Applications of RTCGA package for

The Cancer Genome Atlas data integration

Marcin Kosiński[1], **Przemysław Biecek**[2]
[1]Faculty of Mathematics and Information Sciences, Warsaw University of Technology
[2]Interdisciplinary centre for mathematical and computational modelling, University of Warsaw
kosinskim@student.mini.pw.edu.pl, pbi@icm.edu.pl

Introduction

The Cancer Genome Atlas (TCGA) Data Portal provides a platform for researchers to search, download, and analyze data sets generated by TCGA. It contains clinical information, genomic characterization data, and high level sequence analysis of the tumor genomes[1].

RTCGA package offers download and integration of the variety and volume of TCGA data using patient barcode key, what enables easier data possession. This may have an benefcial infuence on impact on development of science and improvement of patients' treatment. RTCGA is an open-source R package, available to download from Bioconductor[2]. Furthermore, RTCGA package transforms TCGA data to form which is convenient to use in R statistical package. Those data transformations can be a part of statistical analysis pipeline which can be more reproducible with RTCGA.

The key is to understand genomics to improve cancer care.

Finally, we show applications of this software to show how data driven analysis can reveal the therapy practice for the cancer and an example of survival data analysis which used RTCGA to collect data.

[1] https://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp
[2] http://www.bioconductor.org/

RTCGA functionalities

- Check what are available **dates** of TCGA data releases - `availableDates()`
- Check what are available **datasets** for specific cohort - `availableDataSets()`
- Is specific name of dataset available - `checkDataSetsAvailability()`
- **Download** the TCGA data for specific cohort and date - `downloadTCGA()`
- Check what are available **genes' names** for downloaded mutation data - `availableGenesNames()`
- Is specific name of gene avaiable - `checkGenesNamesAvailability()`
- For downloaded genes' mutations types data and clinical data, merge genes' mutation types with clinical data - `mergeTCGA_clinical_mutations()`
- For downloaded genes' expressions data and clinical data, merge genes' expressions with clinical data - `mergeTCGA_clinical_rnaseq()`
- Read clinical data, after download and possible merges, into `data.frame` - `read.delim()`

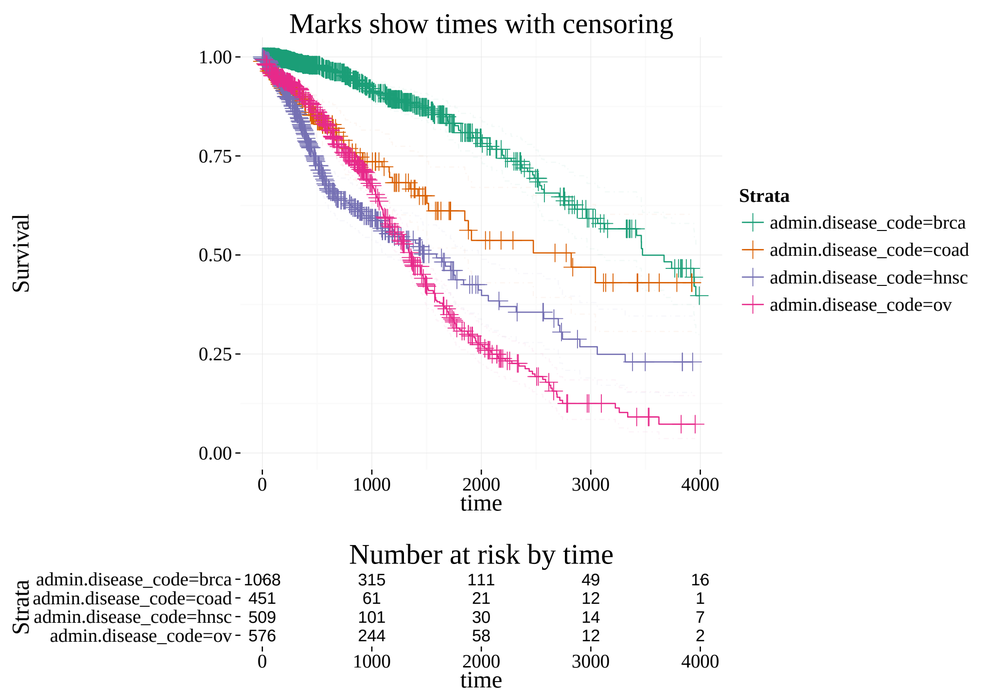
Usage example

```
if (!require(devtools)) { # package installation
install.packages("devtools")
library(devtools) }
install_github("MarcinKosinski/RTCGA")
library(RTCGA) dir.create( "data" )
# data download
downloadTCGA( cancerTypes = c("BRCA", "COAD", "OV", "HNSC"), destDir = "data/")
downloadTCGA( cancerTypes = c("BRCA", "COAD", "OV", "HNSC"),
dataSet = "Mutation_Packager_Calls.Level", destDir = "data/" )
downloadTCGA( cancerTypes = c("BRCA", "COAD", "OV", "HNSC"),
dataSet = Level_3__RSEM_genes_normalized__data.Level", destDir = "data/" )
# untarring data
list.files( "data/") %>% paste0( "data/", .) %>% sapply( untar, exdir = "data/" )
# adding expressions to clinical data
mergeTCGA_clinical_rnaseq( clinicalDir = ,rnaseqDir = ,genes = c("MDM2") )
# adding mutations to clinical data
mergeTCGA_clinical_mutations(clinicalDirHNSC, mutationDirHNSC, gene = "TP53")
# reading data clinicalHNSC <- read.clinical(clinicalDirHNSC)
```

Survival probabilities for cancer types

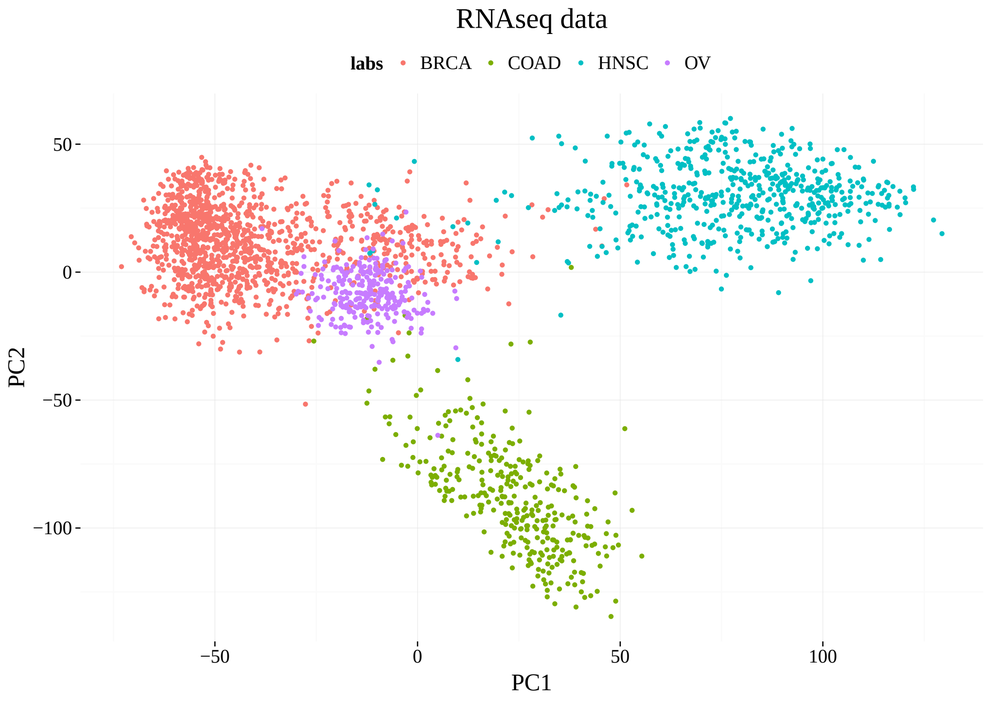
Below are shown Kaplan-Meier probability curves for 4 different cohorts. Used times are variables from clinical datasets:

- `patient.days_to_last_followup`for patients that were known to be alive
- `patient.days_to_death` for patient that were known to pass away



Map of cancers

Example of Principal Component Analysis for genes' expressions data from The Cancer Genome Atlas project. Below are shown 2 first principal components. Colors of the points correspond to the different cohorts.



Codes and Software

https://github.com/MarcinKosinski/RTCGA - source code
http://gdac.broadinstitute.org/ - The Cancer Genome Atlas data sets