Bioinformatics, 34(14), 2018, 2515–2517
doi: 10.1093/bioinformatics/bty124
Advance Access Publication Date: 2 March 2018
Applications Note

OXFORD

Data and text mining

GDCRNATools: an R/Bioconductor package for integrative analysis of lncRNA, miRNA and mRNA data in GDC

Ruidong Li[1,2,†], Han Qu[1,†], Shibo Wang[1], Julong Wei[1,3], Le Zhang[1,2], Renyuan Ma[1,4], Jianming Lu[1,5], Jianguo Zhu[6,*], Wei-De Zhong[5,*] and Zhenyu Jia[1,2,*]

[1]Department of Botany and Plant Sciences, [2]Genetics, Genomics, and Bioinformatics Program, University of California, Riverside, CA 92521, USA, [3]College of Animal Science and Technology, Nanjing Agricultural University, Nanjing, Jiangsu, 210095, China, [4]Department of Mathematics, Bowdoin College, Brunswick, ME 04011, USA, [5]Department of Urology, Guangdong Key Laboratory of Clinical Molecular Medicine and Diagnostics, Guangzhou First People's Hospital, The Second Affiliated Hospital of South China University of Technology, Guangzhou, 510180, China and [6]Department of Urology, Guizhou Provincial People's Hospital, Guizhou, 550002, China

*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: Jonathan Wren

Received on December 6, 2017; revised on February 5, 2018; editorial decision on February 27, 2018; accepted on February 28, 2018

Abstract

Motivation: The large-scale multidimensional omics data in the Genomic Data Commons (GDC) provides opportunities to investigate the crosstalk among different RNA species and their regulatory mechanisms in cancers. Easy-to-use bioinformatics pipelines are needed to facilitate such studies.
Results: We have developed a user-friendly R/Bioconductor package, named *GDCRNATools*, for downloading, organizing and analyzing RNA data in GDC with an emphasis on deciphering the lncRNA-mRNA related competing endogenous RNAs regulatory network in cancers. Many widely used bioinformatics tools and databases are utilized in our package. Users can easily pack preferred downstream analysis pipelines or integrate their own pipelines into the workflow. Interactive *shiny* web apps built in *GDCRNATools* greatly improve visualization of results from the analysis.
Availability and implementation: *GDCRNATools* is an R/Bioconductor package that is freely available at Bioconductor (http://bioconductor.org/packages/devel/bioc/html/GDCRNATools.html). Detailed instructions, manual and example code are also available in Github (https://github.com/Jialab-UCR/GDCRNATools).
Contact: arthur.jia@ucr.edu or zhongwd2009@live.cn or doctorzhujianguo@163.com

1 Introduction

Competing endogenous RNAs (ceRNAs) are RNA molecules that regulate other RNA transcripts by competing for the shared miRNAs. Deregulation of ceRNA networks may lead to human diseases including cancers. Although functions of a few lncRNA-related ceRNAs have been reported to play critical roles in cancer development (Kumar *et al.*, 2014; Liu *et al.*, 2014), the regulatory mechanisms of a large portion of ceRNAs remain to be unraveled.

The Genomic Data Commons (GDC) provides the cancer research community with a repository of standardized genomic and clinical data from multiple programs including The Cancer Genome Atlas (TCGA), Therapeutically Applicable Research to Generate

Effective Treatments (TARGET), as well as data from the Foundation Medicine company. Tools such as *TCGA-Assembler* (Zhu *et al.*, 2014) and *TCGAbiolinks* (Colaprico *et al.*, 2016) that were initially developed for retrieving and analyzing TCGA data from DCC (Data Coordinating Center) have been updated to access GDC data. However, none of them offer a route for integrative analysis of RNA-seq and miRNA-seq data.

Here, we present a new R/Bioconductor package, *GDCRNATools*, for downloading, organizing and integrative analyzing RNA data in GDC (Fig. 1). A newly developed algorithm *spongeScan* (Furió-Tarí *et al.*, 2016) is used to predict miRNA response elements (MREs) in lncRNAs acting as ceRNAs. In addition, databases including *starBase v2.0* (Li *et al.*, 2014), *miRcode* (Jeggari *et al.*, 2012) and *miRTarBase* (Chou *et al.*, 2017) are also integrated and used as evidence basis for miRNA-mRNA and miRNA-lncRNA interactions in the package to identify ceRNAs in cancers. Besides ceRNAs network analysis, many routine analyses can be performed in *GDCRNATools*, including differential gene expression analysis, functional enrichment analysis and univariate survival analysis. *GDCRNATools* allows users to easily perform the comprehensive analysis or integrate their own pipelines such as molecular subtype classification, weighted correlation network analysis (WGCNA; Langfelder and Horvath, 2008) and TF-miRNA co-regulatory network analysis, etc. into the workflow.

2 Implementation and main functions

2.1 Data download and organization

To facilitate utilization of the up-to-date datasets in GDC efficiently, a few simple functions are provided to download and organize the data.

i. **gdcRNADownload** can download HTSeq-Counts data of RNA-seq and isoform quantification data of miRNA-seq automatically by simply specifying project ID and data type.
ii. **gdcParseMetadata** parses metadata associated with downloaded files to facilitate downstream analysis.
iii. **gdcRNAMerge** merges total read counts for 5p and 3p strands of miRNAs in isoform quantification data and HTSeq read counts of gene quantification data to single expression matrices, respectively.

2.2 ceRNAs network analysis

i. **gdcCEAnalysis** uses three criteria to identify competing lncRNA-mRNA pairs: (a) the number and hypergeometric probability of shared miRNAs by a lncRNA-mRNA pair, (b) the strength of positive expression correlation between lncRNA and mRNA and (c) the overall regulation similarity of all shared miRNAs on the lncRNA-mRNA pair.

To identify common miRNAs targeting both lncRNA and mRNA, three miRNA-mRNA interaction databases including *StarBase v2.0* (Li *et al.*, 2014), *miRcode* (Jeggari *et al.*, 2012), and *mirTarBase 7.0* (Chou *et al.*, 2017), as well as three miRNA-lncRNA interaction databases, including *StarBase v2.0* (Li *et al.*, 2014), *miRcode* (Jeggari *et al.*, 2012), and *spongeScan* (Furió-Tarí *et al.*, 2016) are incorporated and used in the *gdcCEAnalysis* function internally. Gene IDs in these databases are updated to the latest Ensembl 90 annotation of human genome, and unified mature miRNA IDs are updated based on the new release miRBase 21. *gdcCEAnalysis* also provides a portal *via* which the user-provided datasets of miRNA-mRNA and miRNA-lncRNA interactions (either predicted using other algorithms or validated through experiments) can be included and utilized for the ceRNAs regulatory network analysis. Pearson's correlation is calculated to measure

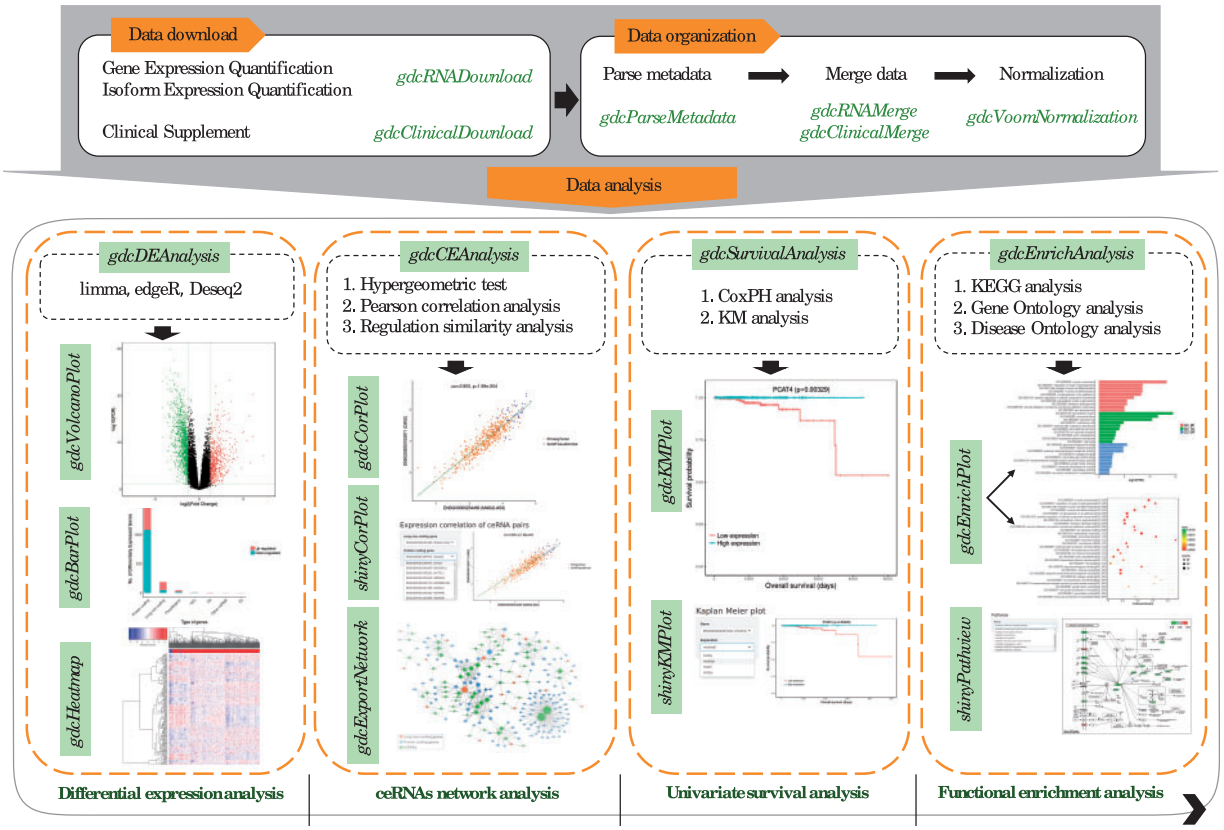

Fig. 1. Workflow of *GDCRNATools*

the expression correlation between lncRNA and mRNA. The overall regulation similarity of all shared miRNAs on the lncRNA-mRNA pair is defined as:

$$\text{Regulation similarity} = 1 - \frac{1}{M}\sum_{k=1}^{M}\left[\frac{|\text{corr}(m_k,l) - \text{corr}(m_k,g)|}{|\text{corr}(m_k,l)| + |\text{corr}(m_k,g)|}\right]^M$$

where M is the total number of shared miRNAs, m_k is the kth shared miRNAs with $k = 1, \ldots, M$, and $\text{corr}(m_k,l)$ and $corr(m_k,g)$ represents the Pearson's correlation between the kth miRNA with lncRNA, and with mRNA, respectively. *gdcCEAnalysis* can also compute sensitivity correlation (the difference between the Pearson's correlation and partial correlation coefficients) for each lncRNA-miRNA-mRNA triplet, defined as:

$$\text{Sensitivity correlation} = \text{corr}(l,g) - \frac{\text{corr}(l,g) - \text{corr}(m_k,l)\text{corr}(m_k,g)}{\sqrt{1 - \text{corr}(m_k,l)^2}\sqrt{1 - \text{corr}(m_k,g)^2}}$$

to measure the contribution of a miRNA in mediating the expression correlation between a lncRNA and mRNA (Paci *et al.*, 2014), where $\text{corr}(l,g)$ is the Pearson's correlation between lncRNA and mRNA.

2.3 Other downstream analyses

i. **gdcDEAnalysis** can implement three most commonly used methods: *limma* (Ritchie *et al.*, 2015), *edgeR* (Robinson *et al.*, 2010) and *DESeq2* (Love *et al.*, 2014) to identify differentially expressed genes (DEGs).
ii. **gdcEnrichAnalysis** performs Gene Ontology (GO) and Kyoto En-cyclopedia of Genes and Genomes (KEGG) functional enrichment analyses using the latest databases through the R/Bioconductor package *clusterProfiler* (Yu *et al.*, 2012). Disease Ontology analy-sis using *DOSE* package (Yu *et al.*, 2015) is also included in the *gdcEnrichAnalysis* function to detect gene-disease associations.
iii. **gdcSurvivalAnalysis** can perform both Cox Proportional-Hazards (CoxPH) regression and Kaplan Meier (KM) survival analyses reporting the hazard ratio, 95% confidence intervals and *P*-value for the tested genes on overall survival.

3 Conclusion

We have developed a novel R/Bioconductor package to conduct advanced analyses of RNA-seq and miRNA-seq data in GDC. This easy-to-use package allows users with little coding experience to perform the entire analysis smoothly. As standardized data from other programs would be submitted to GDC, we believe that *GDCRNATools* will gain ground in cancer research for deciphering the crosstalk among multiple RNA species and their regulatory mechanisms.

Funding

This work was supported by the Faculty Start-up Fund to Z.J. and UC Academic Senate Regents Faculty Fellowship to Z.J., and also partially supported by the grants from National Key Basic Research Program of China [2015CB553706], the Fundamental Research Funds for the Central Universities [2017PY023], the Projects of Guizhou Province [QKHW-G [2014]7004] and the National Natural Science Foundation of China [81571427, 81641102, 81660426, 81360119].

Conflict of Interest: none declared.

References

Chou,C.-H. *et al.* (2017) miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res.*, **46**, D296–D302

Colaprico,A. *et al.* (2016) TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.*, **44**, e71.

Furió-Tarí,P. *et al.* (2016) spongeScan: a web for detecting microRNA binding elements in lncRNA sequences. *Nucleic Acids Res.*, **44**, W176–W180.

Jeggari,A. *et al.* (2012) miRcode: a map of putative microRNA target sites in the long non-coding transcriptome. *Bioinformatics*, **28**, 2062–2063.

Kumar,M.S. *et al.* (2014) HMGA2 functions as a competing endogenous RNA to promote lung cancer progression. *Nature*, **505**, 212–217.

Langfelder,P. and Horvath,S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.

Li,J.-H. *et al.* (2014) starBase v2. 0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res.*, **42**, D92–D97.

Liu,X-h. *et al.* (2014) Lnc RNA HOTAIR functions as a competing endogenous RNA to regulate HER2 expression by sponging miR-331-3p in gastric cancer. *Mol. Cancer*, **13**, 92.

Love,M.I. *et al.* (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.

Paci,P. *et al.* (2014) Computational analysis identifies a sponge interaction network between long non-coding RNAs and messenger RNAs in human breast cancer. *BMC Syst. Biol.*, **8**, 83.

Ritchie,M.E. *et al.* (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47–e47.

Robinson,M.D. *et al.* (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Yu,G. *et al.* (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS J. Integrative Biol.*, **16**, 284–287.

Yu,G. *et al.* (2015) DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics*, **31**, 608–609.

Zhu,Y. *et al.* (2014) TCGA-assembler: open-source software for retrieving and processing TCGA data. *Nat. Methods*, **11**, 599–600.