MLExpResso – NGS, Methylation, Expression, R and a lot of coffee

MLGenSig: Machine Learning Methods for building the Integrated Genetic Signatures

NCN Opus grant 2016/21/B/ST6/02176

Aleksandra Dąbrowska, Alicja Gosiewska, Przemysław Biecek

MLExpResso

Cheat Sheet

Aleksandra Dąbrowska [aut, cre]
Alicja Gosiewska [aut]
Przemysław Biecek [aut, ths]

Introduction

MLExpResso is an R package for integrative analyses and visualization of gene expression and DNA methylation data.

Key functions of this package are:

- identification of genes with affected expression – **calculate_test()** function,
- identification of DMR - differentially methylated regions - **calculate_test()** function,
- identification of regions with changes in expression and methylation - **calculate_comparition_table()** function,
- visualization of identified regions – **plot_gene()** and **plot_volcanoes()** functions.

The joint modeling and visualization of genes expression and methylation improve interpretability of identified signals.

MLExpRessoData

The methodology is supplemented with example applications to The Cancer Genome Atlas data.

MLExpRssoData is an R package which contains information from *The Cancer Genome Atlas (TCGA)* Data Portal. Data sets in this package are based on Bioconductor package **RTCGA**. In examples, we use both, methylation and expression data.

- **BRCA_exp -** It contains information about gene expression: read counts per-gene, computed for genes for 736 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_met -** It contains information about methylation of CpG probes for patients 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 probes. Values inside the table indicate the percentage methylation level of CpG probe for a specified sample.

For aggregation CpG probes to correspond genes we use the *Illumina human methylation* data set from **TxDb.Hsapiens.UCSC.hg18.knownGene** Bioconductor package.

Identification of genes with affected expression

MLExpResso::calculate_test(data, condition, test)

Function **calculate_test()** computes log folds, p-values and means for chosen test for both, methylation and expression data.

```
> BRCA_exp[1:3, 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
```

Example

```
library("MLExpResso")
library("MLExpRessoData")
exp <- BRCA_exp[ ,-1]
gr_exp <- BRCA_exp$SUBTYPE
gr_exp <- ifelse(gr_exp == 'LumA', 'LumA', 'other')
res_exp <- calculate_test(exp, gr_exp, 'lrt')
```

```
> head(res_exp)
     id log2.fold     pval mean_LumA mean_other   mean
1 AURKB       2.3 3.2e-32       539       2324   1485
2  CBX2       2.9 2.8e-26       633       4297   2574
3 KPNA2       1.4 8.6e-24     11547      26427  19434
4 PRR11       3.8 2.3e-22       396       3480   2031
5 BIRC5       2.0 2.0e-21      1957       6658   4449
6  GSG2       1.4 3.5e-21       278        629    464
```

Argument test allows using many different statistic tests for finding differences in expression. All available values are in the table below.

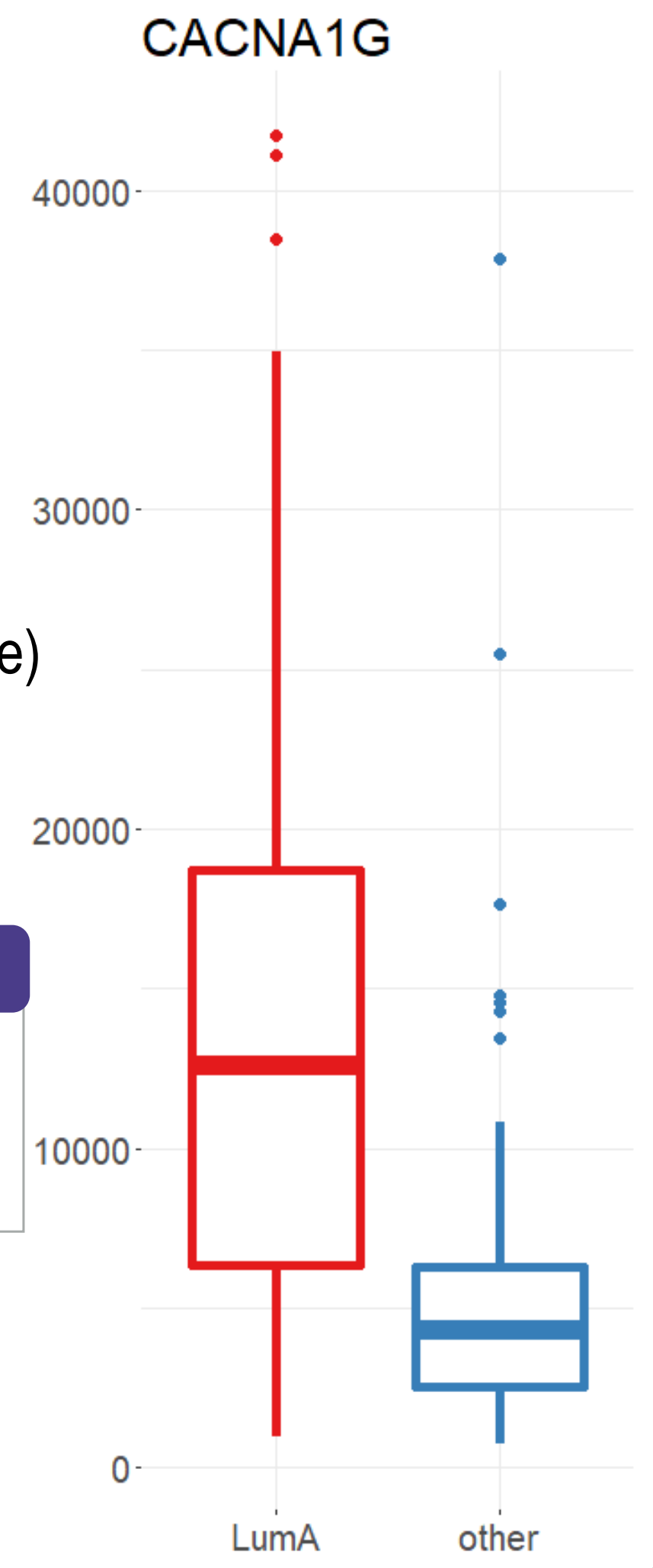
Value	Test
'ttest'	student's t-tets
'nbinom2'	negative binomial test
'lrt'	likelihood-ratio test
'qlf'	quasi-likelihood F-test

MLExpResso::plot_diff_boxplot(data, condition, gene)

Function plot_diff_boxplot() generates a boxplot of values from choosen data frame column with division in groups (two or more).

Example

```
plot_diff_boxplot(data = exp,
                  condition=gr_exp,
                  gene='CACNA1G')
```



Identification of DMR - differentially methylated regions

MLExpResso::calculate_test(data, condition, test)

Argument test allows using two different statistic tests for finding differences in methylation levels. All available values are in the table on the right.

Value	Test
'ttest'	student's t-test
'methyanalysis'	quasi-likelihood F-test

MLExpResso::aggregate_probes(data)

Function **aggregate_probes()** aggregates CpG probes to corresponding genes using, by default, *the Illumina human methylation* data.

```
> BRCA_met[1:3, 1:5]
                             SUBTYPE cg00021527 cg00031162 cg00032227 cg00050312
TCGA-A1-A0SD-01A-11D-A112-05    LumA      0.038       0.79     0.0064      0.024
TCGA-A2-A04N-01A-11D-A112-05    LumA      0.014       0.74     0.0088      0.028
TCGA-A2-A04P-01A-31D-A032-05   Basal      0.014       0.70     0.0094      0.014
```

Example

```
met <- aggregate_probes(BRCA_met)
gr_met <- BRCA_met$SUBTYPE
gr_met <- ifelse(gr_met == 'LumA', 'LumA', 'other')
res_met <- calculate_test(met, gr_met, 'ttest')
```

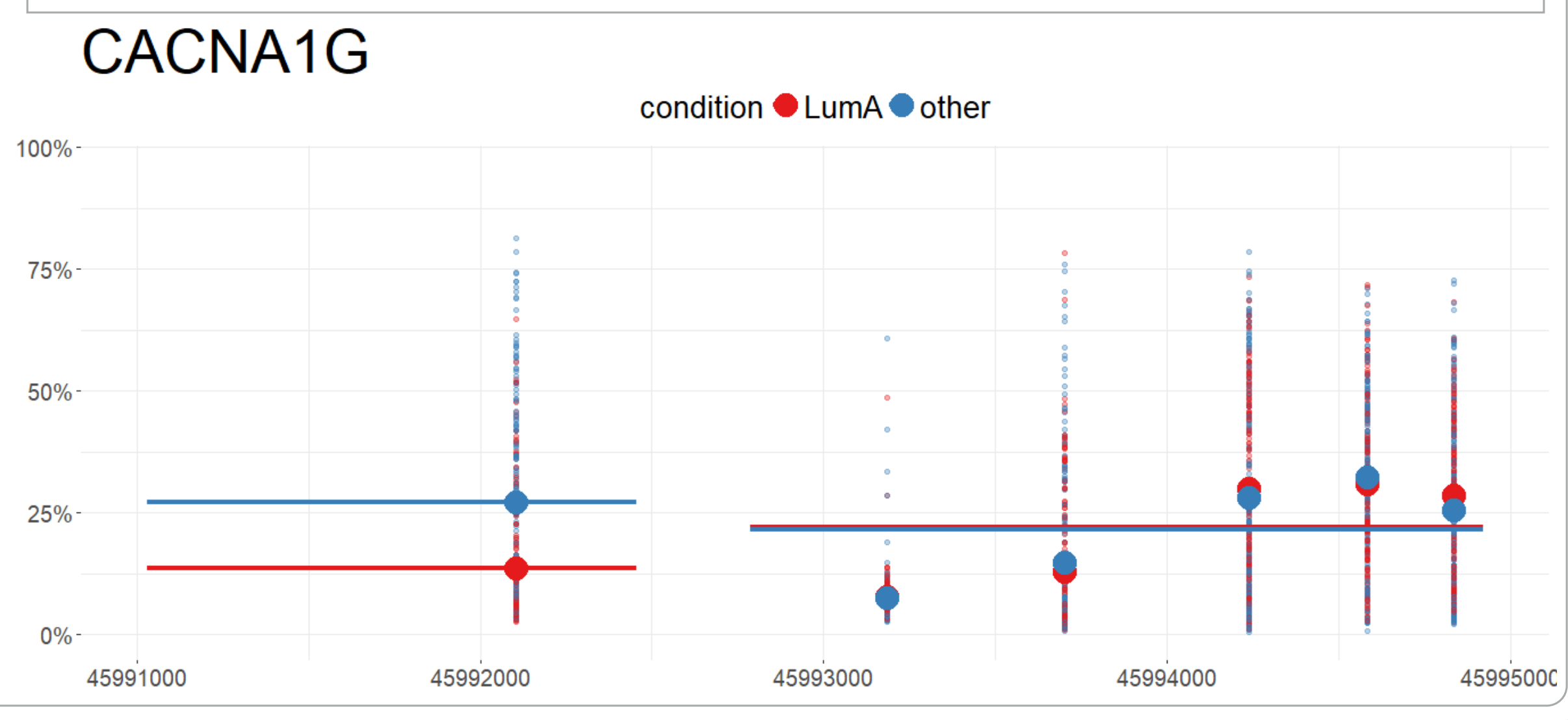
```
> head(res_met)
       id log2.fold    pval mean_LumA mean_other mean
1   ICAM2    -0.152 3.8e-17      0.25       0.41 0.33
2    RILP    -0.051 2.6e-13      0.31       0.36 0.33
3   PIPOX     0.115 5.4e-12      0.42       0.31 0.36
4 TNFSF12    -0.134 5.9e-12      0.18       0.31 0.25
5     CD7     0.098 1.6e-11      0.86       0.77 0.81
6    KSR1     0.200 2.1e-11      0.66       0.46 0.55
```

MLExpResso::plot_methylation_path(data, condition, gene)

Function plot_methylation_path() visualizes a chosen gene with marked CpG probes.
Y axis describes methylation level.
X axis describes a location of the probe on the chromosome.
Horizontal lines show the mean methylation level for each Island in a division to groups. Groups are defined by colors. Large dots symbolize means of methylation level for CpG probes, small dots symbolize methylation levels for each observation.

Example

```
plot_methylation_path(data = BRCA_met ,
                      condition = gr_met,
                      gene = 'CACNA1G',
                      observ = T)
```



Comparing test results

MLExpResso::calculate_comparison_table(data1, data2, condition1, condition2, test1, test2)

Function **calculate_comparison_table()** produces a dataset containing p-values and folds from tests evaluated on two datasets e.g methylation or expression. In addition, it produces an importance ranking column, which is the geometric mean of p-values from both tests and a column with a number of probes related to the gene.

Example

```
calculate_comparison_table(data1 = BRCA_exp[,-1], data2 = BRCA_met_gen,
                           condition1 = condition_exp, condition2 = condition_met,
                           test1 = "nbinom2", test2 = "ttest")
```

```
> head(genes_comparison)
       id nbinom2.log2.fold nbinom2.pval ttest.log2.fold ttest.pval geom.mean.rank no.probes
59  AURKB               2.4      1.7e-37         0.00174    2.1e-01        1.9e-19         2
102  CBX2               2.9      5.4e-31         0.05847    1.2e-06        8.1e-19         2
327 KPNA2               1.5      3.4e-26         0.00121    7.5e-01        1.6e-13         1
277  GSG2               1.4      3.3e-25        -0.00186    2.4e-01        2.8e-13         2
66  BIRC5               2.0      9.5e-24        -0.00054    5.3e-01        2.2e-12         1
334 KRT16               4.3      4.1e-19         0.04868    1.6e-05        2.6e-12         2
```

Visualization of identified regions

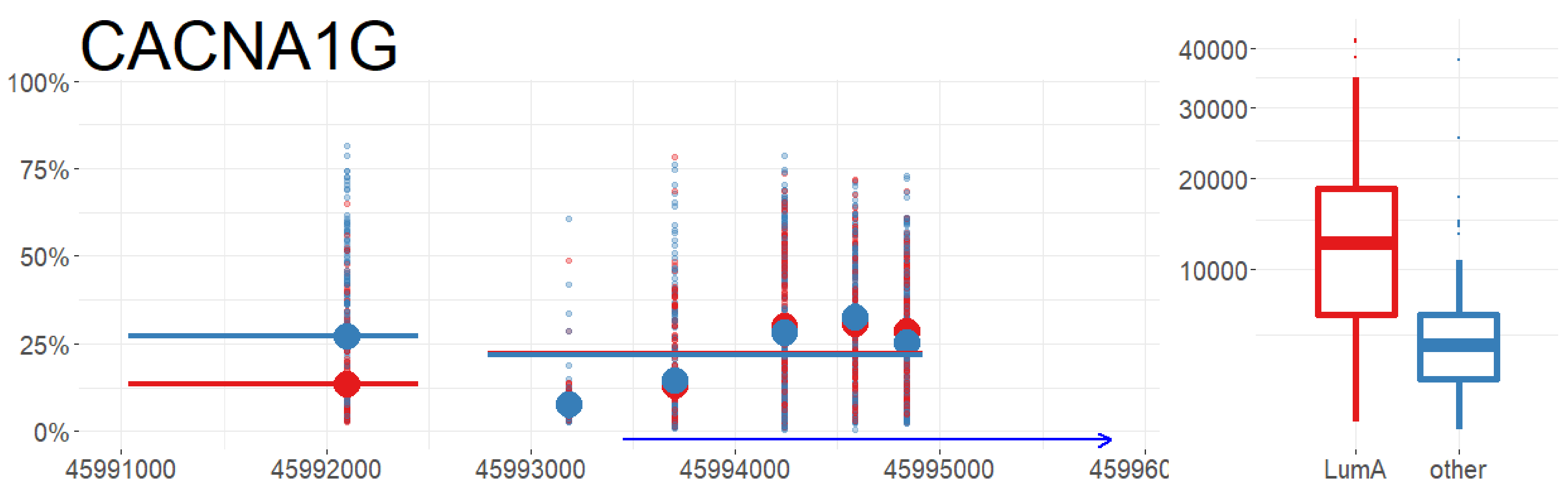
Plot_gene()

MLExpResso::plot_genes(data.m, data.e, condition.m, condition.e, gene)

Function **plot_gene()** generates a dashboard with methylation path for methylation and boxplots for groups for chosen gene.

Example

```
plot_gene(data.m = BRCA_met,
                    data.e = BRCA_exp,
                    condition.m = gr_met,
                    condition.e = gr_exp,
                    gene = "CACNA1G",
                    observ = TRUE,
                    show.gene = TRUE,
                    islands = TRUE)
```



Plot_volcanoes()

MLExpResso::plot_volcanoes(data.m, data.e, condition.m, condition.e, gene, test.m, test.e)

Function **plot_volcanoes()** generates a dashboard with volcano plots for expression and methylation. Also it adds a tables with basic statistics for chosen gene.

Example

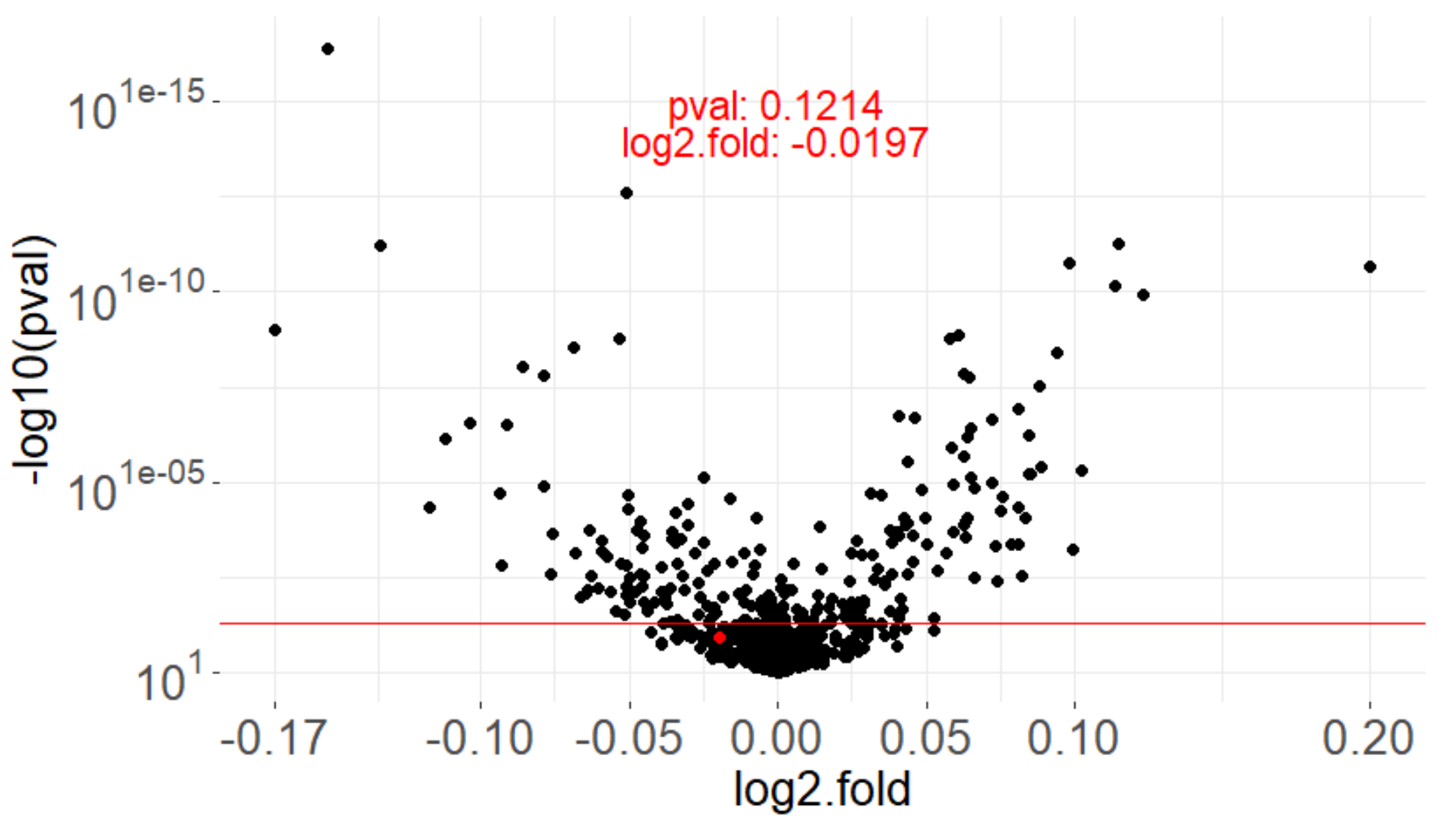
```
plot_volcanoes(data.m = BRCA_met,
               data.e = BRCA_exp,
               condition.m = gr_met,
               condition.e = gr_exp,
               gene = 'CACNA1G',
               test.m = res_met,
               test.e = res_exp)
```

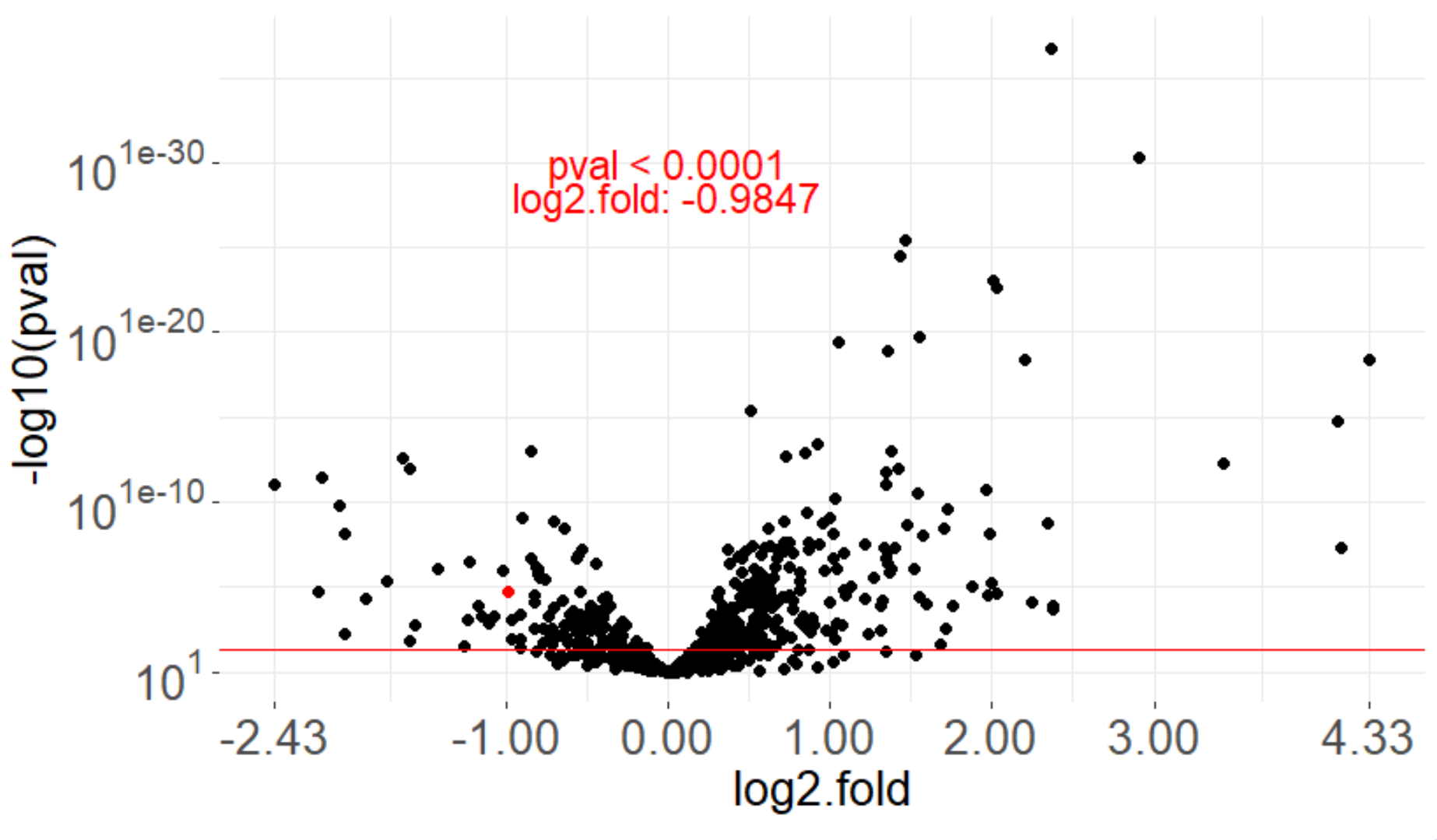
Methylation

	min	1st Q	med	mean	3rd Q	max	count
LumA	0.05	0.14	0.19	0.21	0.27	0.49	155
other	0.04	0.13	0.2	0.23	0.31	0.63	166

Expression (cpm)

	min	1st Q	med	mean	3rd Q	max	count
LumA	977	6351	12628	14302	18716	41717	47
other	752	2556	4360	6185	6314	37883	53


pval: 0.1214
log2.fold: -0.0197
10^{1e-15}
10^{1e-10}
10^{1e-05}
10^{1}
-log10(pval)
log2.fold


pval < 0.0001
log2.fold: -0.9847
10^{1e-30}
10^{1e-20}
10^{1e-10}
10^{1}
-log10(pval)
log2.fold