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,*], Zhenyu Jia[1,*]

[1]Department of Botany and Plant Sciences, University of California, Riverside, CA, USA

[2]Genetics, Genomics, and Bioinformatics Program, University of California, Riverside, CA, USA

[3]College of Animal Science and Technology, Nanjing Agricultural University, Nanjing, Jiangsu, China

[4]Department of Mathematics, Bowdoin College, Brunswick, ME, 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, China

[6]Department of Urology, Guizhou Provincial People's Hospital, Guizhou, China

[¶] Equally contributed

*To whom correspondence should be addressed.

Abstract:

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. We have developed a user-friendly R/Bioconductor package, named *GDCRNATools*, to facilitate downloading, organizing, and analyzing RNA data in GDC with an emphasis on deciphering the lncRNA-mRNA related competing endogenous RNAs (ceRNAs) 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: *GDCRNATools* is an R/Bioconductor package that is freely available at 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 indirectly regulate other RNA transcripts by competing for the shared miRNAs. Deregulation of ceRNA networks may lead to human diseases, such as cancer (Gupta, et al., 2010; Ning, et al., 2015; Schmitt and Chang, 2016). Mounting evidences show that lncRNAs harboring multiple miRNA response elements (MREs) can act as ceRNAs to sequester miRNA activity and thus reduce the inhibition of miRNAs on their target genes (Kallen, et al., 2013; Salmena, et al., 2011). Although 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 and significances 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 National Cancer Institute (NCI) programs including The Cancer Genome Atlas (TCGA) and Therapeutically Applicable Research To Generate Effective Treatments (TARGET). High-quality datasets from non-NCI research programs such as genomic data from the Foundation Medicine company are also maintained in GDC. Advanced tools, such as *TCGA-Assembler* (Zhu, et al., 2014) and *TCGAbiolinks* (Colaprico, et al., 2016) that were initially developed for retrieving, processing and analyzing TCGA data from DCC (Data Coordinating Center) have been updated to access GDC data. However, none of these tools offer a route for a comprehensive analysis of RNA-seq and miRNA-seq data available in GDC.

Here we present a new R/Bioconductor package, named *GDCRNATools* to facilitate downloading, organizing, and integrative analyzing RNA data in the GDC (Fig. 1). Many analyses can be performed using *GDCRNATools* including differential gene expression analysis, ceRNAs regulatory network analysis, univariate survival analysis, and functional enrichment analysis. A newly developed algorithm

spongeScan (Furió-Tarí, et al., 2016) is used to predict MREs in lncRNAs acting as ceRNAs. In addition, databases including *starBase v2.0* (Li, et al., 2013), *miRcode* (Jeggari, et al., 2012) and *miRTarBase 7.0* (Chou, et al., 2017) were also integrated and used as evidence basis for miRNA-mRNA and miRNA-lncRNA interactions. We updated gene IDs in these databases according to the latest Ensembl 90 annotation of human genome, and unified mature miRNA IDs based on the new release miRBase 21. *GDCRNATools* allows users 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.

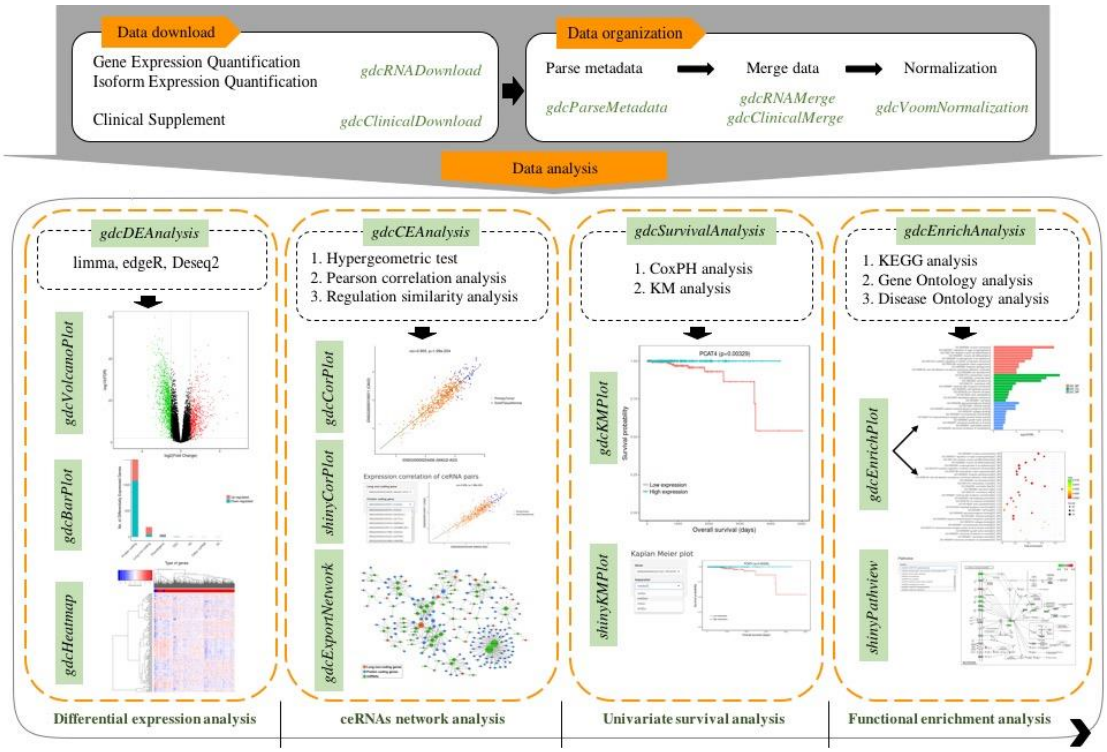

Fig. 1. Workflow of GDCRNATools

2. Implementation and main functions

2.1 Download and process data

Two strategies for downloading RNA expression data are available in the *gdcRNADownload* function: (1) users can simply provide the manifest file that contains Universally Unique Identifiers

(UUIDs) corresponding to data in the GDC cart, or (2) download data automatically by specifying project id and data type. Metadata associated with downloaded files can be easily parsed by *gdcParseMetadata* to facilitate downstream analysis. *gdcRNAMerge* function merges total read counts for 5p and 3p strands of miRNAs (processed from isoform quantification data) and HTSeq read counts of gene quantification data into single expression matrix, respectively. Clinical data can be downloaded and processed by the *gdcClinicalDownload* and *gdcClinicalMerge* functions.

2.2 Differential gene expression analysis

Three most commonly used methods: *limma* (Ritchie, et al., 2015), *edgeR* (Robinson, et al., 2010), and *DESeq2* (Love, et al., 2014) can be implemented in *gdcDEAnalysis* function to identify differentially expressed genes (DEGs). Gene symbols and biotypes based on the Ensembl 90 annotation are reported in the output of *gdcDEReport*.

2.3 ceRNAs regulatory network analysis

Three criteria are used to define competing lncRNA-mRNA pairs: (1) the number of shared miRNAs by a lncRNA-mRNA pair and hypergeometric probability associated with this number, (2) the strength of positive expression correlation between lncRNA and mRNA, and (3) the overall regulation similarity of the lncRNA-mRNAs pair mediated by shared miRNAs, which is defined as:

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

where M is the total number of shared miRNAs, m_k is the kth shared miRNAs with $k = 1, \ldots, M$, and $corr(m_k, l)$ and $corr(m_k, g)$ represents the Pearson's correlation between the kth miRNA with lncRNA, and with mRNA, respectively. Three miRNA-mRNA interaction databases (*StarBase v2.0* (Li, et al., 2013), *miRcode* (Jeggari, et al., 2012), *mirTarBase 7.0* (Chou, et al., 2017)) and three miRNA-lncRNA interaction databases (*StarBase v2.0* (Li, et al., 2013), *miRcode* (Jeggari, et al., 2012),

spongeScan (Furió-Tarí, et al., 2016)) are incorporated and used in the *gdcCEAnalysis* function internally. *gdcCEAnalysis* also allows user-provided data of miRNA-lncRNA or miRNA-mRNA interactions (either predicted using other algorithms or validated through experiments) to be utilized for the ceRNAs regulatory network analysis. The resultant lncRNA-miRNA-mRNA interaction network can be exported by the *gdcExportNetwork* function and then visualized in *Cytoscape* (Shannon, et al., 2003).

2.4 Functional enrichment analysis

gdcEnrichAnalysis performs Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) functional enrichment analyses using the latest databases through the R/Bioconductor package *clusterProfiler* (Yu, et al., 2012). Disease Ontology analysis using *DOSE* package (Yu, et al., 2014) is also included in the *gdcEnrichAnalysis* function to detect gene-disease associations.

2.5 Survival analysis

Two methods are available in *gdcSurvivalAnalysis* function to perform univariate survival analysis: (1) Cox Proportional-Hazards (CoxPH) regression and (2) Kaplan Meier (KM) analysis. *gdcSurvivalAnalysis* takes a list of genes as input and reports the hazard ratio, 95% confidence intervals, and p value of significance test on overall survival of each gene.

2.6 Visualization

In addition to the routine visualization methods, such as volcano plot, heatmap, KM plot, etc., a more attractive visualization feature of *GDCRNATools* is the application of interactive *shiny* web apps, which allow users to view survival curves, expression correlation between lncRNA and mRNA, and enriched pathways by simply selecting genes/pathways of interests on the local webpage.

3. Conclusion

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

References

Chou, C.-H., *et al.* miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Research* 2017.

Colaprico, A., *et al.* TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic acids research* 2016;44(8):e71-e71.

Furió-Tarí, P., *et al.* spongeScan: A web for detecting microRNA binding elements in lncRNA sequences. *Nucleic acids research* 2016;44(W1):W176-W180.

Gupta, R.A., *et al.* Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* 2010;464(7291):1071-1076.

Jeggari, A., Marks, D.S. and Larsson, E. miRcode: a map of putative microRNA target sites in the long non-coding transcriptome. *Bioinformatics* 2012;28(15):2062-2063.

Kallen, A.N., *et al.* The imprinted H19 lncRNA antagonizes let-7 microRNAs. *Molecular cell* 2013;52(1):101-112.

Kumar, M.S., *et al.* HMGA2 functions as a competing endogenous RNA to promote lung cancer progression. *Nature* 2014;505(7482):212-217.

Langfelder, P. and Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics* 2008;9(1):559.

Li, J.-H., *et al.* starBase v2. 0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data. *Nucleic acids research* 2013;42(D1):D92-D97.

Liu, X.-H., *et al.* Lnc RNA HOTAIR functions as a competing endogenous RNA to regulate HER2 expression by sponging miR-331-3p in gastric cancer. *Molecular cancer* 2014;13(1):92.

Love, M.I., Anders, S. and Huber, W. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* 2014;15(12):550.

Ning, S., *et al.* Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers. *Nucleic acids research* 2015;44(D1):D980-D985.

Ritchie, M.E., *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research* 2015;43(7):e47-e47.

Robinson, M.D., McCarthy, D.J. and Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;26(1):139-140.

Salmena, L., *et al.* A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* 2011;146(3):353-358.

Schmitt, A.M. and Chang, H.Y. Long noncoding RNAs in cancer pathways. *Cancer cell* 2016;29(4):452-463.

Shannon, P.*, et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* 2003;13(11):2498-2504.

Yu, G.*, et al.* clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology* 2012;16(5):284-287.

Yu, G.*, et al.* DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics* 2014;31(4):608-609.

Zhu, Y., Qiu, P. and Ji, Y. TCGA-assembler: open-source software for retrieving and processing TCGA data. *Nature methods* 2014;11(6):599-600.