Statistical Modelling Project.
Effect of smoking on the CpG islands methylation

Diego Benito J., Anna Chechenina

December 2022

1 Introduction

Epigenetic modifications constitute one of the most important regulatory mechanisms by which cells modulate gene expression. Tissues, systems, and entire organisms are affected by changes in the expression of genetic information since the molecular machinery by which life functions derives from the expression of genetic material. Understanding the manner by which genes are turned "off" or "on" is of paramount importance in virtually every field that deals with organic systems, and it is of particular interest within medical fields since gene expression plays a role in every type of disease (either as a defense mechanism or a contributor to disease).

This report will focus on one particular DNA epigenetic modification such as DNA methylation within DNA regions called CpG islands. DNA methylation is catalyzed by DNA methyl transferases (DNMTs) and involves the covalent transfer of a methyl group from S-adenosyl methionine (SAM) to the 5' positions of cytosine residues in CG dinucleotides. For the purposes of this report, methylation can be thought of as an "epigenetic signaling tool that cells use to lock genes in the 'off' position". In other words, methylation is a process that prevents the binding of transcription factors to chemically altered (methylated) nucleotides within DNA (see Figure 1) which could affect how proteins interact with the methylated region. The CpG islands are DNA sequences rich in CpG sites (¿50% CpG sites within a 200bp sequence. Since the cytosines are the common methylation targets, the methylation status of the CpG island is one of the measures of chromatin activity in this area. Each gene could contain the promotor and operator regions around the main sequence, their methylation may control the level of gene expression.

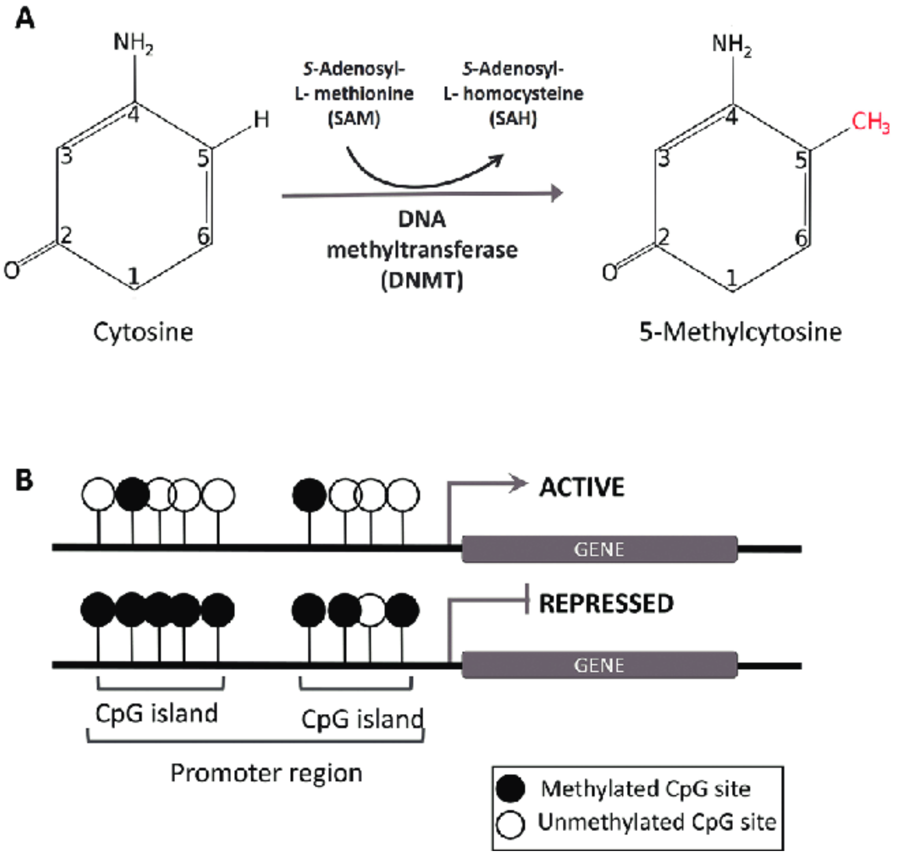


Figure 1: DNA methylation. (A) CpG methylation mechanism is mediated by DNA methyltransferases and consists of the addition of a methyl group to the carbon in the 5th position of cytosines that precedes guanine nucleotides. (B) The CpG islands are DNA sequences rich in CpG sites (¿50% CpG sites within a 200bp sequence). Methylation of CpG islands inside a promoter region may control gene expression [1]

We will particularly focus on the effects of smoking on methylation, which is a stressor that has been shown to affect epigenetic markers [2, 5, 7, 6]. This effect at an epigenetic level is believed to be one of the mechanisms by which tobacco smoke acts as a carcinogenic agent and could cause serious changes in gene expression. There any many experimental papers showing that more than 18760 CpG islands in 7201 annotated genes, representing almost one-third of all known human genes, were differentially methylated in the current smokers versus the number found in the never smokers [3, 7]. It was shown that when the major part of the methylation changes was reversible and returned to the initial state after 5 years of not smoking, some particular methylated genes did not return to the level of the never-smokers even after 30 years of smoking cessation [3].

It was also shown that the expression level of the DNA-methylase in the lungs of smokers was significantly higher compared to non-smokers [4]. This could lead to the rise of epigenetic changes in the related tissues and, consequently, to serious diseases.

In our study, we performed a superficial analysis of the patients' parameters in the data and our main goal was to identify which CpG island has the most significant relation to smoking status.

We would like to highlight the fact that statistical analysis should be treated with caution - the ability to predict and model reality through statistical methods should not be exaggerated. Our analysis will put particular emphasis on how we generated our model and the implications of such a model. In turn, we will tread lightly with the type of causal inferences that we can generate, which is especially important when analyzing data within the field of genetics since there are countless intervening factors.

2 Methods

The raw data was obtained from the open source platform Kaggle. For the data exploration and models construction, we used R language in the R studio environment with the corresponding packages. The data included information on the CpG islands percentage of methylated residues for 621 patients with information on the age, smoking status, and sex of each person.

The full Rmarkdown with all commands we performed could be found on the project github page.

3 Data Exploration

Before we start modeling we need to have a look at the data parameters and their distribution. Our data contains information on the of the 20 CpG islands which were located on the different chromosomes and some parameters of the patients such as smoking status, gender, and age.

First, we explored the distribution of the age parameter, which showed a big shift to the right with the mean around 50 (see Figure 2).

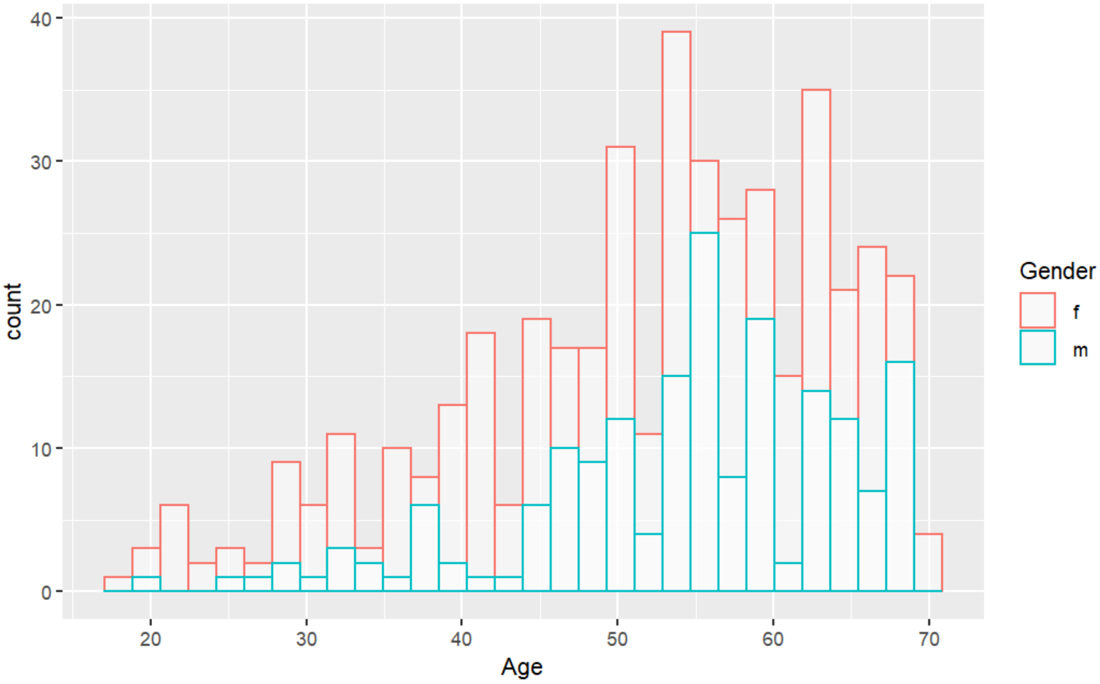


Figure 2: The distribution of the age parameter depending on the sex of the patient

An important point to note is that females are overrepresented within our dataset. Out of 621 individuals, 440 are female whilst 181 are male.

Our initial focus lied on the relationship between smoking and methylation, so we decided to explore how the average methylation between the smoking and non-smoking individuals varied across the CpG islands. The following graph illustrates the difference between average methylation of smokers vs non-smokers per CpG island (see Figure 3).

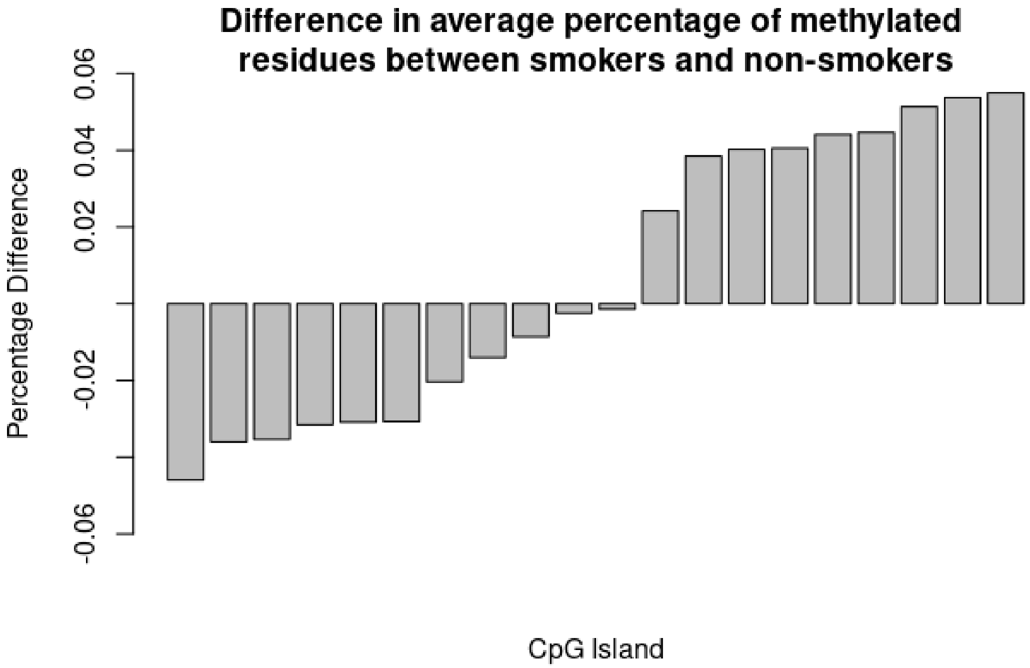


Figure 3: The difference in average methylation depending on the smoking status

If we compare the values for different islands, we can see that the majority of islands have a bimodal distribution with a higher density of around 0.5 and around either 0.8 or 0.1 (see Figure 4).

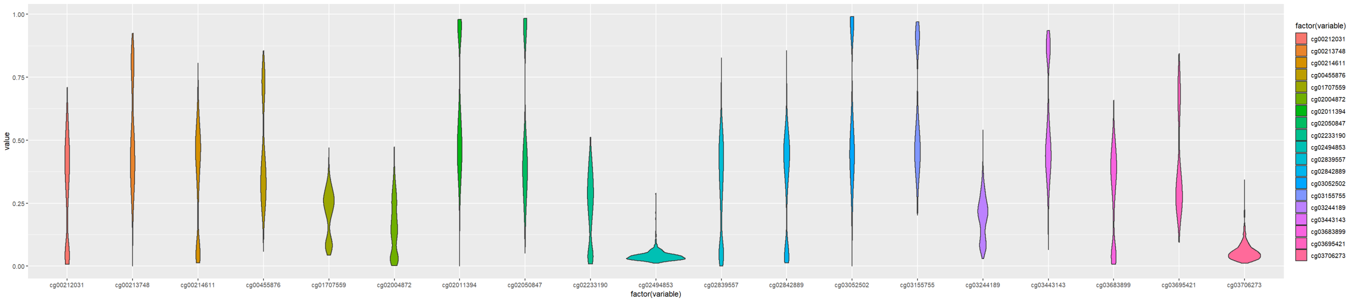


Figure 4: The distribution of the methylation values depending on the island

Then we also built the barplot for each island to compare the methylation level between smokers and non-smokers. As we can see in Figure 5, the difference between smoking and non-smoking people is so that usually only mean stays the same but quartiles are changing. Thus for non-smokers, we have more narrow distributions in most cases compared to smokers.

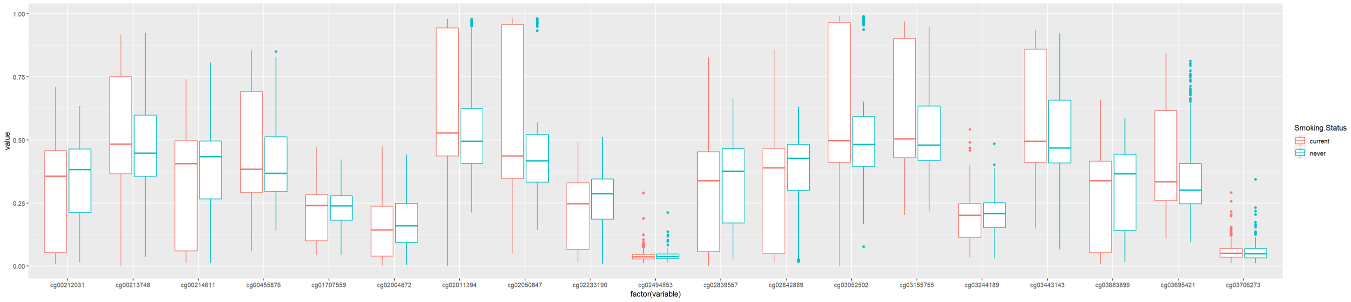


Figure 5: The distribution of the methylation values depending on the island and a smoking status

4 Models construction and comparison

We can convert the methylation percentage in a CpG island into a binary response variable by ascribing some comparison point. By constructing this binary response variable, we can easily implement a logistic regression model which aids in the statistical analysis of our data. Two methods were implemented.

First, we utilized the median degree of methylation for a randomly selected island and compared the observations with said median, and then repeated the process by averaging the degree of methylation amongst the 20 CpG islands.

Additionally, we decided to generalize the methylation status for each patient and build the model for this new parameter to see the bigger picture. In order to do this we performed several steps. First, for all CpG islands, we calculated the mean value among all non-smokers to get the baseline to compare to. Then for all CpG islands for all patients, we calculated new values, taking the difference between the initial value and the mean value for non-smokers for the current island. Thus if all values would be similar, we would expect to get values around zero. Since we do not know how the methylation and gene expression would affect health in the case of each position, we calculated the absolute values of these differences. In the next step, we took the mean values for all CpG islands along the samples and get the new column with the mean values for each patient.

To use these values in the logistic regression model we coded values with absolute values of more than one standard deviation as 1 and others with 0. The classical statistic approach would be to calculate this confidence interval as mean - 3*std, however, in this case, all our data would be insignificant ($std = 0.04, 3 * std = 0.1315776$, when the distribution as on the Figure 6).

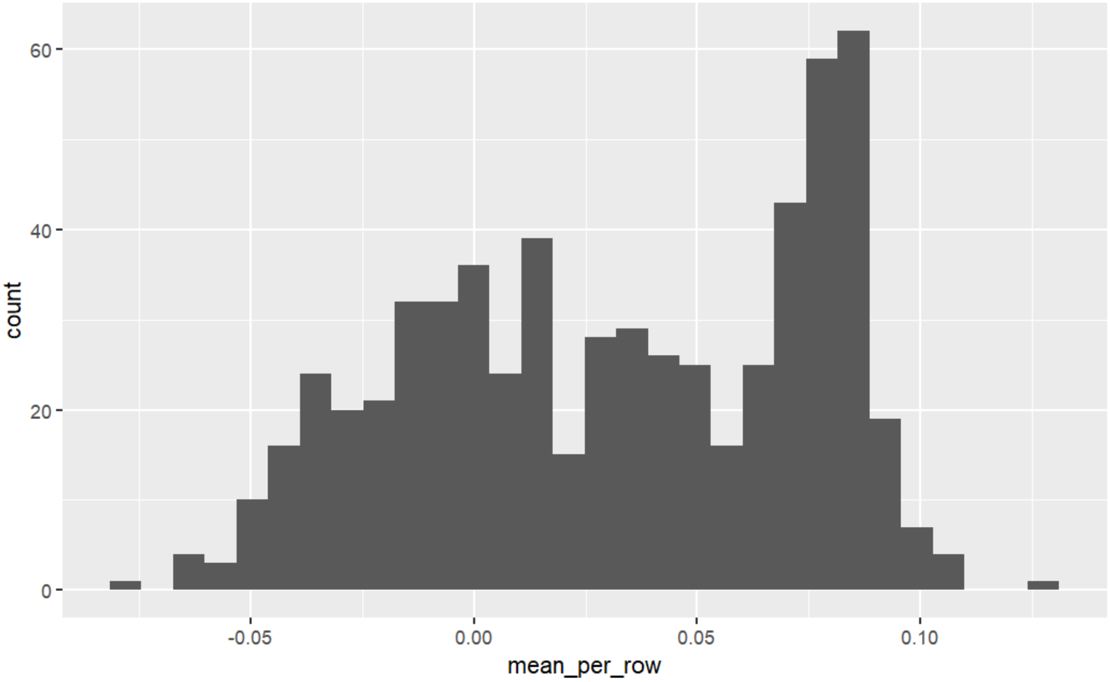
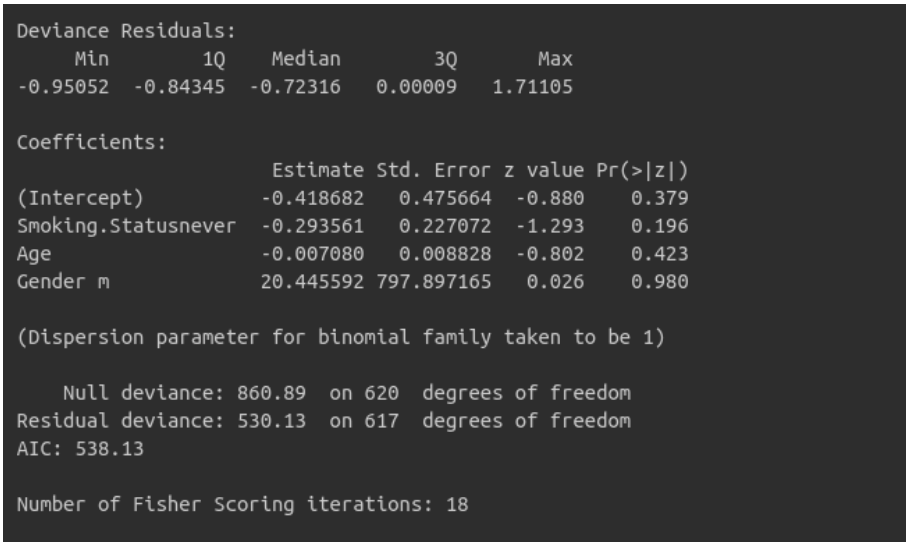


Figure 6: The distribution of the absolute mean differences for each patient

Then, to code all other CpG islands we used the same approach: calculate the difference with a non-smoker mean, take the absolute value, and code it as 1 if the value is more than 1 standard deviation and as 0 if not.

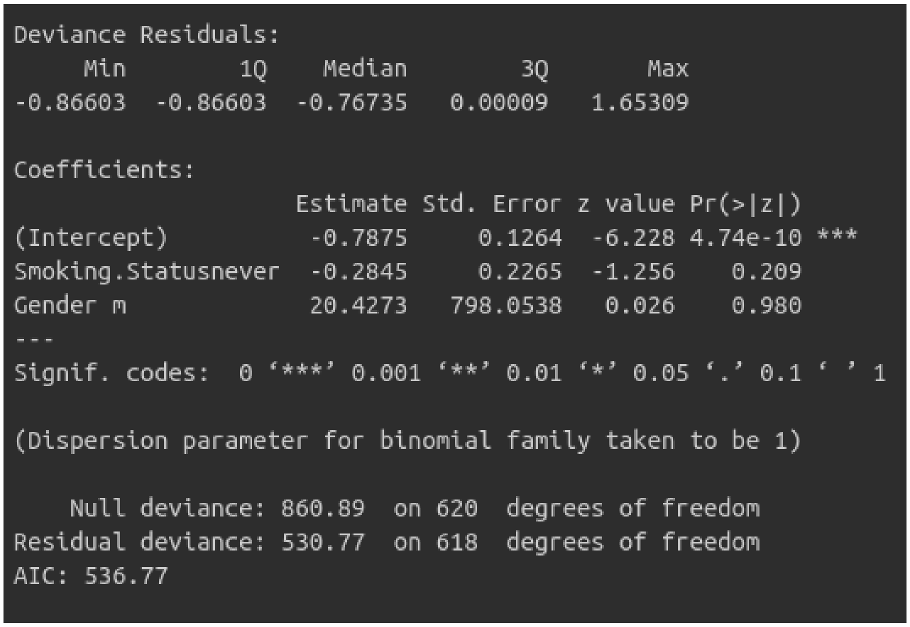
4.1 Comparison to Median as a Binary Response Variable

As mentioned previously, we considered generating a binary response by considering the median as comparison point for each island. Specifically, we take the median percentage of methylation for a given island and compare all observations with said measurement, identifying observations that are **above** the median and those that are **at or below** the median. In turn, we can generate a logistic model that utilizes a combination of predictors (Age + Sex + Smoking Status) to see what models would be a good fit for the data. In this case, we will generate a logistic regression, and for the sake of clarity and brevity we will pool the data by taking an average along the 20 CpG islands to identify which regression model could provide insight into the relationship between smoking and methylation onto the averaged dataset.



```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.95052  -0.84345  -0.72316   0.00009   1.71105

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -0.418682   0.475664  -0.880    0.379
Smoking.Statusnever  -0.293561   0.227072  -1.293    0.196
Age                  -0.007080   0.008828  -0.802    0.423
Gender m             20.445592 797.897165   0.026    0.980

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 860.89  on 620  degrees of freedom
Residual deviance: 530.13  on 617  degrees of freedom
AIC: 538.13

Number of Fisher Scoring iterations: 18
```

Figure 7: The estimates for a logistic regression encompassing all available predictors: Age, Sex, and Smoking Status



```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.86603  -0.86603  -0.76735   0.00009   1.65309

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -0.7875     0.1264  -6.228 4.74e-10 ***
Smoking.Statusnever  -0.2845     0.2265  -1.256    0.209
Gender m             20.4273   798.0538   0.026    0.980
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 860.89  on 620  degrees of freedom
Residual deviance: 530.77  on 618  degrees of freedom
AIC: 536.77
```

Figure 8: The estimates for a logistic regression encompassing all available predictors: Age, Sex, and Smoking Status

Our initial model for the pooled CpG island data utilized all of the predictors, which generated the model shown on the Figure 7.

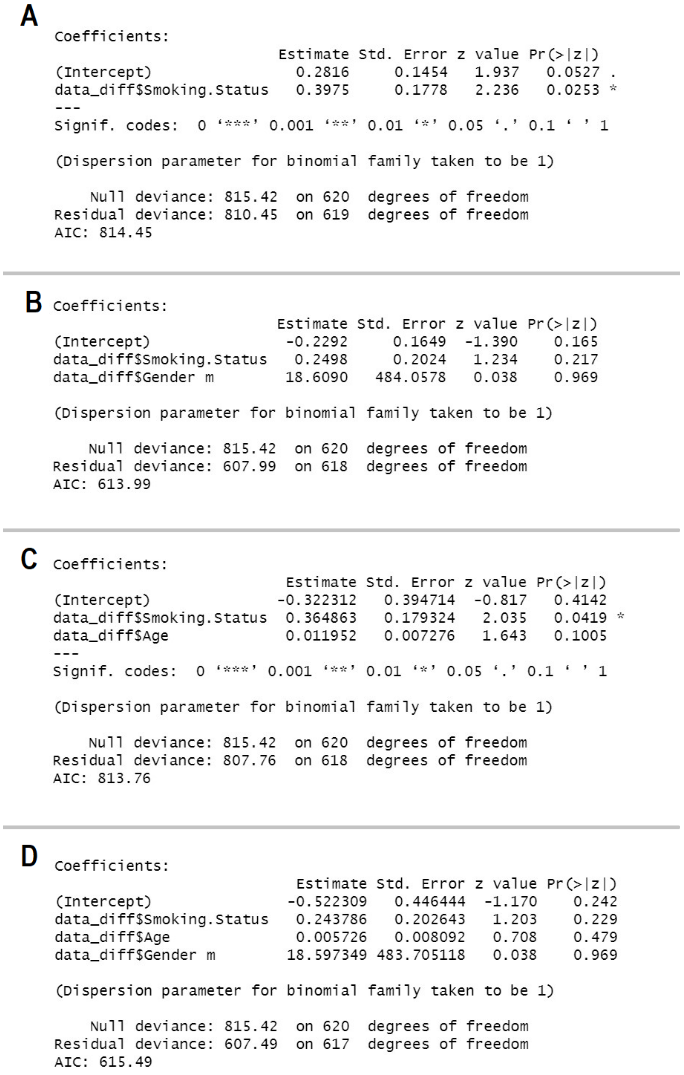
As we can observe, none of the predictors have a statistically significant relationship with the response variable. However, if we remove Age, we generate a model that would be considered a better fit as measured by its AIC statistic shown in the figure below, which is approximately 1.3 units smaller and would indicate that Age could be excluded from the model. Only considering Sex and Smoking Status, our model parameters acquire the following values as shown in the Figure 8.

We've interpreted the estimated values for the parameters as follows: neither Gender nor Non-Smokers are statistically significant parameters with respect to the response variable (as seen by the non-significant z-values) when considering the pooled data, whilst Smokers have a very statistically significant correlation with the response variable. These relationships imply that knowing someone's gender or knowing that they do not smoke would not be informative as to the methylation of the considered CpG islands when compared to the average individual. Meanwhile, we would expect smoker's CpG island's to be udermethylated, as shown by the negative value of the estimate for the parameter.

If we were to give a gross oversimplification, we could say that not smoking isn't an assurance that genetic material has a high or low degree of methylation, whilst smoking **does** correlate with a lower degree of methylation within the investigated CpG islands.

4.2 Comparison to mean with corrections as a Binary Response Variable

In the same way as mentioned before, we performed the modeling for the CpG values obtained using the difference with mean and standard deviation. The results could be found in figure 10. We can see that here Smoking status is a significant parameter in the model with only Smoking status as a predictor and in the model with Smoking status and Age. However, if we look at the BIC criterion, we will observe that when we are using a Gender predictor, it significantly decreases. The logistic curve for the generalized column with mean for patients could be found on figure **??** in Supplementary.



```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                 0.2816     0.1454   1.937   0.0527 .
data_diff$Smoking.Status    0.3975     0.1778   2.236   0.0253 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 815.42  on 620  degrees of freedom
Residual deviance: 810.45  on 619  degrees of freedom
AIC: 814.45
```

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                -0.2292     0.1649  -1.390    0.165
data_diff$Smoking.Status    0.2498     0.2024   1.234    0.217
data_diff$Gender m         18.6090   484.0578   0.038    0.969

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 815.42  on 620  degrees of freedom
Residual deviance: 607.99  on 618  degrees of freedom
AIC: 613.99
```

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -0.322312   0.394714  -0.817   0.4142
data_diff$Smoking.Status  0.364863   0.179324   2.035   0.0419 *
data_diff$Age             0.011952   0.007276   1.643   0.1005
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 815.42  on 620  degrees of freedom
Residual deviance: 807.76  on 618  degrees of freedom
AIC: 813.76
```

```
Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               -0.522309   0.446444  -1.170    0.242
data_diff$Smoking.Status   0.243786   0.202643   1.203    0.229
data_diff$Age              0.005726   0.008092   0.708    0.479
data_diff$Gender m        18.597349 483.705118   0.038    0.969

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 815.42  on 620  degrees of freedom
Residual deviance: 607.49  on 617  degrees of freedom
AIC: 615.49
```

Figure 9: The estimates for a logistic regression encompassing (A) only Smoking status (B) Smoking status and Gender (C) Smoking status and Age (D) all available predictors: Age, Sex, and Smoking Status

4.3 Application to all CpG islands in the dataset

This same analysis can be applied to specific CpG islands, which is shown in the tables 1 and 2 in the Supplementary section. Each particular island would require separate analysis (some have a particularly interesting correlation between gender and methylation) but the breadth of such an analysis might be beyond the scope of this report. However, we should mention that the interpretations may not necessarily correspond to the generalized trend that has been observed regarding smoking as a significant predictor of methylation when considering the islands as a pooled total. Nevertheless, we can say that the most significant islands according to our results are cg00050873, cg01707559, and cg03695421 (p-value less than 0.5 for both approaches). They are might be considered for further investigation.

5 Conclusion

As has been shown across our models, being a smoker correlates with the undermethylated regions of DNA, specifically the CpG islands under investigation. This seems to be consistent with the literature, which highlights tobacco's carcinogenic properties and, more importantly, highlights the existence of known mechanisms by which the inhalation of tobacco smoke and its constituent chemical components may inflict genetic modifications. By manipulating the data (to facilitate the implementation of statistical analysis) we have shown a statistically significant correlation between smoking and under-methylated DNA regions, primarily through applying logistic regressions to analyze our dataset. In particular, we have shown that islands cg00050873, cg01707559, and cg03695421 shown a promising correlation with the smoking status of the patient. From that, we can assume the possible role of the genes located in essential cellular processes.

References

[1] Alejandra Alarcón et al. "Epstein-Barr Virus–Associated Gastric Carcinoma: The Americas' Perspective". In: *Gastric Cancer*. InTech, Sept. 2017. DOI: 10.5772/intechopen.70201. URL: https://doi.org/10.5772/intechopen.70201.

[2] K. M. Bakulski et al. "DNA methylation signature of smoking in lung cancer is enriched for exposure signatures in newborn and adult blood". In: *Scientific Reports* 9.1 (Mar. 2019). DOI: 10.1038/s41598-019-40963-2. URL: https://doi.org/10.1038/s41598-019-40963-2.

[3] Roby Joehanes et al. "Epigenetic Signatures of Cigarette Smoking". In: *Circulation: Cardiovascular Genetics* 9.5 (Oct. 2016), pp. 436–447. DOI: 10.1161/circgenetics.116.001506. URL: https://doi.org/10.1161/circgenetics.116.001506.

[4] Young-Mi Kwon et al. "Different susceptibility of increased DNMT1 expression by exposure to tobacco smoke according to histology in primary non-small cell lung cancer". In: *Journal of Cancer Research and Clinical Oncology* 133.4 (Oct. 2006), pp. 219–226. DOI: 10.1007/s00432-006-0160-2. URL: https://doi.org/10.1007/s00432-006-0160-2.

[5] Ken W. K. Lee and Zdenka Pausova. "Cigarette smoking and DNA methylation". In: *Frontiers in Genetics* 4 (2013). DOI: 10.3389/fgene.2013.00132. URL: https://doi.org/10.3389/fgene.2013.00132.

[6] Pei-Chien Tsai et al. "Smoking induces coordinated DNA methylation and gene expression changes in adipose tissue with consequences for metabolic health". In: *Clinical Epigenetics* 10.1 (Oct. 2018). DOI: 10.1186/s13148-018-0558-0. URL: https://doi.org/10.1186/s13148-018-0558-0.

[7] Dandan Zong et al. "The role of cigarette smoke-induced epigenetic alterations in inflammation". In: *Epigenetics &amp Chromatin* 12.1 (Nov. 2019). DOI: 10.1186/s13072-019-0311-8. URL: https://doi.org/10.1186/s13072-019-0311-8.

6 Supplementary

Table 1: Results for the approach using mean and std

	Coding using mean and std			
CpG island	*AIC*	*Interciept p-value*	*Smoking p-value*	*Gender p-value*
cg00050873	**84.002**	**0.990**	**0.231**	**6.57e-07**
cg00212031	413.39	1.17e-13	0.818180	0.000388
cg00213748	139.75	2.44e-13	0.741	<2e-16
cg00214611	329.43	3.79e-15	0.967	0.982
cg00455876	97.123	1.18e-08	0.52	<2e-16
cg01707559	**391.78**	**3.79e-15**	**0.106**	**0.982**
cg02004872	471.97	1.37e-09	0.635	0.982
cg02011394	19.343	0.995	0.995	0.995
cg02050847	17.833	0.996	0.996	0.996
cg02233190	404.2	1.78e-13	0.798	0.982
cg02494853	380.59	<2e-16	0.79012	0.00401
cg02839557	390.71	2.18e-14	**0.474**	0.982
cg02842889	240.26	8.42e-16	0.962	0.983
cg03052502	6	0.999	0.999	0.999
cg03155755	6	0.999	0.999	0.999
cg03244189	425.55	5.99e-12	**0.46090**	0.00033
cg03443143	19.343	0.996	0.997	0.995
cg03683899	367.06	4.25e-14	0.827	0.982
cg03695421	**93.78**	**0.987**	**0.454**	**0.986**
cg03706273	387.15	1.86e-15	0.879182	0.000276

Table 2: Results for the approach using meadian

	Coding comparing to the median			
CpG island	*AIC*	*Interciept p-value*	*Smoking p-value*	*Gender p-value*
cg00050873	**538**	**0.05266237**	**0.21456206**	**0.97958025**
cg00212031	555	1.648533e-01	6.301802e-01	1.949770e-09
cg00213748	540	0.1717266	0.6946430	0.9795857
cg00214611	540	0.004006945	0.794421452	0.979596075
cg00455876	538	0.4378879	0.6515679	0.9795145
cg01707559	**549**	**5.970484e-01**	**6.337717e-02**	**9.971540e-10**
cg02004872	541	0.1254751	0.6481004	0.9796037
cg02011394	532	0.390419290	**0.453475780**	0.979394570
cg02050847	554	1.466287e-02	7.674247e-01	1.877125e-09
cg02233190	563	2.277574e-01	**4.431326e-02**	.216606e-13
cg02494853	854	0.015640701	0.731873092	0.005680288
cg02839557	541	0.2211755	0.7997601	0.9795887
cg02842889	535	0.81512750	**0.02357497**	0.97948000
cg03052502	539	0.06436991	**0.46691829**	0.97956190
cg03155755	538	0.5955553	0.3807696	0.9795423
cg03244189	587	1.904836e-01	7.449937e-01	4.062910e-19
cg03443143	540	0.0309409	0.8924868	0.9796004
cg03683899	541	0.08652745	0.66901688	0.97958524
cg03695421	**538**	**0.05367708**	**0.21443292**	**0.97958030**
cg03706273	766	9.327707e-05	6.560935e-01	1.828494e-17

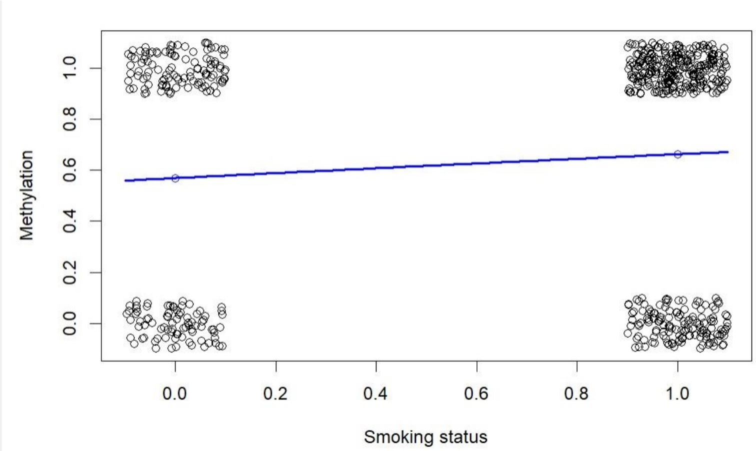


Figure 10: The logistic curve for the model with only Smoking status as a predictor on the generalized data