

CpG methylation data analysis results

Kim Reijntjens

2022-10-03

Introduction

DNA methylation is a process plays a part in gene regulation. Methylation involves the addition or removal of methyl groups to or from the bases and sugars in DNA. methylation occurs on the cytosine of CG locations in the DNA. methylatable CG sequences also called CpG sequences are concentrated in CpG-rich regions called CpG islands, located at the 5'ends of genes. Methylation prevents the formation of certain base pairs, and thereby the accessibility for interactions with other components. methylation adds a hydrophobic character to some of the tRNA regions. that may be important for their interaction with proteins that regulate synthesis. (Klug,William,Cummings,Michael,Spencer,Charlotte,Michael,Palladino) (Tymoczko, 2015)

Relavance of the project

Because alterations in the genome and activity of the genes are associated with common diseases such as cancer or asthma, it is important to know if smoking can be the cause of this. If this is the case than in further research there can be looked at which CpG sites are affected by smoking, and if there is a relation with the affected CpG sites and the ones associated with diseases like cancer and asthma. (Klug,William,Cummings,Michael,Spencer,Charlotte,Michael,Palladino)

1.1 Goal

Can CpG methylation show a relation with smoking, based on a prediction whether a patient is smoking or non-smoking using differences in CpG site values.

Data explanation

For this we use the dataset of 683 patients. the dataset has 683 rows and 24 columns containing the patients; age, gender, smoking status and 20 CpG site values. The dataset was already compressed to 20 CpG site values sites per sample. The original dataset was composed for a study titled “Differential DNA methylation in Rheumatoid arthritis” (NCBI series GSE42861) Where the original dataset contained 485577 rows with methylation data per sample (NCBI)(KAGGLE)

Loading the data

Table 1: An overview from the fist lines of the data (continued below)

GSM	Smoking.Status	Gender	Age	cg00050873	cg00212031
GSM1051525	current	f	67	0.6076	0.4228
GSM1051526	current	f	49	0.3451	0.5687
GSM1051527	current	f	53	0.3213	0.3609
GSM1051528	current	f	62	0.2773	0.3044
GSM1051529	never	f	33	0.4136	0.1313
GSM1051530	current	f	59	0.6229	0.5017

Table 2: Table continues below

cg00213748	cg00214611	cg00455876	cg01707559	cg02004872	cg02011394
0.3725	0.6216	0.2908	0.2671	0.1791	0.4803
0.5006	0.4986	0.3746	0.1903	0.156	0.4181
0.3527	0.3738	0.2307	0.3147	0.1057	0.6151
0.4752	0.4863	0.2952	0.2958	0.1113	0.301
0.3675	0.7612	0.2358	0.2505	0.1691	0.393
0.2632	0.4157	0.4752	0.2539	0.2608	0.5098

Table 3: Table continues below

cg02050847	cg02233190	cg02494853	cg02839557	cg02842889	cg03052502
0.3276	0.2411	0.06707	0.247	0.4692	0.4002
0.3465	0.1755	0.04694	0.2367	0.3075	0.377
0.2375	0.2464	0.03824	0.2446	0.3578	0.305
0.3045	0.177	0.02672	0.001641	0.4457	0.2715
0.3062	0.3017	0.03702	0.3343	0.395	0.3266
0.4052	0.3853	0.02583	0.3092	0.3219	0.5334

cg03155755	cg03244189	cg03443143	cg03683899	cg03695421	cg03706273
0.415	0.2214	0.4758	0.2077	0.2092	0.13
0.3974	0.2171	0.5445	0.1844	0.1938	0.09853
0.5213	0.185	0.5371	0.3931	0.268	0.04025
0.4345	0.1654	0.5079	0.2812	0.2179	0.1015
0.4301	0.1811	0.4055	0.3108	0.2801	0.07786
0.5716	0.211	0.3778	0.4694	0.3433	0.04578

The first four rows are information about the patient and the remaining 20 are genetic information about the CpG sites of the patients. GSM is an identification number per patient, NCBI shows the full data that was collected for that patient. Smoking status tell us whether a patient is currently smoking or not. Gender and age are self-explanatory. The following 20 column are the CpG site values, these values show a percentage of methylation in these rows. Which can result in a repressed gene activity. We deleted the GMS column because is hold unique row identifiers. Fitting a tree with unique row identifiers in a dataset will split every single row in one node, which will give you a high predictive value. > This will cause overfitting.

Table 5: Summary with basic statistics about the data columns
(continued below)

GSM	Smoking.Status	Gender	Age
Length:683	Length:683	Length:683	Min. :18.00
Class :character	Class :character	Class :character	1st Qu.:47.00
Mode :character	Mode :character	Mode :character	Median :56.00
NA	NA	NA	Mean :53.82
NA	NA	NA	3rd Qu.:62.00
NA	NA	NA	Max. :80.00
NA	NA	NA	NA

Table 6: Table continues below

cg00050873	cg00212031	cg00213748	cg00214611
Min. :0.1186	Min. :0.00695	Min. :0.0000	Min. :0.01247
1st Qu.:0.4131	1st Qu.:0.06317	1st Qu.:0.3635	1st Qu.:0.06946
Median :0.5052	Median :0.36554	Median :0.4713	Median :0.41575
Mean :0.5600	Mean :0.30960	Mean :0.5191	Mean :0.34106
3rd Qu.:0.8144	3rd Qu.:0.45981	3rd Qu.:0.7278	3rd Qu.:0.49745
Max. :0.8989	Max. :0.70999	Max. :0.9236	Max. :0.80606
NA's :62	NA's :62	NA's :62	NA's :62

Table 7: Table continues below

cg00455876	cg01707559	cg02004872	cg02011394
Min. :0.05917	Min. :0.04333	Min. :0.00262	Min. :0.0000
1st Qu.:0.29300	1st Qu.:0.11080	1st Qu.:0.04284	1st Qu.:0.4261
Median :0.37968	Median :0.23873	Median :0.14933	Median :0.5157
Mean :0.44718	Mean :0.21435	Mean :0.15542	Mean :0.6058
3rd Qu.:0.66283	3rd Qu.:0.28061	3rd Qu.:0.24263	3rd Qu.:0.9412
Max. :0.85443	Max. :0.46999	Max. :0.47384	Max. :0.9792
NA's :62	NA's :62	NA's :62	NA's :62

Table 8: Table continues below

cg02050847	cg02233190	cg02494853	cg02839557
Min. :0.05234	Min. :0.00863	Min. :0.01162	Min. :0.00000
1st Qu.:0.33963	1st Qu.:0.08838	1st Qu.:0.02865	1st Qu.:0.06384
Median :0.42754	Median :0.25982	Median :0.03695	Median :0.35042
Mean :0.54369	Mean :0.23250	Mean :0.04077	Mean :0.30088
3rd Qu.:0.95558	3rd Qu.:0.33702	3rd Qu.:0.04677	3rd Qu.:0.45786
Max. :0.98320	Max. :0.51173	Max. :0.28947	Max. :0.82739
NA's :62	NA's :62	NA's :62	NA's :62

Table 9: Table continues below

cg02842889	cg03052502	cg03155755	cg03244189
Min. :0.01346	Min. :0.0000	Min. :0.2020	Min. :0.02972
1st Qu.:0.05483	1st Qu.:0.4025	1st Qu.:0.4245	1st Qu.:0.11976
Median :0.39757	Median :0.4940	Median :0.4962	Median :0.20397
Mean :0.32362	Mean :0.5907	Mean :0.5895	Mean :0.19552
3rd Qu.:0.47385	3rd Qu.:0.9631	3rd Qu.:0.8988	3rd Qu.:0.24921
Max. :0.85625	Max. :0.9902	Max. :0.9696	Max. :0.54074
NA's :62	NA's :62	NA's :62	NA's :62

cg03443143	cg03683899	cg03695421	cg03706273
Min. :0.06496	Min. :0.00788	Min. :0.0949	Min. :0.01120
1st Qu.:0.40963	1st Qu.:0.06159	1st Qu.:0.2566	1st Qu.:0.03413
Median :0.48314	Median :0.34422	Median :0.3208	Median :0.04961
Mean :0.56841	Mean :0.28442	Mean :0.3978	Mean :0.05769
3rd Qu.:0.85436	3rd Qu.:0.41866	3rd Qu.:0.5965	3rd Qu.:0.06916
Max. :0.93589	Max. :0.65925	Max. :0.8433	Max. :0.34380
NA's :62	NA's :62	NA's :62	NA's :62

When we look at the summary of the patient data we see in the column of the methylation data that there are 62 missing values. If we scroll true the data quickly we see that there are some rows who miss all CpG values, those are the ones that show up as 62 missing in every CpG values. these rows will be deleted because it gives us no information for the CpG sites.

Data exploratory analysis

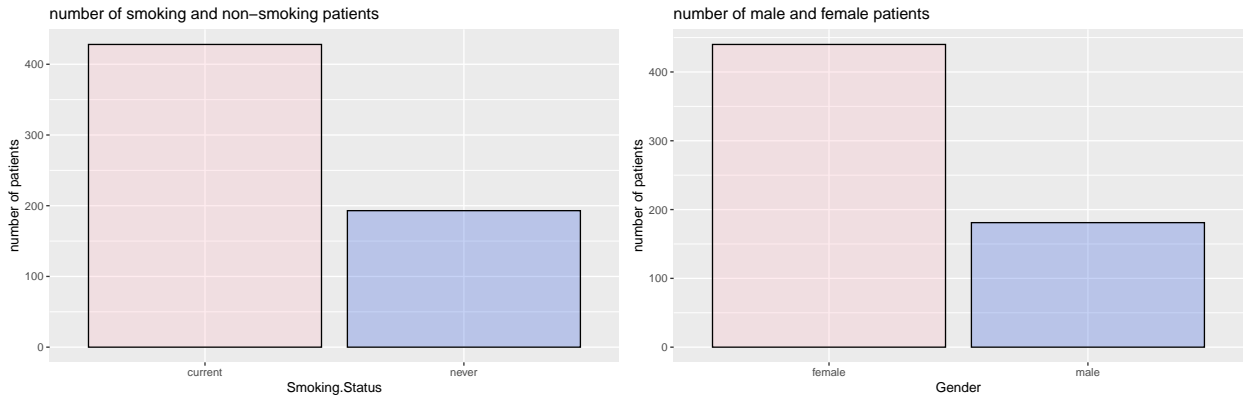


Figure 1: Comparison distribution smoking status and gender

In the figure above we see a distribution overview of the gender and smoking status of the patients. We see a high number of patients are female and the majority is smoking. We will not remove data to get an even distribution of the data, but we do need to keep this in mind when using data mining.

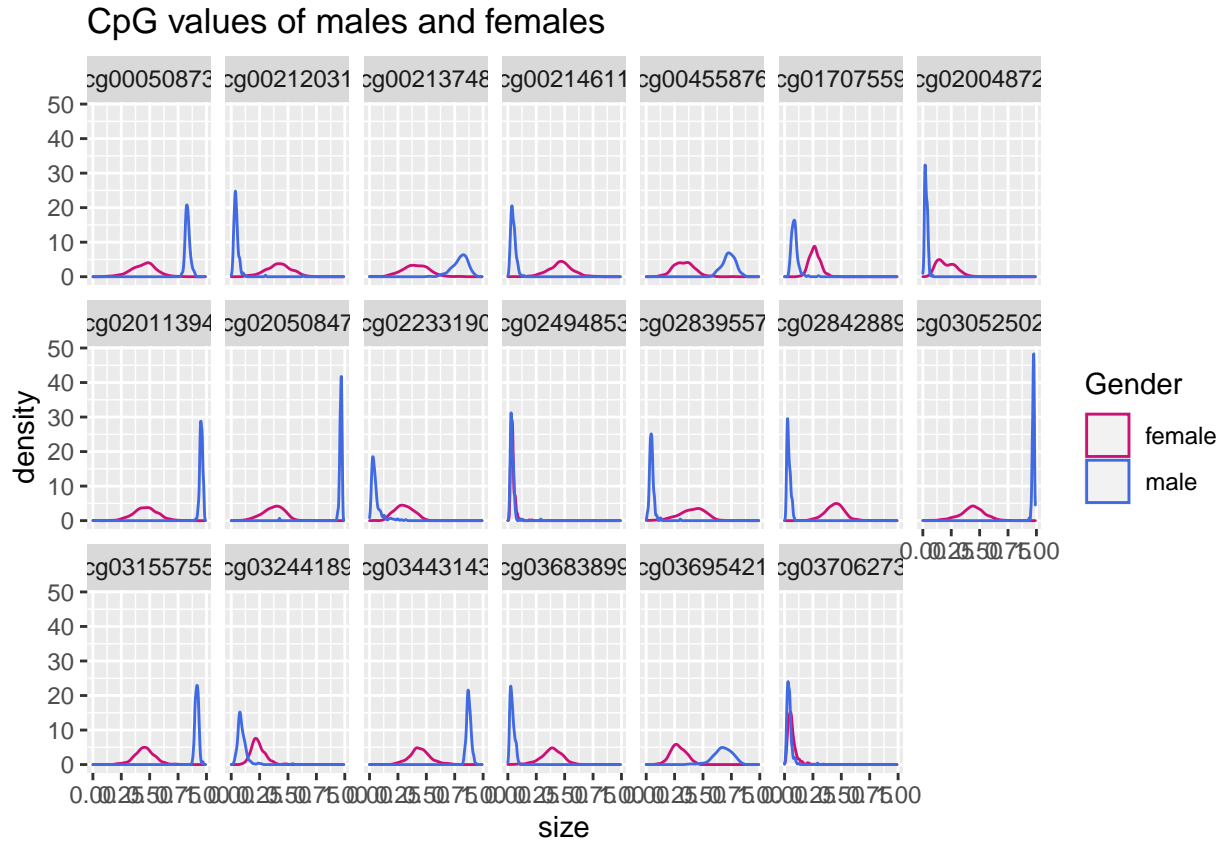


Figure 2: CpG values in percentages comparison of males and females

CpG values of smoking and non smoking patients

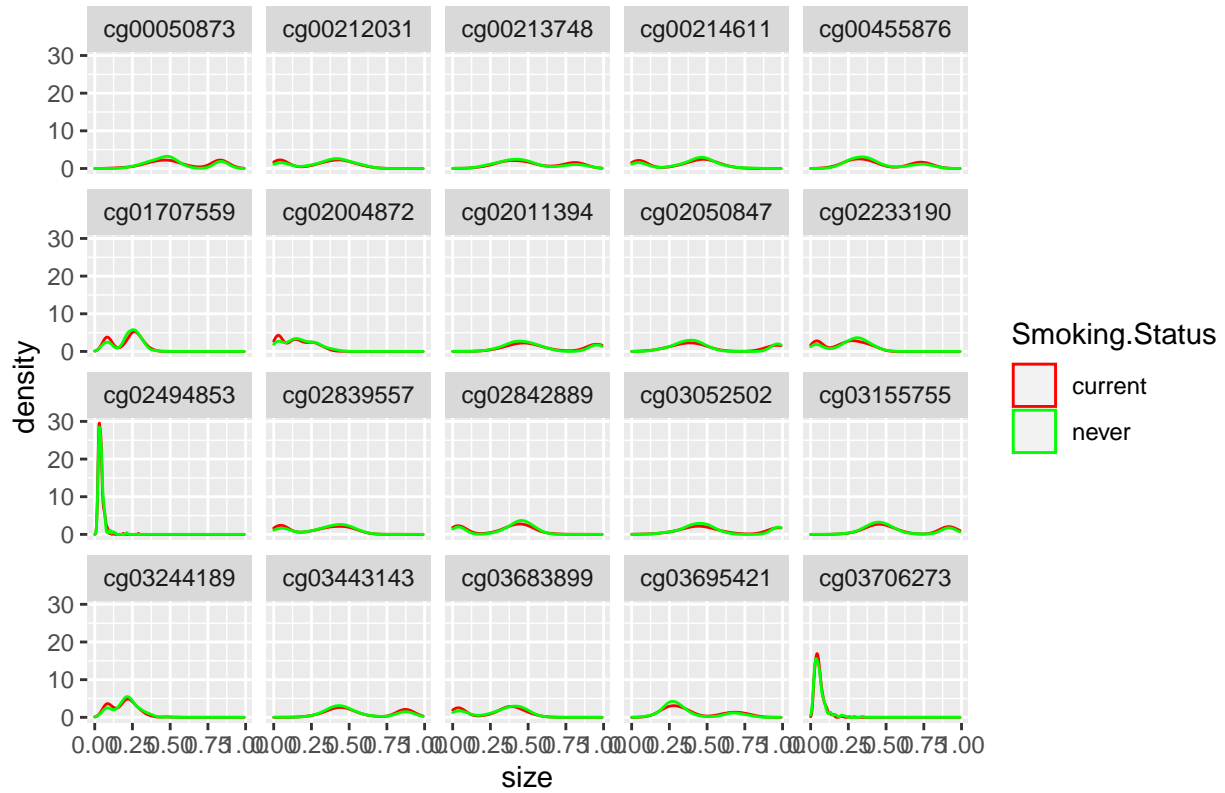
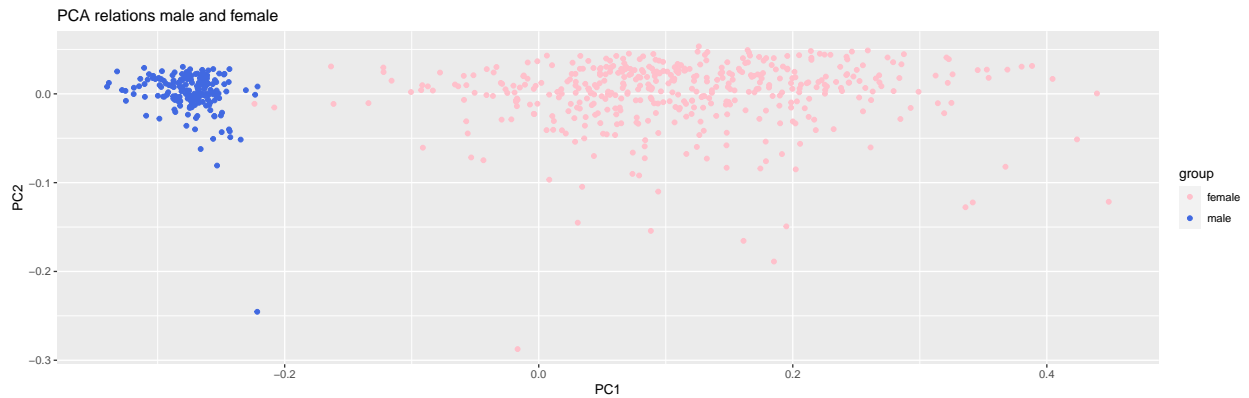
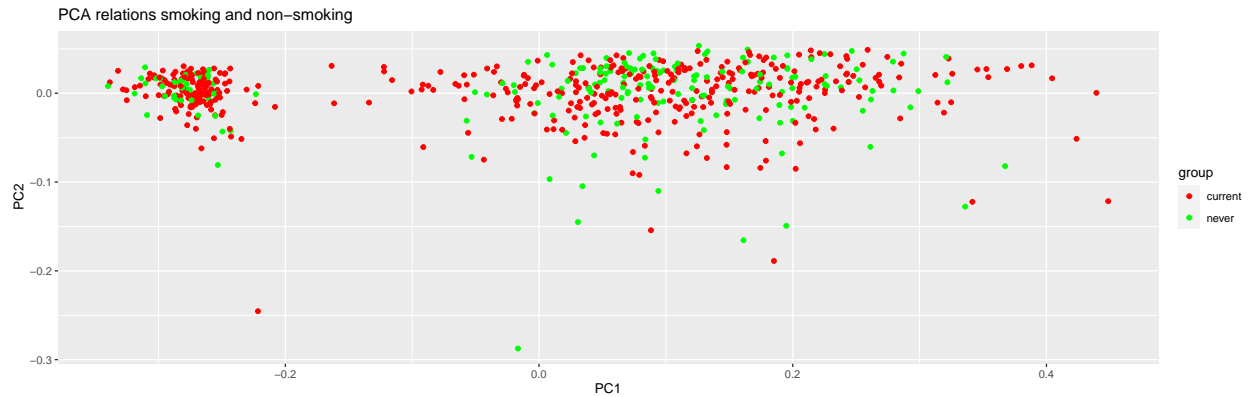


Figure 3: CpG values in percentages comparison of smoking and non smoking patients

As you can see in figure 2, the male and female CpG values are very different from each other. the green and red curve are on top of each other which means that there is not much difference in the CpG values for smoking and non smoking patients. This gives us already a lot of information for the research goal. The bimodal double curve in this graph stands for male and female, we observed that in figure 2 where we compared the males and female.





A principal component analysis gives an extra confirmation of the underlying groups that we saw in the CpG value differences. The PCA aims to show pattern in multivariate data. In figure 4 and 5 we see two groups that show a relation in the data. There is a clear difference in males and females. The smokers and non smokers are distributed in both these groups.

Discussion

Male and female methylation is shown to be different, we cant go deeper in detail in explaining these differences because we do not know the associated genes to the CpG locations in our dataset

Conclusion

The goal was to understand the given dataset and to clean the data. it is yet to discover if CpG methylation can show a relation with smoking, based on a prediction whether a patient is smoking or non-smoking using differences in CpG site values. The data shows good prospect for the use of machine learning because of the the patterns and correlation that is found in the data. We already saw that there is a difference in CpG values for gender. Smoking status had not yet shown a clear relation with the CpG values. This will be researched further using machine learning.

Referances

- GSE42861. (n.d.). Retrieved from ncbi: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE42861> Klug, W. S., Cummings, M. R., Spencer, C. A., & Palladino, M. A. (n.d.). concepts of genetics.
- thomaskonstantin. (n.d.). Retrieved from kaggle: (<https://www.kaggle.com/datasets/thomaskonstantin/cpg-values-of-smoking-and-non-smoking-patients>)
- Tymoczko, J. L. (2015). Biochemistery: A short course. macmilian publishers.