Practical 2: The epigenetic clock: Variable ranking and multiple testing

Verena Zuber, Jelena Besevic, Alpha Forna, and Saredo Said

31/1/2019

Part 1: The epigenetic clock: Epigenetic marks associated with ageing

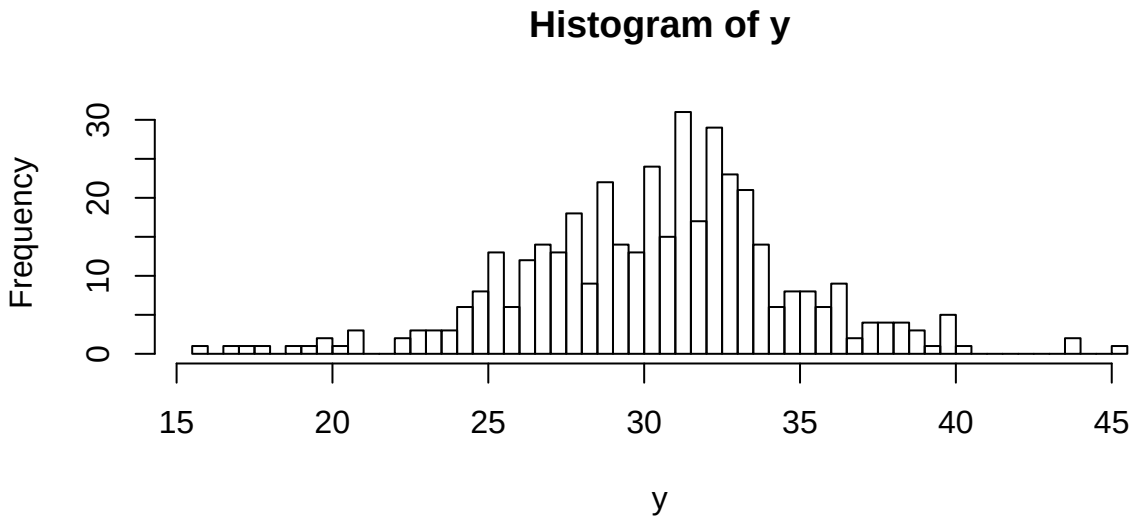
Our epigenome is highly impacted by environmental changes. One particular interesting aspect is ageing and how epigentic marks such as methylation are affected by ageing. Scientists have shown that there exist specific methylation sites that correlate with age, an observation that has been made in humans, chimpanzees, mice, or rats. Based on our knowledge which methylation sites correlate with healthy ageing we can use these epigenetic marks as biomarkers to predict the "actual biological age" of an individual. For example, if an individual has been exposed to pollutants or suffered from stress, its "actual biological age" of its body might be much older than its true age.

If you want to read more on the topic, there is a Nature Feature on the scientist Steve Horvath who first proposed to use methylation to measure the biological age

https://www.nature.com/news/biomarkers-and-ageing-the-clock-watcher-1.15014

In this practical we consider data on $n = 409$ healthy mice and methylation of $p = 3,663$ conserved methylation sites. Load the dataset, that contains the methylation matrix as predictor matrix and the age of the mice (in months) stored in the vector y. Familiarise yourself with the dataset using the following commands

```
load("data_epigenetic_clock_control")
#alternatively try load("data_epigenetic_clock_control.rds")
y = control_mice$y_control
hist(y,breaks=50)
```




```
x = control_mice$x_control
dim(x)
```

```
## [1]  409 3663
```

The first part is concerned with performing a ranking of methylation sites that have the strongest association with ageing.

Question 1.1

Compute a linear regression of the first methylation site against the age of the mice. Note in order to access the first column of a matrix, use square bracket like this [,1] and for the jth variable use [,j]. Figure out which element in the $coefficients matrix contains the p-value of the regression coefficient. Use again the squared brackets to index only the $p-$value.

Question 1.2

In order to compute the massively univariate linear regression estimate, we need to automate this computation for all $p = 3,663$ methylation sites. First initiate a vector where to save your $p-$values

```
pvec = rep(NaN, 3663)
```

Write a 'for loop' to iterate through all variables. In case you are not familiar with the 'for loop', use this practical here for help https://www.r-bloggers.com/how-to-write-the-first-for-loop-in-r/ . In each iteration j save the $p-$value of the respective regression into the pvec vector at position pvec[j].

Question 1.3

Rank the methylation sites according to their $p-$values and show the top 10 methylation sites that are associated with ageing.

Part 2: FDR control of the variable ranking

The next step is to correct the p-values for multiple testing. Since we performed $p = 3,663$ tests we need to make sure we do control the rate of false positive discoveries. There are different approaches to correct for multiple testing. Here, we perform the key methods and discuss the results in the end.

Question 2.1 Plot a histogram and discuss which distribution this vector of $p-$values has.

Question 2.2 Perform the Bonferroni correction on your $p-$value vector. Use $\alpha = 0.05$ as your significance threshold. How many methylation sites would be considered as Non-Null after Bonferroni correction?

Question 2.3 Perform the Benjamini-Hochberg Fdr correction on your $p-$value vector. Use $\alpha = 0.05$ as your significance threshold. How many methylation sites would be considered as Non-Null after Benjamini-Hochberg Fdr correction?

Question 2.4 Perform the q-value Fdr correction on your $p-$value vector. The qvalue function is provided in the qvalue package on Bioconductor. Make sure you have the latest version of R (>3.5) installed. Follow the instructions on the homepage for installation https://www.bioconductor.org/packages/release/bioc/html/qvalue.html and call the package using

```
library(qvalue)
```

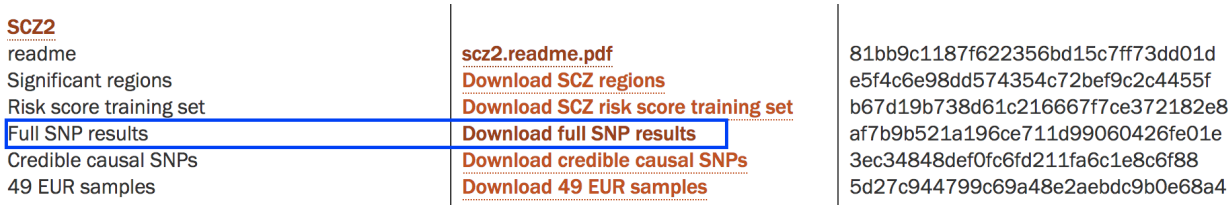
Use $\alpha = 0.05$ as your significance threshold. How many methylation sites would be considered as Non-Null after $q-$value Fdr correction?

Question 2.5 Perform the local fdr correction on your $p-$value vector using the fdrtool R package. Use the option statistic='pvalue' and extract the local fdr estimate in the $\$lfdr$ value. How many methylation sites would be considered as Non-Null after local fdr correction at a level of $\alpha = 0.2$?

Question 2.6 Look at the top graph of the mixture models that the fdrtool package has as ouput. Where is the Null distribution and where is the Non-Null distribution? Would you consider this to be a good model fit?

Question 2.7 The fdrtool estimates the proportion of Null variables. Access this information in the $param value. What is the proportion of Null variables?

Question 2.8 Compare the 4 ways of how to perform multiple testing correction and discuss which one is the most conservative.



SCZ2		
readme	scz2.readme.pdf	81bb9c1187f622356bd15c7ff73dd01d
Significant regions	Download SCZ regions	e5f4c6e98dd574354c72bef9c2c4455f
Risk score training set	Download SCZ risk score training set	b67d19b738d61c216667f7ce372182e8
Full SNP results	Download full SNP results	af7b9b521a196ce711d99060426fe01e
Credible causal SNPs	Download credible causal SNPs	3ec34848def0fc6fd211fa6c1e8c6f88
49 EUR samples	Download 49 EUR samples	5d27c944799c69a48e2aebdc9b0e68a4

Figure 1: Screenshot PGC download. In section SCZ2 download Full SNP results.

Part 3 (optional): Random p-values

In order to evaluate the Null distribution of $p-$values of a linear regression model we perform a simulation study.

Question 3.1 First simulate both x and y from the normal-distribution with mean 0 and variance 1, for each variable draw a 1000 random draws using the rnorm() function in R using

```
set.seed(1234)
x=rnorm(1000)
y=rnorm(1000)
```

Compute the linear regression of x on y and look at the $p-$value of the regression coefficient.

Question 3.2 Repeat this operation 10,000 times using a 'for loop'. Save your $p-$values of each iteration in an object pvec defined as

```
pvec = rep(NaN, 10000)
```

Question 3.3 Plot a histogram and discuss which distribution this vector of $p-$values has.

Part 4 (Optional): FDR correction on GWAS summary data on Schizophrenia

Go to the webpage of the Psychiatric Genetics Consortium (PGC) https://www.med.unc.edu/pgc/results-and-downloads and download in the section SCZ2 the full SNP results by clicking on the link 'Download full SNP results' (See Figure 1).

After downloading and un-zipping the file we can load it into R using and check the dimension of the dataset

```
data=read.csv("ckqny.scz2snpres", sep="\t")
dim(data)
```

```
## [1] 9444230      10
```

The file includes genome-wide data on over 9m SNPs. Looking at the first few lines of the data we find the column providing the $p-$values for the SNPS and save it as a vector pvec.

```
head(data)
```

```
##   hg19chrc       snpid a1 a2     bp  info      or     se      p ngt
## 1     chr1   rs4951859  C  G 729679 0.631 0.97853 0.0173 0.2083   0
## 2     chr1 rs142557973  T  C 731718 0.665 1.01949 0.0198 0.3298   0
## 3     chr1 rs141242758  T  C 734349 0.666 1.02071 0.0200 0.3055   0
## 4     chr1  rs79010578  A  T 736289 0.649 0.98748 0.0193 0.5132   0
## 5     chr1 rs143225517  T  C 751756 0.853 0.99681 0.0164 0.8431   0
## 6     chr1   rs3094315  A  G 752566 0.881 0.99601 0.0149 0.7870  36
```

```
pvec = data$p
N=length(pvec)
N
```

```
## [1] 9444230
```

Question 4.1 Plot a histogram of the $p-$value vector pvec. Discuss if this distribution can be modelled using a mixture distribution.

Question 4.2 Perform Bonferroni correction of the $p-$value vector pvec using the function p.adjust. How many SNPs are significant at a Bonferroni adjusted level of 0.05?

Question 4.3 Perform the Benjamini Hochberg FDR correction of the p-value vector pvec using the function p.adjust. How many SNPs are significant at a Benjamini Hochberg FDR adjusted level of 0.05?

Question 4.3 Perform the q-value FDR correction of the p-value vector pvec using the function qvalue in the qvalue package on bioconductor. Follow the instructions on the homepage for installation https://www.bioconductor.org/packages/release/bioc/html/qvalue.html and call the package using

```
library(qvalue)
```

How many SNPs are significant at a $q-$value level of 0.05? What is the proportion of Null SNPs?

Question 4.4 Perform the local fdr correction on your $p-$value vector using the fdrtool R package. Use the option statistic='pvalue' and extract the local fdr estimate in the $\$lfdr$ value. How many SNPs would be considered as Non-Null after local fdr correction at a level of $\alpha = 0.2$?

Question 4.5 Look at the graph of the mixture models that the fdrtool package has as ouput. Would you consider this to be a good model fit? Further, what is the proportion of Null variables?