

Ekspresja genów (RTCGA.mRNA)

by Emilia Momotko, Martyna Śpiewak, Mikołaj Waśniewski

Abstract Celem projektu jest identyfikacja czynników, które różnicują prognozy pacjentów w rokowaniach czasu przeżycia w zależności od tła genetycznego. Platformą, która opisuje wybrane przez nas dane genetyczne jest platforma RTCGA.mRNA, przedstawiająca informacje dotyczące ekspresji genów wśród pacjentów u których zaobserwowano nowotwór.

Dane

Wszelkie zbiory danych użyte w analizie pochodzą z repozytorium [RTCGA](#).

Analizie poddaliśmy następujące zbiory danych, każdy z nich odpowiada innemu rodzajowi nowotworu:

- **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.

Identyfikacja istotnych biomarkerów

Do uzyskania informacji o biomarkerach i ich wpływie na prognozę leczenia posłużyliśmy się [testem logrank](#). Dzięki funkcji `survdiff` z pakietu `survMisc` ([Dardis, 2015](#)) byliśmy w stanie wyznaczyć wartość p-value tegoż testu.

Dla każdego rodzaju nowotworu i dla każdego genu oceniliśmy różnicę w czasie przeżycia. Pacjenci zostali podzieleni na **2 grupy**. W **pierwszej** znaleźli się ci, których wartość ekspresji genów była poniżej mediany z dostępnych wartości ekspresji (grupa została oznaczona jako `lower`), natomiast do **drugiej** grupy zostali przydzieleni pozostali uczestnicy badania, czyli ci, dla których wartości ekspresji były wyższe lub równe medianie (grupa oznaczona jako `higher`).

W pliku `biomarkers.csv` znajduje się lista najbardziej istotnych markerów dla ekspresji genów. Markery zostały wybrane w ten sposób, że dla każdego z 13 powyższych typów raka wyznaczyliśmy 100 najbardziej istotnych markerów, posługując się wartością p-value z testu logrank (innymi słowy, wybraliśmy 100 markerów z najmniejszą wartością p-value). Ostateczny zbiór istotnych markerów powstał jako suma istotnych markerów po wszystkich typach raka. Otrzymaliśmy w ten sposób 1089 istotnych markerów.

Aplikacja Shiny

W celu przedstawienia wyników naszej analizy, stworzyliśmy aplikację Shiny ([Chang et al., 2015](#)). Aplikacja znajduje się pod adresem <http://mi2.mini.pw.edu.pl:8080/RTCGA/MMM/shiny/> lub można ją pobrać z [repozytorium](#).

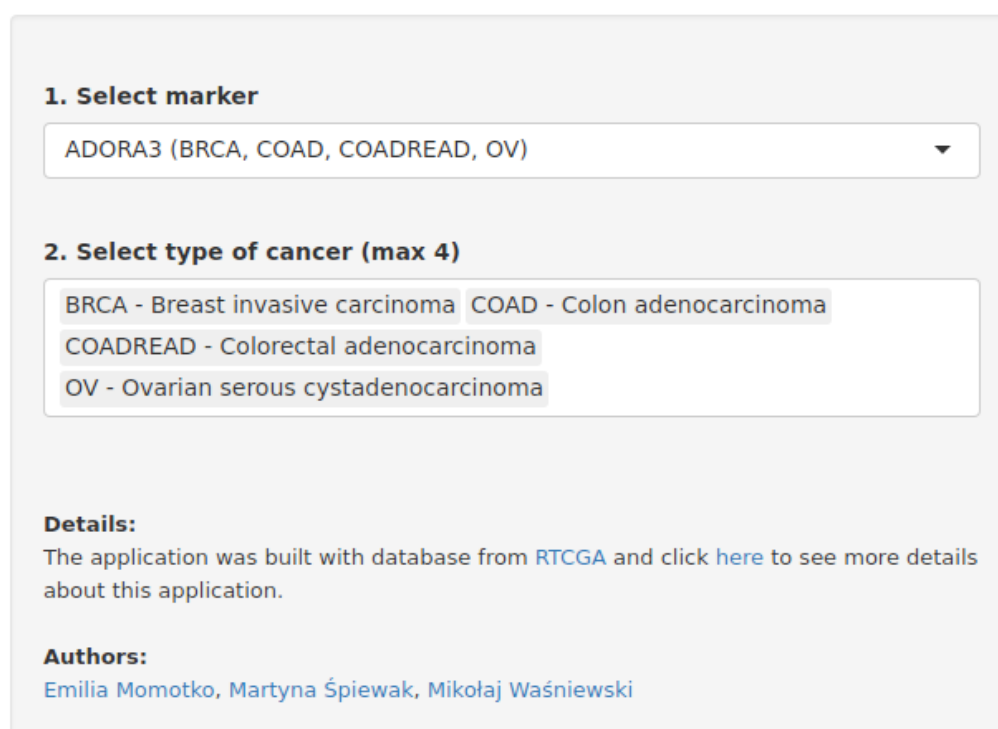
Głównym zamysłem naszej aplikacji jest możliwość porównaniu istotności pojedynczego markera względem różnych typów raka.

Jak korzystać z aplikacji?

Na początku wybieramy marker, dla którego chcemy przeanalizować wartości ekspresji przy różnych typach raka. W tym przypadku założyliśmy, że możemy wybrać dokładnie jeden marker. Ponadto, przy każdej nazwie markera w nawiasie znajdują się nazwy rodzajów raka, dla których ten marker okazał się istotnym czynnikiem w analizie przeżycia. Na Rysunku 1 widzimy, że marker `ADORA3` jest

istotny statystycznie (względem testu logrank) dla następujących typów raka: BRCA, OV, COAD, COADREAD.

Następnie, wybieramy typy raków, dla których chcemy przeprowadzić analizę porównawczą względem wybranego markera. Tutaj użytkownik w celu analizy może wybrać maksymalnie 4 typy.



The screenshot shows a web application interface with a light gray background. It contains two main sections for selection:

- 1. Select marker**: A dropdown menu with the text "ADORA3 (BRCA, COAD, COADREAD, OV)" and a downward arrow.
- 2. Select type of cancer (max 4)**: A list of four cancer types, each in a light gray box:
 - BRCA - Breast invasive carcinoma
 - COAD - Colon adenocarcinoma
 - COADREAD - Colorectal adenocarcinoma
 - OV - Ovarian serous cystadenocarcinoma

Below these sections, there is a **Details:** section with text stating the application was built with a database from RTCGA and a link to see more details. At the bottom is an **Authors:** section listing Emilia Momotko, Martyna Śpiewak, and Mikołaj Waśniewski.

Figure 1: Panel boczny: wybór markera i typów raka

Sama aplikacja składa się z czterech zakładek: instrukcja obsługi korzystania z aplikacji, krzywa przeżycia Kaplana Meiera, rozkład wartości markera, zbiór danych przygotowany do pobrania.

Krzywa przeżycia Kaplana Meiera

W tym przypadku, modeluje czas do zgonu, mianowicie szacujemy jaki jest odsetek pacjentów, którzy przeżyli określony okres czasu (ograniczamy się w analizie do okresu 10 lat), w każdej z dwóch grup wyznaczonych przez odpowiednio dobrany próg związany z wartością ekspresji genu (jak zostało opisane powyżej).

Do porównania krzywych przeżycia w grupach w okresie 10 lat, wykorzystaliśmy **test logrank**. Wynik tegoż testu znajduje się (w postaci wartości p-value) w dolnym lewym rogu ryciny. Kolor czerwony danego p-value podkreśla istotność wybranego markera.

Widzimy, że wartość p-value z Rysunku 2 wynosi $6e - 4$. Wówczas, przy ustalonych poziomie istotności $\alpha = 0.05$, mamy podstawy do odrzucenia hipotezy zerowej o równości krzywych przeżycia. Innymi słowy, możemy sądzić, że czas przeżycia różni się istotnie w obu grupach, pacjenci z wyższą wartością markera żyją dłużej niż pacjenci z wartościami niższymi niż mediana tych wartości. Za-uważmy ponadto, że po okresie 5 lat obserwowania pacjenta po stwierdzeniu u niego raka piersi, w grupie higher przeżywalność jest na poziomie ok. 60%, natomiast w grupie lower aż na poziomie 87%.

Drugą użytą miarą porównania czasu przeżycia jest **iloraz szans** między grupą lower a higher (ang. *odds ratio*) w określonym punkcie czasowym. W tym celu na panelu bocznym pojawił się suwak, który pozwala na wybór punktu odcięcia, domyślnie jest on ustawiony w połowie analizowanego czasu przeżycia, czyli 5 lat.

Na powyższych wykresie, iloraz szans wynosi 3.28, to znaczy, że szansa przeżycia 5 lat od stwierdzenia nowotworu dla osoby z grupy lower jest ponad trzykrotnie większa niż dla osoby z grupy higher.

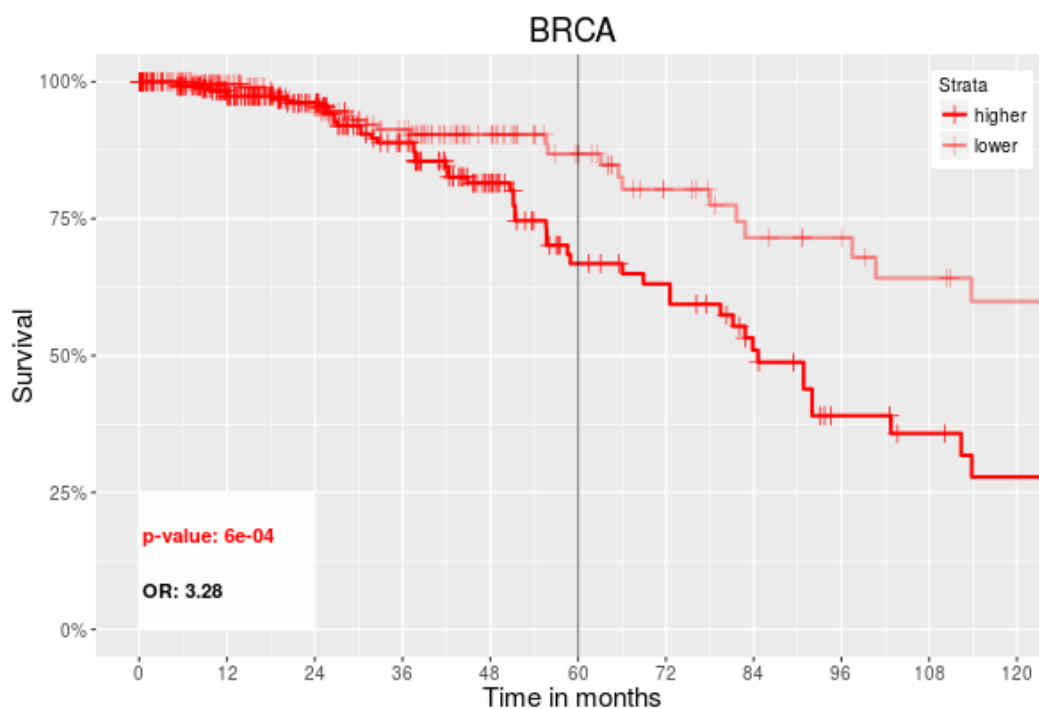


Figure 2: Krzywa przeżycia Kaplana - Meiera dla markera: ADORA3

Rozkład wartości wybranego markera

Kolejna zakładka przedstawia porównanie rozkładu wartości markera. Do tego celu użyliśmy wykresów skrzypcowych (ang. *violin plot*) oraz wykresów skrzynkowych (ang. *box plot*). Plusem takiego połączenia jest jednocześnie porównywanie wielu charakterystyk rozkładu, czy rozkład jest symetryczny, czy skośny, jak zachowuje się w ogonach, ponadto obserwujemy wartość mediany i pozostałych kwartyli. Możemy również porównywać powyższe własności między różnymi typami raków.

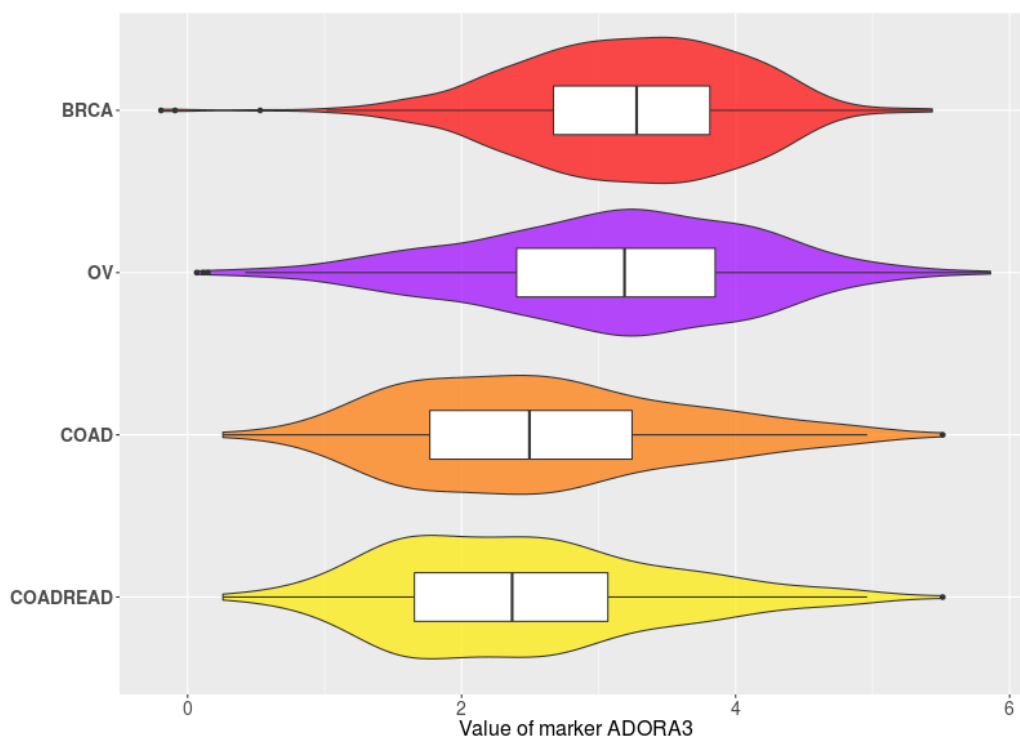


Figure 3: Rozkład wartości ekspresji genu

Co widzimy?

Przed wszystkim wartości mediany różnią się względem typów raka. Dla raka **BRCA** oraz **OV** wartości mediany i pozostałych kwartyli są zbliżone. Podobnie w grupie **COAD** oraz **COADREAD**. Ponadto pierwsze dwa wykresy skrzypcowe sugerują, że rozkład wartości markera jest prawoskośny, natomiast kolejne dwa wykresy wskazują na lewoskośność rozkładu. Zatem rozkład wartości ekspresji genu różni się między rodzajami raka.

Zbiór danych

Przewidzieliśmy sytuację, w której użytkownik chciałby kontynuować analizę istotności ekspresji genów dla różnych typów raka. W tym celu stworzyliśmy zakładkę Data. W zakładce znajduje się zbiór danych użyty do wykonania krzywych przeżycia oraz wykresów skrzypcowych i wykresów skrzynkowych. Dane odnoszą się tylko do wybranego markera i typów raka. Po naciśnięciu przycisku Download, zbiór zostanie pobrany na komputer użytkownika. Od tej pory, mamy możliwość samodzielnej pracy z danymi. Natomiast, jeśli użytkownik chciałby pracować na całym zbiorze danych, może odwiedzić repozytorium [RTCGA](#), gdzie znajdują się pełne dane.

Podsumowanie

Reasumując, aplikacja Shiny **Expression Gene (RTCGA.mRNA)** służy do porównania wpływu biomarkerów na efekt leczenia ze względu na typ raka. Pomocnicza lista, dana przy każdym markerze, pozwala skupiać się użytkownikowi na tych typach raków, dla których wybrany biomarker daje wyniki istotne statystycznie testu logrank. W celu stwierdzenia wpływu biomarkerów na efekt leczenia, w aplikacji zostały zamieszczone krzywe przeżycia Kaplana-Meiera oraz rozkład wartości ekspresji genu.

Zatem, aplikacja służy jako narzędzie pomocnicze, dające jedynie wstępne, obrazowe wyniki prowadzonej analizy istotności biomarkerów. W celu głębszej analizy, użytkownik może pobrać dane dotyczące danej konfiguracji biomarker–typ raka.

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 Politechnika Warszawska
Koszykowa 75
Warszawa, Polska
momotkoe@student.mini.pw.edu.pl

Martyna Śpiewak
MiNI Politechnika Warszawska
Koszykowa 75
Warszawa, Polska
spiewakm2@student.mini.pw.edu.pl

Mikołaj Waśniewski
MiNI Politechnika Warszawska
Koszykowa 75
Warszawa, Polska
wasniewskim@student.mini.pw.edu.pl