Applications of Statistical Methods
to the Data from The Cancer Genome Atlas Study
Available in the RTCGA.data Family of R Packages

Marcin Kosiński[1,2], Marta Gładych[3,4], Maciej Wiznerowicz[3,4], Przemysław Biecek[1,2]

[1]Faculty of Mathematics and Information Sciences, Warsaw University of Technology
[2]Faculty of Mathematics, Informatics and Mechanics, University of Warsaw
[3]Gene Therapy Laboratory, Department of Cancer Immunology, Chair of Medical Biotechnology, Poznan University of Medical Sciences
[4]Gene Therapy Laboratory, Department of Cancer Immunology, Greater Poland Cancer Center

M.P.Kosinski@gmail.com   Software: https://github.com/mi2-warsaw/RTCGA.tools

Introduction

The following poster presents **RTCGA.data**: a family of R packages with data from The Cancer Genome Atlas Project (TCGA) study. TCGA is a comprehensive and coordinated effort to accelerate our understanding of the molecular basis of cancer through the application of genome analysis technologies, including large-scale genome sequencing[a]. We converted selected datasets from this study into few separate packages that are hosted on Bioconductor[b,c,d,e,f]. These R packages make selected datasets easier to access and manage. Data sets in RTCGA.data packages are large and cover complex relations between clinical outcomes and genetic background.

[a]http://cancergenome.nih.gov/
[b]http://bioc.ism.ac.jp/packages/devel/data/experiment/html/RTCGA.clinical.html
[c]http://bioc.ism.ac.jp/packages/devel/data/experiment/html/RTCGA.mutations.html
[d]http://bioc.ism.ac.jp/packages/devel/data/experiment/html/RTCGA.rnaseq.html
[e]http://bioc.ism.ac.jp/packages/devel/data/experiment/html/RTCGA.cnv.html
[f]http://bioc.ism.ac.jp/packages/devel/bioc/html/RTCGA.html

These packages will be useful for at least three audiences: biostatisticians that work with cancer data; researchers that are working on large scale algorithms, for them RTGCA data will be a perfect blasting site; teachers that are presenting data analysis method on real data problems. In this poster we present applications of the statistical modeling methods to the cancer's subtypes basing on the available cancer data of over 8,000 patients.

The key is to understand genomics to improve cancer care.

The Principal Components for RNAseqs

With data from RTCGA.data family and functions from **RTCGA.tools** it is simple to perform Principal Components Analysis of the RNAseq data for various cancer types and to plot 2 main components. Genes' expressions amongs **Adrenocortical carcinoma (ACC), Cholangiocarcinoma (CHOL), Glioma (GBM), Pheochromocytoma + Paraganglioma (PCPG) and Uveal Melanoma (UVM)** vary and samples group in view of cancer type.

This can be reproduced in R with

```
library(RTCGA.rnaseq)
rnaseqBiplot(c("ACC", "CHOL", "GBM", "PCPG", "UVM"))
```

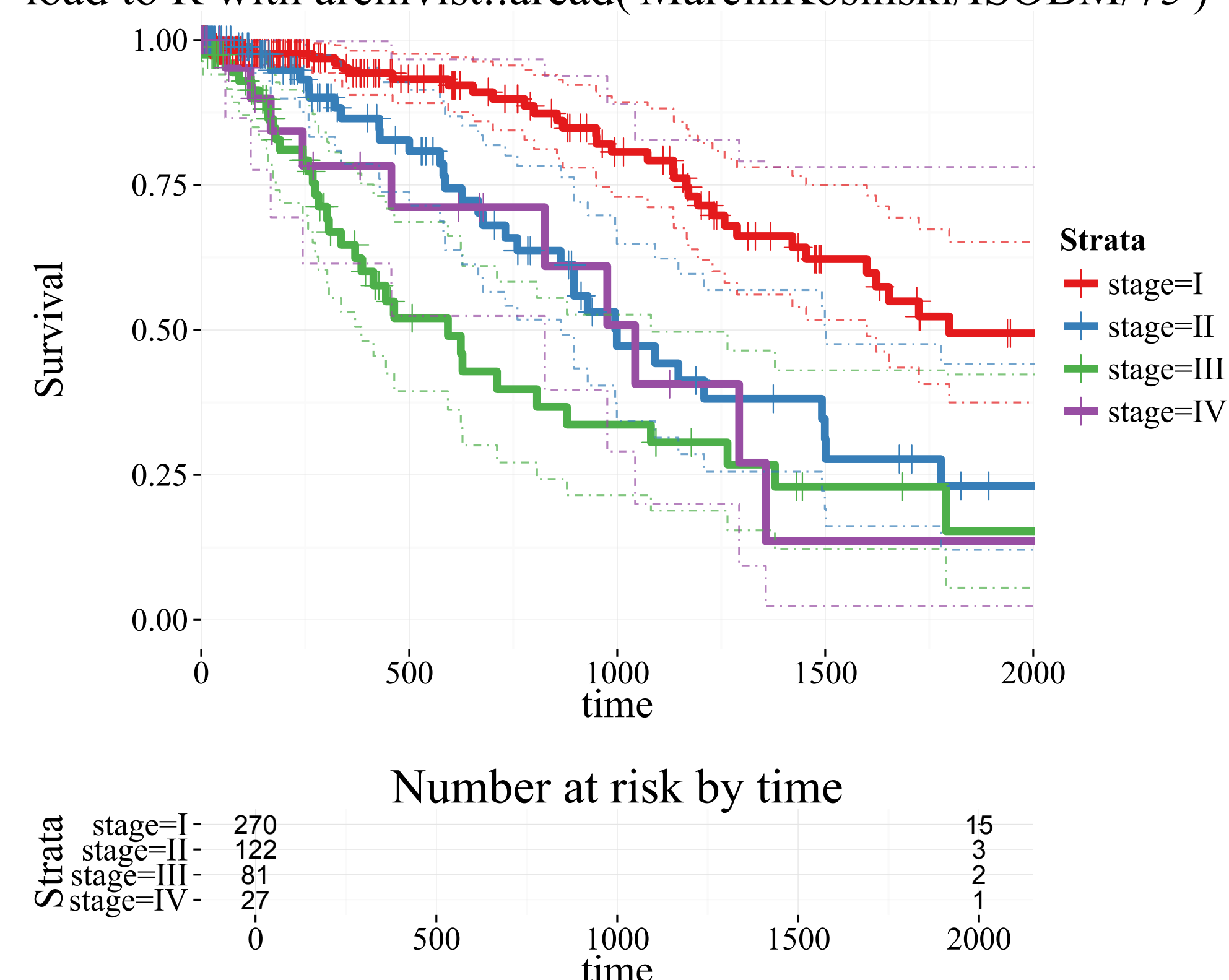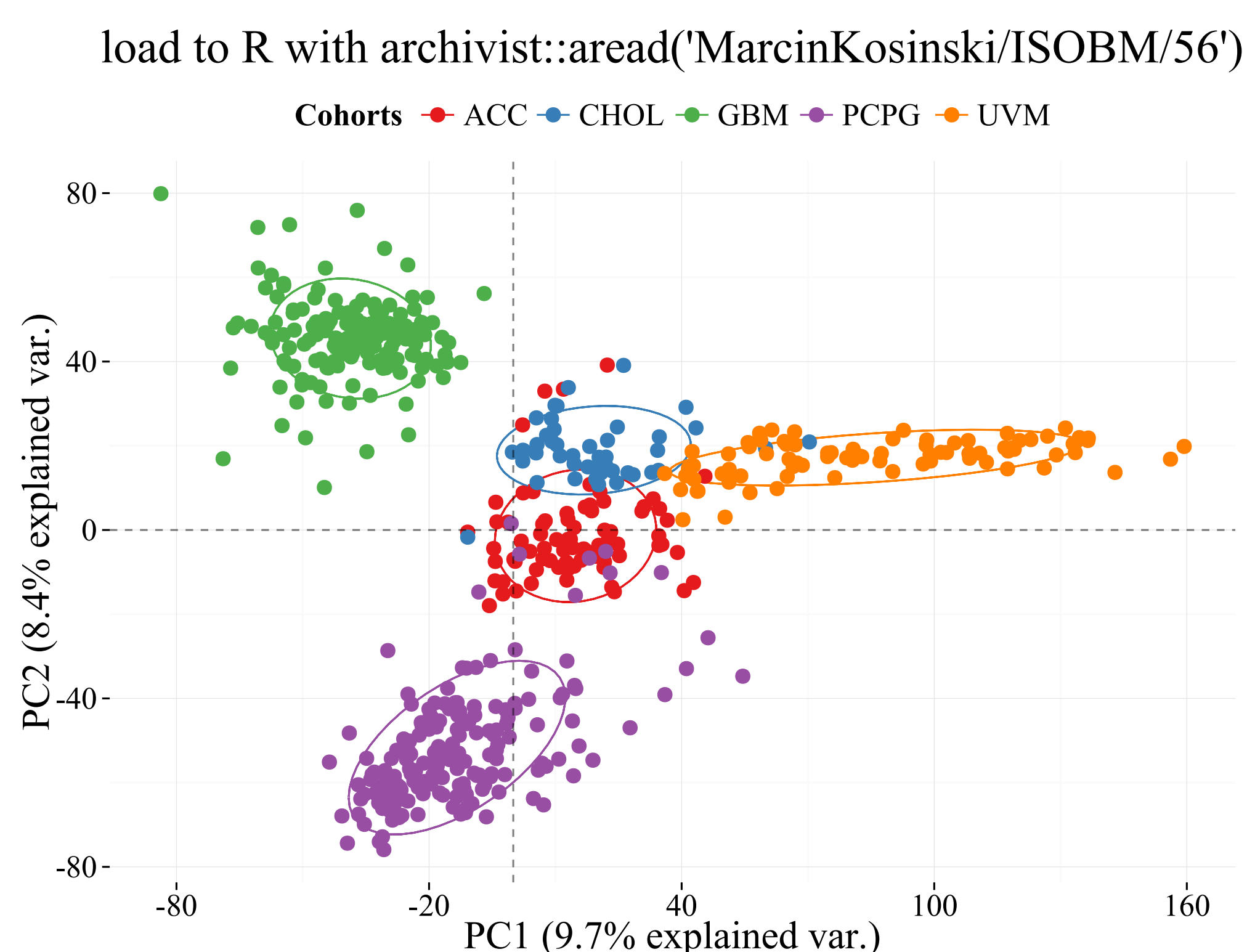load to R with archivist::aread('MarcinKosinski/ISOBM/56')

The Kaplan-Meier survival curves

RTCGA.data family is excellent when one researches in a field of survival analysis and genomics. Survival times are included in clinical datasets. This example plots Kaplan-Meier estimates of the survival functions for patients suffering from **Lung adenocarcinoma (LUAD)**, divided into stages of the cancer.

This can be reproduced in R with

```
library(RTCGA.clinical)
clinicalStageSurvival(LUAD.clinical, xlims = c(0,2000))
```

load to R with archivist::aread('MarcinKosinski/ISOBM/75')

Boxplots of Genes' Expressions

RTCGA.data family and RTCGA.tools package provide an easy to acces set of tools that allow to create useful figures in a simple R command. Below is an exmaple of boxplots for logarithm transformation for **ETF1** gene expression, divided on cancer types and 3 most popular levels of mutations in gene **TP53**.

This can be reproduced in R with

```
library(RTCGA.mutations);library(RTCGA.rnaseq)
mutationsBox(c('BRCA','HNSC','LUSC','PRAD'),'TP53','ETF1')
```

load to R with archivist::aread('MarcinKosinski/ISOBM/3b')