Usage - survival status case

Aleksandra Dąbrowska, Alicja Gosiewska

2017-05-24

Contents

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_all_surv` data set

In this section, we will work with the methylation level data from TCGA database. Package contains `BRCA_methylation_all_surv` dataset. `BRCA_methylation_all_surv` contains information about methylation of CpG islands 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 column to a survival status, more precisely: `1` corresponds to `Dead`, `0` to `Alive`. We divided this column

in the following way: -patients who have observation time longer than 3 years and any vital status we assign to `0` group -patients who have observation time shorter than 3 years and `Dead` in vital status we assign to `1` group -we disregarded patients not belonging to previous groups.

Other columns correspond to CpG islands. Values inside the table indicate the methylation level of CpG island for specified sample.

```
library(MetExpR)
```

```
##
```

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

```
##                                 sampleID survival_status cg00000292
## TCGA-A2-A04N-01A-11R-A115-07 TCGA-A2-A04N               0  0.7433957
## TCGA-A2-A04P-01A-31R-A034-07 TCGA-A2-A04P               1  0.2897206
## TCGA-A2-A04Q-01A-21R-A034-07 TCGA-A2-A04Q               0  0.7898920
## TCGA-A2-A04T-01A-21R-A034-07 TCGA-A2-A04T               0  0.6512270
##                                cg00002426 cg00003994
## TCGA-A2-A04N-01A-11R-A115-07 0.07044132 0.32317983
## TCGA-A2-A04P-01A-31R-A034-07 0.25927969 0.02402149
## TCGA-A2-A04Q-01A-21R-A034-07 0.63619354 0.10885097
## TCGA-A2-A04T-01A-21R-A034-07 0.27268734 0.03413620
```

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_gen <- map_to_gene(BRCA_methylation_all_surv[,-c(1,2)])
head(BRCA_methylation_gen[,-1])[1:5,1:4]
```

```
##                                 AARSD1       AATF      AATK       ABC1
## TCGA-A2-A04N-01A-11R-A115-07 0.8355825 0.21367129 0.8466190 0.01275865
## TCGA-A2-A04P-01A-31R-A034-07 0.8786166 0.03277413 0.3417919 0.01455092
## TCGA-A2-A04Q-01A-21R-A034-07 0.8819490 0.03460160 0.7264985 0.01283647
## TCGA-A2-A04T-01A-21R-A034-07 0.7739978 0.02501599 0.6276399 0.01278928
## TCGA-A2-A04V-01A-21R-A034-07 0.8006830 0.03056880 0.7214240 0.01753320
```

```
#cos nie tak w mappowaniu(wyrzuca pierwszą kolumne) - poprawic
```

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 survival status.

```
condition <- ifelse(BRCA_methylation_all_surv$survival_status==1, "Dead","Alive")
#zera i jedynki nie sa dobrym pomyslem-
#dostajemy error przy wywolaniu makeContrasts
#Error in makeContrasts(contrasts = forms, levels = design) :
#  The levels must by syntactically valid names in R, see help(make.names).  Non-valid names: 0,1
```

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_gen[,-c(1,2)], 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
## CACNG4     CACNG4 0.18716309 -0.09875430 0.0004089692 0.1978177
## ZNF287     ZNF287 0.05124223 -0.04622191 0.0004860387 0.1978177
## PHOSPHO1 PHOSPHO1 0.04674227 -0.06235858 0.0007391206 0.2005481
## SCARF1     SCARF1 0.88038042  0.02309974 0.0031414465 0.5469656
## ATP1B2     ATP1B2 0.02798455 -0.03252621 0.0033597395 0.5469656
## PPP1R1B   PPP1R1B 0.20946946 -0.10474034 0.0063883291 0.7954215
```

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_all_surv` data set

In this section we will use data set `BRCA_mRNAseq_all_surv`, which contains information about gene expression. Rows of this data set correspond to samples taken from patients. First column `SUBTYPE`corresponds to a subtype of BRCA cancer, next column, like in `BRCA_methylation_all_surv` to the survival status, next columns correspond to genes.

```
BRCA_mRNAseq_all_surv[1:5,1:5]
```

```
##                                  sampleID survival_status A1BG A1CF A2BP1
## TCGA-A1-A0SE-01A-11R-A084-07 TCGA-A1-A0SE               0 1341    0     2
## TCGA-A1-A0SF-01A-11R-A144-07 TCGA-A1-A0SF               0  836    1     0
## TCGA-A1-A0SH-01A-11R-A084-07 TCGA-A1-A0SH               0 1126    1     4
## TCGA-A1-A0SK-01A-12R-A084-07 TCGA-A1-A0SK               1  626    1     1
## TCGA-A1-A0SN-01A-11R-A144-07 TCGA-A1-A0SN               0  244    0     1
```

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_all_surv$survival_status==1,"Dead","Alive")
head(condition,8)
```

```
## [1] "Alive" "Alive" "Alive" "Dead"  "Alive" "Alive" "Dead"  "Alive"
```

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_all_surv[,-c(1,2)], 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 117 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:

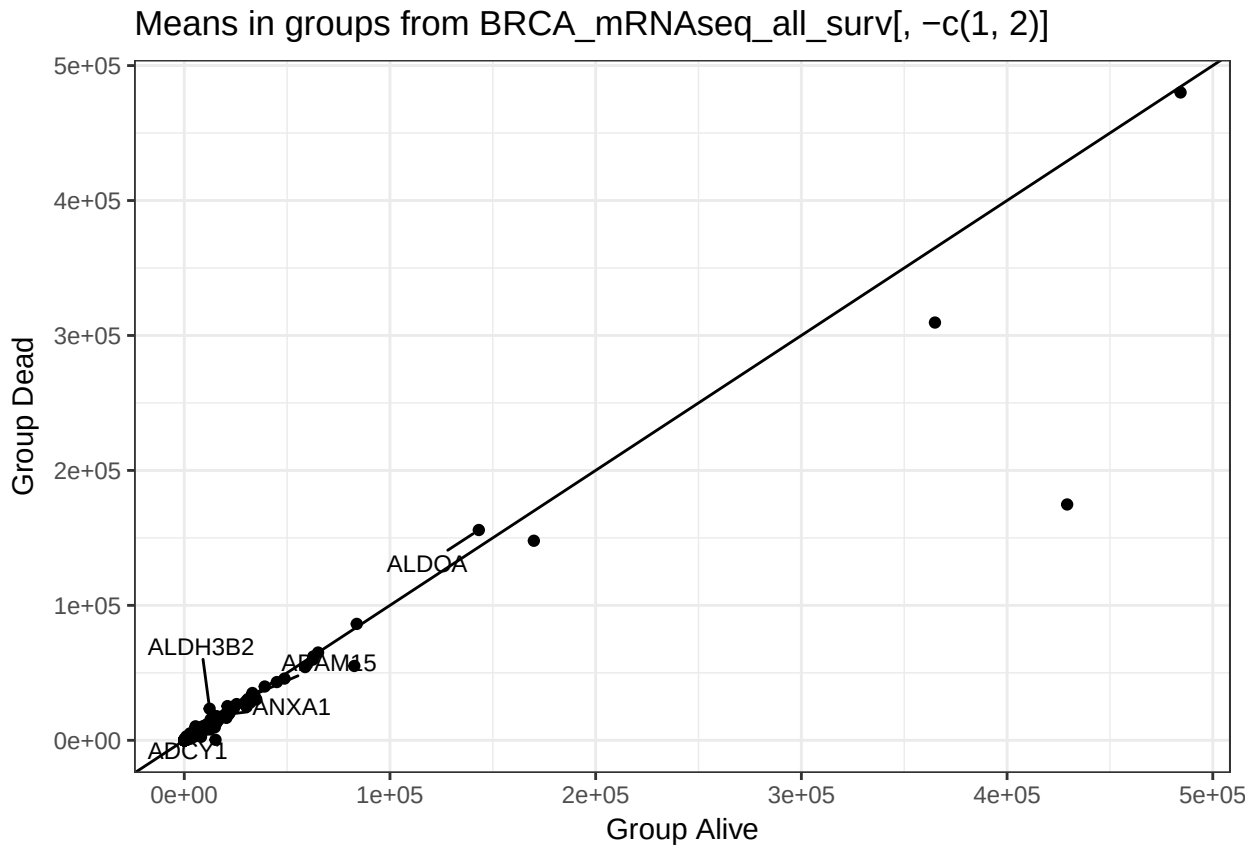
```
head(test.expr)
```

```
##          id         mean   log2.fold         pval        padj
## A1BG   A1BG   724.692096 -0.16354329 0.2911276062 0.631469658
## A1CF   A1CF     1.968352  0.28570801 0.3917060374 0.705690802
## A2BP1 A2BP1     7.340081 -0.71778707 0.0558100821 0.256663316
## A2LD1 A2LD1   477.736721 -0.02174672 0.8592587582 0.946562959
## A2ML1 A2ML1  1647.065939  1.63341375 0.0000125719 0.001042968
## A2M     A2M 76975.290989 -0.53103106 0.0045937626 0.054990040
```

Visualization

em_plot

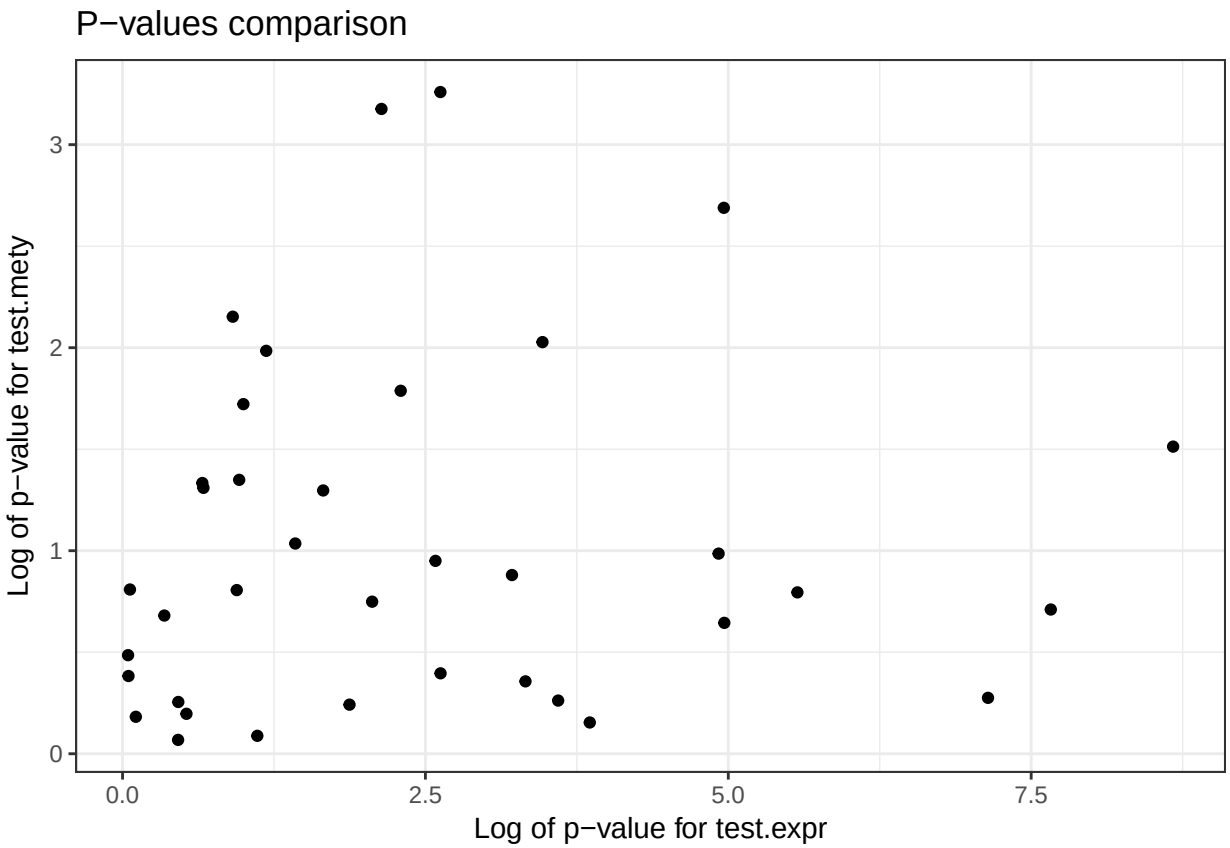
```
em_plot(BRCA_mRNAseq_all_surv[,-c(1,2)], condition, names=5)
```



log-log p-value

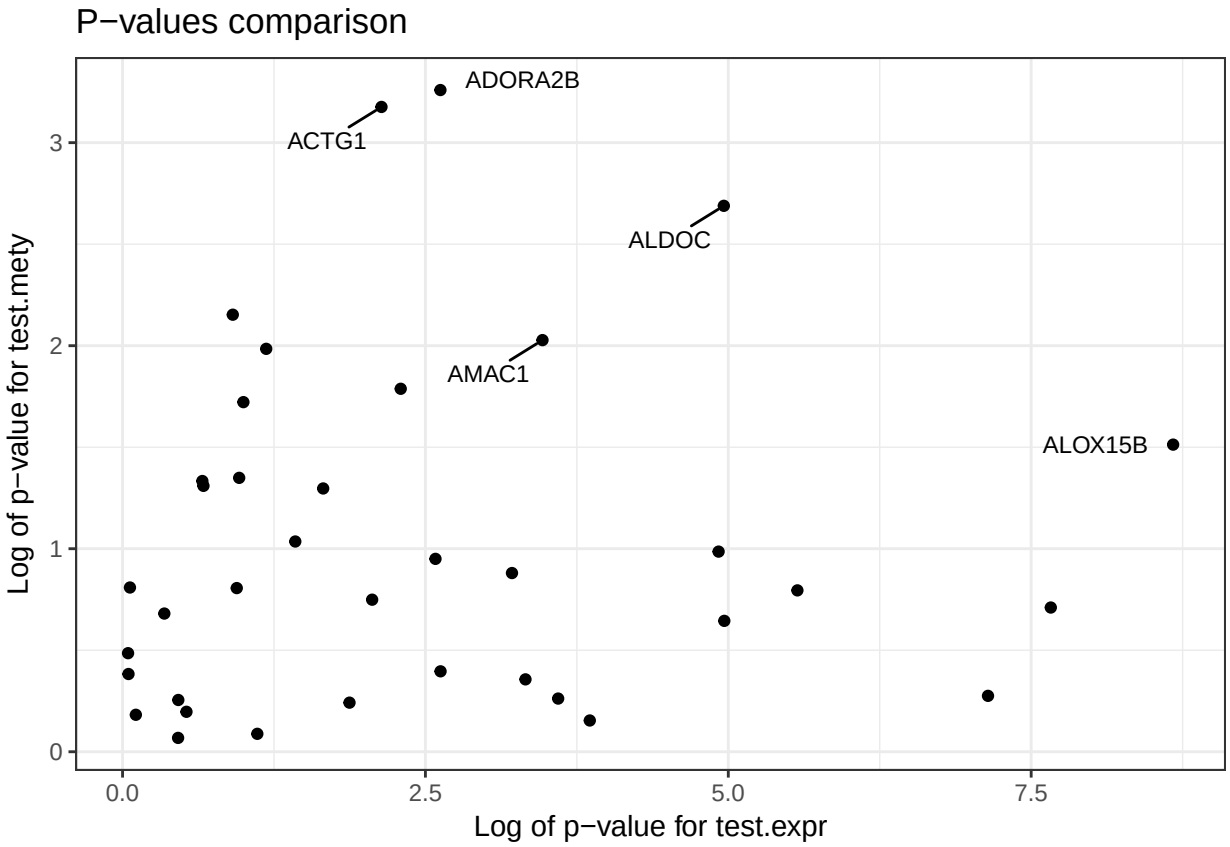
Firstly, we want to visualise the p-values for expression and methylation from negative binomial test and t-test respectively.

```
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)
```

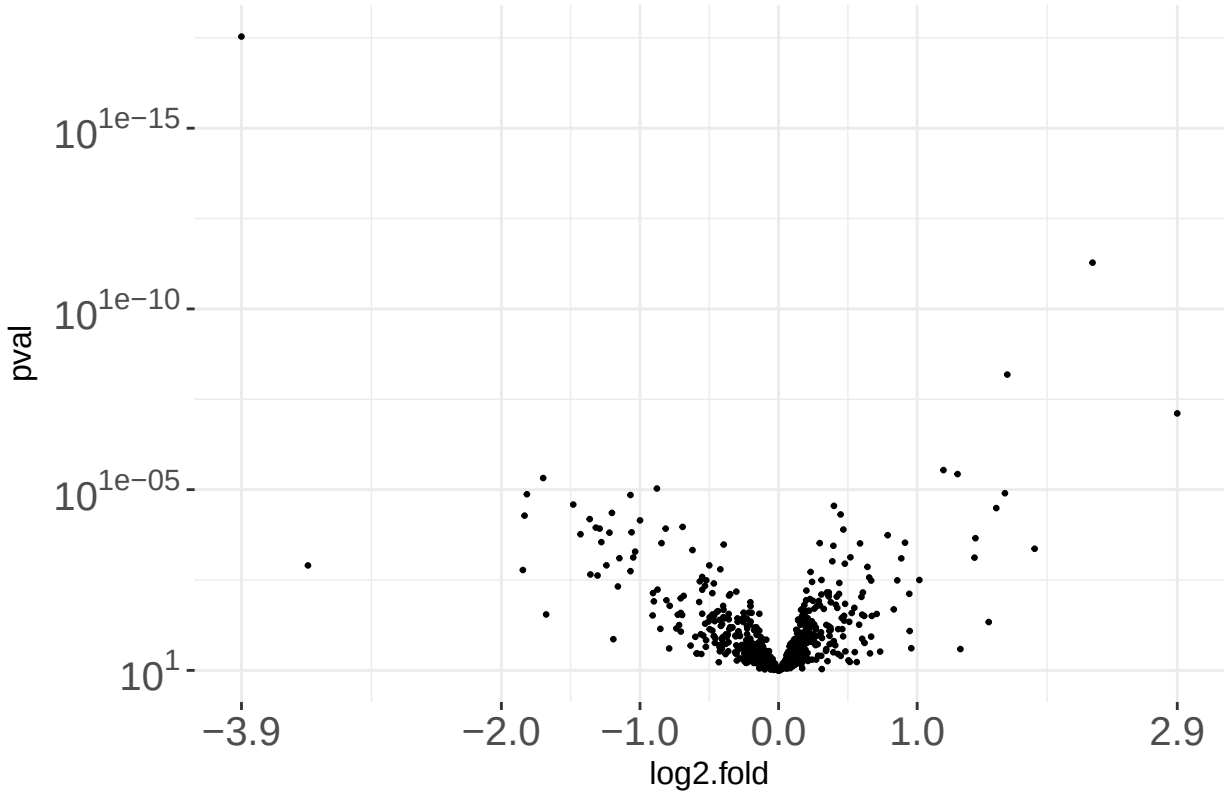


To read more about `p_values_plot` (e.g other ways to labeling genes) see help page for that function.

Volcano plot

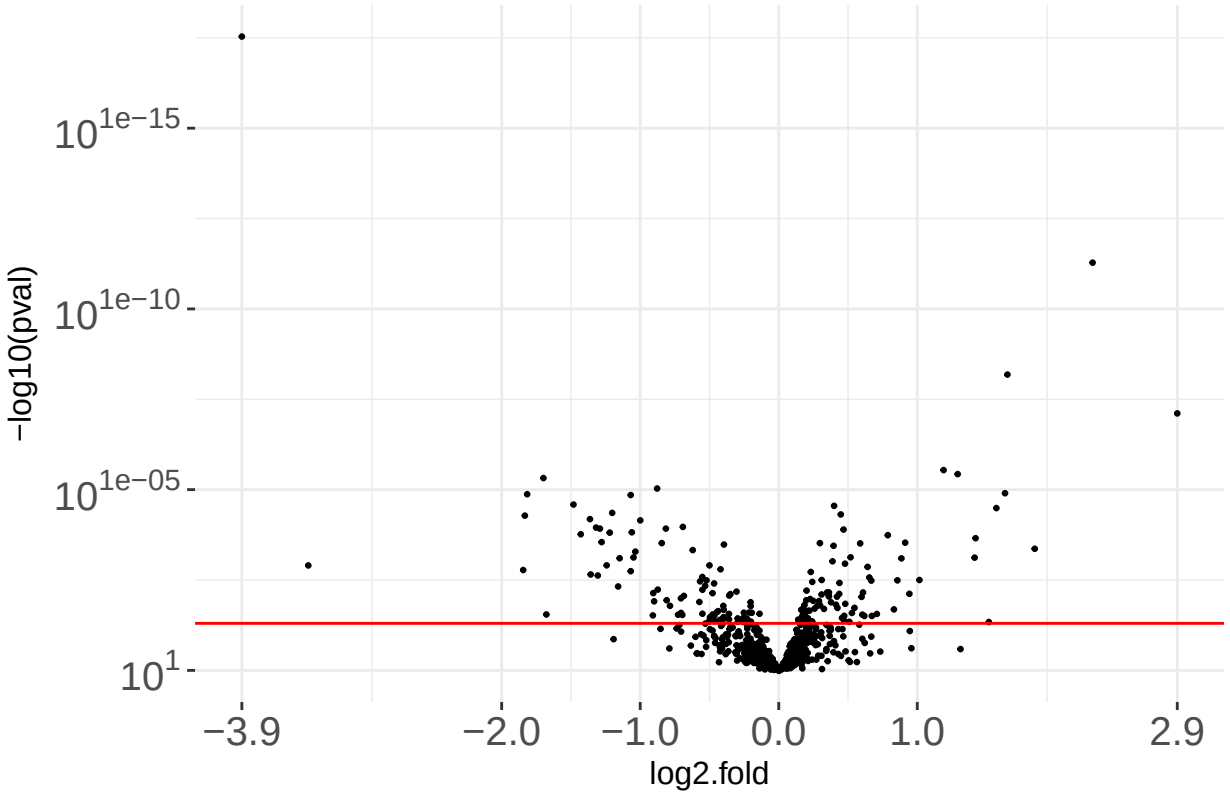
For identify changes in our data sets we use a volcano plot - some type of scatter-plot. It plots logarithm of p-value versus logarithm of fold-change on the y and x axes, respectively.

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

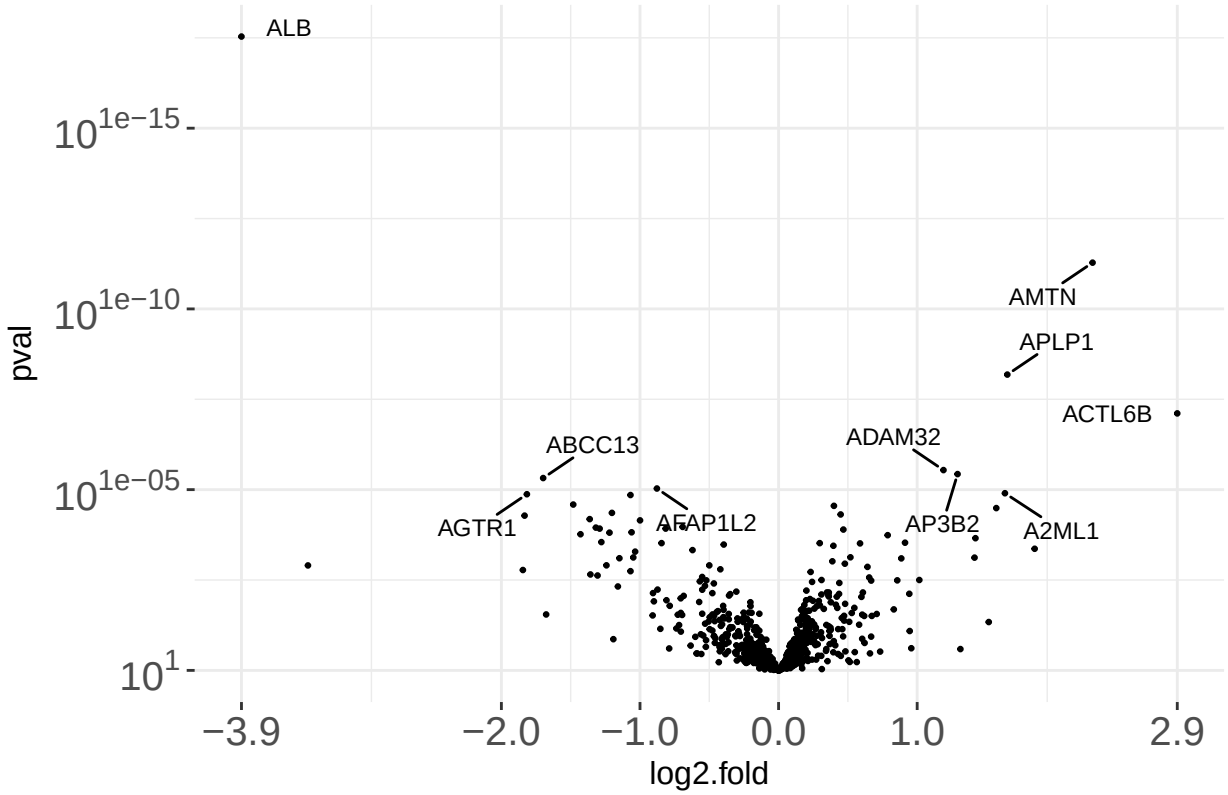


Function `volcano_plot` has parameters that allow to better analyze the results: `line` and `names`. The `line` parameter allows to set the horizontal line on plot on selected value. The `names` parameter signs choosen number of genes with the lowest p-value.

```
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.

```
gen <- colnames(BRCA_methylation_gen)[10]
BRCA1_gene <- CpG_mean(BRCA_methylation_all_surv, gen)
BRCA1_gene
```

```
##             Name HG18_coord Symbol CPG_ISLAND CPG_ISLAND_LOCATIONS
## 95    cg00081975   46067369  ABCC3       TRUE 17:46066791-46067798
## 20579 cg20633883   46067050  ABCC3       TRUE 17:46066791-46067798
##             mean
## 95    0.08790436
## 20579 0.07532627
```

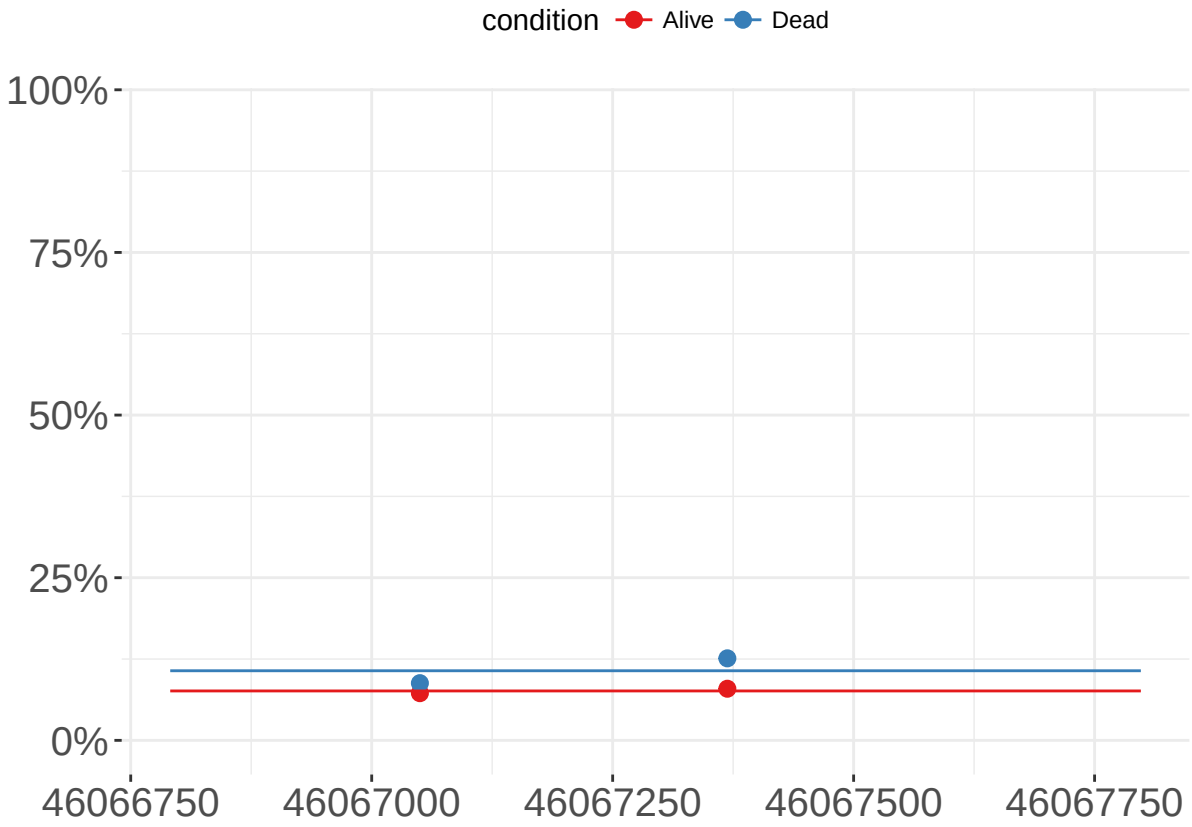
Methylation and expression in groups.

Two subtype groups of cancer in one plot.

```
condition <- ifelse(BRCA_methylation_all_surv$survival_status==1,"Dead", "Alive")
genereg_vs_met(BRCA_methylation_all_surv, condition, gen)
```

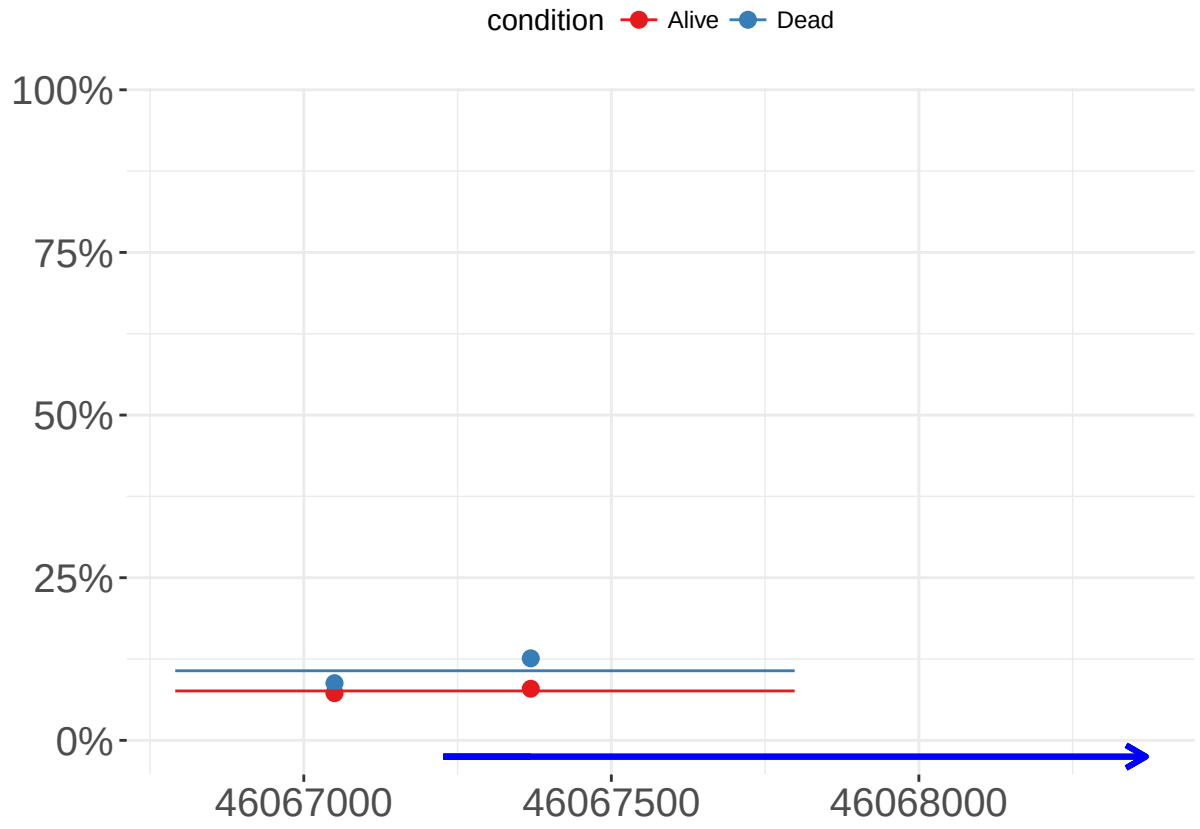
```
## 'select()' returned 1:1 mapping between keys and columns
```

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

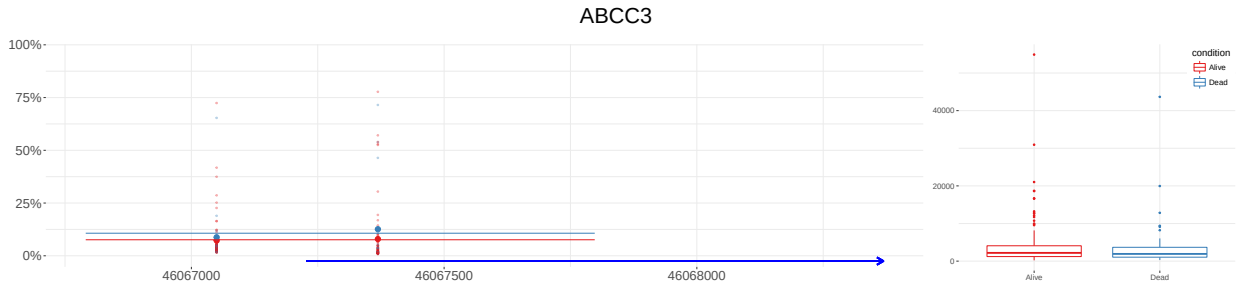



```
genereg_vs_met(BRCA_methylation_all_surv, condition, gen, show_gen=TRUE)
```

```
## '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_all_surv[,-c(1,2)],BRCA_mRNAseq_all_surv[,-c(1,2
```

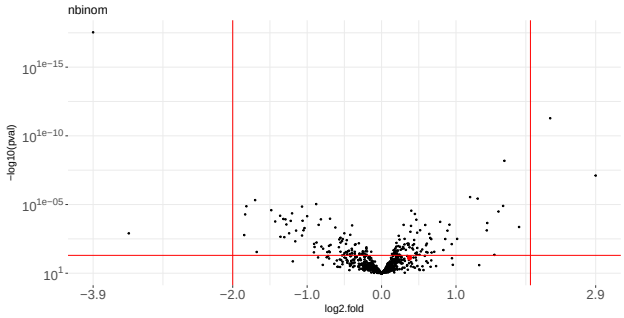



```
visual_volc(condition.e, condition.m, BRCA_methylation_all_surv[,-c(1,2)],BRCA_mRNAseq_all_surv[,-c(1,2
```

ABCC3

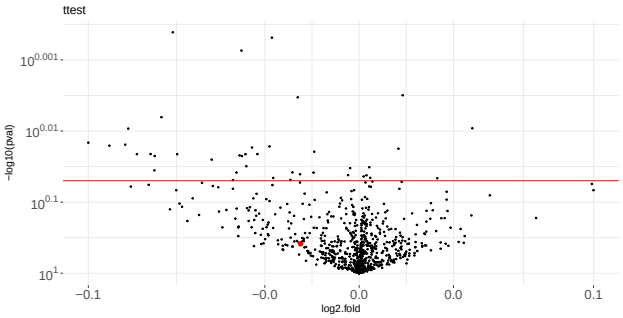
Expression (cpm)

	min	1st Q	med	mean	3rd Q	max	count
Alive	219.8	1213.5	2196	3643.3	4085	54896.7	224
Dead	293.5	1050.6	1920.6	4086.5	3677.1	43662.1	45



Methylation

	min	1st Q	med	mean	3rd Q	max	count
Alive	0	0	0	0.1	0.1	0.8	76
Dead	0	0	0	0.1	0.1	0.7	17




```
visual_volc(condition.e, condition.m, BRCA_methylation_all_surv[,-c(1,2)],BRCA_mRNAseq_all_surv[,-c(1,2
```

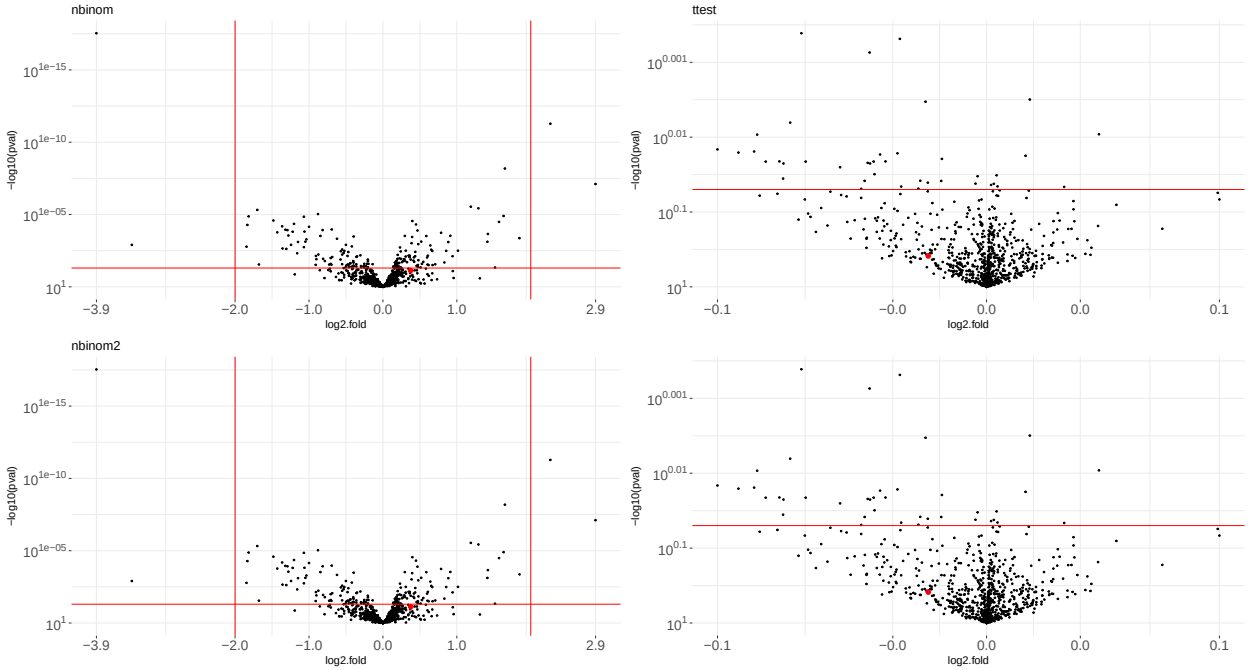
ABCC3

Expression (cpm)

	min	1st Q	med	mean	3rd Q	max	count
Alive	219.8	1213.5	2196	3643.3	4085	54896.7	224
Dead	293.5	1050.6	1920.6	4086.5	3677.1	43662.1	45

Methylation

	min	1st Q	med	mean	3rd Q	max	count
Alive	0	0	0	0.1	0.1	0.8	76
Dead	0	0	0	0.1	0.1	0.7	17



Other data set

We consider a data set from lung cancer - `LUSC`

`LUSC_methylation_all_surv` data set

Like in a BRCA case we have a data set containig methylation level for CpG islands.

We compute the `t-test` for this dataset

```r
condition <- ifelse(LUSC_methylation_all_surv$survival_status==1,"Dead","Alive")

LUSC_methylation_gen <- map_to_gene(LUSC_methylation_all_surv[,-c(1,2)])

test.mety.lusc <- test_diff(LUSC_methylation_gen,condition,"ttest")

head(test.mety.lusc)
```

```
##              id       mean    log2.fold        pval      padj
## PIGS       PIGS 0.02080026 -0.002705409 0.001705489 0.8371246
## ARRB2     ARRB2 0.01856907 -0.002351529 0.008073617 0.8371246
## TRIM25   TRIM25 0.01455957 -0.002845108 0.008329885 0.8371246
## PRPSAP1 PRPSAP1 0.01420674 -0.003073094 0.008543977 0.8371246
## MAFG       MAFG 0.01260102 -0.003442610 0.009355024 0.8371246
## LSM12     LSM12 0.01846602 -0.004669461 0.010067948 0.8371246
```

LUSC_mRNAseq_all_surv data set

Like in a BRCA case we have a data set containig expression for genes.

We compute the `nbinom` for this dataset

```r
condition <- ifelse(LUSC_mRNAseq_all_surv$survival_status==1,"Dead","Alive")
#tutaj krzyczy ze ma ujemne wartości
test.nbinom.lusc <- test_diff(LUSC_mRNAseq_all_surv[,-c(1,2)],condition,"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 91 genes
## -- DESeq argument 'minReplicatesForReplace' = 7
## -- original counts are preserved in counts(dds)

## estimating dispersions

## fitting model and testing
```