Bioinformatics, 33(19), 2017, 3131–3133
doi: 10.1093/bioinformatics/btx378
Advance Access Publication Date: 12 June 2017
Applications Note

Systems biology

CancerSubtypes: an R/Bioconductor package for molecular cancer subtype identification, validation and visualization

Taosheng Xu[1,†], Thuc Duy Le[2,3,*,†], Lin Liu[2], Ning Su[1], Rujing Wang[1], Bingyu Sun[1], Antonio Colaprico[4,5], Gianluca Bontempi[4,5] and Jiuyong Li[2,*]

[1]Institute of Intelligent Machines, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei 230031, China, [2]School of Information Technology and Mathematical Sciences, University of South Australia, Adelaide, SA 5095, Australia, [3]Centre for Cancer Biology, University of South Australia, Adelaide, SA 5000, Australia, [4]Interuniversity Institute of Bioinformatics in Brussels (IB)2 and [5]Machine Learning Group (MLG), Department d'Informatique, Université Libre de Bruxelles (ULB), 1050 Brussels, Belgium

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Inanc Birol

Received on February 27, 2017; revised on May 14, 2017; editorial decision on June 6, 2017; accepted on June 8, 2017

Abstract

Summary: Identifying molecular cancer subtypes from multi-omics data is an important step in the personalized medicine. We introduce *CancerSubtypes*, an R package for identifying cancer subtypes using multi-omics data, including gene expression, miRNA expression and DNA methylation data. *CancerSubtypes* integrates four main computational methods which are highly cited for cancer subtype identification and provides a standardized framework for data pre-processing, feature selection, and result follow-up analyses, including results computing, biology validation and visualization. The input and output of each step in the framework are packaged in the same data format, making it convenience to compare different methods. The package is useful for inferring cancer subtypes from an input genomic dataset, comparing the predictions from different well-known methods and testing new subtype discovery methods, as shown with different application scenarios in the Supplementary Material.
Availability and implementation: The package is implemented in R and available under GPL-2 license from the Bioconductor website (http://bioconductor.org/packages/CancerSubtypes/).
Contact: thuc.le@unisa.edu.au or jiuyong.li@unisa.edu.au
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

With the accumulation of a wealth of data from large-scale cancer genomic projects, e.g. The Cancer Genome Atlas (TCGA), identifying cancer molecular subtypes is becoming an important topic in cancer biology and a crucial step towards personalized treatment practices. Distinguishing molecular subtypes for a given cancer greatly assists cancer therapy by moving away from the 'one-size-fits-all' approach to patient care.

There have been several computational methods for identifying cancer subtypes from multi-omics data and they have been shown to be effective in grouping patients into different subtypes with distinct survival patterns. Despite their usefulness, most subtype discovery methods only demonstrated the analyses for some particular datasets used in their publications, making it difficult to compare the results with different studies. Moreover, each individual method requires different formats of the input dataset, e.g. a matrix of

features × *samples* in one method and a matrix of *samples* × *features* in another. They also generate the outputs with different formats, making it inconvenient for downstream analyses. There is a need to have a suite of cancer subtype discovery methods that can be repeated and compared for a wide range of datasets and generate outputs with the same format for follow-up analyses. Here we present the *CancerSubtypes R* package that implements the well-known cancer subtype discovery methods (four highly cited, one combined method, and one in-house method listed in the following section) using the same format of input and output. Moreover, we provide the pipelines for processing data and the feature selection methods to prepare the input for the subtyping algorithms. The package also includes the functions for validating and comparing the predictions and visualizing the results.

CancerSubtypes can be used in a wide range of scenarios, such as identifying cancer subtypes for different TCGA datasets using the different methods and comparing the results. We discuss the detailed usage and analysis process in four of the typical scenarios in the Supplementary Material. With the explosion of biological data generation, we hope that *CancerSubtypes* will speed up the research into computational approaches for disease subtype discovery and designing personalized treatments.

2 Implementation and main functions

2.1 Data pre-processing and feature selection

CancerSubtypes provides the basic data pre-process functions: distribution check, imputation and normalization. In most cases, genomic datasets are high-dimensional, and contain noise and missing values. Feature dimension reduction is needed to remove irrelevant features and reduce noises. We implement four feature selection methods: Variance (Var), median absolute deviation (MAD), COX model (David, 1972) and principal component analysis (PCA). Var, MAD and PCA are commonly used in many cancer genomic studies. Meanwhile, using the Cox model enables us to select important survival related features specific to cancer subtype studies. The output of data pre-processing and feature selection steps is a matrix that is ready for the downstream analyses.

2.2 Cancer subtype identification methods

CancerSubtypes contains four highly cited computational methods and a combined method for cancer subtype identification using genomic data.

- Consensus clustering (CC) (Monti *et al.*, 2003) is an unsupervised clustering method, which is frequently used and has several successful applications in cancer subtype discovery.
- Consensus non-negative matrix factorization (CNMF) (Brunet *et al.*, 2004) is an effective dimension reduction method used for finding molecular patterns from high-dimensional datasets.
- Integrative clustering (iCluster) (Shen *et al.*, 2009) uses a joint latent variable model for iCluster of multi-omics data.
- Similarity network fusion (SNF) (Wang *et al.*, 2014) is a method using SNF for aggregating multi-omics data to discover the similarities between patients.
- We propose a new method, SNF-CC to combine SNF and CC together to take the advantages of both for cancer subtype identification.
- Weighted SNF (WSNF) (Xu *et al.*, 2016) is similar to SNF but it takes the level of importance of genes into consideration. The gene weights are calculated based on the number of links the genes have in the miRNA-Transcription Factor-mRNA regulatory network.

Although CC and CNMF are designed for single-genomic datasets (e.g. gene expression datasets), iCluster, SNF, SNF-CC and WSNF focus on multi-omics data analysis. All these methods are repackaged to have the same format of the inputs and outputs for easy comparison of the methods.

2.3 Results validation and visualization

Validating and interpreting the identified cancer subtype results are important for understanding the significance of the findings and gaining insights into the causes of the subtypes. To evaluate and visualize the results, *CancerSubtypes* provides the following four statistical methods to present a credible computing and biology interpretation.

- Survival analysis is used to evaluate how well a method discriminates the survival patterns from different subtypes.
- Differential expression tests the expression difference between each subtype and a reference group (e.g. a set of normal samples).
- Statistical significance of clustering (Liu *et al.*, 2008) tests the significance of the difference in data distribution between subtypes.
- Silhouette width (Rousseeuw, 1987) is used to measure how well a sample is matched to its identified subtype compared to other subtypes. A high Silhouette value indicates that the sample is well matched.

Figure 1 shows a group analysis result of the evaluation functions (Survival analysis and Silhouette width) using the TCGA glioblastoma multiforme dataset (contained in the Bioconductor R package *CancerSubtypes*).

3 Applications, future works and conclusion

We present in the Supplementary Material four typical applications of the package, including identifying cancer subtypes using TCGA datasets, applying different features (genes) selection methods for subtype discovery, comparing the results and the performance of different

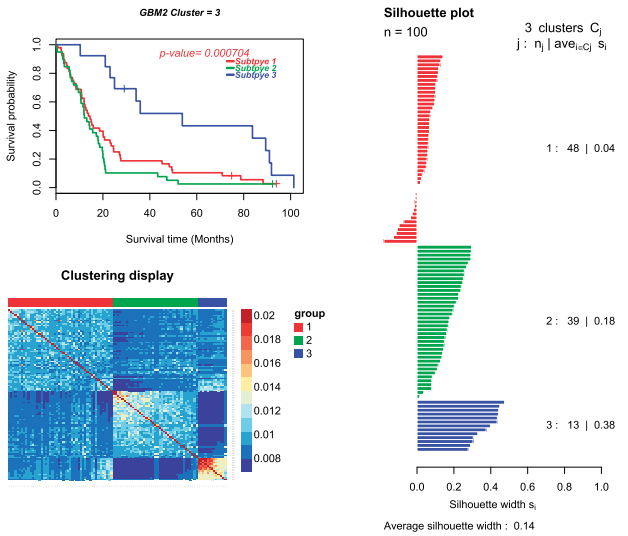

Fig. 1. The survival curves, heatmap of the sample similarity matrix and Silhouette width plots of the identified cancer subtypes for the TCGA glioblastoma multiforme dataset. The three subtypes are obtained by using the COX model for feature selection and SNF for cancer subtype identification

commonly used cancer subtype discovery methods, and examining the impact of different genomic data types in cancer subtype identification.

Although the *CancerSubtypes* gives the users the flexibility of customizing and running the codes in their local computers, our future works will provide tools for non-technical users. The first possible future work is to create a web-server to allow users explore cancer subtypes using the built-in datasets, e.g. TCGA datasets. Users would not need to upload the input data, and they have the option to choose different computational methods of identifying cancer subtypes in combination with different feature selection methods. However, this application will be limited to the datasets we stored in our server. The second direction is to make our package available (as an application) to Cancer Genomic Cloud http://www.cancergenomicscloud.org/, where the TCGA and other datasets have been made available for access from the applications and the computing resources are much more abundant than a private single server, and therefore making it more convenient for users to use our application.

In conclusion, we have developed the Bioconductor package, *CancerSubtypes* to provide a suite of cancer subtype analysis tools and embed the analyses in a standardized framework. The package will serve a wide range of users, including biologists, bioinformaticians and computational biologists, in analyzing cancer subtypes from data.

Funding

This work was supported by Australian Research Council (http://www.arc.gov.au/) Discovery Project [DP140103617]. T.D.L. was supported by NHMRC [grant ID 1123042]. Binyu Sun was supported by National Natural Science Foundation of China [ID 31371340]. Antonio Colaprico and Gianluca Bontempi were supported by the BridgeIRIS project, (http://mlg.ulb.ac.be/BridgeIRIS) funded by INNOVIRIS, Region de Bruxelles Capitale, Brussels, Belgium and by GENomic profiling of Gastrointestinal Inflammatory-Sensitive CANcers (GENGISCAN) http://mlg.ulb.ac.be/GENGISCAN, Belgian FNRS PDR [T100914F to G.B.].

Conflict of Interest: none declared.

References

Brunet,J.-P. *et al.* (2004) Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. USA*, **101**, 4164–4169.

David,C.R. (1972) Regression models and life tables (with discussion). *J. R Stat. Soc.*, **34**, 187–220.

Liu,Y. *et al.* (2008) Statistical significance of clustering for high-dimension, low–sample size data. *J. Am. Stat. Assoc.*, **103**, 1281–1293.

Monti,S. *et al.* (2003) Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn.*, **52**, 91–118.

Rousseeuw,P.J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, **20**, 53–65.

Shen,R. *et al.* (2009) Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, **25**, 2906–2912.

Wang,B. *et al.* (2014) Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods*, **11**, 333–337.

Xu,T. *et al.* (2016) Identifying cancer subtypes from mirna-tf-mrna regulatory networks and expression data. *PloS One*, **11**, e0152792.