MLExpResso: differential expression and methylation analysis

Case study using RTCGA data

Aleksandra Dąbrowska, Alicja Gosiewska

Contents

1 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 `MLExpResso` provides methods to test for differential expression and methylation by use of the negative binonial distribution and t-test. Additionaly 'MExpResso allows to visualize results in a simple way.

2 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.

2.1 Methylation

2.1.1 `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 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 islands. Values inside the table indicate the methylation level of CpG island for specified sample.

```
library(MLExpResso)
```

```
library(MLExpRessodata)
```

```
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
```

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

```
BRCA_methylation_gen <- aggregate_probes(BRCA_methylation_chr17[,-1])
head(BRCA_methylation_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
```

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 a vector, which each element corresponds to a sample. Our division into this two groups relies on numbers of occurences of each subtype. The `LumA` subtype is the most common, in case of breast cancer.

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

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

2.1.2 Testing

```
res_met <- calculate_test(BRCA_methylation_chr17[,-1], condition_met, test="ttest")
head(res_met)
```

```
##          id      mean  log2.fold         pval
## 1 cg10275770 0.3330801 -0.1515132 3.795188e-17
## 2 cg06144905 0.4111616  0.1738623 9.908386e-14
## 3 cg04049033 0.6462602 -0.1027159 1.513031e-13
## 4 cg12045829 0.2485025 -0.1341285 5.910592e-12
## 5 cg02473123 0.8127112  0.0982269 1.653082e-11
## 6 cg05246522 0.5549808  0.1997340 2.069380e-11
```

2.2 Expression

2.2.1 `BRCA_mRNAseq_chr17` data set

Data set `BRCA_mRNAseq_chr17` contains information about gene expression. This data set contains per-gene read counts 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_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
```

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

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

2.2.2 Testing

```
res_exp <- calculate_test(BRCA_mRNAseq_chr17[,-1], condition_exp, test="lrt")
head(res_exp)
```

```
##         id      mean log2.fold          pval
## 59   AURKB   1485.01  2.339920 3.191000e-32
## 105   CBX2   2574.48  2.895062 2.834335e-26
## 336  KPNA2 19433.77  1.447288 8.551812e-24
## 505  PRR11   2030.69  3.822148 2.286874e-22
## 66   BIRC5   4448.76  1.988998 1.953941e-21
## 283   GSG2    464.31  1.405039 3.527773e-21
```

2.3 Comparing test results

```
genes_comparison <- calculate_comparison_table(BRCA_mRNAseq_chr17[ ,-1], BRCA_methylation_gen,
                               condition_exp, condition_met, test1="nbinom2", test2="ttest")
head(genes_comparison)
```

```
##        id nbinom2.log2.fold nbinom2.pval ttest.log2.fold   ttest.pval
## 1   AANAT        0.36342624   0.22109578      0.07237890 1.026904e-05
## 2 AARSD1       -0.07435954   0.45027777      0.01114861 5.052981e-01
## 3   AATF        0.13308688   0.18597163      0.01360521 2.136847e-01
## 4   AATK        0.02308606   0.92313513      0.06619897 1.427292e-05
## 5  ABCA5       -0.07034249   0.73314806      0.00510806 6.353608e-01
## 6  ABCA6       -0.44685708   0.08266088     -0.03402781 4.393248e-02
##    genes_rank
```

```
## 1 0.001506798
## 2 0.476995295
## 3 0.199347171
## 4 0.003629853
## 5 0.682505353
## 6 0.060261904
```

2.4 Choosing most different genes

Sorting . . .

```
genes_comparison_sorted <- genes_comparison[order(genes_comparison$ttest.pval), ]
head(genes_comparison_sorted)
```

```
##            id nbinom2.log2.fold nbinom2.pval ttest.log2.fold    ttest.pval
## 302    ICAM2         0.2077055 2.102511e-01     -0.15151320 3.754116e-17
## 519     RILP         0.3139026 5.252448e-02     -0.05073691 2.575168e-13
## 466    PIPOX         0.3589979 4.371416e-02      0.11505558 5.360053e-12
## 652 TNFSF12        -0.4357410 3.486764e-04     -0.13412855 5.867083e-12
## 133      CD7         1.3728093 9.099866e-07      0.09822690 1.641919e-11
## 341     KSR1         0.4261993 1.658465e-02      0.19973400 2.054467e-11
##       genes_rank
## 302 2.809461e-09
## 519 1.163011e-07
## 466 4.840560e-07
## 652 4.522956e-08
## 133 3.865391e-09
## 341 5.837174e-07
```
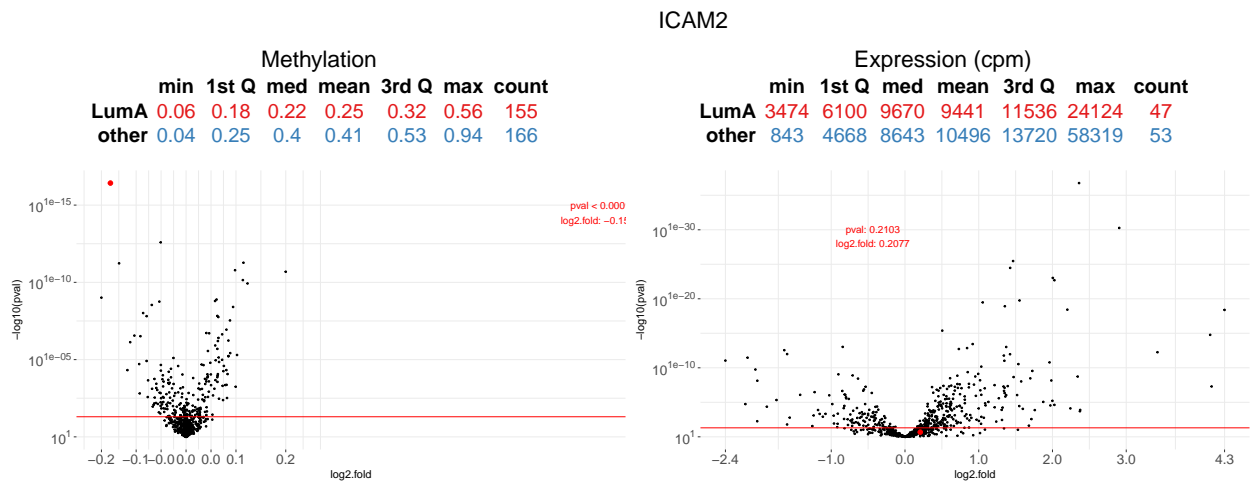
ICAM2!

2.5 Visualization

Visualizing chosen gene - IGFALS.

```
test_exp <- genes_comparison[ ,c(1,2,3)]
test_met <- genes_comparison[ ,c(1,4,5)]
```

```
plot_volcanoes(BRCA_methylation_chr17[,-1],BRCA_mRNAseq_chr17[,-1],condition_met, condition_exp,
               "ICAM2",
               test_exp, test_met,
               values=TRUE)
```

ICAM2

Methylation

	min	1st Q	med	mean	3rd Q	max	count
LumA	0.06	0.18	0.22	0.25	0.32	0.56	155
other	0.04	0.25	0.4	0.41	0.53	0.94	166

Expression (cpm)

	min	1st Q	med	mean	3rd Q	max	count
LumA	3474	6100	9670	9441	11536	24124	47
other	843	4668	8643	10496	13720	58319	53

Note that `plot_gene` methylation require data frame with cpg islands, not genes.

```
plot_gene(BRCA_methylation_chr17,BRCA_mRNAseq_chr17,condition_met, condition_exp, "ICAM2")
```


5