MLExpResso

a tool for integrative analyses and visualization of gene expression and DNA methylation data

Alicja Gosiewska
Warsaw University of Technology,
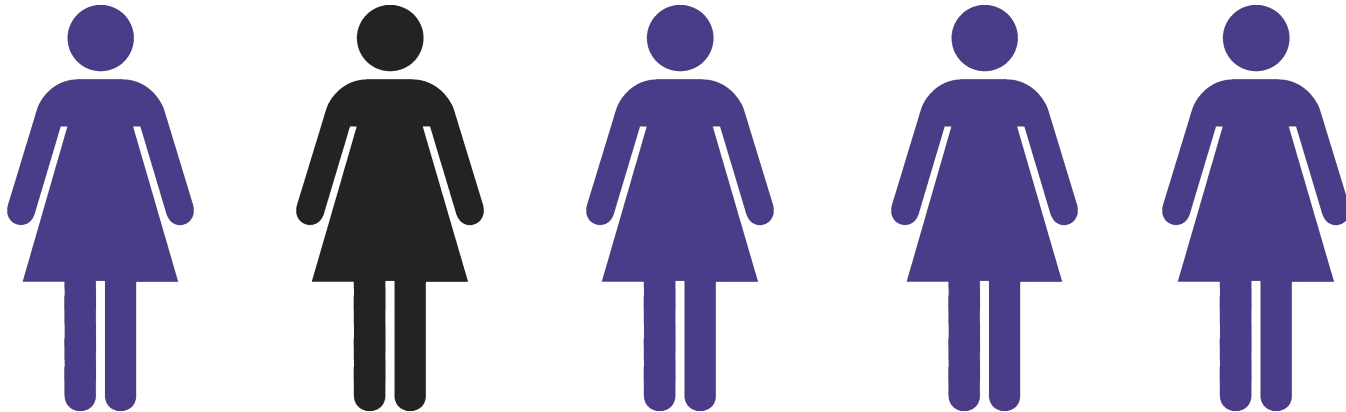**Faculty of Mathematics and Information Science**
MI^2

Berlin, 14.12.2017

About

one **in five**

women diagnosed with breast cancer worldwide will have

HER2 positive breast cancer.

MI

MLExpResso

1) **Identification of differentially expressed genes**

2) **Identification of differentially methylated regions**

3) **Identification of regions with changes in expression and methylation**

4) **Visualization of identified regions**

MI

Expression of genes

```
> BRCA_expression[1:8, 1:12]
                         AANAT AARSD1  AATF AATK ABCA5 ABCA6 ABCA8 ABCA9 ABCC3 ABI3  ABR ACACA
TCGA-A1-A0SB-01A-11R-A144-07     9   2354  2870  317  1071   170   548   322   118  124 4960  5036
TCGA-A1-A0SD-01A-11R-A115-07     2   1846  5656  312  2107   735  1060   649  2786  823 6531 10428
TCGA-A1-A0SE-01A-11R-A084-07    11   3391  9522  736  1600   849   704  1176  2814  805 7302 20787
TCGA-A1-A0SF-01A-11R-A144-07     0   2169  4625  169   615   442   377   543  1925  500 5029  9562
TCGA-A1-A0SG-01A-11R-A144-07     1   2273  3473   92   249   547   417   413  1227  474 6982  8636
TCGA-A1-A0SH-01A-11R-A084-07     2   3113  7669  396  2337  1149   667   875  3910  929 9999 42456
TCGA-A1-A0SI-01A-11R-A144-07     1   1089  2870   97   556   647  1061   457  1572  635 5590  5298
TCGA-A1-A0SJ-01A-11R-A084-07     8   1857 11175  656  4443  1159  2648   922  2304  805 8935 11828
```

Methylation level of CpG probes

```
> BRCA_methylation[1:8, 1:6]
                         cg00021527 cg00031162  cg00032227 cg00050312 cg00053292 cg00063144
TCGA-A1-A0SD-01A-11D-A112-05 0.03781858  0.7910348 0.006391233 0.02356901 0.01806759  0.9192175
TCGA-A2-A04N-01A-11D-A112-05 0.01437552  0.7359370 0.008752293 0.02770303 0.01711573  0.9186985
TCGA-A2-A04P-01A-31D-A032-05 0.01360124  0.6967802 0.009442039 0.01402589 0.02012769  0.9186004
TCGA-A2-A04Q-01A-21D-A032-05 0.01525656  0.5341244 0.014674247 0.01510194 0.02155129  0.8985550
TCGA-A2-A04T-01A-21D-A032-05 0.01167384  0.7378100 0.012251559 0.01174021 0.02682894  0.9200222
TCGA-A2-A04U-01A-11D-A112-05 0.04266864  0.7428112 0.008710026 0.01303019 0.01765034  0.9525430
TCGA-A2-A04V-01A-21D-A032-05 0.02172694  0.5373882 0.011088346 0.01917646 0.01983831  0.9442921
TCGA-A2-A04W-01A-31D-A112-05 0.07874157  0.7497528 0.010024020 0.01340650 0.02791492  0.8860238
```

MI

Identification of significant features - expression
LumA subtype vs other subtypes of breast cancer

```
> library(MLExpResso)
> res_expression <- calculate_test(
+    data = BRCA_expression,
+    condition = condition_expression,
+    test = "nbinom2"
+ )
> head(res_expression)
      id log2.fold              pval mean_LumA mean_other      mean
1  AURKB   2.303668 1.705520e-37   539.0426  2323.8868   1485.01
2   CBX2   2.777812 5.490481e-31   632.5106  4296.6038   2574.48
3  KPNA2   1.446017 3.398752e-26   11547.36  26427.38 19433.77
4   GSG2   1.407218 3.318054e-25   278.2128   629.3396    464.31
5  BIRC5   1.948513 9.512118e-24  1957.085  6658.358   4448.76
6  PRR11   1.967561 2.054430e-23    396.383  3479.981   2030.69
```

MI

Identification of significant features - methylation
LumA subtype vs other subtypes of breast cancer

```
> BRCA_methylation_gen <- aggregate_probes(data = BRCA_methylation)
> head(BRCA_methylation_gen)[1:5, 1:5]
                                AANAT     AARSD1       AATF       AATK       ABC1
TCGA-A1-A0SD-01A-11D-A112-05 0.7148533 0.8625816 0.24294092 0.7835302 0.01401806
TCGA-A2-A04N-01A-11D-A112-05 0.5850106 0.8355825 0.21367129 0.8466190 0.01275865
TCGA-A2-A04P-01A-31D-A032-05 0.4495537 0.8786166 0.03277413 0.3417919 0.01455092
TCGA-A2-A04Q-01A-21D-A032-05 0.7120650 0.8819490 0.03460160 0.7264985 0.01283647
TCGA-A2-A04T-01A-21D-A032-05 0.6010397 0.7739978 0.02501599 0.6276399 0.01278928

> res_methylation <- calculate_test(
+ data = BRCA_methylation_gen,
+ condition = condition_methylation,
+ test = "ttest"
+ )
> head(condition_methylation)
[1] "LumA"  "LumA"  "other" "other" "other" "other"           mean
1    ICAM2 -0.15151320 3.754116e-17 0.2547275  0.4062407 0.3330801
2     RILP -0.05073691 2.575168e-13 0.3079069  0.3586438 0.3341447
3    PIPOX  0.11505558 5.360053e-12 0.4242804  0.3092248 0.3647812
4 TNFSF12 -0.13412855 5.867083e-12 0.1791401  0.3132686 0.2485025
5      CD7  0.09822690 1.641919e-11 0.8635077  0.7652808 0.8127112
6     KSR1  0.19973400 2.054467e-11  0.658270   0.458536 0.5549808
```
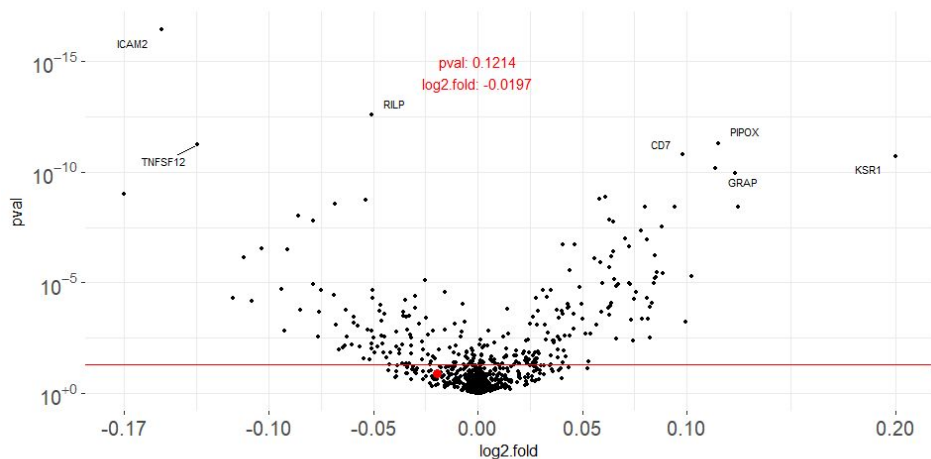
MI

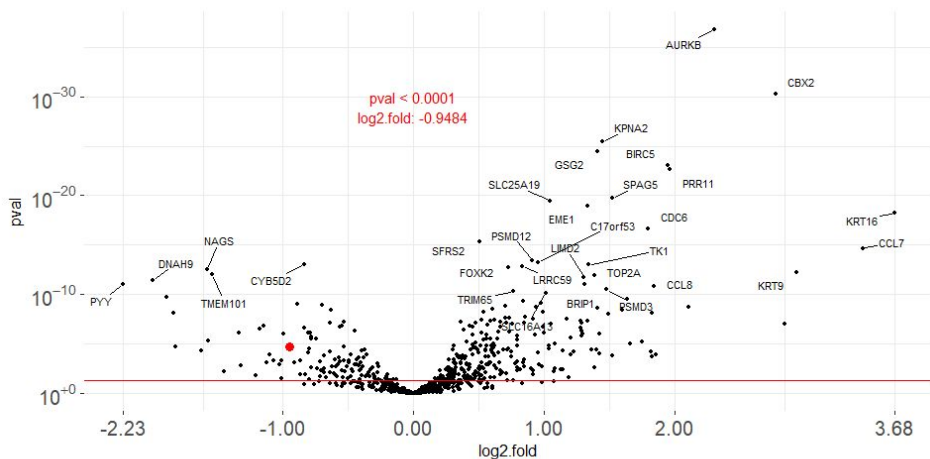# Identification of significant features
plot_volcanoes()

Integration of methylation and expression

plot_gene()

MLExpResso

Cheat Sheet

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

MI
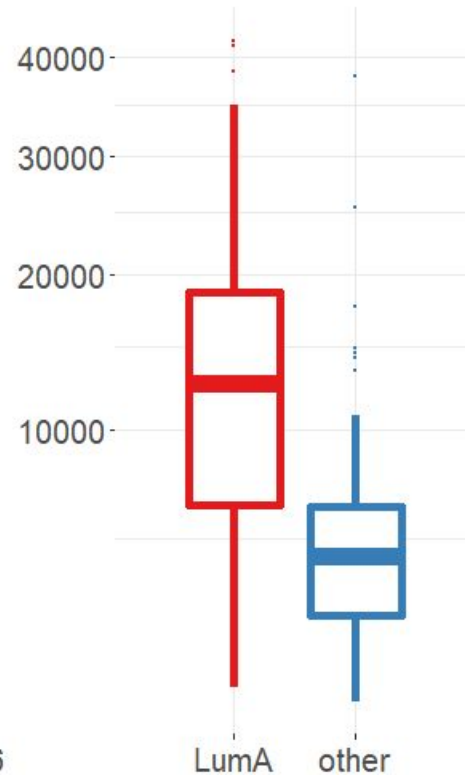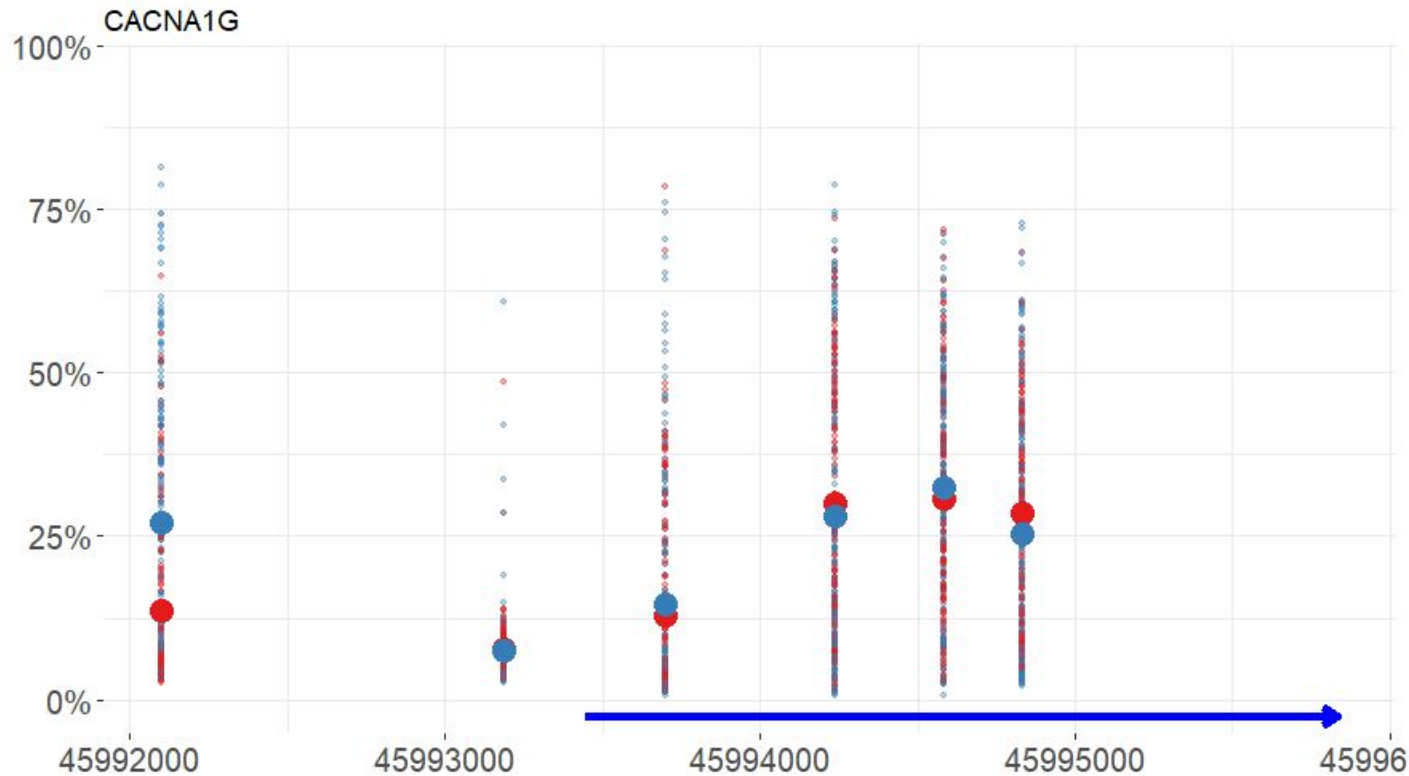
Introduction

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.

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
```

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

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
```

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')
```


CACNA1G

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 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)
```


CACNA1G

MI
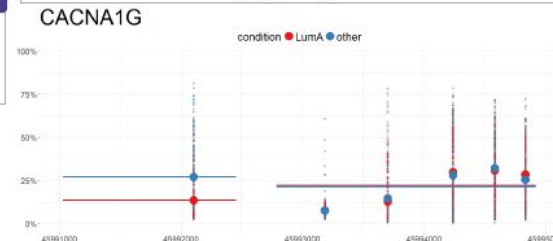
"MLGenSig: Machine Learning Methods for building the Integrated Genetic Signatures"
NCN Opus grant 2016/21/B/ST6/02176

GitHub https://github.com/geneticsMiNIng/MLGenSig

agosiewska

@ alicjagosiewska@gmail.com