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

Przemysław Biecek, Alicja Gosiewska, Aleksandra Grudziąż

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

Introduction

Technological progress has led to increasing amounts of data in molecular biology. As a result, statistical methods are used to support the analysis of these data. **MLExpResso** is a tool that simplifies the analysis of DNA data. To identify changed DNA regions it applies statistical methods such as t-student test, negative binomial test, and others.

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

Gene expression is the process by which information from a gene is used in the synthesis of a functional gene products, such as proteins.

DNA methylation is a process by which methyl groups are added to the DNA molecule. It can change the activity of a DNA segment without changing the sequence.

Key functions of MLExpResso are:

- identification of differentially methylated regions based on RRBS data,
- identification of differentially expressed genes based on RNAseq data,
- identification of regions with changes in expression and methylation between two conditions,
- visualization of identified regions.

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

MLExpResso uses methods implemented in various R packages available in Bioconductor for analysis of expression (i.e. DESeq2 [2], edgeR [3]) and methylation (i.e. MethyAnalysis [4]). In addition, it supports visualization of identified regions. The developed solution can be used to better understand the interdependence of expression and methylation and their joint effect on the selected feature.

References:

[1] MLExpResso, ,,Machine Learning for Genetic Signatures" R package [https://github.com/geneticsMiNIng/MLGenSig].

[2] Love M, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, Genome Biology 2014 15:550.

[3] Robinson M, McCarthy D, Smyth G. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data, Bioinformatics 2010 Jan 1; 26(1): 139–140.

[4] Du P, Bourgon R. (2017). methyAnalysis: an R package for DNA methylation data analysis and visualization.

[5] Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemporary Oncology. 2015;19(1A):A68-A77.

The aim of the Analysis and Data

The aim of the analysis is to increase the knowledge about cancer genomics. Subtypes of cancer may be identified by changes in the DNA expression or methylation level. Such konwledge is important for the medicine. It leads to develop better methods of identifying and treating cancer.

In this analysis we compare **LumA** subtype of a **breast cancer (BRCA)** and other subtypes of this type of cancer.

The methodology is supplemented with example applications to **The Cancer Genome Atlas (TCGA)** data. TCGA is a collaboration between the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) that has generated comprehensive, multi-dimensional maps of the key genomic changes in 33 types of cancer.

In examples, we use methylation and expression data from the Bioconductor package **RTCGA**.

- **BRCA_exp** data set 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.

	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

- **BRCA_met** data set contains information about methylation of DNA regions (**CpG probes)** for 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 CpG probes. Values inside the table indicate the percentage methylation level of CpG probe for a specified sample.

	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

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

In this analysis we will focus on the gene **CACNA1G** located on chromosome 17.

Identification of Differentially Methylated Regions (DMR)

MLExpResso::aggregate_probes(data) **MLExpResso::calculate_test(data, condition, test)**

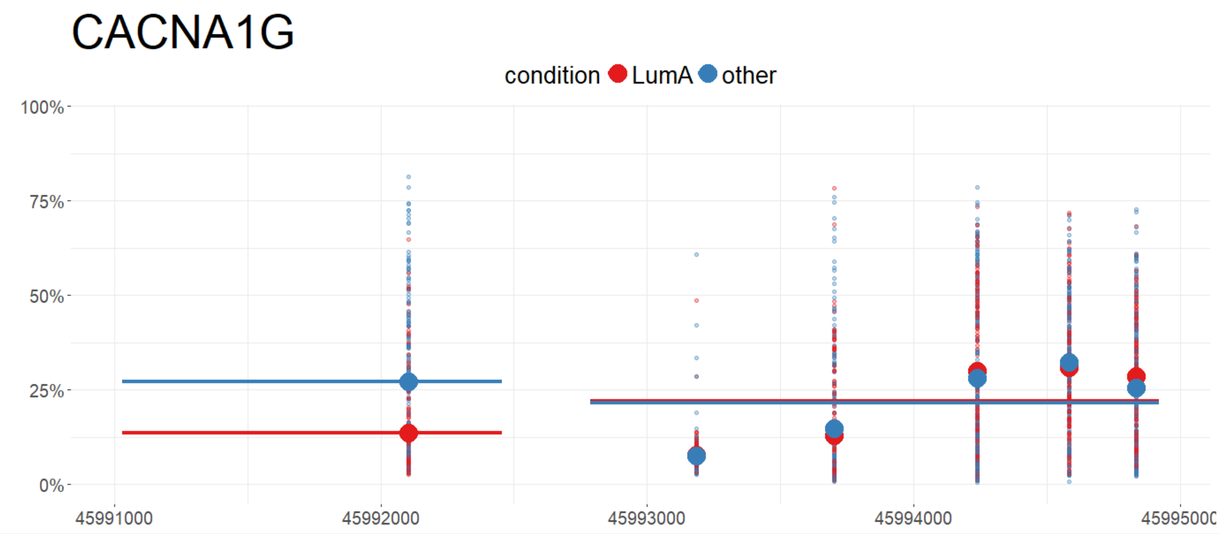
First step of the methylation analysis is aggregating CpG probes to corresponding genes by function *aggregate_probes()*. Next step is identifying genes with differences in mean methylation levels. Function *calculate_test()* provides two statistical tests for this purpose: Student's t-test and quasi-likelihood test. This function computes log-fold changes, p-values and means for both tests .

Result is a table with rows corresponding to genes and columns corresponding to test p-value, group means and overall mean. Below we present results for first six genes. Results are calculated for tests for differences methylation between LumA breast cancer subtype and other subtypes.

	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)

A very usefull functionality of MLexpResso is function *plot_methylation_path()*, which is used to visualization of identified genes and coresponding CpG probes.
Y axis describes methylation level. X axis describes a location of the CpG probe on the chromosome.
Horizontal lines show the mean methylation level for CpG islands (groups of CpG probes) with division into groups. Groups are defined by colors. Large dots symbolize means of methylation level for CpG probes, small dots symbolize methylation levels for each observation.
In this plot we can see that the there is a group of the CpG probes near CACNA1G gene that have different methylation levels for LumA breast cancer subtype than for other subtypes.



Identification of Genes with Affected Expression

MLExpResso::calculate_test(data, condition, test)

Function *calculate_test()* provides four statistical tests for identifying differences in gene expression: Student's t-test, negative binomial test, likelihood-ratio test, quasi-likehood F-test.

This function computes log-fold changes, p-values and means for chosen test.

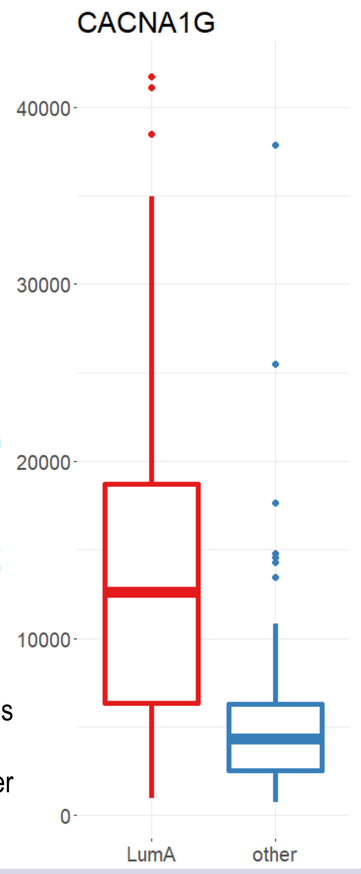
Result is a table with rows corresponding to genes and columns corresponding to test p-value, group means and overall mean. Below we present results for expression of gene CACNA1G.

	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 with values of gene expression with division into groups (two or more).

In the plot we can see that expression of CACNA1G is higher for patients with LumA subtype of breast cancer.



Comparing Test Results

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

This function produces a table containing p-values for two tests. It supports comparing results of tests for expression and methylation for the same gene. In addition, it produces an importance ranking column, which is the geometric mean of p-values from tests. This ranking supports the identification of genes with changed expression and methylation.

	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

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 tables with basic statistics for chosen gene.

Volcano plot is a scatter plot to visual identification of statistically significant data. It displays fold change on the log2 scale against negative log10 of p-value.
Each point on the plot represents the statistical result for a single gene. The horizontal line represents threshold of p-value.
Chosen gene is marked by red dot.

This dashboard is another tool that supports identification of genes with differences in expression and methylation for different subtypes of breast cancer.

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