miRCancerdb: a database for correlation analysis between microRNA and gene expression in cancer

Mahmoud Ahmed†, Huynh Nguyen†, Trang Lai† and Deok Ryong Kim*†

Abstract

Objectives: microRNAs regulate expression of target genes by specifically binding to their transcripts, subsequently leading to translational inhibition or mRNA degradation. Gene regulation by microRNAs has been implicated in a wide range of physiological and pathological conditions. Here, we leverage the use of public-access data and the available genomic annotations to pre-calculate the correlation of the expression of a large number of microRNAs with gene at the mRNA and protein level in the context of cancers.

Results: Expression data of miRNAs, mRNAs and proteins in cancer patients from The Cancer Genome Atlas along with TargetScan miRNAs-target annotations were used to calculate the expression correlations between miRNAs and features (mRNAs/proteins) in a number of cancer studies. We then packed the output of this analysis into a database and made it available through an interactive web application. The miRCancerdb is an easy-to-use database to investigate the microRNAs-dependent regulation of target genes involved in development of cancer.

Keywords: miRCancerdb, Cancer, MicroRNA, TCGA

Introduction

miRNA (microRNA) are involved, among other functions, in the regulation of gene expression in a wide range of physiological and pathological conditions [1, 2]. Identification of miRNAs and their gene targets is challenging in both bioinformatics and experimental biology. The sequence-based search of gene targets for miRNAs usually yields a high number of false positives [3]. To experimentally confirm these hits, biologist have to go through long lists of potential targets and use laborious laboratory techniques to isolate the meaningful functional associations among them and the miRNAs. The problem is only exacerbated by the diverse expression profiles and functions of miRNAs across tissues and conditions [4].

An attempt to tackle the problem of miRNA target identification while maintaining a reasonable high-throughput is to use a combination of sequence-based search methods and high-throughput expression profiling such as microarrays and RNA sequencing [5, 6]. The process starts by manipulating the expression of a particular miRNA in a certain condition followed by surveying the gene expression by arrays or sequencing. The correlation between the expression of the miRNA and its potential target genes are then calculated [7]. Genes with sequences that match the miRNA seed and correlate strongly in their expression with the miRNA constitute higher chances to be true positives.

Here, we leverage the use of public-access data and the available annotations to pre-calculate the correlation values between the expression of a large number of miRNAs and genes in the context of cancer. We used TCGA (The Cancer Genome Atlas) [8] to obtain miRNA, gene and protein expression profiles in 34 type of cancer and the TargetScan annotation database [9] to calculate the expression correlation for miRNAs and their targets at
Main text

Data sources

Expression data of miRNAs, genes and proteins were obtained from TCGA using the R package RTCGA [10]. RTCGA provides the processed counts of these three genetic elements from patient samples using miRNA sequencing, RNA sequencing and Reverse Phase Protein Arrays (RPPA), respectively. Human miRNA target annotations were obtained using the R package targetscan.Hs.eg.db [11]. The miRNA miRBase, gene ENTREZ ID, official gene symbols and manufacturer antibody IDs were used appropriately to connect the data from different assays and sources.

Analysis pipeline

Expression data from TCGA and target annotations were first processed to a tidy format and different identifiers were translated to common ones. miRNAs and features (mRNA/proteins) were excluded when their expression values were less than 1. Pearson's correlations of miRNA expression matrix and gene expression or protein expression matrices were calculated in each of the TCGA studies (n = 34). The following equation was used as implemented in the R base cor function to calculate the Pearson's correlation (\( \rho \)) [12].

\[
\rho_{X,Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}
\]

where, \( E \) is the expectation, \( \mu \) is the mean and \( \sigma \) is the standard deviation.

Missing values (NA) and correlation values less than 0.1 were omitted to bring the data to a manageable size. Correlation data for each feature in each study were then merged in a single table to build a database file. In addition, a target table of each miRNA and its corresponding targets was added to the database file. Finally, a tidy table of miRNAs expression profiles (counts) in TCGA studies was added to the database file. Figure 1a show the workflow of the analysis and the corresponding output views.

Database file

The database file is a sqlite3 instance. It contains four tables: cor_rnaseq for miRNA and mRNA expression correlations, cor_rppa for miRNA and protein expression correlations, targets for the miRNA-feature targets and profiles for the miRNA expression in TCGA studies.

Web interface and data presentation

We built an interactive shiny application on top of the database to provide an easy-to-use user interface for miRCancerdb [13]. The web application is available at https://mahshaaban.shinyapps.io/miRCancerdb/. The application main homepage consists of main and control panels. The control panel (Fig. 1b) is where the user can input a search query and customize the results. The results are shown on the main panel in different representations/views (Table 1). In addition, we provide an explanatory text for getting started, how-to and about sections to give an overview and detailed user's instruction on the website.

Building and using the database locally

The source files to build and use miRCancerdb locally are available on https://github.com/MahShaaban/miRCancerdb. It consists of R scripts to build the database file and launch an interactive browser application locally. The commands to build the application are wrapped in a makefile that makes it easy to trigger the local build with only a single command.

To do that, download the github repository, navigate to the miRCancerdb directory and run make.

```
git clone https://github.com/MahShaaban/miRCancerdb
cd miRCancerdb
make
```

the miRCancerdb directory and run make.

To only run the browser application after the initial build, use the make section, run the R script from the command-line or use within Rstudio.

```
make launch_app
R -e "shiny::runApp('app.R',launch.browser="",TRUE)"
```

More details about building and using the database locally is available at the database github repository.

Use case

To illustrate querying the database using the web interactive application, we provide an example of a simple query. Here, we used the query options in the control panel to enter the TCGA identifiers for adrenocortical carcinoma (ACC) and bladder urothelial carcinoma (BLCA) studies, and the miRBase IDs for has-let-7b and has-let-7c. All other options are set to the defaults. Figure 2 show the outputs from the miRCancerdb views.

miRCancerdb provides multiple views to output and represent the results of each query. In this case we show the results from only two. The dot and the profile views. Figure 2a shows the output of the previous query in a dot graph where the queried miRNA on the x-axis and the correlated genes on the y-axis, the figures are faceted by
the queried studies. Each dot represent a recorded correlation with the size corresponding to its value and the color to the direction (positive or negative). Figure 2b is a point graph with the miRNAs on the x-axis and the log count in each study on the y-axis. Together, the dot and profile views show the correlated genes, the magnitude, direction of the correlation and the distribution of the count of each miRNA in each of the queried studies. To reproduce the exact figures while using the database locally, one needs to build the database first as described above then run the following code in an R session.

We start by loading the required libraries. Then, we connect to the database SQLite file and filter the table `cor_rnaseq` to get the correlations of the query of interest.
### Table 1. miRCancerdb query results representations/views.

| View       | Content                                                                                                                                                                                                 |
|------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Table view | This is a tidy tabular presentation of the search results. The table consists of four columns: mirna_base, miRNA miRBase ID; feature (mRNA/protein) hits; study, TCGA study identifier; and cor, calculated correlation.     |
| Dot view   | This is a point graph representation of the search results. Columns represent each query miRNA in each TCGA study and rows represent feature (mRNA/protein) hits. Each point is scaled by the correlation value and colored by the direction of the correlation; positive or negative. This is a bar graph representation of the search results. Panels represent each query miRNA in each TCGA study. On the x-axis are the feature (mRNA/protein) hits and on the y-axis are the correlation values. |
| Bar view   | This is a bar graph representation of the search results. Panels represent each query miRNA in each TCGA study. On the x-axis are the feature (mRNA/protein) hits and on the y-axis are the correlation values.                                                            |
| Profile view | This is a point graph representation of the expression profiles of query miRNAs. Each panel represents a TCGA study. On the x-axis are the query miRNAs and on the y-axis are the log count transformation values. Points are jittered on the x-axis for better visibility.                                |
| Summary view | This is a summary report about the query miRNAs and TCGA studies. It includes three sections. First, pie charts showing the percentage of positive and negative correlated features (mRNA/protein) with each query miRNA in each TCGA study. Second, a table with the total number of targetscan known targets for each miRNA. Finally, density plots showing the distribution of the correlation values of all and target features (mRNA/protein). Each panel represents a TCGA study and curves are colored by the corresponding miRNAs. |
| Selection view | A tool to show the available TCGA studies and miRNAs                                                                                                                                          |

---

**Fig. 2** Dot and profile views from a miRCancerdb query use case. A use case of miRCancerdb database to query the miRNAs; has-let-7b and has-let-7c in ACC and BLCA cancer studies. **a** A dot view of the top 5 hits of the expression correlation between the miRNAs and genes. Points are scaled by the correlation value and colored by the direction of the correlation; positive (green) or negative (red). **b** A profile view of the expression (log count) of the miRNAs in TCGA studies.
Utility and discussion
Identification of the regulatory elements (transcription factors and miRNAs) at complex pathological conditions like cancer is a very important step to understand their pathophysiological mechanisms [1, 4]. These regulatory elements often work as a complex network and drive many processes such as cancer growth, differentiation and metastasis [2]. Cistrome Cancer is a database that provides information about transcription factors correlation and regulatory potentials to gene expression [15].

miRCancerdb can be deployed in a wide variety of research questions. Biologists interested in studying certain miRNAs can use it for their initial surveys. Mainly, the database provides calculated correlations values of these elements to either gene or protein expression in a large number of cancer studies. In addition, it is possible to integrate the available sequence-based annotations to limit the number of highly correlated genes to predicted targets. Also, those who are interested in larger questions can query the database to find out all miRNAs involved in certain types of cancers and compare them to each other. This can be used to discover and compare regulation patterns in neoplastic conditions.

The database file was built with several points in mind. First, we filtered out the less useful information to bring the file size to a manageable size. Second, the file is a stand-alone database which can be accessed programmatically using any SQLite client. Finally, the source code for building the database and launching the browser application is open sourced. This, in one hand leverages the reproducibility and makes updating the database very easy when needed. On the other, the scripts can be used to build the database on local machines and integrate it in different analysis pipelines.

To our knowledge, no available public database provide a data-driven correlation/co-expression of miRNA and genes in cancer. We developed an R package called cRegulome (unpublished) that interfaces the miRCancerdb programmatically from R environment. In addition to the miRNA correlations, the package provide access to transcription factor-gene correlations from the cistrome project [14]. miRCancer is another cancer association database of miRNA based on text mining the literature [15].

Limitations
In this kind of integrative analysis based on public access data, we are limited by the available data and annotations. The current miRNA annotations are far from complete. The plan is to keep up with updates from the two major data and annotation sources, namely the TCGA and the TargetScan projects. That is through rebuilding the database with each future updates of the RTCGA and targetscan.Hs.eg.db Bioconductor packages.

In the current version of the database, we treated each TCGA study as one entity. Another plausible way is to consider the cancer heterogeneity and individual variations. This is to stratify the analysis with histological and/or clinical information about the samples. However, this would complicate the pipeline and increase
the database size dramatically, so we might address this issue in a later version.

In addition, improving the user experience on the web application and providing graphics that can handle bigger queries is a priority. Finally, providing a similar application and contributed to writing the manuscript. DRK contributed

Acknowledgements Not applicable.

Competing interests The authors declare that they have no competing interests.

Availability of data and material The datasets used in this integrative analysis included TCGA public access data and the miRNA target TargetScan annotations. We accessed the TCGA data using the R package RTCGA. We accessed the TargetScan annotations using the R package targetscan.Hs.eg.db.

Consent for publication Not applicable.

Ethics approval and consent to participate Not applicable.

Funding This study was supported by the Basic Science Research Program through the National Research Foundation (NRF) of Korea funded by the Ministry of Education (2015R1D1A01019753) and by the Ministry of Science, ICT and Future Planning (NRF-2015R1A5A2008833).

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 8 December 2017 Accepted: 11 January 2018 Published online: 07 February 2018

References

1. Hayes J, Peruzzi PP, Lawler S. MicroRNAs in cancer: biomarkers, functions and therapy. Trends Mol Med. 2014;20(8):460–9. https://doi.org/10.1016/j.molmed.2014.06.005.

2. He L, Hannan GJ. Correction: MicroRNAs: small RNAs with a big role in gene regulation. Nat Rev Genet. 2004;5(8):631. https://doi.org/10.1038/nrg1415.

3. Thompсон DW, Bracken CP, Goodall GJ. Experimental strategies for microRNA target identification. Nucleic Acids Res. 2011;39(16):6845–53. https://doi.org/10.1093/nar/gkr330.

4. Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mark RH, Ferrando AA, Downing JR, Jacks T, Horvitz HR, Golub TR. MicroRNA expression profiles classify human cancers. Nature. 2005;435(7043):834–8. https://doi.org/10.1038/nature03702.

5. Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, Castle J, Bartel DP, Linsley PS, Johnson JM. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature. 2005;433(7027):769–73. https://doi.org/10.1038/nature03151.

6. Grimson A, Farh KKH, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol Cell. 2007;27(1):91–105. https://doi.org/10.1016/j.molcel.2007.06.017.

7. Karginov FV, Conaco C, Xuan Z, Schmidt BH, Parker JS, Mandel G, Hannan GJ. A biochemical approach to identifying microRNA targets. Proc Natl Acad Sci. 2007;104(49):19291–6. https://doi.org/10.1073/pnas.0709711004.

8. International Cancer Genome Consortium, Hudson TJ, Anderson W, Artez MA, HN and TL performed the analysis built the database, the interactive application and contributed to writing the manuscript. DRK contributed
Futreal PA, Campbell PJ, Vincent-Salomon A, Richardson AL, Reis-Filho JS, van de Vijver M, Thomas G, Masson-Jacquemier JD, Aparicio S, Borg A, Berresen-Dale AL, Caldas C, Foekens JA, Strunnenberg HG, van’t Veer L, Easton DF, Spellman PT, Barker AD, Chinn L. International network of cancer genome projects. Nature. 2010;464(7291):993–8. https://doi.org/10.1038/nature08987.

9. Agarwal V, Bell GW, Nam JW, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. Elife. 2015;4:1–1. https://doi.org/10.7554/elif e.05005.

10. Kosinski M, Biecek P. RTCGA: the cancer genome atlas data integration. 2016. https://rtcga.github.io/RTCGA. Accessed 8 Dec 2017.

11.Csardi G. Targetscan Hs.e.g.db: TargetScan miRNA target predictions for human. 2013.

12. R Development Core Team. R Internals. Vienna, Austria. 2015. doi: 3-900051-14-3. https://cran.r-project.org/doc/manuals/r-release/R-ints.html. Accessed 8 Dec 2017.

13. Chang W, Cheng J, Alifara J, Xie Y, McPherson J. shiny: Web Application Framework for R. 2016. https://cran.r-project.org/package=shiny. Accessed 8 Dec 2017.

14. Liu T, Ortiz JA, Taing L, Meyer CA, Lee B, Zhang Y, Shin H, Wong SS, Ma J, Lei Y, Pape UJ, Poidinger M, Chen Y, Yeung K, Brown M, Turpaz Y, Liu XS. Cistrome: an integrative platform for transcriptional regulation studies. Genome Biol. 2011;12(8):83. https://doi.org/10.1186/gb-2011-12-8-r83.

15. Xie B, Ding Q, Han H, Wu D. miRCancer: a microRNA-cancer association database constructed by text mining on literature. Bioinformatics. 2013;29(5):638–44. https://doi.org/10.1093/bioinformatics/btt014.