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  log2.fold         pval mean_LumA mean_other      mean
## 1 cg10275770 -0.1515132 3.795188e-17 0.2547275  0.4062407 0.3330801
## 2 cg06144905  0.1738623 9.908386e-14 0.5010717  0.3272094 0.4111616
## 3 cg04049033 -0.1027159 1.513031e-13 0.5931423  0.6958582 0.6462602
## 4 cg12045829 -0.1341285 5.910592e-12 0.1791401  0.3132686 0.2485025
## 5 cg02473123  0.0982269 1.653082e-11 0.8635077  0.7652808 0.8127112
## 6 cg05246522  0.1997340 2.069380e-11  0.658270   0.458536 0.5549808
```

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 SUBTYPEcorresponds 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 log2.fold         pval mean_LumA mean_other     mean
## 1 AURKB  2.339920 3.191000e-32  539.0426  2323.8868  1485.01
## 2  CBX2  2.895062 2.834335e-26  632.5106  4296.6038  2574.48
## 3 KPNA2  1.447288 8.551812e-24  11547.36   26427.38 19433.77
## 4 PRR11  3.822148 2.286874e-22   396.383   3479.981  2030.69
## 5 BIRC5  1.988998 1.953941e-21  1957.085   6658.358  4448.76
## 6  GSG2  1.405039 3.527773e-21  278.2128   629.3396   464.31
```

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")
```

```
## Warning in sqrt(result[, 2] * result[, 4]): wyprodukowano wartości NaN
```

```
## Warning: Column `id` joining character vector and factor, coercing into
## character vector
```

```
head(genes_comparison)
```

```
##           id nbinom2.log2.fold nbinom2.pval ttest.log2.fold ttest.pval
## 354    LSM12       0.056616564    0.4954925    9.234253e-05 0.87094671
## 579  SMARCE1      -0.023092119    0.8318721   -2.377235e-04 0.60251281
## 77  C17orf61      -0.006548806    0.9585559   -1.027549e-03 0.10515489
## 367      MED9      -0.003427851    0.9677837   -2.615948e-03 0.03653572
## 482  PRKAR1A      -0.003404099    0.9791404   -4.099641e-03 0.08891401
## 273      GRB2      -0.068323804    0.5412281   -2.148793e-04 0.85842993
##     geom.mean.rank No.probes
## 354    0.002286507         2
## 579    0.002342976         2
## 77     0.002594074         2
## 367    0.002994508         1
## 482    0.003735717         2
## 273    0.003831628         2
```

2.4 Choosing most different genes

Sorting …

```r
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
##     geom.mean.rank No.probes
## 302            NaN         2
## 519            NaN         2
## 466      0.2032356         2
## 652      0.2417546         2
## 133      0.3672149         1
## 341      0.2917644         1
```

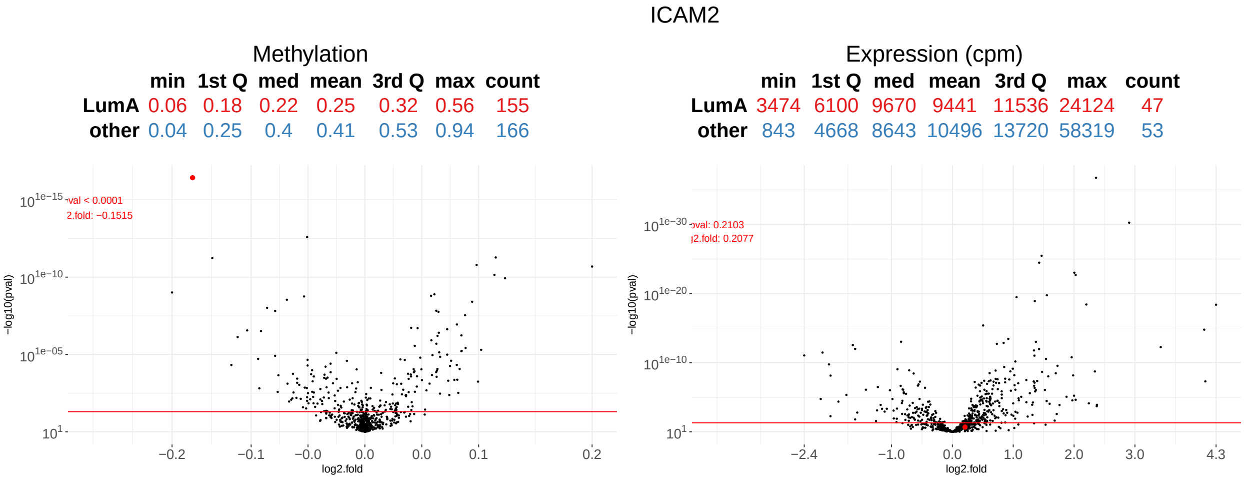
ICAM2!

2.5 Visualization

Visualizing chosen gene - IGFALS.

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

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

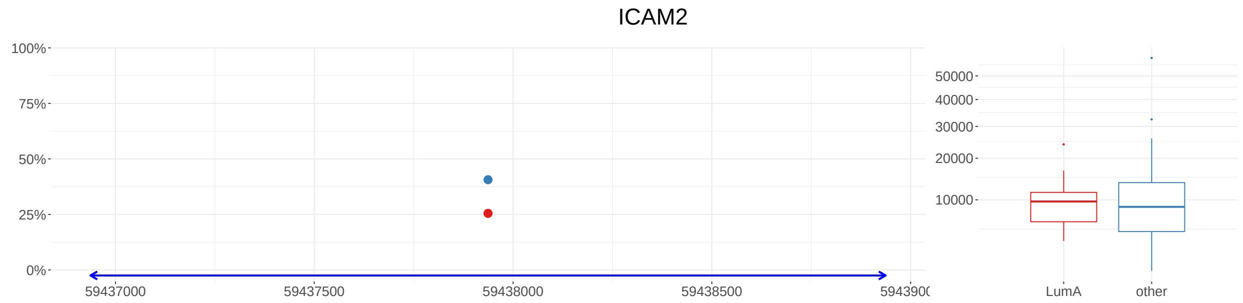


	min	1st Q	med	mean	3rd Q	max	count
LumA	0.06	0.18	0.22	0.25	0.32	0.56	155
other	0.04	0.25	0.4	0.41	0.53	0.94	166

	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")
```


ICAM2
594390(
LumA
other