Gene expression (RTCGA.mRNA)

by Emilia Momotko, Martyna Śpiewak, Mikołaj Waśniewski

Abstract Project goals keep the focus on identification of genetic factors that display differences in prognosis survival time depending on genetic background. The platform describes selected genetic data is `RTCGA.mRNA`. It contains information about gene expression among patients whose cancer was observed.

Data

All data sets used in the analysis come from the repository RTCGA .

We analyzed the following data sets, each of them corresponding to the type of tumor:

- **BRCA** - Breast invasive carcinoma;
- **COAD** - Colon adenocarcinoma;
- **COADREAD** - Colorectal adenocarcinoma;
- **GBMLGG** - Glioblastoma multiforme;
- **KIPAN** - Pan-kidney cohort (KICH + KIRC + KIRP);
- **KIRC** - Kidney renal clear cell carcinoma;
- **KIRP** - Kidney renal papillary cell carcinoma;
- **LGG** - Lower Grade Glioma;
- **LUAD** - Lung adenocarcinoma;
- **LUSC** - Lung squamous cell carcinoma;
- **OV** - Ovarian serous cystadenocarcinoma;
- **READ** - Rectum adenocarcinoma;
- **UCEC** - Uterine Corpus Endometrial Carcinoma.

Identification of significant biomarkers

To obtain information on biomarkers and their impact on the prognosis of treatment we used logrank test. Using function `survdiff` from `survMisc`(Dardis, 2015) package we were able to determine the p-value of this test.

We also examined differences in survival time between each type of cancer and each marker. The patients were divided into two groups. The first group corresponds to patients whose the value of gene expression was below the global median of whole values of marker (`lower`) and other were assigned to the second group, that is to say that patients whose the value of biomarker is greater than the global median (`higher`).

The file `biomarkers.csv` includes the list of the most significant biomarkers. Markers were selected in such a way that for each of the 13 types of cancer we determined the set of the 100 most significant biomarkers, using the p-value of logrank test. In other words, for each type of cancer we chose 100 biomarkers with the smallest p-value. The final dataset with the most significant biomarkers was determined as the union of sets with the most significant biomarkers for each type of cancer. Eventually, this set contains 1089 biomarkers.

Shiny application

In order to present the results of our analysis, we have created a Shiny application(Chang et al., 2015). The application can be found at `http://mi2.mini.pw.edu.pl:8080/RTCGA/MMM/shiny/` or you can download it from the repository.

The main aim of this application is to present qualities and prognostic abilities for particular gene marker. Running the comparison of several markers was not in our intention.

How to use?

Firstly, select the marker you want to examine in the first place. We assumed that the user can choose only one from markers' list. Moreover, in brackets are located the names of the types of cancer for which the selected marker proved to be a significant factor in the survival analysis.

For example, the figure 1 shows that the marker `ADORA3` is statistical significant (according to logrank test) for the following types of cancer: BRCA, OV, COAD, COADREAD.

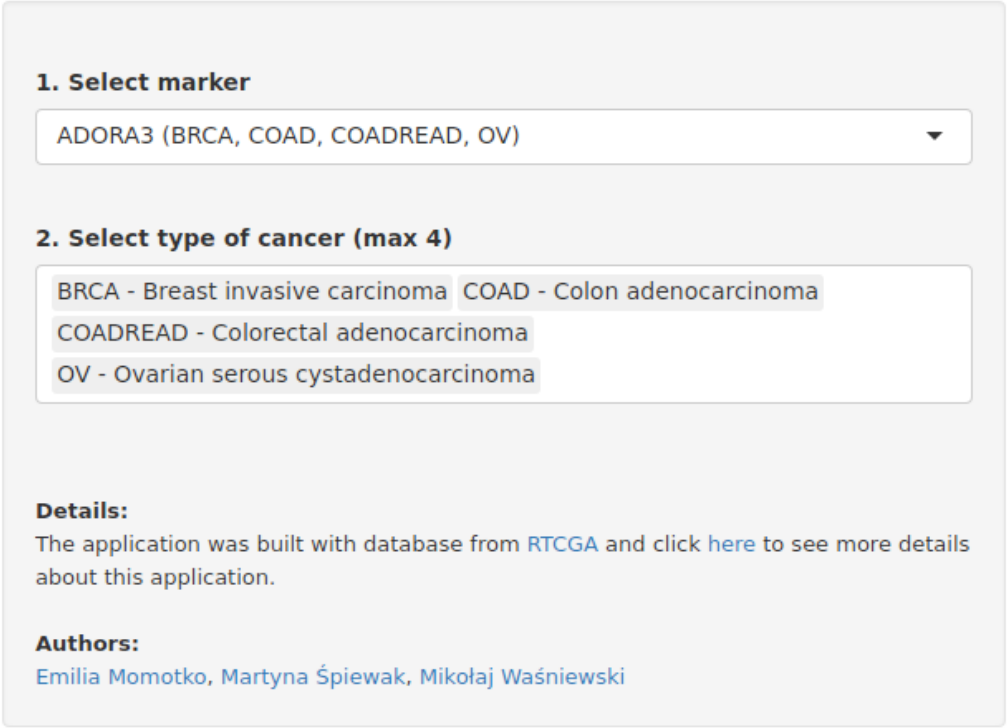

Figure 1: Side panel: select marker and types of cancer

Next, select the type of cancer. Basically, we assumed that the user can choose only four different types of cancer to comparison.

The application consists of four tabs: application manual, the Kaplan-Meier survival curve, the distribution of the marker values and the data set ready for download.

Kaplan-Meier survival curve

Modeling time to death is main object of this tab. Namely, we would like to estimate what is the proportion of patients who experienced a certain period of time (limit our analysis to the period of 10 years after cancer was observed) in each two groups designated by the threshold depending on the median of marker values (as described above).

To comparision survival curve in two group during the period of 10 years we used logrank test. The result of this test (showed as p-value of logrank test) is located in the bottom left corner of the plot. If p-value is coloured in red, it emphasizes the significance of selected marker.

As we can see the p-value in Figure 2 equals $6e-4$. It means we can reject null hypothesis about equality of survival curves (if the significant level is given, we assumed $\alpha = 0.05$). In other words, we can consider that survival time is different between two groups: `lower` and `higher`, the patients whose the value of marker was below the global median live on longer than the patients whose the value of marker was above the global median.

We note, after the period of five years after observing cancer survival is at approx. 60 % in the group `higher`, while up at 87 % in the group `lower`.

The second measurement used to comparison survival time is odds ratio between lower and higher group in selected point of time can also be the point of the interest (labeled as OR). The side panel contains slider specially for this tab. The additional vertical line, which appears on the plot, marks the time selected in the prompt number 3 (slider). The default value of this point equals half of whole analyzing time period, it means 5 year.

In Figure 2 odds ratio equals 3.28, so the chance of survival the period of 5 years after observing cancer for the person from the `lower` is more than three times higher than for the person from the `higher`.

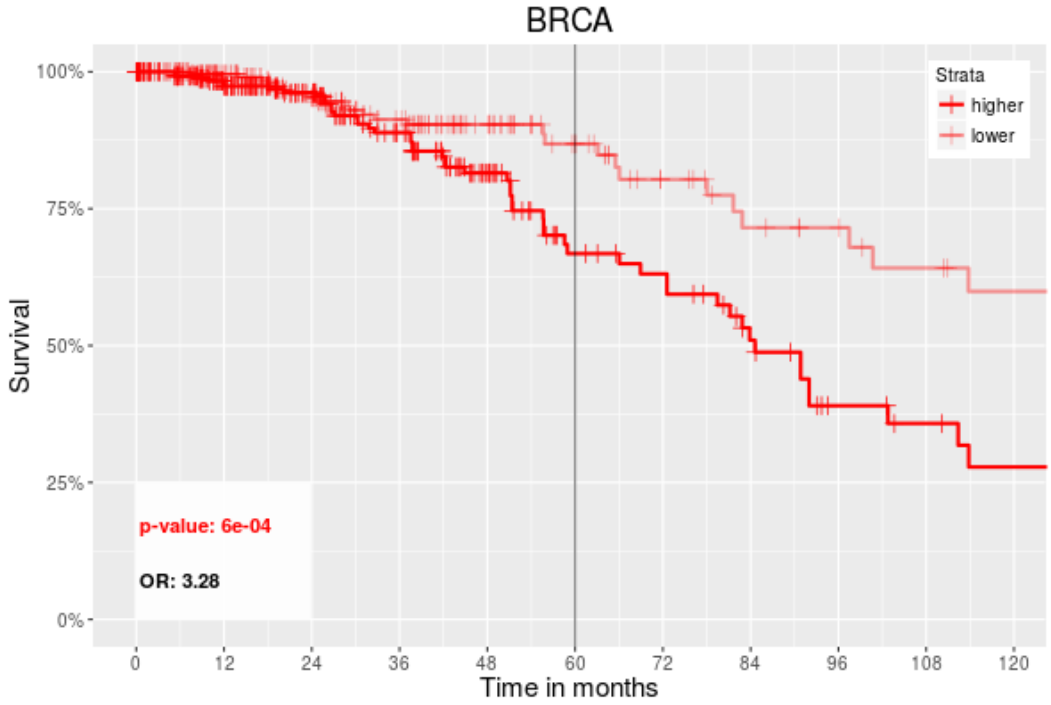

Figure 2: The Kaplan-Meier survival curve for marker ADORA3

Density & Box plot

Next tab shows a comparison of the distribution of the marker values for different types of cancer. The violin plot and box plot were used for this purpose. The advantage of this combination is the simultaneous comparison of multiple distribution characteristics. We can easily determine the symmetry or the skewness of the distribution, the tails behaviour, etc. Besides, the plot shows median and quartile values. Thus, that combination of two type of plot eases the comparison of the behaviour of selected marker across all cancers.

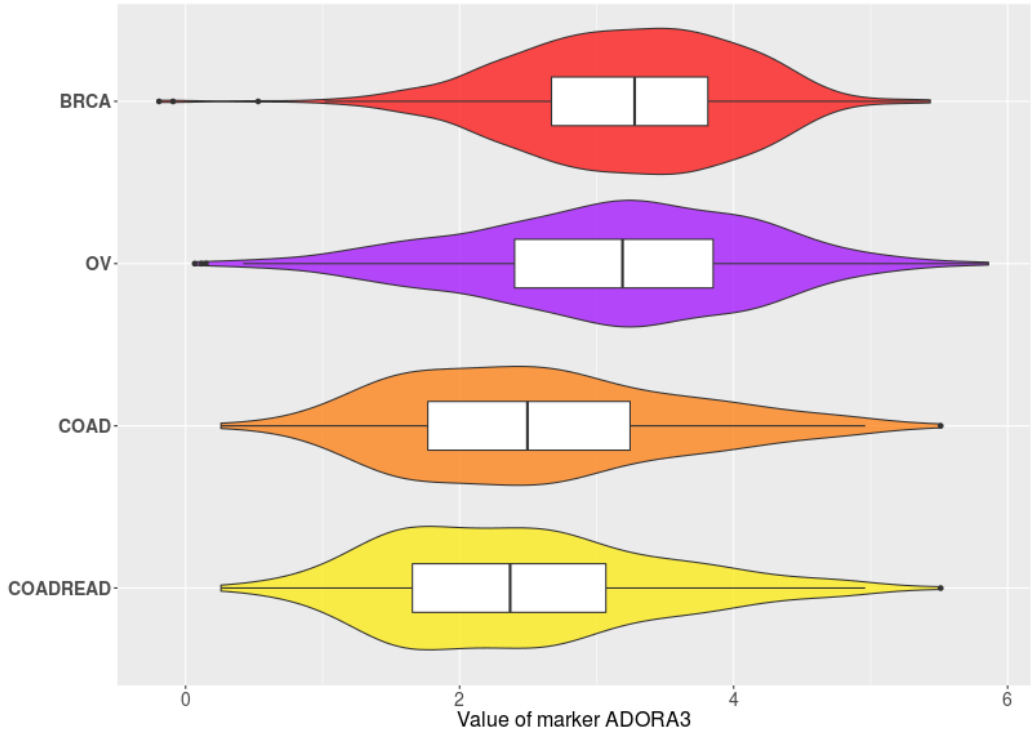

Figure 3: The distribution of the marker ADORA3 values

What conclusions can we draw?

First of all, the median values are different between the types of selected cancer. For cancer **BRCA** and **OV** the median values and the quartiles are similar. The same property holds for the second group **COAD** and **COADREAD**. Furthermore, the first two violin plots suggest that the distribution of the marker values is a right-skewed, while the next two plots indicate a left-skewed. Thus, the distribution of marker values are different across the types of cancer.

Download

The last tab contains the table with the relevant data. By clicking the button `Download` you can transfer this data into your disk and then perform another analysis. On the top of the page (just above the Download button), you will find the link which transfer you to the Wikipedia where information about chosen marker are included. On the other hand, if you would like to work on the entire database, you may visit the repository RTCG, where there are full details.

Summary

To sum up, the Shiny application **Gene Expression (`RTCG.mRNA`)** was built to compare the impact of biomarkers on treatment effect across the types of cancer. Secondary list was given at each marker allows the user to focus on these types of cancers for which the selected biomarker gave statistical significance results of logrank test. Additionally, in order to determine the effect of treatment in the application were given Kaplan-Meier survival curves and the distribution of marker values.

Then, the application is only simple tool that gives only primary results. If you would like to continue analysis you can feel free to download prepared dataset from `Download` tab in that you can analyze set-up data biomarker–type of cancer.

Bibliography

W. Chang, J. Cheng, J. Allaire, Y. Xie, and J. McPherson. *shiny: Web Application Framework for R*, 2015. URL https://CRAN.R-project.org/package=shiny. R package version 0.12.2. [p1]

C. Dardis. *survMisc: Miscellaneous Functions for Survival Data*, 2015. URL https://CRAN.R-project.org/package=survMisc. R package version 0.4.6. [p1]

Emilia Momotko
MiNI, The Warsaw University of Technology
Koszykowa 75
Warsaw, Poland
momotkoe@student.mini.pw.edu.pl

Martyna Śpiewak
MiNI, The Warsaw University of Technology
Koszykowa 75
Warsaw, Poland
spiewakm2@student.mini.pw.edu.pl

Mikołaj Waśniewski
MiNI, The Warsaw University of Technology
Koszykowa 75
Warsaw, Poland
wasniewskim@student.mini.pw.edu.pl