RTCGA.methylation Explorator documentation

by Katarzyna Sobiczewska

Introduction

The RTCGA package (Kosinski and Biecek, 2015) provides accessible and easy way to analyse data from TCGA project (visit this page: `http://cancergenome.nih.gov/` to get more information about the project).

The RTCGA.methylation Explorator allows you to go into the data (source: `https://github.com/RTCGA/RTCGA.methylation`) and visualize interesting connections between DNA characteristics and the length of survival time of patiens stricken with different types of cancer.

Data specifications

The application allows you to explore 9 different types of cancer which are collected in different datasets with number of observations given in braces:

- Breast invasive carcinoma - BRCA (343)
- Colon adenocarcinoma - COAD (202)
- Glioblastioma multiforme - GBM (283)
- Pan-kidney cohort - KIPAN (973)
- Kidney renal clear cell carcinoma - KIRC (439)
- Acute Myeloid Leukemia - LAML (194)
- Lung adenocarcinoma - LUAD (89)
- Lung squamous cell carcinoma - LUSC (160)
- Uterine Corpus Endometrial Carcinoma - UCEC (118)

For each observation we had almost **300,000 features** which were expressed by different loci on DNA. Data values are from unit interval and they describe percentage of methylation for each biomarker.

To select the most important biomarkers the RTCGA.methylation dataset was merged with clinical information about each patient and time of survival was used to extract features which are significant. Using **Kaplan-Meier's estimator** we compared times of survival in strata indicated by median of methylation value. A biomarker was important if the differenece for survival was signifiant, thus almost 300,000 tests was made (more informaction about used methods you can find here (Therneau and Lumley)).

In this way we extracted over **13,000 biomarkers** that are significant for survival. They are listed in `Gene names` panel.

Application details

The main goal of this application is to visualize survival influences in time for given biomarkers. Insight into distribution of biomarkers is also provided. The application has three main panels:

1. Survival

Here you can inquire the influence of particular biomarkers on survival and differences between them in different types of cancer. The Kaplan-Meier methods was used (Goel et al., 2010).

Output details

- **Survival curves** with possibility to customise strata by setting up methylation cut-off.
- **Significance level** (p-value) for given biomarker that has an influence on survival in strata indicated **by median**.
- **Odds ratio** for time interval and startum set up by user. Note, that there is no possibility to calculate odds ratio with single stratum. In that case change the methylation cut-off so as to obtain two groups to compare with each other. If you are not sure about appropriate threshold value, take advantage of `Biomarkers distribution` panel.

2. Biomarkers distribution

In this place you can find what is the distribution of methylation values for the biomarkers in different types of cancer. Are there any similarities between them?

Output details

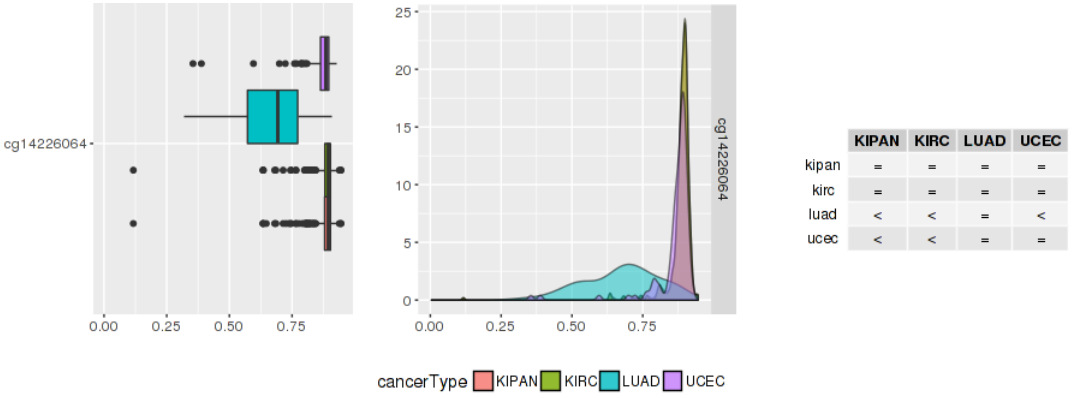
- **Boxplots** and **density plots** for each biomarker.
- **Kolmogorow-Smirnov test** results from comparing a biomarker distribution in each two types of cancer. One-sided test was used in this place and the results are gathered as a table. You can choose if you want to see results as p-values or decisions (one from "<", "="). Decision are made with $\alpha = 0.05$. Distributions are recognized as equal if we have decision: "=" on both sides of diagonal.

On the following example we can observe some similarities in distribution between cancer types, however Kolmogorov-Smirnov test:

H0: cancer=CANCER

H1: cancer<CANCER

reveal that only KIRC and KIPAN distributions of `cg14226064` are the same (KIRC=KIPAN vs KIRC<KIPAN hypothesis was accepted as well as KIPAN=KIRC vs KIPAN<KIRC).



	KIPAN	KIRC	LUAD	UCEC
kipan	=	=	=	=
kirc	=	=	=	=
luad	<	<	=	<
ucec	<	<	=	=

Figure 1: Biomarkers distribution example

3. Biomarkers list

If gene names is more preferable way for you to explore this data, please use the `Biomarkers list` panel to find biomarkers that are suitable to your genes list.

Output details

- **Biomarker name** - names used in this application.

- **Gene name** (in accordance with given locus). Notice, that one locus may have more than one accordant gene as well as one gene may have more than one suitable loci.
- **Common for (w/ p.val for survival)** - here the p-value is the measure of biomarker significancy for survival where strata are defined by median.
- **Number of common cancers** - number of cancer types for which a biomarker is significant for survival.

Additional tips

The application gives you possibility to download results and modify on your computer. The table is write as csv file. Plot outputs are saving as `ggplot` objects and they might be modify only with `ggplot2` R package.

Bibliography

M. K. Goel, P. Khanna, and J. Kishore. Understanding survival analysis: Kaplan-meier estimate. *Int J Ayurveda Res.*, pages 274–278, 2010. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3059453/. [p1]

M. Kosinski and P. Biecek. *RTCGA package*, 2015. URL https://github.com/RTCGA/. [p1]

T. M. Therneau and T. Lumley. *R: Test Survival Curve Differences*. URL https://stat.ethz.ch/R-manual/R-devel/library/survival/html/survdiff.html. [p1]

Katarzyna Sobiczewska

fk.katarzyna@gmail.com