CancerMIRNome: an interactive analysis and visualization database for miRNome profiles of human cancer

Ruidong Li[1,2,*], Han Qu[1], Shibo Wang[1], John M. Chater[1], Xuesong Wang[1,2], Yanru Cui[3], Lei Yu[1,2], Rui Zhou[1], Qiong Jia[1,2], Ryan Traband[1], Meiyue Wang[4], Dongbo Yuan[5], Jianguo Zhu[5,*], Wei-De Zhong[6,*], Zhenyu Jia[1,2,*]

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

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

[3] College of Agronomy, Hebei Agricultural University, Baoding, China

[4] Department of Anesthesiology, Perioperative and Pain Medicine, Stanford University School of Medicine, Stanford, CA, USA

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

[6] Department of Urology, Guangdong Key Laboratory of Clinical Molecular Medicine and Diagnostics, Guangzhou First People's Hospital, School of Medicine, South China University of Technology, Guangzhou, China

*To whom correspondence should be addressed: Ruidong Li at rli012@ucr.edu. Correspondence may also be addressed to Zhenyu Jia at arthur.jia@ucr.edu, Wei-De Zhong at zhongwd2009@live.cn, or Jianguo Zhu at doctorzhujianguo@163.com.

ABSTRACT

MicroRNAs (miRNAs), which play critical roles in gene regulatory networks, have emerged as promising biomarkers for a variety of human diseases, including cancer. In particular, circulating miRNAs that are secreted into circulation exist in remarkably stable forms, and have enormous potential to be leveraged as non-invasive diagnostic biomarkers for early cancer detection. The vast amount of miRNA expression data from tens of thousands of samples in various types of cancers generated by The Cancer Genome Atlas (TCGA) and circulating miRNA data produced by many large-scale circulating miRNA profiling studies provide extraordinary opportunities for the discovery and validation of miRNA signatures for cancer diagnosis and prognosis. Novel and user-friendly tools are desperately needed to facilitate the data mining of such valuable cancer miRNome datasets. To fill this void, we developed CancerMIRNome, a comprehensive database for the interactive analysis and visualization of cancer miRNome data based on TCGA and public circulating miRNome datasets. A series of cutting-edge bioinformatics tools and functions have been packaged in CancerMIRNome, allowing for a pan-cancer analysis of a miRNA of interest across multiple cancer types and a comprehensive analysis of cancer miRNome profiles to identify diagnostic and prognostic signatures. The CancerMIRNome database is publicly available at http://bioinfo.jialab-ucr.org/CancerMIRNome.

INTRODUCTION

miRNAs are a class of small endogenous non-coding RNAs of ~22nt in length that negatively regulate the expression of their target protein-coding genes (1). It has been reported that miRNAs are involved in many biological processes, such as cell proliferation, differentiation, and apoptosis (2–5). Mounting evidence has demonstrated that miRNAs are dysregulated in various types of human cancers (6–8), which may be leveraged as expression signatures for cancer diagnosis and prognosis. Circulating miRNAs represent the miRNAs that are secreted into extracellular body fluids, where they are incorporated in extracellular vesicles (EVs), such as shed microvesicles (sMVs) and exosomes, or in apoptotic bodies, or form complexes with RNA binding proteins, such as Argonates (AGOs). These protected circulating miRNAs remain in remarkably stable forms, rendering potential cancer biomarkers for non-invasive early detection or tissue-of-origin localization (9–11).

The vast amount of miRNA expression data in TCGA as well as many large-scale circulating miRNA profiling datasets are readily available for discovery and validation of cancer miRNA biomarkers (12–14). An online tool OncomiR has been developed to explore dysregulated miRNAs associated with tumor development and progression based on the TCGA data resource (15). While the analytical functions provided by OncomiR are useful, many functions are lacking for the comprehensive analysis of cancer miRNome data. In addition, OncomiR is only designed for TCGA data analysis and it doesn't support data visualization and exportation, which constrains its wide application. Sophisticated and user-friendly web tools are desperately needed to facilitate the data mining of the valuable cancer miRNome repository and promote translational application of miRNAs in cancer. To fill this void, we developed CancerMIRNome, a web application for the interactive analysis and visualization of cancer miRNome data based on 10,998 samples from 33 TCGA projects and 21,993 samples of 32 cancer types from 40 circulating miRNA profiling studies (Figure 1).

CancerMIRNome provides a suite of advanced functions for (1) a pan-cancer characterization of a miRNA of interest across multiple cancer types, and (2) a comprehensive miRNome analysis to identify diagnostic and prognostic signatures. Advanced visualizations are supported to produce publication-quality vector images in PDF format which can be easily downloaded. All the processed data deposited in

CancerMIRNome, including the normalized miRNA expression data and metadata for each dataset can be easily downloaded, allowing for further analysis by the end users (Figure 1).

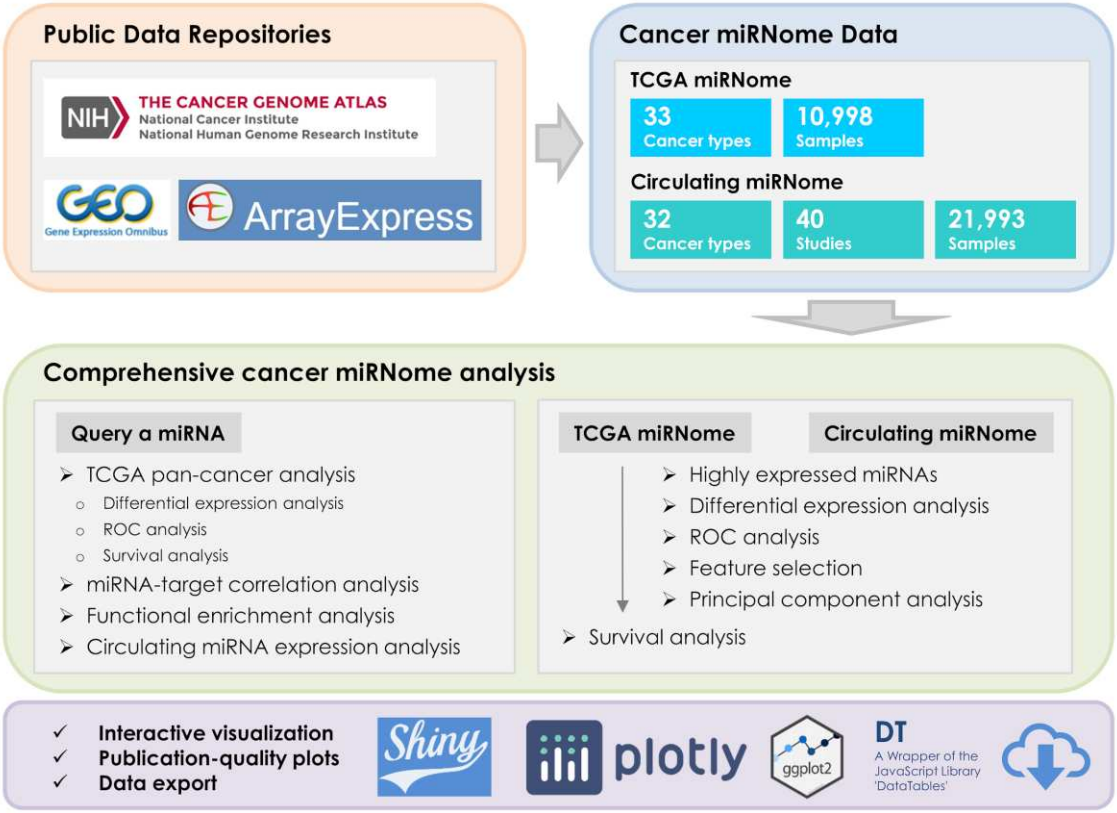

Figure 1. Overview of the CancerMIRNome database

DATA COLLECTION AND PROCESSING

TCGA miRNome datasets

The R/Bioconductor package GDCRNATools was used to download and process the miRNA expression data and clinical data of 33 cancer types in TCGA (16). Isoform expression quantification data of miRNA-Seq were downloaded from National Cancer Institute (NCI) Genomic Data Commons (GDC) using the *gdcRNADownload* function. The expression data from the same project were merged to a single expression matrix using the *gdcRNAMerge* function in the R package GDCRNATools, followed by a normalization with the Trimmed Mean of M-values (TMM) normalization method implemented in the R package edgeR (17). Clinical information including age, tumor

stages, overall survival, etc. were retrieved from the XML file of each sample with the *gdcClinicalMerge* function in GDCRNATools.

Circulating miRNome datasets

An extensive search for circulating miRNA expression data in cancer was performed in public databases, including NCBI Gene Expression Omnibus (GEO) and ArrayExpress. A total of 40 public circulating miRNA expression datasets with over 1000 miRNAs and more than 10 samples in each dataset were identified by searching the keywords 'circulating', 'whole blood', 'serum', 'plasma', 'extracellular vesicle', and 'exosome', in combination with 'miRNA' or 'microRNA', and with 'cancer', 'tumor', and 'carcinoma'. Both expression data and metadata were downloaded by the *getGEO* function in the R package GEOquery (18). Metadata of the samples were processed using custom scripts with manual inspections.

DATABASE CONTENT AND USAGE

A series of cutting-edge bioinformatics tools and functions have been packaged in CancerMIRNome, allowing for a pan-cancer analysis of a miRNA of interest across multiple cancer types and a comprehensive analysis of cancer miRNome profiles.

Query a miRNA of interest

Users can query a miRNA of interest by typing the miRNA accession number, miRNA ID of miRBase release 22.1 (19), or previous miRNA IDs in the 'Search a miRNA' field and selecting this miRNA from the dropdown list. In addition to the general information including IDs and sequence of the queried miRNA, links to five miRNA-target databases including ENCORI (20), miRDB (21), miTarBase (22), TargetScan (23), and Diana-TarBase (24) are also provided.

The single miRNA analysis modules include: (1) pan-cancer differential expression (DE) analysis, receiver operating characteristic (ROC) analysis, and Kaplan Meier (KM) survival analysis in TCGA; (2) miRNA-target correlation analysis; (3) functional enrichment analysis of miRNA targets; and (4) circulating miRNA expression analysis (Figure 2).

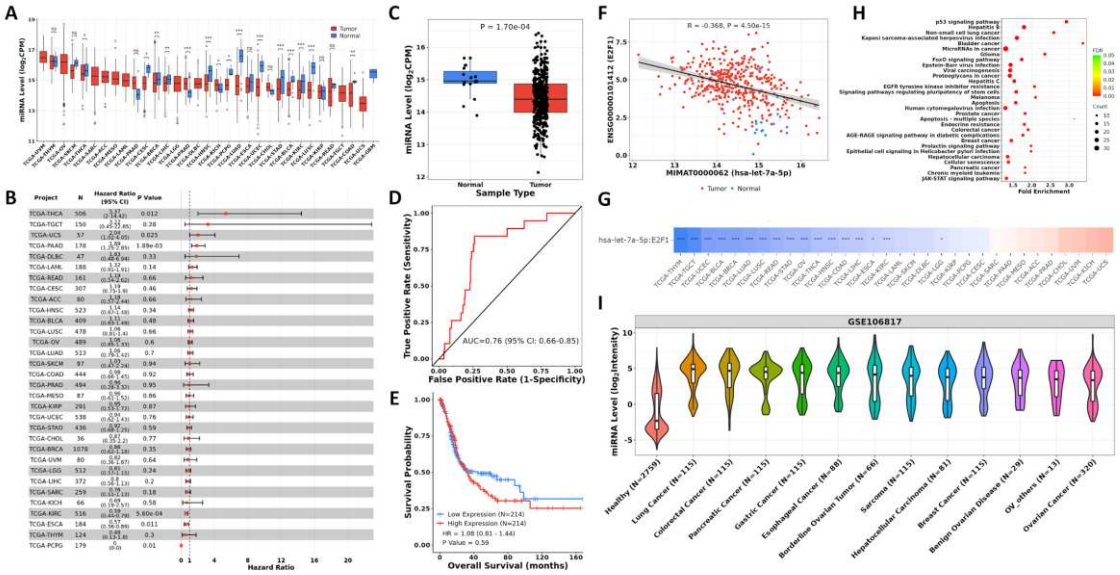

Figure 2. CancerMIRNome outputs from the query of a miRNA of interest. (**A**) Pan-cancer differential expression analysis across all TCGA projects. (**B**) A forest plot visualizing pan-cancer survival analysis across all TCGA projects. (**C**) Boxplot of the miRNA expression in tumor and normal samples from the selected TCGA project. (**D**) An ROC curve illustrating the diagnostic ability of the miRNA in the selected TCGA project. (**E**) Kaplan Meier analysis of overall survival between tumor samples with high and low expression of the miRNA of interest defined by its median expression value in the selected TCGA project. (**F**) Correlation analysis of the miRNA with one of its targets in a TCGA project. (**G**) An interactive heatmap visualizing miRNA-target correlations across all TCGA projects. (**H**) A bubble plot visualizing the functional enrichment of target genes for the miRNA. (**I**) A violin plot visualizing the circulating miRNA expression in a selected cancer circulating miRNome dataset.

1. *Pan-cancer analysis*

Pan-cancer DE analysis and ROC analysis can be performed for 33 cancer types in TCGA to evaluate the diagnostic ability of any miRNA in differentiating tumor samples from normal samples. The expression levels and the statistical significance of the miRNA in all the TCGA projects can be visualized in a box plot (Figure 2A). Prognostic ability of a miRNA can be assessed by KM analysis of overall survival (OS) in patients with high *versus* low expression levels of this miRNA in tumor samples. A forest plot displaying the number of tumor samples, hazard ratio (HR), 95% confidence interval (CI) of the HR, and p value for each cancer type in TCGA is used to visualize the result of pan-cancer survival analysis (Figure 2B). These pan-cancer functions can be implemented for a selected TCGA project when needed. A box plot with miRNA

expression, an ROC curve, and a KM survival curve for the selected project will be displayed (Figure 2 C-E).

2. *miRNA-target correlation analysis*

Pearson correlation between a miRNA and its targets can be analyzed in the TCGA datasets. The miRNA-target interactions are based on miRTarBase 2020 – an experimentally validated miRNA-target interactions database (22). The expression correlations between a miRNA and all of its targets in a selected TCGA project are listed in an interactive data table. Users can select a miRNA-mRNA pair to visualize the scatter plot (Figure 2F). An interactive heatmap is also available to visualize and compare the strength of miRNA-target correlations across TCGA projects (Figure 2G).

3. *Functional enrichment analysis of miRNA targets*

In CancerMIRNome, functional enrichment analysis of the target genes for a miRNA can be conducted using clusterProfiler (25), with support of many pathway/ontology knowledgebases, including Kyoto Encyclopedia of Genes and Genomes (KEGG) (26), Gene Ontology (GO) (27), Reactome (28), Disease Ontology (DO) (29), Network of Cancer Gene (NCG) (30), DisGeNET (31), and Molecular Signatures Database (MSigDB) (32). A data table will be created to summarize the significantly enriched pathways/ontologies and the top pathways/ontologies can be visualized as bar plot and bubble plot (Figure 2H).

4. *Circulating miRNA expression*

Expression of an interested miRNA in whole blood, serum, plasma, extracellular vesicles, or exosomes from both healthy and cancer patients can be conveniently explored in the 40 circulating miRNome datasets. Users can select one or more datasets for an analysis, through which violin plots are displayed for visualization and comparison (Figure 2I).

Comprehensive miRNome analysis

Comprehensive miRNome analysis can be performed for each of the 33 TCGA projects and the 40 circulating miRNome datasets in CancerMIRNome, including (1) identification of highly expressed miRNAs; (2) DE analysis between two user-defined subgroups; (3) ROC analysis between tumor and normal samples in a TCGA dataset or between cancer patients and healthy volunteers in a circulating miRNome dataset; (4) identification of diagnostic miRNA markers based on a machine learning algorithm;

(5) principal component analysis for dimensionality reduction; and (6) selection of prognostic miRNA biomarkers and construction of prognostic models using the TCGA data (Figure 3).

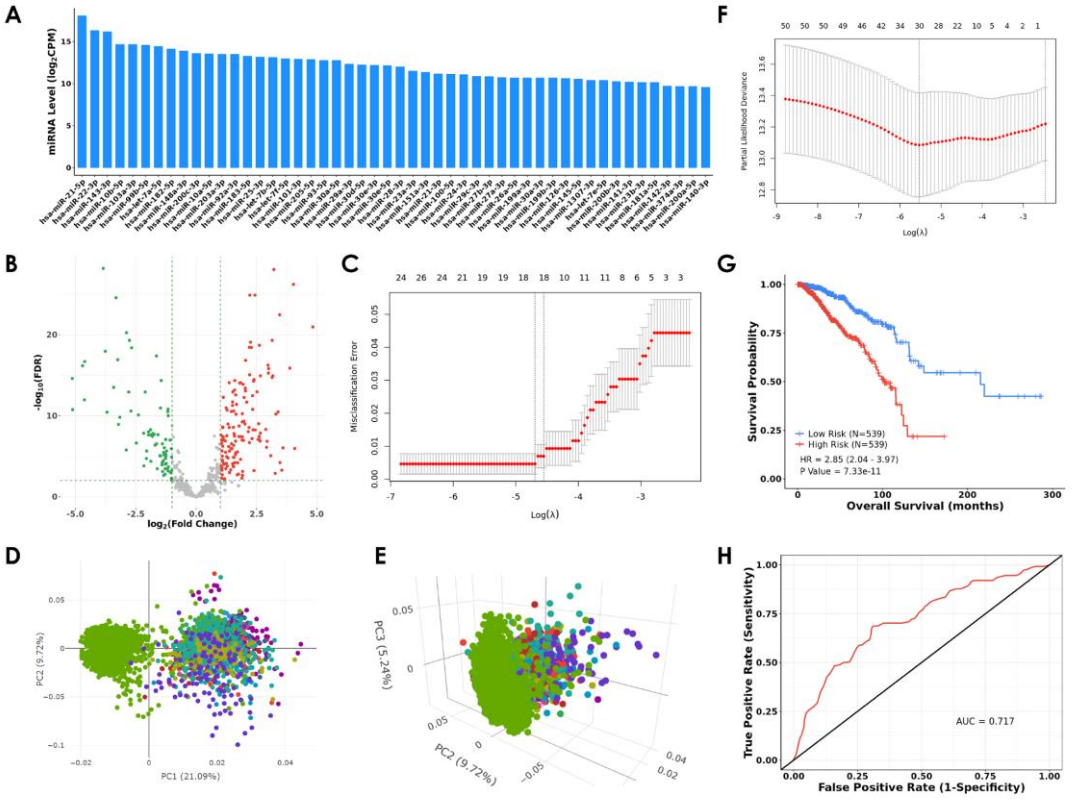

Figure 3. CancerMIRNome outputs from the comprehensive miRNome data analysis of an interested dataset. (**A**) Bar plot of highly expressed miRNAs. (**B**) A volcano plot visualizing the differentially expressed miRNAs between two user-defined groups. (**C**) Selection of diagnostic miRNA biomarkers using LASSO. (**D**) 2D interactive visualization of principal component analysis result using the first two principal components. (**E**) 3D interactive visualization of principal component analysis result using the first three principal components. (**F**) Selection of prognostic miRNA biomarkers using a regularized Cox regression model with LASSO penalty to develop a prognostic model. (**G**) Kaplan Meier survival analysis evaluating the prognostic ability of the miRNA-based prognostic model. (**H**) Time-dependent ROC analysis evaluating the prognostic ability of the miRNA-based prognostic model.

1. *Highly expressed miRNAs*

Highly expressed miRNAs are identified if their counts per million (CPM) are greater than 1 in more than 50% of the samples in a TCGA project or if their abundances are ranked in the top 500 miRNAs in a circulating miRNome dataset. The top 50 highly expressed miRNAs are visualized in a bar plot based on their median expression values (Figure 3A).

2. *DE analysis*

The highly expressed miRNAs in a dataset can be compared between two user-defined subgroups for identifying the significantly DE miRNAs (Figure 3B). Clinical variables, including sample type, tumor stages, gender, etc., may be utilized for grouping samples. For example, the DE analysis can be performed not only between tumor and normal samples, but also between patients at early and late tumor stages. The R package limma (33) is used for the DE analysis.

3. *ROC analysis*

The ROC analysis can be carried out to screen the highly expressed miRNAs for the diagnostic biomarkers, which can distinguish tumor samples from normal samples in a TCGA dataset, or distinguish cancer patients from healthy volunteers in a circulating miRNome dataset, and therefore may be leveraged for non-invasive early cancer detection. All the miRNAs are ranked in a data table based on their AUC values.

4. *Machine learning-based feature selection*

For any dataset, the least absolute shrinkage and selection operator (LASSO) (34, 35), a machine-learning algorithm, can be used to detect the miRNAs with diagnostic power and develop a classification model using these miRNA signatures for cancer diagnosis (Figure 3C).

5. *Principal component analysis*

Principal component analysis can be utilized to analyze the highly expressed miRNAs in any dataset such that all patient samples may be visualized in a 2D or 3D interactive plot using the first two or three principal components, respectively (Figure 3D and Figure 3E).

6. *Survival analysis*

Both Cox Proportional-Hazards (CoxPH) regression analysis and Kaplan-Meier (KM) survival analysis are supported in CancerMIRNome to identify prognostic miRNA biomarkers in any TCGA projects. Significant miRNAs in the univariate CoxPH analysis will be jointly analyzed using a regularized Cox regression model with LASSO penalty to develop a prognostic model (35) (Figure 3F). The prognostic model, which is a linear combination of the finally selected miRNA variables with the LASSO-derived regression coefficients, will be used to calculate a risk score for each patient. All the patients will be divided into either a high-risk group or a low-risk group based on the

median risk value for the cohort. The KM survival analysis and time-dependent ROC analysis can be performed to evaluate the prognostic ability of the miRNA-based prognostic model (Figure 3G and Figure 3H).

Data download

All the processed data, including the 33 TCGA miRNome datasets, the 40 circulating miRNome datasets in human cancers, and the integrated miRNA annotation data can be downloaded easily on the 'Download' page of CancerMIRNome. The ExpressionSet class is used for the miRNA expression data and metadata of the miRNome datasets. The miRNA annotation data includes the miRNA ID, miRNA name, and miRNA sequence from the latest miRBase release 22.1, and the previous miRNA names from miRBase release 10.0 to release 21. The data are downloaded as RDS files, which can be easily imported into R.

SUMMARY AND FUTURE DIRECTIONS

In this study, we present to the cancer research community a user-friendly web tool, CancerMIRNome, for the interactive analysis and visualization of cancer miRNome by leveraging 10,998 tumor and normal samples from 33 TCGA projects and 21,993 samples of 32 cancer types from 40 public circulating miRNA profiling studies. A suite of well-designed functions is provided to facilitate data mining and analysis at both the miRNA level and the miRNome level (or dataset level). Advanced visualizations are supported in CancerMIRNome and the publication-quality vector images can be easily created and downloaded. Moreover, all the data and results are exportable, allowing for further local analyses by the end users. While CancerMIRNome is diligently serving the cancer research community, we are open to any feedback from users and will constantly maintain and improve this database. New datasets, analytical methods, and visualization functions will be included in CancerMIRNome as soon as they are available. We expect that CancerMIRNome would become a valuable online resource for a comprehensive analysis of cancer miRNome data not only for experimental biologists, but also for bioinformatics scientists in the field.

AVAILABILITY

The CancerMIRNome database is publicly available at http://bioinfo.jialab-ucr.org/CancerMIRNome. The source code for processing miRNome data and building the database is available at https://github.com/rli012/CancerMIRNome. All the processed data deposited in CancerMIRNome can be downloaded easily on the 'Download' page of the database.

FUNDING

This work was supported by Z.J.'s UC Riverside Faculty Start-up Fund and UC Cancer Research Coordinating Committee Competition Award. D.Y. and J.Z. were supported by the National Natural Science Foundation of China (81660426), Science and Technology Project of Guizhou Province in 2017 ([2017]5803), the High-level innovative talent project of Guizhou Province in 2018 ([2018]5639), and the Science and Technology Plan Project of Guiyang in 2019 [2019]2-15. R. Z. and W.Z. were supported by the grants from National Natural Science Foundation of China (82072813, 8157142), Guangzhou Municipal Science and Technology Project (201803040001).

CONFLICT OF INTEREST

The authors declare that they have no competing interests.

REFERENCES

1. Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.

2. Lu,Y., Thomson,J.M., Wong,H.Y.F., Hammond,S.M. and Hogan,B.L.M. (2007) Transgenic over-expression of the microRNA miR-17-92 cluster promotes proliferation and inhibits differentiation of lung epithelial progenitor cells. *Dev. Biol.*, **310**, 442–453.

3. Wang,Y., Baskerville,S., Shenoy,A., Babiarz,J.E., Baehner,L. and Blelloch,R. (2008) Embryonic stem cell-specific microRNAs regulate the G1-S transition and

promote rapid proliferation. *Nat. Genet.*, **40**, 1478–1483.

4. Chang,T.-C., Wentzel,E.A., Kent,O.A., Ramachandran,K., Mullendore,M., Lee,K.H., Feldmann,G., Yamakuchi,M., Ferlito,M., Lowenstein,C.J., *et al.* (2007) Transactivation of miR-34a by p53 broadly influences gene expression and promotes apoptosis. *Mol. Cell*, **26**, 745–752.

5. Vidigal,J.A. and Ventura,A. (2015) The biological functions of miRNAs: lessons from in vivo studies. *Trends Cell Biol.*, **25**, 137–147.

6. Lu,J., Getz,G., Miska,E.A., Alvarez-Saavedra,E., Lamb,J., Peck,D., Sweet-Cordero,A., Ebert,B.L., Mak,R.H., Ferrando,A.A., *et al.* (2005) MicroRNA expression profiles classify human cancers. *Nature*, **435**, 834–838.

7. Thomson,J.M., Newman,M., Parker,J.S., Morin-Kensicki,E.M., Wright,T. and Hammond,S.M. (2006) Extensive post-transcriptional regulation of microRNAs and its implications for cancer. *Genes Dev.*, **20**, 2202–2207.

8. He,L., He,X., Lim,L.P., de Stanchina,E., Xuan,Z., Liang,Y., Xue,W., Zender,L., Magnus,J., Ridzon,D., *et al.* (2007) A microRNA component of the p53 tumour suppressor network. *Nature*, **447**, 1130–1134.

9. Schwarzenbach,H., Nishida,N., Calin,G.A. and Pantel,K. (2014) Clinical relevance of circulating cell-free microRNAs in cancer. *Nat. Rev. Clin. Oncol.*, **11**, 145–156.

10. Mitchell,P.S., Parkin,R.K., Kroh,E.M., Fritz,B.R., Wyman,S.K., Pogosova-Agadjanyan,E.L., Peterson,A., Noteboom,J., O'Briant,K.C., Allen,A., *et al.* (2008) Circulating microRNAs as stable blood-based markers for cancer detection. *Proc. Natl. Acad. Sci. U. S. A.*, **105**, 10513–10518.

11. Xu,R., Rai,A., Chen,M., Suwakulsiri,W., Greening,D.W. and Simpson,R.J. (2018) Extracellular vesicles in cancer - implications for future improvements in cancer care. *Nat. Rev. Clin. Oncol.*, **15**, 617–638.

12. Sudo,K., Kato,K., Matsuzaki,J., Boku,N., Abe,S., Saito,Y., Daiko,H., Takizawa,S., Aoki,Y., Sakamoto,H., *et al.* (2019) Development and Validation of an Esophageal Squamous Cell Carcinoma Detection Model by Large-Scale MicroRNA Profiling. *JAMA Netw Open*, **2**, e194573.

13. Yokoi,A., Matsuzaki,J., Yamamoto,Y., Yoneoka,Y., Takahashi,K., Shimizu,H., Uehara,T., Ishikawa,M., Ikeda,S.-I., Sonoda,T., *et al.* (2018) Integrated extracellular microRNA profiling for ovarian cancer screening. *Nat. Commun.*, **9**, 4319.

14. Shimomura,A., Shiino,S., Kawauchi,J., Takizawa,S., Sakamoto,H., Matsuzaki,J., Ono,M., Takeshita,F., Niida,S., Shimizu,C., *et al.* (2016) Novel combination of serum

microRNA for detecting breast cancer in the early stage. *Cancer Sci.*, **107**, 326–334.

15. Wong,N.W., Chen,Y., Chen,S. and Wang,X. (2018) OncomiR: an online resource for exploring pan-cancer microRNA dysregulation. *Bioinformatics*, **34**, 713–715.

16. Li,R., Qu,H., Wang,S., Wei,J., Zhang,L., Ma,R., Lu,J., Zhu,J., Zhong,W.-D. and Jia,Z. (2018) GDCRNATools: an R/Bioconductor package for integrative analysis of lncRNA, miRNA and mRNA data in GDC. *Bioinformatics*, **34**, 2515–2517.

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

18. Davis,S. and Meltzer,P.S. (2007) GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*, **23**, 1846–1847.

19. Kozomara,A., Birgaoanu,M. and Griffiths-Jones,S. (2019) miRBase: from microRNA sequences to function. *Nucleic Acids Res.*, **47**, D155–D162.

20. Li,J.-H., Liu,S., Zhou,H., Qu,L.-H. and Yang,J.-H. (2013) 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.

21. Chen,Y. and Wang,X. (2020) miRDB: an online database for prediction of functional microRNA targets. *Nucleic Acids Res.*, **48**, D127–D131.

22. Huang,H.-Y., Lin,Y.-C.-D., Li,J., Huang,K.-Y., Shrestha,S., Hong,H.-C., Tang,Y., Chen,Y.-G., Jin,C.-N., Yu,Y., *et al.* (2019) miRTarBase 2020: updates to the experimentally validated microRNA–target interaction database. *Nucleic Acids Res.*, **48**, D148–D154.

23. Agarwal,V., Bell,G.W., Nam,J.-W. and Bartel,D.P. (2015) Predicting effective microRNA target sites in mammalian mRNAs. *Elife*, **4**.

24. Karagkouni,D., Paraskevopoulou,M.D., Chatzopoulos,S., Vlachos,I.S., Tastsoglou,S., Kanellos,I., Papadimitriou,D., Kavakiotis,I., Maniou,S., Skoufos,G., *et al.* (2017) DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA–gene interactions. *Nucleic Acids Res.*, **46**, D239–D245.

25. Yu,G., Wang,L.-G., Han,Y. and He,Q.-Y. (2012) clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS*, **16**, 284–287.

26. Kanehisa,M., Furumichi,M., Tanabe,M., Sato,Y. and Morishima,K. (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**, D353–D361.

27. The Gene Ontology Consortium (2019) The Gene Ontology Resource: 20 years

and still GOing strong. *Nucleic Acids Res.*, **47**, D330–D338.

28. Fabregat,A., Jupe,S., Matthews,L., Sidiropoulos,K., Gillespie,M., Garapati,P., Haw,R., Jassal,B., Korninger,F., May,B., *et al.* (2018) The Reactome Pathway Knowledgebase. *Nucleic Acids Res.*, **46**, D649–D655.

29. Schriml,L.M., Mitraka,E., Munro,J., Tauber,B., Schor,M., Nickle,L., Felix,V., Jeng,L., Bearer,C., Lichenstein,R., *et al.* (2019) Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res.*, **47**, D955–D962.

30. Repana,D., Nulsen,J., Dressler,L., Bortolomeazzi,M., Venkata,S.K., Tourna,A., Yakovleva,A., Palmieri,T. and Ciccarelli,F.D. (2019) The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. *Genome Biol.*, **20**, 1.

31. Piñero,J., Ramírez-Anguita,J.M., Saüch-Pitarch,J., Ronzano,F., Centeno,E., Sanz,F. and Furlong,L.I. (2020) The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.*, **48**, D845–D855.

32. Liberzon,A., Subramanian,A., Pinchback,R., Thorvaldsdóttir,H., Tamayo,P. and Mesirov,J.P. (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739–1740.

33. Ritchie,M.E., Phipson,B., Wu,D., Hu,Y., Law,C.W., Shi,W. and Smyth,G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47.

34. Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc.*, **58**, 267–288.

35. Friedman,J., Hastie,T. and Tibshirani,R. (2010) Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, **33**, 1–22.