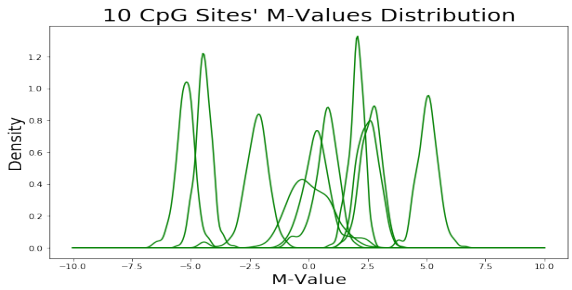


Fig. 4: M-Value Distribution of 10 Random Sites

V. Methodology

The methylation data is provided as beta values, which is bound between 0 and 1. Prior to generating the regression models, the beta values were converted to M-values, which can take any real value, using the formula:

$$Beta_i = \frac{2^{M_i}}{2^{M_i}+1}$$

$$M_i = \log 2(\frac{Beta_i}{1-Beta_i})$$

The M-value is recommended for conducting differential methylation analysis.

For each of the 450k CpG sites, a linear regression model was trained using the metadata factors - age, BMI and smoking status. Since the metadata from 322 samples is used repetitively for the 450k models, multiple test correction is applied. Two different correction methods are tested - Bonferroni and False Discovery Rate Benjamini-Hochberg (FDR-BH). We then selected the CpG sites that are statistically significant at the level of 0.01 after correction. Since we are interested in the differential methylation of these sites across samples, variance method is used to filter these sites further. Sites are sorted in descending order of standard deviation, then the top sites with standard deviation larger than 5.0% are selected and used for further analysis.

VI. Results

A. Model

We performed linear regressions using age, BMI, and smoking status as the predictor variables. First, correlation between the factors was checked. Since there is low correlation between factors, as shown in Fig. 5, the effect of each factor can be meaningfully interpreted from the regression models.

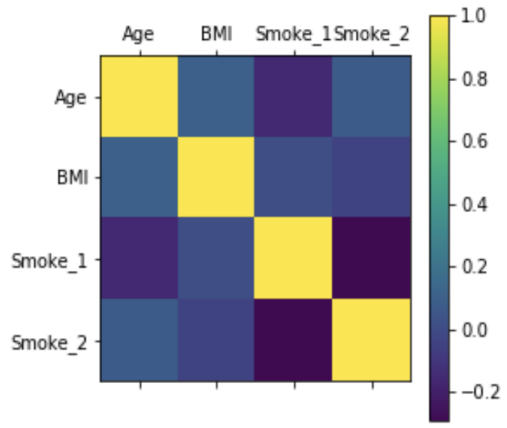


Fig. 5: Correlation plot for factors Age, BMI, Smoking status 1 (quit smoking), Smoking status 2 (smoking)

To perform linear regression on all the variables, we first convert the smoking status value from categorical variables to one-hot encoded vectors. The process allows the representation of categorical data in a discrete form. It transforms an integer encoded vector (a vector consists 0, 1 and 2 in our case) into a matrix with a dimension of integer vector length times the number of categories. For example, a vector of [0, 2, 1, 2, 1, 0] converts into the following form:

$$[1,0,0],[0,0,1],[0,1,0],[0,0,1],[0,1,0],[1,0,0]$$

. However, the one-hot encoded vectors have linear dependency, which makes us drop the first column. In other words, category zero of the final vector. We then perform linear regression using the smoking status 1 and 2. If a sample has smoking status 0, the two variables' coefficients would be multiplied by zero.

After fitting the 322 samples' age, BMI and one-hot encoded smoking status on m-values for each of the 450k CpG sites, we obtained factor coefficients of the CpG sites. To judge whether a predictor has a significant linear relationship with the CpG site it is fitting, we performed t-tests with FDR-Benjamini adjusted significance value. The method downsized the number of sites for further inspection, as discussed in next section.

1) Bonferroni correction: Since 450k regression models were generated from metadata of 322 samples, we have 450k hypotheses being tested on a small dataset. Any factors having significance value smaller than 0.01 may have false positive error just by the virtue of definition of p-value. To eliminate this problem, multiple testing correction was applied on the p-values. In the first pass, we used Bonferroni correction. The results shown in Fig.6 indicate that of possible 450k sites only 3601 sites are significant with age, 803 are significant with BMI, 3 with smokers, and 1 with ex-smokers, at a significance level of 0.01.

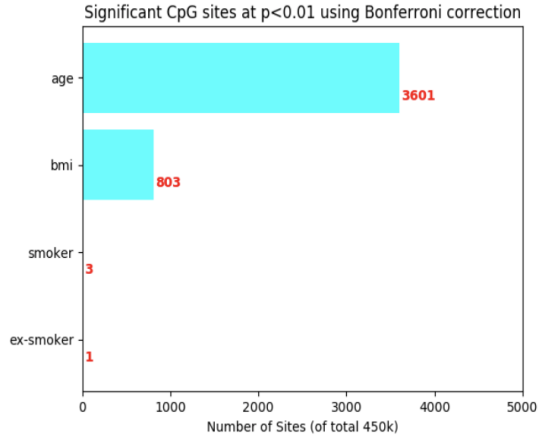


Fig. 6: Bonferroni correction is severe in reducing false positive error. It selected only 3601 sites of possible 450k sites as significant with age at $p < 0.01$

2) False Discovery Rate Benjamini-Hochberg correction: We selected FDR-BH correction method, which is a recommended method for genomics. Setting an upper bound on false discovery rate at 1%, we found factors significant at p = 0.01. The results in Fig.7 indicate that this method is less severe than Bonferroni. There are 38,913 sites significant with age, 30,328 with BMI, 5 with smokers and 1 with ex-smokers at significance level of 0.01.

B. Differentially Methylated Sites

One of the biological questions of interest is to find the differentially methylated CpG sites between various phenotypes, including smoking, age and bmi. Other researchers have performed similar studies to detect the association between differential methylation with these phenotypes with or without adjusting other confounding factors. The novelty of our research is that the dataset was collected from skin samples and the models were built while considering other covariates.

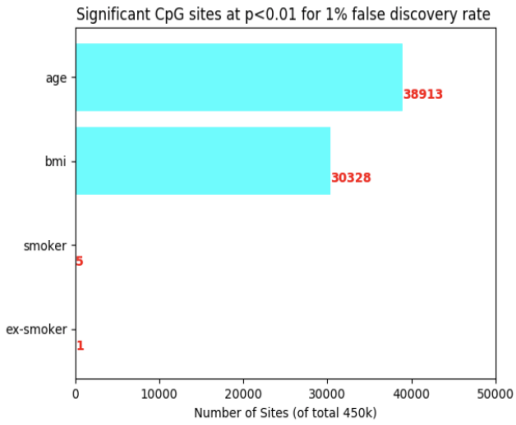


Fig. 7: FDR Benjamini-Hochberg correction method applied to group the sites by factors. Since FDR is not as severe as Bonferroni correction, we get more sites in each group

1) Age-related Differential Methylation: Of the CpG sites shortlisted, we want to identify the sites that are differentially methylated over the sample space. Using correlation coefficient as the metric, all correlation coefficients for Age were found to be positive. The sites were sorted in descending order using absolute values of correlation coefficient. Top such sites are shown in Fig.8., which closely match those reported in literature [4-9]. Since number of significant sites for Smokers and Ex-smokers are only five and one respectively, all were considered.

Age

CpG	corr coeff
cg10729426	0.597562
cg23606718	0.533302
cg20899581	0.526647
cg23995914	0.512989
cg00699993	0.502731
cg24938830	0.476900
cg06385324	0.468221

Fig. 8: Top differentially methylated CpG sites sorted in descending order of absolute value of correlation coefficient. Shown for factors Age

We also plot the methylation levels of Top 4 differentially methylated CpG sites vs Age. As we can see, there is a visible positive linear relationship between methylation level and age. Importantly, the positive relationship between methylation and age has been mentioned by other researchers[4-9]. We will do further validation analysis in the validation section of this paper.

2) BMI-related Differential Methylation: Similarly, the top 4 differentially methylated CpG sites related to BMI are plotted here. The most differentially methylated one is negatively related to BMI. The rest of the three have obvious positive relationship. The validation procedures will be stated in the later section of the report.

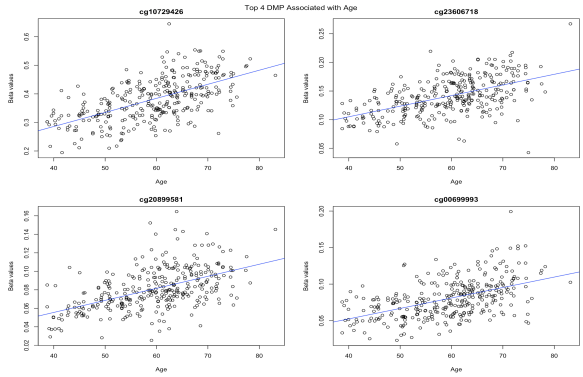


Fig. 9: Top 4 Differentially Methylated Sites (Methylation level vs Age)

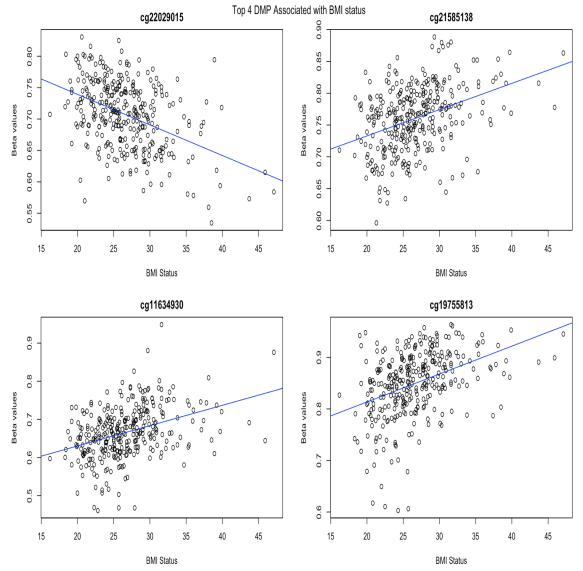
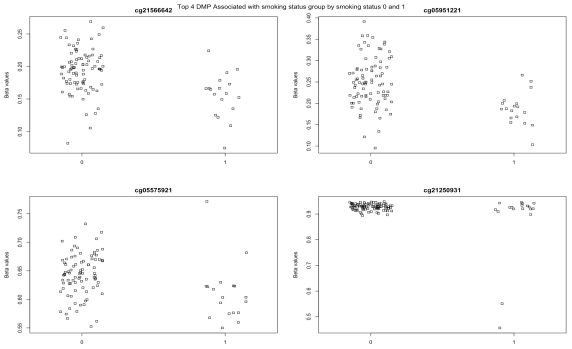


Fig. 10: Top 4 Differentially Methylated Sites (Methylation level vs BMI)

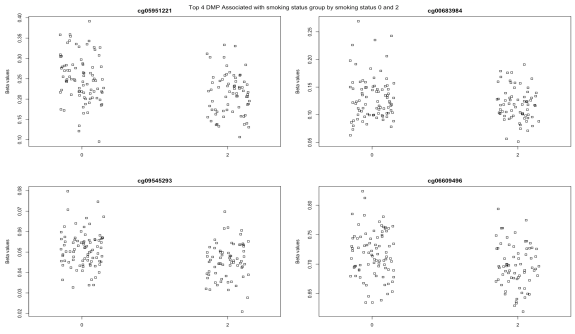
3) Smoking-related differential methylation: For detecting differentially methylated CpG sites, pairwise comparisons were made among smoking status levels, and methylation levels of 322 samples of the CpG sites associated with smoking status were plotted, grouped by smoking level. In Fig. 11, we can observe different methylation levels between smoking groups. The validation of these CpG sites will be discussed in the later section of this repoty.

C. Grouping CpG sites by significant factor

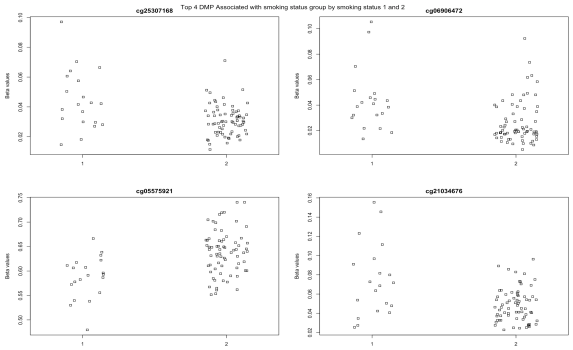
The developed linear regression models give the dependency of M-value of each site on the factors. The CpG sites were then grouped based on the factor(s) they are significant with. The purpose is to identify if the CpG sites in a group represent certain gene regions or whether they are CpG islands, which is addressed in a later section. The interaction



(a) Top 4 DMP (Methylation level group by smoking status 0 and 1)



(b) Top 4 Differentially Methylated Sites (Methylation level group by smoking status 0 and 2)



(c) Top 4 Differentially Methylated Sites (Methylation level group by smoking status 1 and 2)

Fig. 11: Top 4 Differentially Methylated Sites (Methylation level with Smoking Status)

parameters have currently not been included in the models in this study, but can , however, be part of the future steps.

1) Overlapping CpG Sites: The sites identified above have overlaps between different factors e.g. there are 3344 sites that are significantly affected by age as well as BMI, as shown in the Venn diagram in Fig.12, drawn using Venny 2.1. Age and BMI individually have significant effect on 36945 and 27692 sites respectively. It is interesting to note that smoking comes up as a significant factor in five sites only. It may be relevant to explore in future work whether the methylated sites in smokers get unmethylated in ex-smokers.

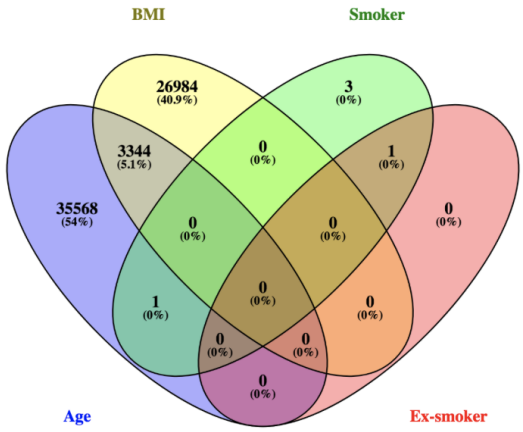


Fig. 12: Venn diagram of CpG sites in different factor groups - age, BMI, Smoker, Ex-smoker

2) Interaction Among Factors: To identify interactions among different factors, we apply the multiple linear regression model on the 450k CpG sites. A multiple regression has the following variables: age, BMI, one-hot vectorized smoking status 1 and 2, age*BMI, age*smoker, age* ex-smoker, BMI * smoker, and BMI* ex-smoker. The reason not to include any interaction term between smoking statuses is to avoid linearity among vectors. The FDR-BH adjusted p-values with the significance value of 0.1 identify cg13485335 as the only site having significant age*BMI factor. The interaction term has a coefficient of approximately -0.007. One way of interpreting the model is for a subject at age 50, one unit of increment in BMI would decrease methylation by 0.35. Another interpretation is that for a person having BMI of 20, every year the m-value at cg13485335 decreases by 0.14. In addition to the interaction term between age and BMI, that of age and smoker has two significant sites. BMI * smoker has 18 significant sites, and BMI *ex-smoker has one site.

D. Differentially Methylated Regions (DMR)

Although a probe-wise analysis is useful and informative, it is also interesting to study whether some proximal CpGs are collectively differentially methylated. Therefore, in this part, we identify the differentially methylated regions. Here, to identify the DMRs, we first identify the methylation status of each probe and group them into regions. There are several Bioconductor packages that can identify DMR from 450k data. Here, we would like to use dmrcate to perform our analysis since it is based on limma, the package which is used to find the Differentially Methylated Sites. In our case, there are 322 individuals in total, and it is possible to use an epigenome-wide association study (EWAS) method to detect differentially methylated DNA regions with phenotypes.

Additionally, each probe is annotated with relevant genome information such as the genomic position, gene annotation and etc. By default, the annotation process is accomplished

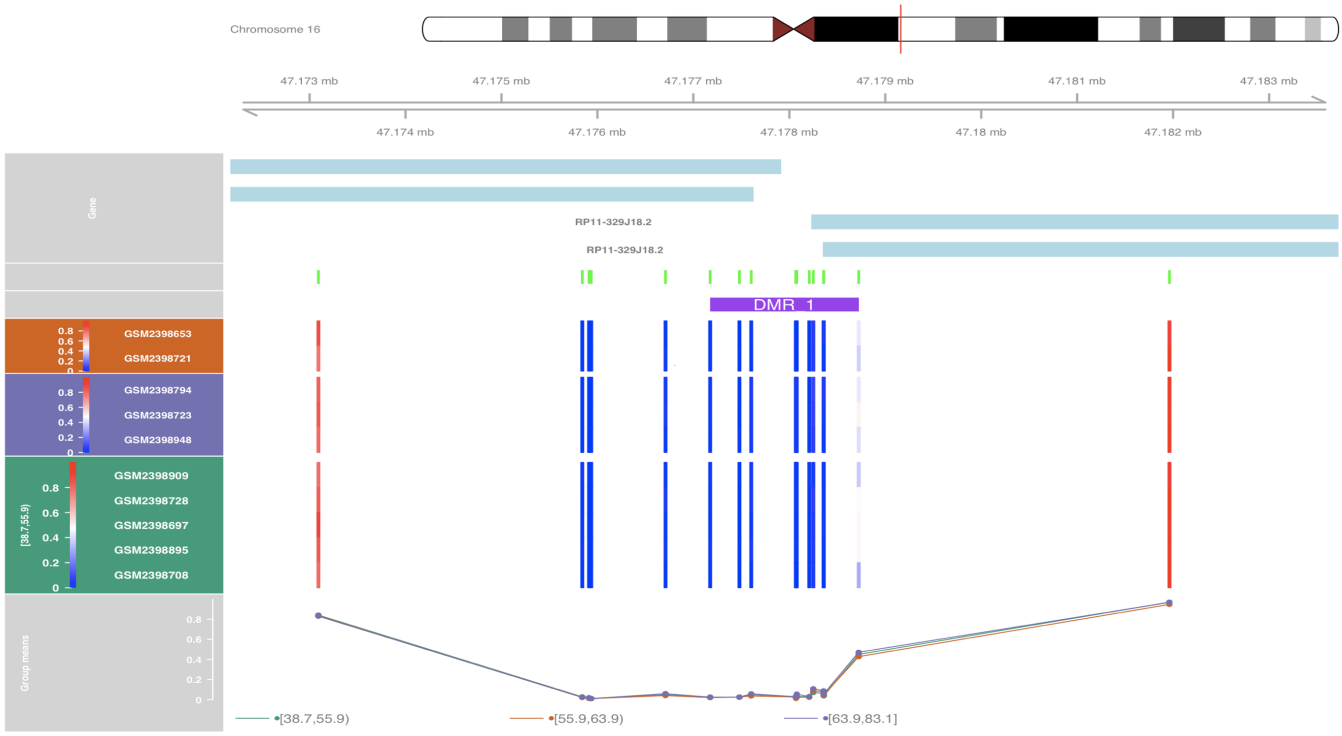


Fig. 13: DMRs with gene name and loc in Chromosome with methylation levels. The top bar represents the corresponding chromosome of the methylated regions and light blue arrow represents the gene name. The following spectrums is the heatmap of methylation level for each CpG sites grouped by smoking status. The bottom lines connects the methylation level of each CpG sites group by smoking status.

using ilmn12.hg19. Here, we plot the differentially methylated regions, which are distinguished by the comparison between smoking status 0 and smoking status 1. These DMRs are in the Chromosome 16 in Fig 13.

E. Clustering CpG Sites

After identifying significant CpG sites associated with each of the three factors - age, BMI and Smoking, we try to identify distinct clusters of CpG sites within each group by clustering them based on their methylation values across samples. The purpose of this analysis is to identify distinct groups of CpG sites that show the same characteristics with respect to change in their methylation values across different factors observed in this study. To find such distinct clusters of CpG sites and their characteristic behaviour, we perform cluster analysis using a three step process outlined below that consists of scaling, clustering and analyzing cluster groups.

1) Scaling Methylation Values: The first step of the clustering process is to pre-process the methylation values in order to bring them to an appropriate scale. Since we want to identify clusters that change in the same manner across different samples, irrespective of their absolute mean values, we standardize each CpG site's m-value across different samples to a mean of zero and standard deviation of one.

2) DBSCAN Clustering: After converting the m-values of all the CpG sites across samples to a standard scale, we perform clustering using the scaled m-values across samples as features. We use the Density-based spatial clustering of applications with noise (DBSCAN) proposed by Ester et al [14]. DBSCAN relies on a density-based notion of clusters which is designed to discover clusters of arbitrary shape by grouping together points that are close to each other based on a distance measurement (Euclidean distance in our case) - epsilon, and a minimum number of points (5 in our case). It also marks the points that are in low-density regions as noise.

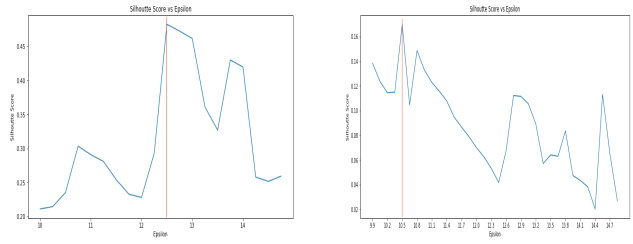


Fig. 14: Variation of silhouette score for different values of epsilon for age(left) and BMI(right) related CpG sites

For optimizing the value of epsilon, we use a metric called silhouette score. The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). We, thus, pick the epsilon value

that maximizes the silhouette score for our clusters. From Fig. 14, we can see that for age related CpG sites, the optimum value of epsilon is 10.5, while that for BMI related CpG, the optimum value of epsilon is 12.5.

3) Analyzing Clusters: Using the optimum epsilon values, we then run the DBSCAN algorithm to identify clusters of CpG sites within each group - age related and BMI related. As can be seen from Table I, the DBSCAN clustering results in 3 clusters for age related CpG sites with 133, 9 and 5 samples respectively in the 3 clusters, while marking 3853 CpG sites as noise. For BMI related CpG sites in Table II, we see that the algorithm extracts two clusters with 1856 and 520 CpG sites in each, and marks 1624 CpG sites as noise.

We also visualize the clusters by plotting the non-noise samples across their principal components 1 and 2 in Fig. 15.

Now that we have the list of CpG sites in each cluster, for future steps, we can evaluate their biological significance by annotating them with relevant genomic information.

Cluster ID	No. of CpG Sites
1	133
2	9
3	5
Noise	3853

TABLE I: Distribution of sites in different clusters for Age related CpG Sites

Cluster ID	No. of CpG Sites
1	1856
2	520
Noise	1624

TABLE II: Distribution of sites in different clusters for BMI related CpG Sites

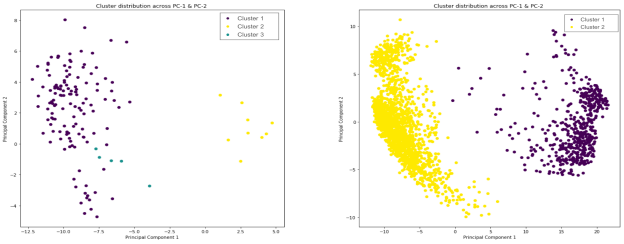


Fig. 15: Visualizing clusters of CpG sites for Age (left) and BMI (right) across Principal Components 1 and 2.

VII. Validation

In order to further validate our result, we compare our findings with the reported CpG sites of other researchers. Through validation, we can observe the difference in results using skin tissue compared to using blood cells.

A. Age-related Differential Methylation Validation

The first CpG site, cg10729426, was reported by Zhang 2012 under a patent application for predicting age and identifying agents that induce or inhibit premature aging. The second CpG site, cg23606718, was reported as a marker in methylation analysis [10]. The third CpG site, cg20899581, was reported as one of CpG sites with DNA methylation levels significantly altering with age in adipose tissue[11]. The fourth one, cg00699993, was reported in another patent application [14].

B. BMI-related Differential Methylation Validation

The CpG site, cg21585138, was mentioned by a researcher in a genome-wide analysis of DNA methylation variance. Other three CpG sites have not been reported, but the cell type of samples could impact the methylation level of CpG sites since our data was collected from skin sample rather than common blood cells. Therefore, it is reasonable to believe that others are real differentially methylated CpG sites which have not been reported by researchers yet.

C. Smoking-related Differential Methylation Validation

The study by Baglietto [15] has identified CpG sites cg05951221, cg21566642, and cg05575921 having significant association with smoking and lung cancer risk. Another study by Shenker has identified cg21566642 having association with smoking as well. The other two sites have not been reported by researches yet.

Acknowledgment

We want to thank our mentor at DSI, Columbia University - Prof. Andreas Mueller, our domain experts and industry partners at Unilever - Dr. David Gunn, Melissa Matzke and Sheila Rocha, and TA for capstone project - Yogesh Garg for their kind cooperation and guidance.

References

[1] Gyorffy B, Gyorffy A and Tulassay Z, *The problem of multiple testing and its solutions for genome-wide studies* Orv Hetil, 2005, 146(12):559-563

[2] Figueroa, M., Genomic assays and the multiple testing problem

[3] Du P., Zhang X., Huang C., et al., Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis, BMC Bioinformatics 2010, 11:587

[4] Langevin SM, Houseman EA, Christensen BC, Wiencke JK, Nelson HH, Karagas MR, et al. The influence of aging, environmental exposures and local sequence features on the variation of DNA methylation in blood. Epigenetics. 2011;6(7):90819.

[5] Bell JT, Tsai PC, Yang TP, Pidsley R, Nisbet J, Glass D, et al. Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. PLoS Genet. 2012;8(4):e1002629.

[6] Horvath S, Zhang Y, Langfelder P, Kahn RS, Boks MP, van Eijk K, et al. Aging effects on DNA methylation modules in human brain and blood tissue. Genome Biol. 2012;13(10):R97.

[7] Fraga MF, Esteller M. Epigenetics and aging: the targets and the marks. Trends Genet. 2007;23(8):4138.

[8] Bollati V, Schwartz J, Wright R, Litonjua A, Tarantini L, Suh H, et al. Decline in genomic DNA methylation through aging in a cohort of elderly subjects. Mech Ageing Dev. 2009;130(4):2349.

[9]]Florath I, Butterbach K, Mller H, Bewerunge-Hudler M, Brenner H. Cross-sectional and longitudinal changes in DNA methylation with age: an epigenome-wide analysis revealing over 60 novel age-associated CpG sites. Hum Mol Genet. 2014;23(5):1186201.

[10] Hannum, Gregory et al. Genome-Wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. Molecular cell 49.2 (2013): 359367. PMC. Web. 4 Dec. 2017.

[11] Tina Rnn, Petr Volkov, Linn Gillberg, Milana Kokosar, Alexander Perfilyev, Anna Louisa Jacobsen, Sine W. Jrgensen, Charlotte Brns, Per-Anders Jansson, Karl-Fredrik Eriksson, Oluf Pedersen, Torben Hansen, Leif Groop, Elisabet Stener-Victorin, Allan Vaag, Emma Nilsson, Charlotte Ling; Impact of age, BMI and HbA1c levels on the genome-wide DNA methylation and mRNA expression patterns in human adipose tissue and identification of epigenetic biomarkers in blood, Human Molecular Genetics, Volume 24, Issue 13, 1 July 2015, Pages 37923813

[12] Eura et al. Arthritis Rheumatism, 2009, vol. 60, pp. 1416-1426

[13] Martin Ester, Hans-Peter Kriegel, Jrg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96), Evangelos Simoudis, Jiawei Han, and Usama Fayyad (Eds.). AAAI Press 226-231.

[14] Martin Ester, Hans-Peter Kriegel, Jrg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96), Evangelos Simoudis, Jiawei Han, and Usama Fayyad (Eds.). AAAI Press 226-231.

[15] Baglietto, Laura, et al. DNA methylation changes measured in pre-Diagnostic peripheral blood samples are associated with smoking and lung cancer risk. International Journal of Cancer, vol. 140, no. 1, Nov. 2016, pp. 5061., doi:10.1002/ijc.30431.

[16] Shenker, Natalie S., et al. Epigenome-Wide association study in the European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking. Human Molecular Genetics, vol. 22, no. 5, 2012, pp. 843851., doi:10.1093/hmg/dds488.