Vignette Title

Aleksandra DĂ„â€¦browska, Alicja Gosiewska

2017-05-13

Contents

Package (Abstract?)

It is considered that the result of increased methylation is decreased gene expression. While, recent studies suggest that the relationship between methylation and expression is more complex than was previously thought. The package . . . provides methods to test for differential expression and methylation by use of the negative binonial distribution and t-test. Additionaly package . . . allows to visualize results in a simple way.

Standard Workflow

In this vignette we will work with the data sets containing information about gene expression and methylation for patients with breast cancer. We will analyze differences in methylation and expression for patients with different subtypes of BRCA cancer.

Function test_diff

The main function of the package is `test_diff`. It allows to find differences between genes methylation or expression, taking into account additional information about samples.

Methylation

Methylation is a process by which methyl groups are added to the DNA molecule. It can change the activity of a DNA without changing the sequence. DNA methylation typically acts to repress gene transcription. But there exist situations in which adding the methyl groups intensify it. DNA methylation is associated with a lots of key processes including genomic imprinting, repression of transposable elements, aging and carcinogenesis. In our work we want to bind methylation process and carcinogenesis.

BRCA_methylation_chr17 data set

In this section, we will work with the methylation level data from TCGA database. Package contains `BRCA_methylation_chr17` dataset. `BRCA_methylation_chr17` contains information about methylation of CpG islands located on 17th chromosome for patienst with breast cancer. Rows of this data set correspond to patients, more precisely, to samples taken from patients. First column `SUBTYPE`corresponds to a subtype of BRCA cancer, next columns correspond to CpG islands. Values inside the table indicate the methylation level of CpG island for specified sample.

```
library(MetExpR)
```

```
##
```

```
head(BRCA_methylation_chr17)[1:5,1:4]
```

```
##                              SUBTYPE cg00021527 cg00031162  cg00032227
## TCGA-A1-A0SD-01A-11D-A112-05    LumA 0.03781858  0.7910348 0.006391233
## TCGA-A2-A04N-01A-11D-A112-05    LumA 0.01437552  0.7359370 0.008752293
## TCGA-A2-A04P-01A-31D-A032-05   Basal 0.01360124  0.6967802 0.009442039
## TCGA-A2-A04Q-01A-21D-A032-05   Basal 0.01525656  0.5341244 0.014674247
## TCGA-A2-A04T-01A-21D-A032-05   Basal 0.01167384  0.7378100 0.012251559
```

Data preparation

In this analysis we would like to find genes with different methylation level. At first we need to use function `map_to_gene`, which generates new data frame with CpG islands mapped to genes.

```
BRCA_methylation_chr17_gen <- map_to_gene(BRCA_methylation_chr17[,-1])
head(BRCA_methylation_chr17_gen)[1:5,1:4]
```

```
##                                  AANAT    AARSD1       AATF      AATK
## TCGA-A1-A0SD-01A-11D-A112-05 0.7148533 0.8625816 0.24294092 0.7835302
## TCGA-A2-A04N-01A-11D-A112-05 0.5850106 0.8355825 0.21367129 0.8466190
## TCGA-A2-A04P-01A-31D-A032-05 0.4495537 0.8786166 0.03277413 0.3417919
## TCGA-A2-A04Q-01A-21D-A032-05 0.7120650 0.8819490 0.03460160 0.7264985
## TCGA-A2-A04T-01A-21D-A032-05 0.6010397 0.7739978 0.02501599 0.6276399
```

Function `test_diff` allows us to test for differences between the base means for two or more conditions.

In this case we have two conditions, connected with subtypes of breast cancer.

Before we go to the testing, we need to define condition values for each sample. We would like to test for differences between `LumA` subtype and `other` subtypes of breast cancer, so we create vector, which each element corresponds to a sample. Our division into this two group relies on numbers of occurences of each subtype. The `LumA` subtype is the most common, in case of breast cancer.

```
condition <- ifelse(BRCA_methylation_chr17$SUBTYPE=="LumA","LumA", "other")
head(condition,8)
```

```
## [1] "LumA"  "LumA"  "other" "other" "other" "other" "LumA"  "other"
```

T-test

One of the most used tools for testing differences between values is t-test. The null hypothesis we have consider, is that means in two groups are equal. To use it in `test_diff` function, we set value of parameter `test` on “ttest”.

```
test.mety <- test_diff(BRCA_methylation_chr17_gen, condition, test="ttest")
```

As a result we obtain a data frame with columns corresponds to: id of gene, mean, logarithm of fold change, p-value for t-test, adjusted p-value (BH method). For more information about customizing this function see the help page for `test_diff`.

```
head(test.mety)
```

```
##              id      mean   log2.fold         pval         padj
## ICAM2     ICAM2 0.3330801 -0.15151320 3.754116e-17 3.063359e-14
## RILP       RILP 0.3341447 -0.05073691 2.575168e-13 1.050668e-10
## PIPOX     PIPOX 0.3647812  0.11505558 5.360053e-12 1.196885e-09
## TNFSF12 TNFSF12 0.2485025 -0.13412855 5.867083e-12 1.196885e-09
## CD7         CD7 0.8127112  0.09822690 1.641919e-11 2.679612e-09
## KSR1       KSR1 0.5549808  0.19973400 2.054467e-11 2.794075e-09
```

Expression

Gene expression is the process by which information from a gene is used in the synthesis of proteins. The process of gene expression is used by all known life.

`BRCA_mRNAseq_chr17` data set

In this section we will use data set `BRCA_mRNAseq_chr17`, which contains information about gene expression. This data set contains per-gene read counts computed for genes from 17th chromosome for 100 patients with breast cancer. Rows of this data set correspond to samples taken from patients. First column `SUBTYPE`corresponds to a subtype of BRCA cancer, next columns correspond to genes.

```
BRCA_mRNAseq_chr17[1:5,1:5]
```

```
##                              SUBTYPE AANAT AARSD1 AATF AATK
## TCGA-A1-A0SB-01A-11R-A144-07  Normal     9   2354 2870  317
## TCGA-A1-A0SD-01A-11R-A115-07    LumA     2   1846 5656  312
## TCGA-A1-A0SE-01A-11R-A084-07    LumA    11   3391 9522  736
## TCGA-A1-A0SF-01A-11R-A144-07    LumA     0   2169 4625  169
## TCGA-A1-A0SG-01A-11R-A144-07    LumA     1   2273 3473   92
```

Negative binomial test

Negative binomial test, which uses negative binomial distribution is an another tool for finding differencial expression between our conditions.

As in the t-test we also need a description of the samples, which we keep in a vector, whose elements correspond to different gorups.

In our example we will test for differential expression between groups with LumA breast cancer subtype and other subtypes of that cancer. Again we will use vector `conditions`, which consist of two values corresponds to subtype of breast cancer: LumA and other.

```
condition<-ifelse(BRCA_mRNAseq_chr17$SUBTYPE=="LumA","LumA","other")
head(condition,8)
```

```
## [1] "other" "LumA"  "LumA"  "LumA"  "LumA"  "LumA"  "other" "LumA"
```

For using negative binomial test, in function `test_diff` we set value “nbinom2” for parameter `test`. (negative binomial test from DESeq2 package)

```
test.expr <- test_diff(BRCA_mRNAseq_chr17[,-1], condition, test="nbinom2")
```

```
## converting counts to integer mode

## estimating size factors

## estimating dispersions

## gene-wise dispersion estimates

## mean-dispersion relationship

## final dispersion estimates

## fitting model and testing

## -- replacing outliers and refitting for 71 genes
## -- DESeq argument 'minReplicatesForReplace' = 7
## -- original counts are preserved in counts(dds)

## estimating dispersions

## fitting model and testing
```

As a result we obtain the following data frame:

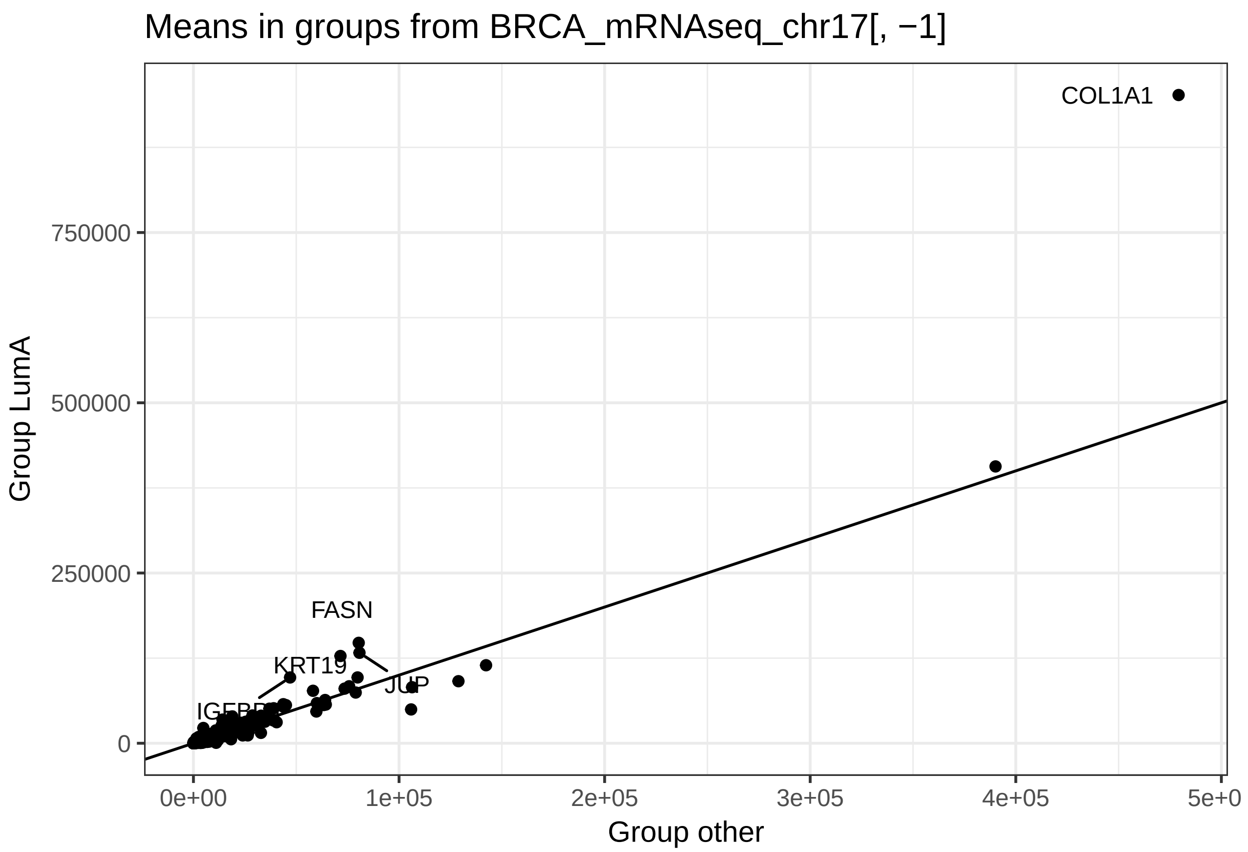
```
head(test.expr)
```

```
##            id        mean   log2.fold       pval      padj
## AANAT   AANAT    3.455436  0.36342624 0.22109578 0.3202379
## AARSD1 AARSD1 2779.448414 -0.07435954 0.45027777 0.5590183
## AATF     AATF 6750.269650  0.13308688 0.18597163 0.2781953
## AATK     AATK  352.805108  0.02308606 0.92313513 0.9531664
## ABCA5   ABCA5 1933.257431 -0.07034249 0.73314806 0.8079895
## ABCA6   ABCA6  689.547294 -0.44685708 0.08266088 0.1429916
```

Visualization

em_plot

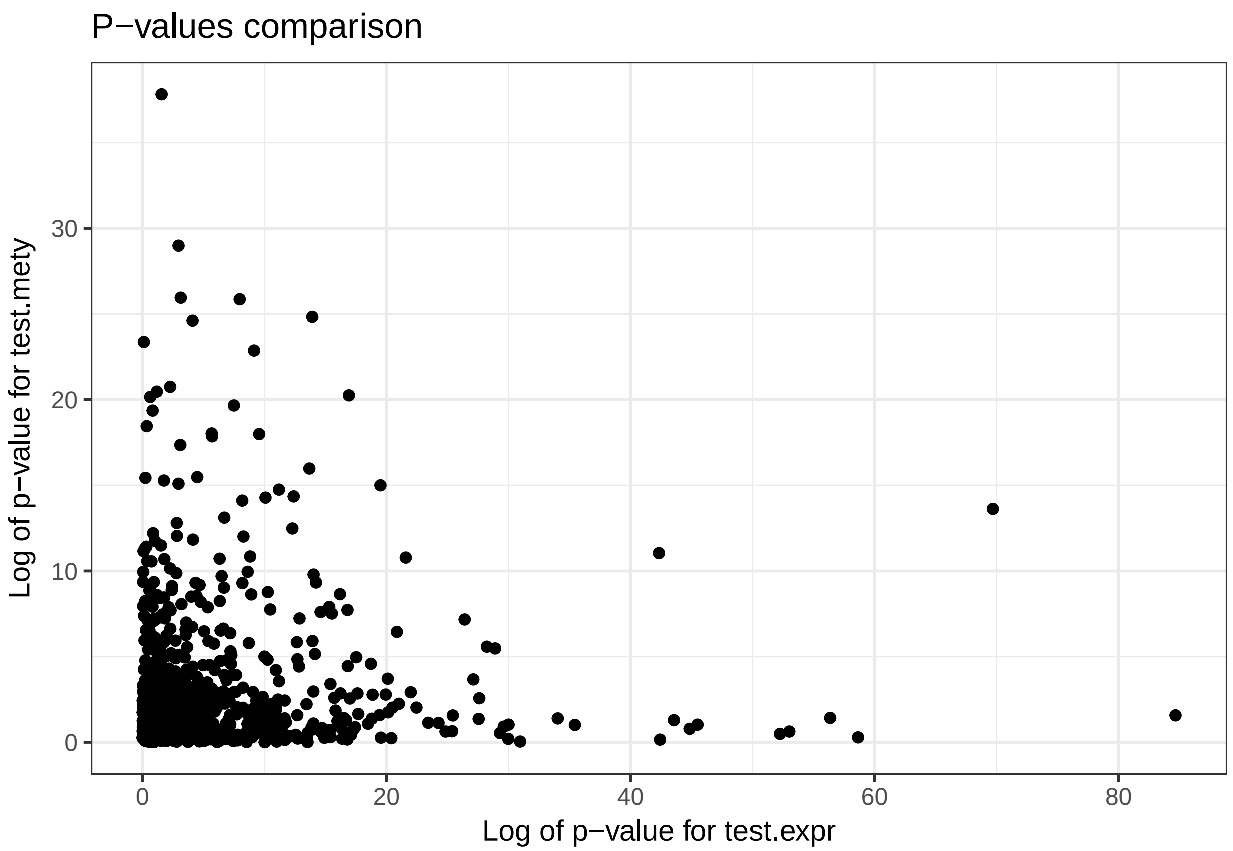
```
em_plot(BRCA_mRNAseq_chr17[,-1], condition, names=5)
```



log-log p-value

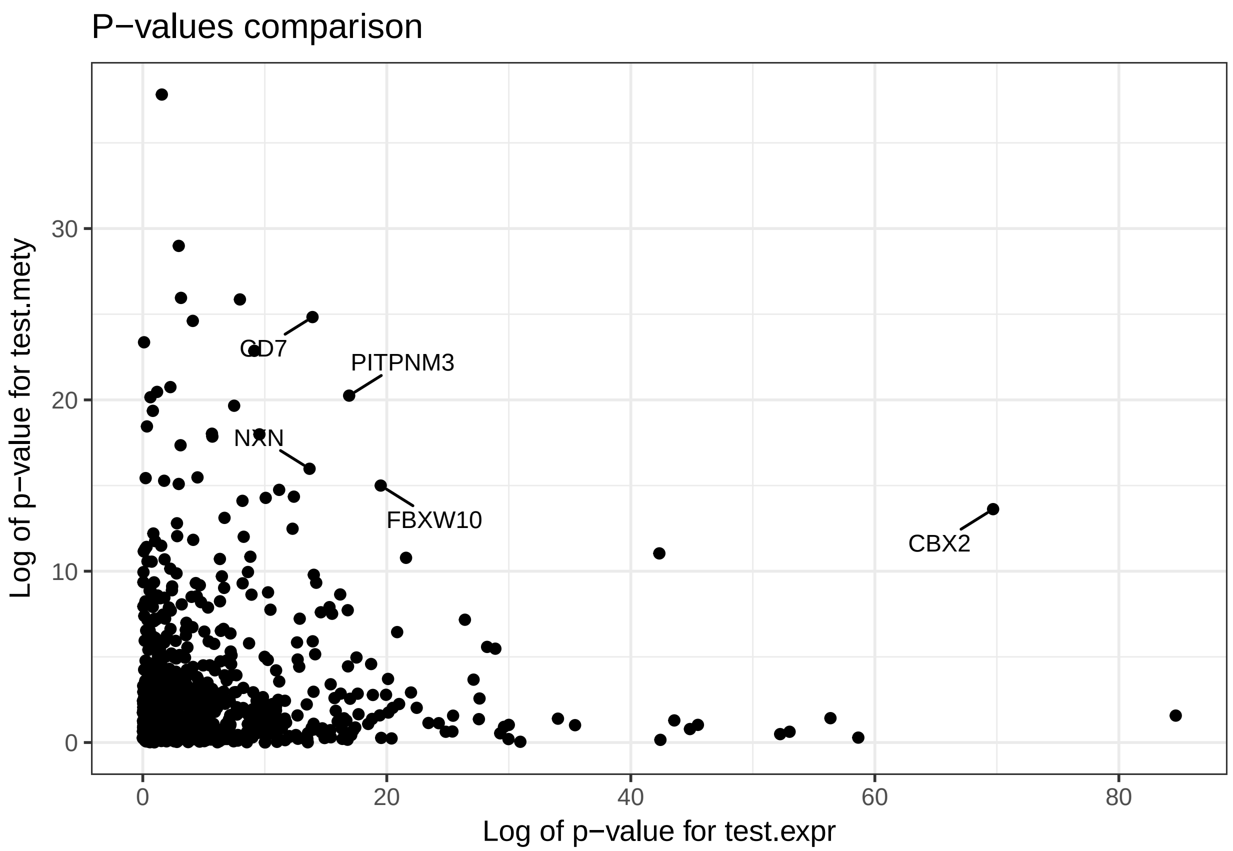
Firstly, we want to visualise the p-values for expression and methylation from negative binomial test and t-test respectively.

```
p_values_plot(test.expr, test.mety)
```



Additionally, `names` parameter allows to mark genes with sum of p-values for methylation and expression, lower than given value. Value of parameter `names` defines, number of genes to label.

```
p_values_plot(test.expr, test.mety, names = 5)
```

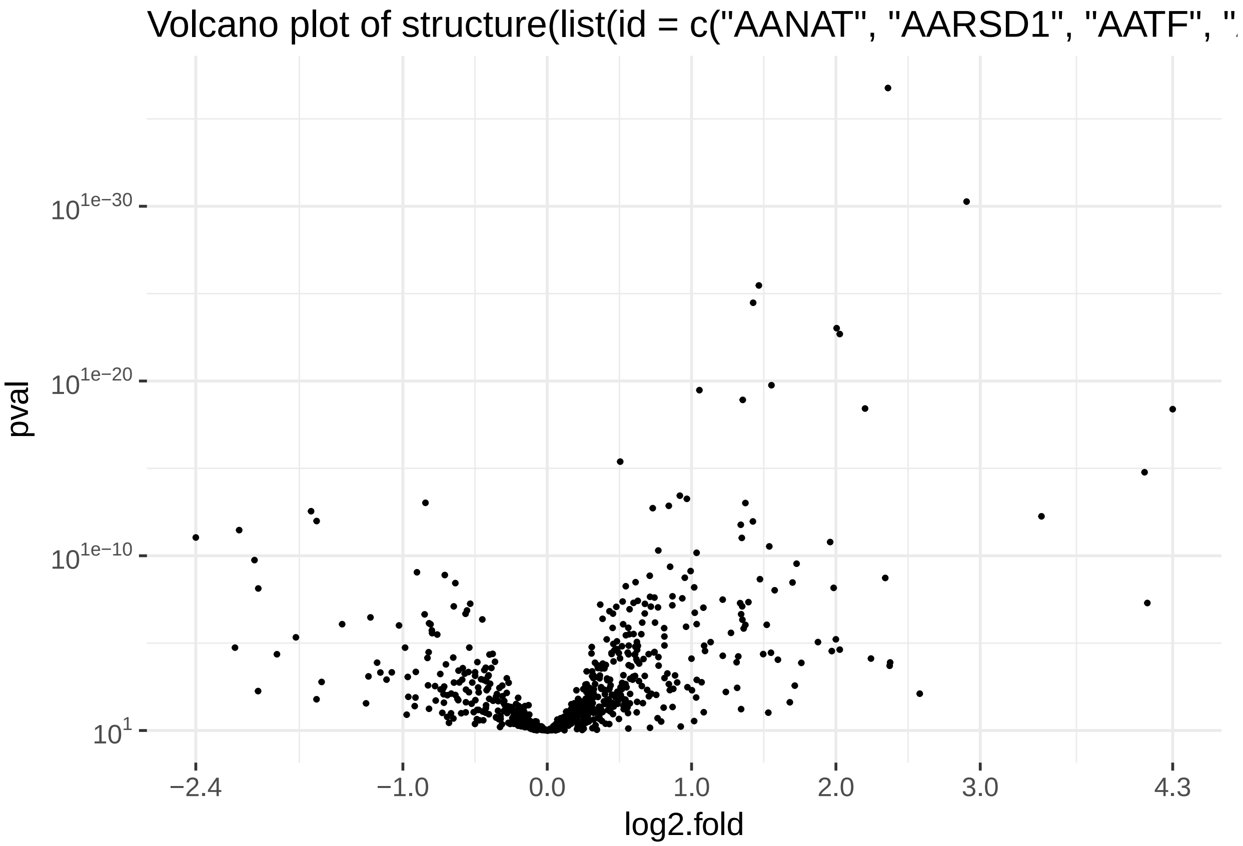


To read more about `p_values_plot` (e.g other ways to labeling genes) see help page for that function.

Volcano plot

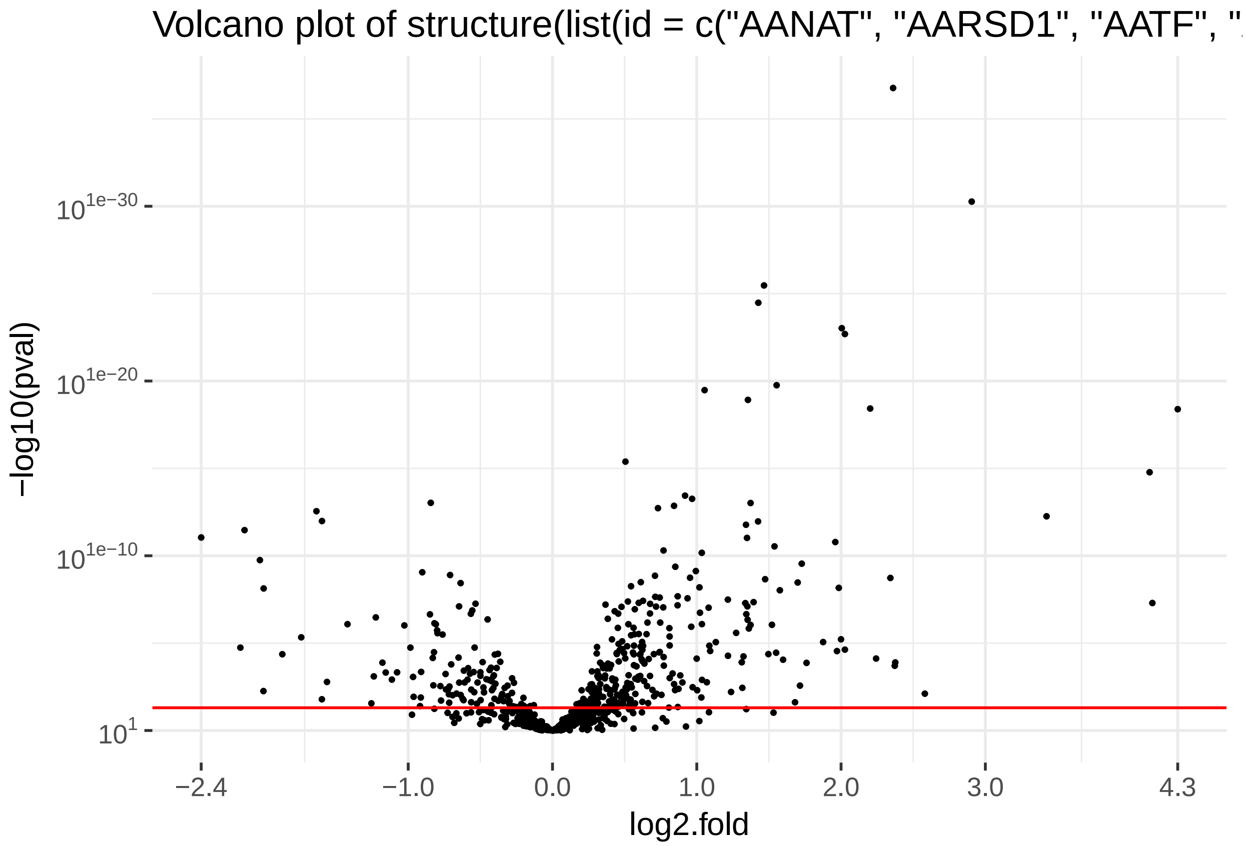
For identify changes in our data sets we use a volcano plot - some type of scatter-plot. It plots logarithm of p-value versus logarithm of fold-change on the y and x axes, respectively.

```
volcano_plot(test.expr)
```

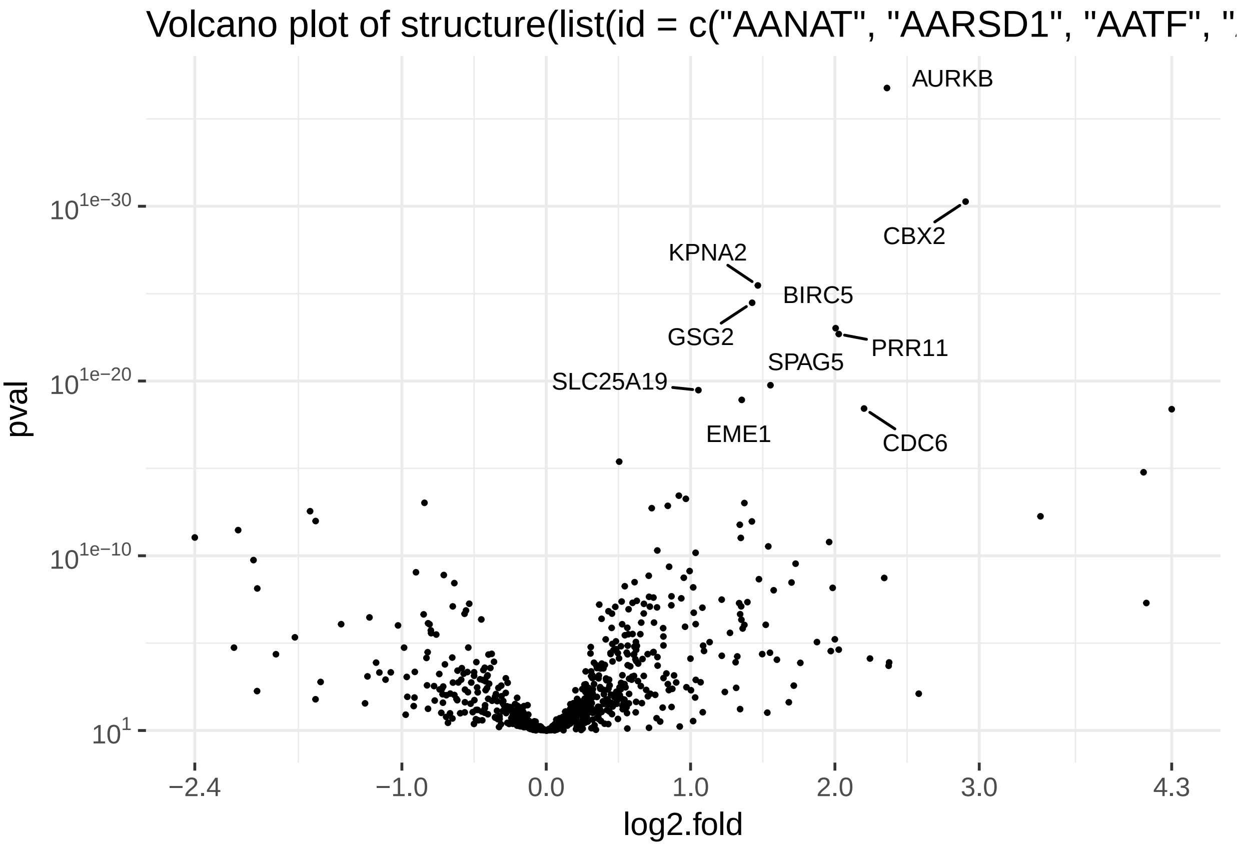


Function `volcano_plot` has parameters that allow to better analyze the results: `line` and `names`. The `line` parameter allows to set the horizontal line on plot on selected value. The `names` parameter signs choosen number of genes with the lowest p-value.

```
volcano_plot(test.expr, line = 0.05)
```



```
volcano_plot(test.expr, names = 10)
```



Methylation and expression for one gene.

In the end we want to present the distribution of methylation and expression for choosen genes.

Function `CpG_mean` computes methylation means of CpG islands for choosen gene. In this case: “BRCA1”

```
BRCA1_gene <- CpG_mean(BRCA_methylation_chr17, "BRCA1")
BRCA1_gene
```

```
##             Name  MapInfo Symbol CPG_ISLAND CPG_ISLAND_LOCATIONS
## 4712  cg04658354 38530970  BRCA1       TRUE 17:38530194-38531162
## 6961  cg06973652 38532148  BRCA1       TRUE 17:38531525-38532730
## 7047  cg07054526 38526034  BRCA1       TRUE 17:38525979-38526990
## 8929  cg08993267 38530848  BRCA1       TRUE 17:38530194-38531162
## 11917 cg11964474 38532181  BRCA1       TRUE 17:38531525-38532730
## 19075 cg19088651 38530739  BRCA1       TRUE 17:38530194-38531162
## 19527 cg19531713 38530585  BRCA1       TRUE 17:38530194-38531162
##             mean
## 4712  0.05802403
## 6961  0.83772148
## 7047  0.93138624
## 8929  0.06280625
## 11917 0.79200965
## 19075 0.11090400
## 19527 0.10361921
```

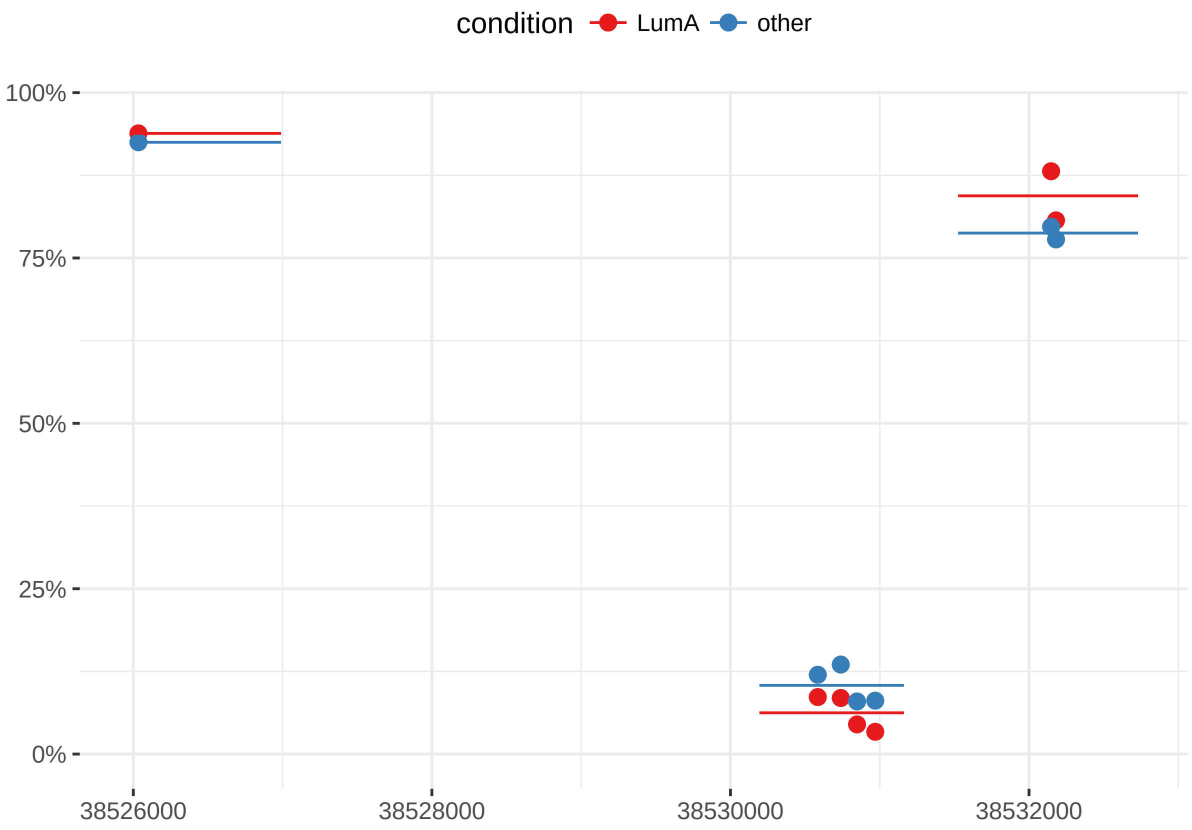
Methylation and expression in groups.

Two subtype groups of cancer in one plot.

```
condition <- ifelse(BRCA_methylation_chr17$SUBTYPE=="LumA","LumA", "other")
genereg_vs_met(BRCA_methylation_chr17, condition, "BRCA1")
```

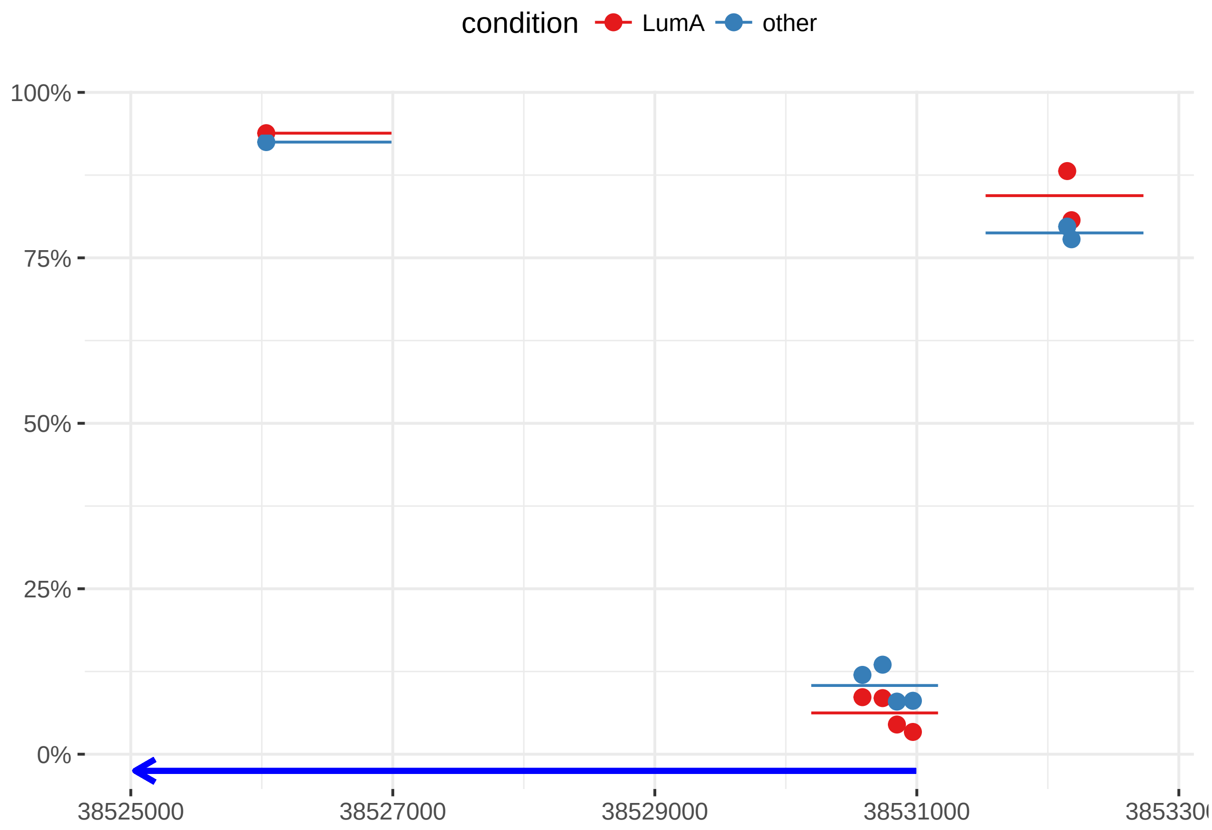
```
## 'select()' returned 1:1 mapping between keys and columns
```

```
## 'select()' returned 1:many mapping between keys and columns
```

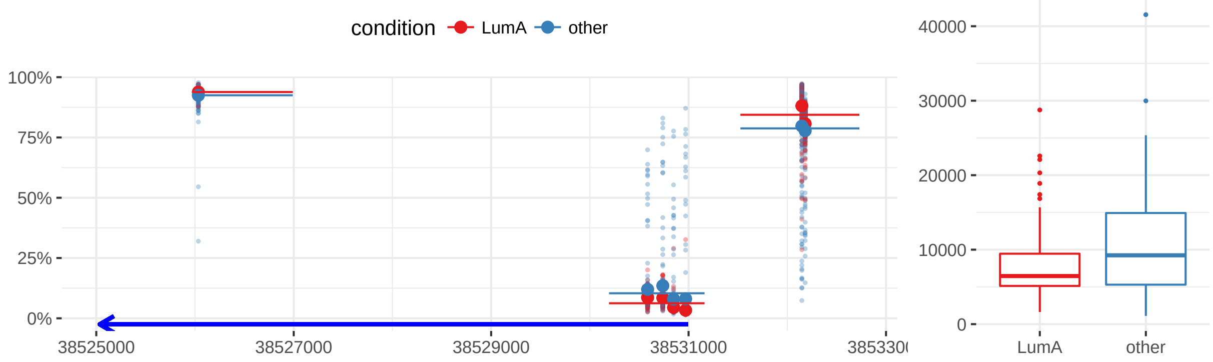


```
genereg_vs_met(BRCA_methylation_chr17, condition, "BRCA1", show_gen=TRUE)
```

```
## 'select()' returned 1:1 mapping between keys and columns
## 'select()' returned 1:many mapping between keys and columns
```

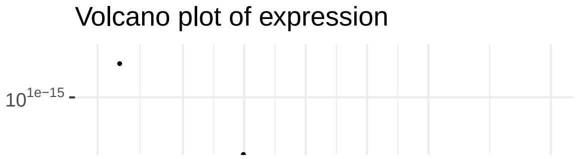


```
visual_gene(condition.e, condition.m, BRCA_methylation_chr17[,-1],BRCA_mRNAseq_chr17[,-1], "BRCA1", tes
```

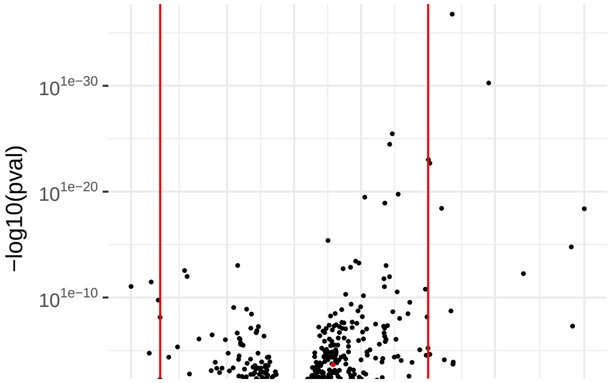


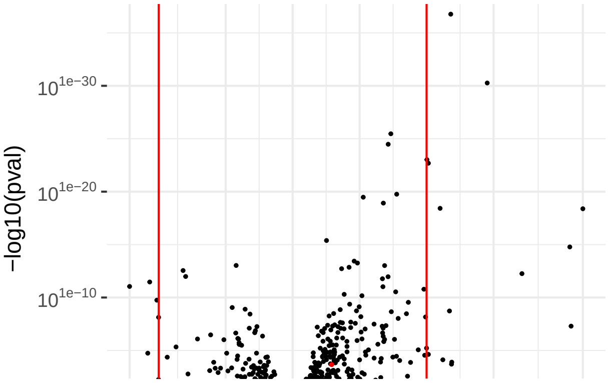
```
visual_volc(condition.e, condition.m, BRCA_methylation_chr17[,-1],BRCA_mRNAseq_chr17[,-1], "BRCA1", lis
```

Volcano plot of expression
10^1e−30



```
visual_volc(condition.e, condition.m, BRCA_methylation_chr17[,-1],BRCA_mRNAseq_chr17[,-1], "BRCA1", lis
```


10^1e−30
10^1e−20
10^1e−10
−log10(pval)


10^1e−30
10^1e−20
10^1e−10
−log10(pval)