Data Wrangling: Crohn Disease methylome dataset

Maksym Vaskin

ETSIAAB, Technical University of Madrid, Madrid, Spain

E-mail: biomedmax@gmail.com

27-Dec-2018

1. Introduction

Methylation is one of epigenetic modifications of eukaryotic DNA. It happens mostly on cytosine and guanine rich regions of the genome (CpG islands). These regions are usually correlated with regulatory regions of the genome and methylated CpG are generally correlated with gene silencing. This methylation needs to be interpreted carefully as other epigenetic factors such as chromatin accessibility and histone marks present an interplay with methylation, so methylated CpG is not necessarily associated to inactive genes.

Complex organisms might have hundreds of cell types that are genetically identical despite sharing exactly the same gene content. This differentiation can occur thanks to differential gene activation/inactivation in different cell types that are in part regulated by methylation. Similarly, in some diseases, methylation pattern is altered leading to disease progression, resistance to particular treatments etc… Therefore, studying regulatory mechanisms of gene expression such as methylation might be informative to explain disease occurrence, progression and heterogeneity.

2. Dataset and Software

This dataset is taken from GEO repository (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE99788) and presents a data of bisulphite sequencing of ileal fibroblasts of patients with Crohn Disease (a bowel disease) and healthy individuals performed by Illumina Infinium EPIC Human Methylaton Beadchip. The samples of affected individuals can be further classified in non-inflamed, (7 samples), inflamed (2 samples) and stenotic (4 samples) Crohn disease. It also has a quality of methylation assessment for each CpG, representing the "confidence" of every specific CpG methylation call, which will not be used in this analysis. Software used was R version 3.4.4 and dplyr, stingr and tibble libraries.

3. Results

3.1 Sub-setting and data cleaning

Firstly, only a subset of CpG islands was selected, as the number of instances was too high. Only the first 10000 rows were selected. The resulted data set thus consisted of 10000 CpG islands for each of the 18 samples. Each sample had 3 columns: methylated signal, unmethylated signal and detection.pval (associated with the "confidence" of methylated call) – a total of 54 columns.

The columns indicating pvalues of methylation calls were removed and the resulting dataset had then 36 columns.

3.2 Data transformation

According to the literature (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5120494/), methylation at CpG island is quantified by beta values, given by $\beta=M/(M+U+a)$, being M methylate signal, U being unmethylated signal and "a" being the offset usually equal to 100. Essentially, this formula calculates the ratio of methylated to unmethylated signals and normalizes it, giving each CpG a value from

0 to 1, where 0 is fully unemthylated and 1 fully methylated.

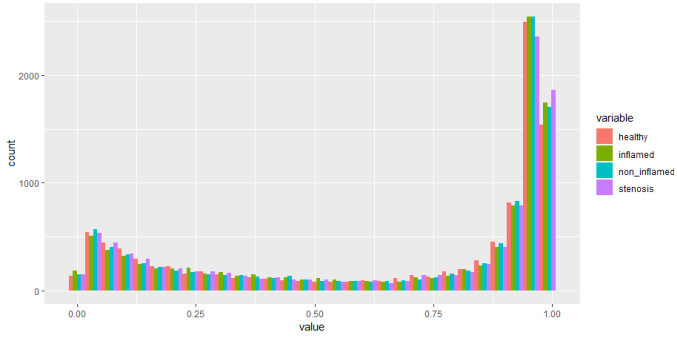
Beta values formula was applied to all the dataset, creating a new dataframe of 18 columns where each column was each sample's beta values for each CpG island (rows).

Next, 4 new subsets of data were created. For each new subset, columns of certain disease type were assigned (healthy, stenosis, inflammatory and non-inflammatory).

3.3. Data exploration

The first thing that was done is to calculate the mean and the standard deviation of each CpG island, thus grouping all patients with specific conditions with two values (two columns). Then, this data was joined in one dataset containing all the means of all CpG islands and their respective standard deviations.

Then, we plot an histogram to see the distribution of methylation levels. The result is interesting: CpG islands are mostly either very methylated or almost not methylated at all, very few show intermediate levels of methylation.



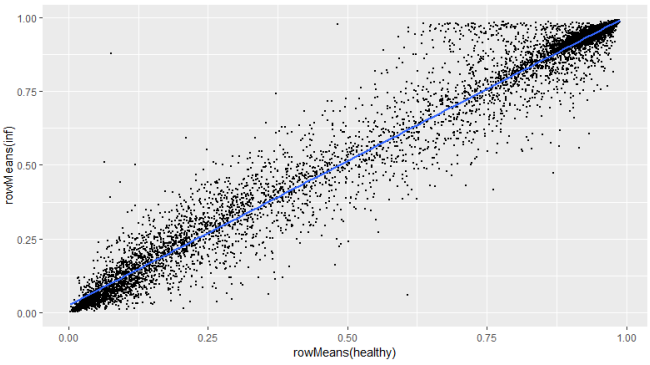
The next thing that can be done is a t.test on each CpG (each row), to compute a p-value of the difference between healthy and sick individuals. Because this function is computationally difficult even with a reduced dataset, it was reduced even more (100 CpGs) and only t.test between two conditions (healthy and inflammatory Crohn Disease was tested). The parameters were left on default and corresponded to a confidence level of 0.95%. After performing t-test, a new dataframe was created where rows were CpG islands and with their respective p-values.

After such a test, it would probably be interesting to retrieve CpG sites with significantly different methylation levels between the two conditions. Because we are testing many CpGs at the same time, it would probably be wise to apply a Bonferroni correction for multiple hypothesis testing (given by 1-0.05/n, where n is the number of hypotheses tested). We are working with 100 CpGs, so the new confidence level should be 0.9995%. Applying this correction and filtering the result, we see that we filtered out 13% of the CpGs. It is to note that we are applying a Bonferroni correction with n=100 as we are working with a small dataset, if we were to apply this to the whole dataset, a much greater percentage of CpG islands would be filtered out.

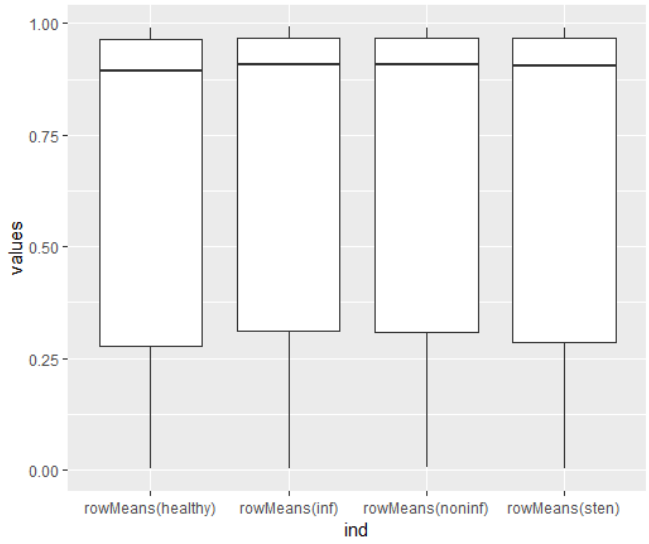
3.4. Plots with ggplot2

This is a very preliminary and incomplete analysis of this dataset, with no quality control or normalization of any kind, so few meaningful plots can be constructed and they will not be informative. Nevertheless, we can do it.

For example, we can draw a plot of mean methylation levels of each CpG for healthy and sick (inflammatory Crohn disease) and draw a correlation line (linear regression). The same could be done for other types of Crohn disease and compared, or even comparing all types of Crohn disease combined VS healthy. Nevertheless, here only healthy VS inflammatoy plot is presented. In this graph, we also see the distribution pattern of methylation profiles skewed towards either 1 or 0, as in the histogram before. Moreover, we see that there is a very high correlation of methylation levels between disease and healthy (except for some outliers seen in the bottom right and top left corners), which is not surprising as we don´t expect many CpG islands to have a very big difference in methylation between two conditions. The differences would most likely be subtle.



Perhaps, if we want to see the global levels of methylation, a more interesting approach would be to build boxplots. It these boxplots, we see that there seems to be no difference in global methylation levels between conditions. All of them have almost the same average and quantile distribution.



4. Discussion and Conclusions

This is a very simplistic and superficial analysis intended to be more of a data wrangling exercise than actual mathylome analysis.

There are no differences in global methylation deregulation between individuals with Crohn disease and healthy individuals which is not surprising, since there is no reason to think this disease should be directly related to methylation deficiency. Instead, a much reasonable approach would be to search for specific CpG that are differentially methylated between two states. Once these CpGs are found, their location on genome should be visualized and biological sense must be queried to establish the hypothesis of why/how it might affect the disease (is it located near a promoter of a gene somehow related with Crohns physiopathology or normal function of bowel?, etc…). Despite that the data cleaning was incomplete and no quality control was made (subtraction of background methylation levels, for example), t-test was performed for each CpG island. The next step would be to perform this test on whole dataset and with all 3 disease conditions against healthy, retrieve those CpGs with the lowest p-value and continue the analysis with them, including biological wet-lab assays.