Bioinformatics, 32(6), 2016, 952–954
doi: 10.1093/bioinformatics/btv677
Advance Access Publication Date: 14 November 2015
Applications Note

OXFORD

Data and text mining

TCGA2STAT: simple TCGA data access for integrated statistical analysis in R

Ying-Wooi Wan[1,2], Genevera I. Allen[3,4] and Zhandong Liu[1,4,*]

[1]Computational and Integrative Biomedical Research Center, [2]Department of Obstetrics and Gynecology, Baylor College of Medicine, Houston, TX, USA, [3]Department of Statistics and Electrical & Computer Engineering, Rice University, Houston, TX, USA and [4]Department of Pediatrics-Neurology, Jan and Dan Duncan Neurological Research Institute, Texas Children's Hospital, Baylor College of Medicine, Houston, TX, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on August 19, 2015; revised on October 21, 2015; accepted on November 11, 2015

Abstract

Motivation: Massive amounts of high-throughput genomics data profiled from tumor samples were made publicly available by the Cancer Genome Atlas (TCGA).

Results: We have developed an open source software package, TCGA2STAT, to obtain the TCGA data, wrangle it, and pre-process it into a format ready for multivariate and integrated statistical analysis in the R environment. In a user-friendly format with one single function call, our package downloads and fully processes the desired TCGA data to be seamlessly integrated into a computational analysis pipeline. No further technical or biological knowledge is needed to utilize our software, thus making TCGA data easily accessible to data scientists without specific domain knowledge.

Availability and implementation: TCGA2STAT is available from the https://cran.r-project.org/web/packages/TCGA2STAT/index.html.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Contact: zhandong.liu@bcm.edu

1 Introduction

Over the last decade, The Cancer Genome Atlas (TCGA) consortium has measured large-scale genomics and clinical profiles of cancer patients so that scientists can study tumor genomes and decipher the genetic underpinnings of cancer. The TCGA data can be downloaded from web portals or via web services, such as the TCGA data portal (https://tcga-data.nci.nih.gov/tcga/), cBio (Cerami *et al.*, 2012; Gao *et al.*, 2013), canEvolve (Samur *et al.*, 2013), or Broad Institute GDAC Firehose (http://gdac.broadinstitute.org/). However, manual download of this massive data is time consuming and web service calls like the `firehose_get` function require additional program installation and technical setup. Most importantly, these two approaches cannot be easily integrated into a framework for statistical analysis. Many extra steps and technical knowledge of molecular platform data formats are needed to wrangle and pre-process the data before it can be statistically analyzed. Further, this process must be repeated when new data versions or additional samples become available, hindering efforts at version-control and reproducible research.

Others have provided software to obtain the TCGA data. cBio, for example, provides an `R` and `Matlab` package but was not designed to be used for genome-scale data analysis. It requires input of a list of genes from users and thus limits the exploratory use of the data. Another `R` package, `RTCGAToolbox` downloads TCGA data from Firehose (Samur, 2014), but the downloaded data is not pre-processed into data formats conducive for multivariate statistical analysis. Further, linking and merging functions necessary for integrated statistical analyses such as sample matching across multiple platforms and merging clinical and molecular data are not available in this package.

Because of these problems, use of the TCGA data can be limited to those with domain expertise, rendering the data inaccessible for general data scientists. In response, we have developed an `R` package

TCGA2STAT that makes the TCGA data in the open access data tier easily accessible to all by downloading, wrangling and pre-processing the data into a data matrix or list of matrices ready for multivariate or integrated statistical analyses. The package imports both molecular profiles and clinical data of more than thirty cancer types profiled with more than eleven high-throughput genomics platforms, including microarray, sequencing, methylation array, SNP array and array-CGH (See a full list in vignette). The imported molecular profiles and clinical data are automatically combined into a matrix for easy supervised analysis. Further, we provide functions for linking and merging samples from different molecular platforms that are necessary for integrated analysis. Most importantly, our package has one simple interface to perform all of these functions. The ease of obtaining this big biological data will encourage computational scientist to mine the TCGA data which in turn will bring new insights and breakthroughs in cancer research.

2 Implementation

The TCGA2STAT package is developed under the R statistical computing environment and is compatible with version 3.0 or later. The package uses HTTP calls to the Firehose site and parses necessary information to download the data. This is achieved via functions implemented through the XML package (Lang, 2013). The data imported is the latest version of all version-stamped standardized data sets hosted and maintained by Firehose, and usage of the data imported via this package constitutes an agreement with the TCGA data usage policy.

For some platforms, statistical analysis of data obtained directly from Firehose can be challenging without further data pre-processing. For example, mutation data in MAF files have patients repeated in multiple rows as multiple mutations are found, and the number of mutations differs across patients; this makes formatting mutation data into a data matrix difficult. Hence, TCGA2STAT filters mutation data based on status and variant classification and then aggregates the filtered data at the gene level. This yields a gene-by-sample data matrix with a value of one in cell (i,j) if a mutation is found in gene i and patient j, and zero otherwise. TCGA2STAT also wrangles and preprocesses data from other platforms such as CGH array and SNP array data into a gene by patient data matrix. Details of all wrangling and preprocessing steps are described in the package vignette (Supplementary Data).

3 Functions and examples

The TCGA2STAT package includes one major function and three additional utility functions:

- **getTCGA**. This is the main function of the package which obtains data from Firehose and processes the data into a matrix that can be used directly for statistical analysis in R. Only two inputs are required: the cancer type desired and the data platform or molecular profiling type desired. Our package supports over thirty different cancer types and eleven different data platforms. For example, `getTCGA(disease="OV", data.type="RNASeq2")` will obtain the RNASeq2 level III RSEM data from TCGA ovarian cancer patients. With the same function, the user can specify the specific types of data for each platform, such as counts instead of RPKM for RNASeq data. Also, users can choose to download the molecular profiles along with clinical data or filter the data by particular clinical covariates.
- **SampleSplit**. This function can be used to split the data imported via getTCGA into groups of samples profiled from the primary tumor, recurrent tumor, or normal/control groups.
- **OMICSBind**. For integrated statistical analysis of patients from two or more data platforms, this function can be used to merge the genomic data sets into coupled matrices with the same patient order.
- **TumorNormalMatch**. This function matches molecular profiles of the samples from primary tumor and normal/control groups in the same order, thus giving a data matrix ready for pair-wise statistical analyses.

These functions can be seamlessly integrated into R scripts for any statistical analysis of high-throughput genomics data. As an example shown in code snippets below, we use the TCGA2STAT package to download RNASeq2 and methylation data from lung squamous cell carcinoma (LUSC) patients (Part I), combine them for integrated analyses (Part II), and then use the combined R object for a simple canonical correlation analysis (Part III). The results are shown in Figure 1. A comprehensive walk-through of this example and code snippets for drawing the results in Figure 1 are given in the package vignette.

```
# Part I: Download NGS expression and methylation data
  for LUSC
methyl <- getTCGA(disease="LUSC",
  data.type="Methylation")
rnaseq2 <- getTCGA(disease="LUSC",
  data.type="RNASeq2", clinical=TRUE)
met.var <- apply(methyl$dat, met.var >=
   quantile(met.var, 0.99, na.rm=T)
   &!is.na(met.var))
rnaseq2.var <- apply(log10(1+rnaseq2$
 dat), 1, var)rnaseq2.var >=
   quantile(rnaseq2.var, 0.99, na.rm=T)
 &!is.na(rnaseq2.var))

# Part II: Merge the two data types for integrated
  analysis
met.rnaseq2 <- OMICSBind(dat1 = rnaseq.data, dat2=
met.data)
```

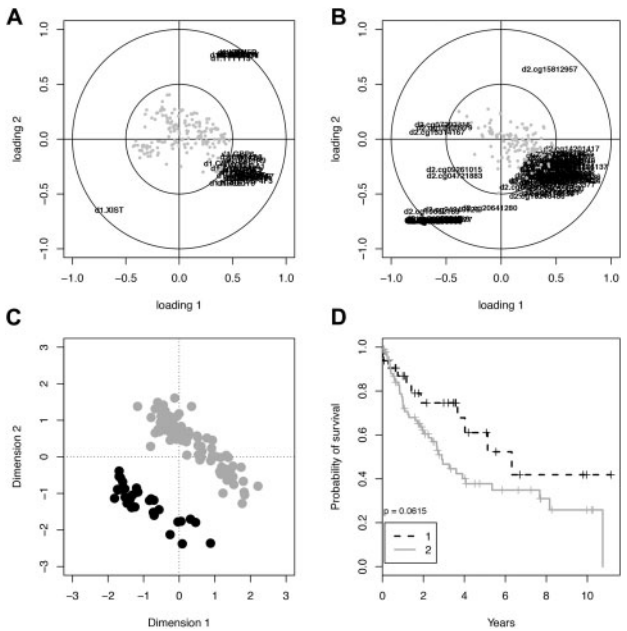

Fig. 1. Results of regularized canonical correlation analysis on gene and methylation expressions of TCGA LUSC patients. We show the canonical loadings of the first two dimensions on genes (**A**) and methylation regions (**B**). By projecting onto the canonical loadings, the LUSC patients can be separated into two groups (**C**) with distinct survival outcomes (**D**)

```
# Part III: Perform CCA on merged data, X and Y
lusc.cc <- rcc(t(met.rnaseq2$X), t(met.rnaseq2$Y),
  0.75025, 0.5005)
```

4 Conclusion

We have developed an R package that seamlessly downloads and preprocesses the TCGA data into objects ready for integrated statistical analysis. An advantage of this package is that users can obtain and maintain the large-scale TCGA data without additional technical knowledge other than R scripting. Our package will thus encourage many data scientists to mine this rich data source, potentially leading to breakthroughs in cancer genomics.

Funding

Z.L. and Y.W. are partially supported by the Houston Endowment and National Science Foundation (NSF) - Division of Mathematical Sciences (DMS) - 1263932; GA is supported by NSF DMS-1264058.

Conflict of Interest: none declared.

References

Cerami,E. *et al*. (2012) The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.*, **2**, 401–404.

Gao,J. *et al*. (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. *Sci. Signal.*, **6**, pl1–pl1.

Lang,D.T. (2013) *XML: Tools for parsing and generating XML within R and S-Plus*. R package version 3.98-1.1.

Samur,M.K. (2014) RTCGAToolbox: a new tool for exporting TCGA Firehose data. *PLoS ONE*, **9**, e106397.

Samur,M.K. *et al.* (2013) canEvolve: a web portal for integrative oncogenomics. *PLoS ONE*, **8**, e56228.