


RESEARCH NOTE

Open Access



miRkit: R framework analyzing miRNA PCR array data

Maria Tsagiopoulou¹, Anastasis Togkousidis¹, Nikolaos Pechlivanis^{1,4}, Maria Christina Maniou¹, Aristeia Batsali², Angelos Matheakakis^{2,3}, Charalampos Pontikoglou^{2,3} and Fotis Psomopoulos^{1*} 

Abstract

Objective: The characterization of microRNAs (miRNA) in recent years is an important advance in the field of gene regulation. To this end, several approaches for miRNA expression analysis and various bioinformatics tools have been developed over the last few years. It is a common practice to analyze miRNA PCR Array data using the commercially available software, mostly due to its convenience and ease-of-use.

Results: In this work we present miRkit, an open source framework written in R, that allows for the comprehensive analysis of RT-PCR data, from the processing of raw data to a functional analysis of the produced results. The main goal of the proposed tool is to provide an assessment of the samples' quality, perform data normalization by endogenous and exogenous miRNAs, and facilitate differential and functional enrichment analysis. The tool offers fast execution times with low memory usage, and is freely available under a MIT license from <https://bio.tools/mirkit>. Overall, miRkit offers the full analysis from the raw RT-PCR data to functional analysis of targeted genes, and specifically designed to support the popular miScript miRNA PCR Array (Qiagen) technology.

Keywords: miRNA, qPCR, RT-PCR, PCR array, GO, KEGG

Introduction

MicroRNAs are small non-coding RNA molecules with a critical role in gene expression regulation [1]. They are implicated in mRNA post-transcriptional modulation in the cell as well as released into circulation and transferred to other target cells [2]. For this reason, and beyond their key role in intracellular pathways [2], miRNAs have emerged as biomarkers in clinical medicine [3] and are thought to represent appealing novel therapeutic modalities [4]. Also, the expression levels of miRNAs are known to be deregulated in diseases and malignancies [5, 6].

Various approaches have been used to profile the expression of miRNAs [4] such as RT-PCR arrays,

microarrays, small RNA-seq [7]. Quantitative real-time PCR (RT-PCR) assays are sensitive and specific in detecting and quantifying the expression of miRNAs in the human miRNA genome (miRNome) [4]. Within this context, the commercially available human miRNome miScript miRNA PCR Array (Qiagen) can be used to profile the 1066 most abundantly expressed and best characterized miRNA sequences in the human miRNome, as annotated in miRBase Release 16 (www.miRBase.org).

Raw RT-PCR data are typically analyzed using the manufacturer's software. In fact, several open-source packages analyze RT-PCR data with the traditional Ct (threshold cycle) quantification approach [8], ignoring more sophisticated and publicly available methods to analyze the expression profiles.

Currently, and going beyond miRNA data, there are plenty of options for differential expression analysis in which the user can employ a larger range of more sophisticated methods, ending up with values for logFC

*Correspondence: fpsom@certh.gr

¹ Institute of Applied Biosciences, Centre of Research and Technology Hellas, 57001 Thessaloniki, Greece

Full list of author information is available at the end of the article



and adj-pvalue. To assess the potential benefits of analyzing this data with an open source tool, we developed miRkit, a framework written in R, specific for miScript miRNA PCR Array (Qiagen) technology. The proposed toolset offers the whole analysis of the raw RT-PCR data completely automated, including quality control of the samples, normalization of endogenous and exogenous controls, differential expression analysis and functional analysis of targeted genes. Finally, the package has fast execution time and uses very low memory.

Main text

Implementation

Input data

The main R script of the workflow reads the input data from three distinct files stored in a folder:

1. Count table: This is the main data file, containing the different samples on columns and the measurement of each well on the rows. The proposed tool is applicable on miScript miRNA PCR Array (Qiagen) which contains 384 wells and examines 372 miRNAs, 12 controls. Specifically, each well of 372/384 contains a miScript Primer Assay for a miRNome or pathway/disease/functionally-related mature RNA. Moreover, 2 wells contain replicate *C. elegans* miR-39 miScript

Primer Assays and can be used as an alternative normalizer for array data (Ce), 6 wells contain an assay for a different snoRNA/snRNA that can be used as a normalization control for the array data. Finally, there are two wells which contain replicate miRTC Primer Assays (RTC) and two wells that contain positive PCR controls (PPC).

2. Metadata: This file includes a list of sample IDs and the corresponding group e.g. normal/tumor
3. Annotation of miRNAs well: A file that links the information of the well with the examined miRNA.

Workflow

The framework is implemented into three distinct phases, as shown in Fig. 1; [1] QC and normalization, [2] differential analysis and [3] functional analysis. Specifically:

1. QC and normalization.

The quality control process examines two different aspects:

- a. the maximum percentage of not detected or not available values (NA's) in each column. This quality threshold is defined by the user. If a sample

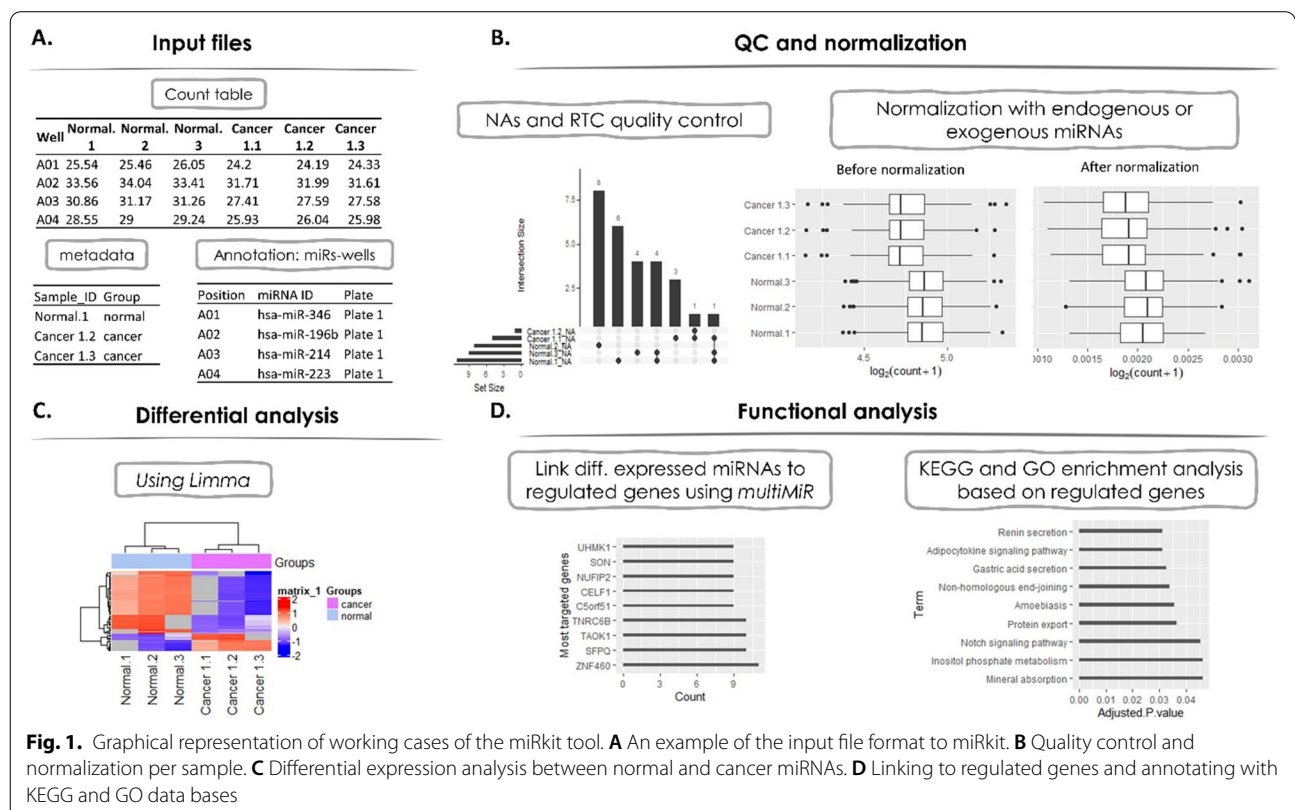


Fig. 1. Graphical representation of working cases of the miRkit tool. **A** An example of the input file format to miRkit. **B** Quality control and normalization per sample. **C** Differential expression analysis between normal and cancer miRNAs. **D** Linking to regulated genes and annotating with KEGG and GO data bases

fails the comparison then it is excluded from the rest of the analysis.

- b. the ratio between reverse transcription control (RTC) assay, which detects an artificial RNA template, and positive PCR controls (PPC), which monitor for PCR inhibitors, is calculated and a standard threshold is used to validate the reverse transcription efficiency which described by Qiagen. Specifically, if the ratio is <5 the sample is passed this quality step.

The data normalization module includes the option of endogenous and exogenous miRNA approach. Normalization of miRNAs using the endogenous controls corrects for factors that could influence the quantification such as different quantities of input RNA, RNA degradation, presence of inhibitors, errors in sample handling. Exogenous controls are typically used on difficult samples such as plasma/serum or other biofluids. Many exogenous controls are not present in humans so it is a good exogenous control for human samples. MiScript PCR Controls are primers designed to quantify a panel of 5 snoRNAs (SNORD61, SNORD68, SNORD72, SNORD95, and SNORD96A) and the snRNA RNU6B (RNU6-2) as endogenous controls and cel-miR-39 as exogenous.

The output of this step is the normalized data matrix that includes the samples which passed the NA's criterion. Additionally, a visualization option is available, which allows to generate figures that are automatically stored within the analysis folder, and include an upset plot for the NA's distribution and boxplots with counts before and after the normalization.

2. Differential analysis

This module is performed using the limma package in R [9]. The output includes the differentially expressed miRNAs using a user-defined adjusted p-value as a threshold. Moreover, a hierarchical cluster analysis is performed at this stage and a corresponding heatmap is constructed and stored in the analysis directory.

3. Functional analysis

The downstream analysis links the differentially expressed miRNAs with the regulated genes using the multiMIR package, which includes several databases such as mirtarbase [10], tarbase [11], diana_microt [12] etc. for both predicted and validated targets. Moreover, the targeted genes of the differentially expressed miRNAs are used for KEGG and

Gene ontology (GO) enrichment analysis, as facilitated by the *enrichR* package [13]. Finally, barplots that present the results of the enrichment analysis are stored in the analysis folder.

At the end of the entire process miRkit produces a list of tables containing all significant events along with the corresponding plots, separated into different folders. Additionally, a report file is automatically exported. The report contains information of the particular execution process, including the user-defined criteria, the rationale for the excluded samples, the overall time required for execution and the total memory usage.

Case study

The miRkit was applied on artificial data provided by Qiagen. The selected NAs percentage threshold was set to 10% and all samples passed the control of NAs and RTC. The data normalization step was performed using the endogenous miRNAs. The total execution time was 18 min and 51.02 s (Table 1) and the memory usage was 274.7 MB. Detailed instructions are available in the repository of the tool (<https://bio.tools/mirkit>) as well as the sample input and the output presented above.

Qualitative comparison to HTqPCR tool

We used HTqPCR [14], a software toolbox for dealing with RT-PCR data in order to compare its functionalities with miRkit since they have several of them in common. Also, HTqPCR is a well-known package for analyzing RT-PCR data. Both of them are written in R including quality control, normalization, clustering and differential analysis. We highlighted the clear advantages of miRkit such as the automatic process of all phases, the linking of the statistically significant miRNAs with public databases. The main differences are listed in Table 2. A clear advantage of miRkit is the automatic process of all phases. Moreover, the linking of the statistically significant miRNAs with public databases gives the user the opportunity to complete the functional analysis within the frame of miRkit.

Overall, miRkit implementation supports a fast execution with low memory usage in order to (i) perform quality control of the samples and data normalization, (ii) identify significant differences on the expression profiles

Table 1 Table with the execution times in each phase

Phase	Execution time
Quality control	10.42 s
Differential analysis	2.32 s
Functional analysis	18 min 38.25 s

Table 2 Comparison of functionalities offered by miRkit and HTqPCR

	miRkit	HTqPCR
Language	R	R
Input	Standard output from miScript miRNA PCR Array (Qiagen) technology/ Handles data from multiple plates	Data preprocessing is required/ Only single-plate data, consisting of either 96 or 384 wells
Usage	Automatic	Manual (users need to write their one code)
Quality control	Yes	Yes
Data filtering	Yes Automatically excluding samples based on NAs and on RTC	No No standard way implemented
Normalization	Yes Endogenous/exogenous genes	Yes Scaling up the values or changing the total distribution of values
Clustering	Yes	Yes
Differential analysis	Yes	Yes
Link miRNAs with databases	YES (mirtarbase, tarbase, diana_microt, etc.)	No
Enrichment analysis of deregulated genes	Yes (KEGG and GO)	No

of miRNAs, and (iii) link the significant miRNAs with the targeted genes and biological processes. In each step of the process, the tool produces also the relevant visual representations of the results.

Compared to the traditional commercial software for analyzing RT-PCR data, miRkit aims to become fully aligned to the FAIR principles (Findable, Accessible, Interoperable, Reusable) for Research Software [15]. With the exception of open source and a free to use tool, miRkit has the distinct advantage in terms of usage, such as the completely automated process, the data filtering based on the data quality, the linking of the differentially expressed miRNAs to genes through different databases, and the GO/KEGG enrichment analysis of deregulated genes. Finally, this package has fast execution time and uses very low memory.

Although the miRkit is focusing on one platform technology, there are many examples of commercial array platforms for which researchers developed tools in R in order to analyze their data with open source tools that are more automated and provide a better workflow for complete analysis. Using the array of 450 K Illumina as an example, there are plenty of options in R such as minfi, RnBeads, shinyMethyl, etc. Having alternative options to analyze the data, especially through workflows supporting completely automated processes from raw data to the association with genes and pathways like the proposed tool, is a critical element for the scientific community.

It is freely available on GitHub and is accompanied by detailed documentation and examples, in order to facilitate the reproducibility of the presented results. Our method provides a new perspective towards analyzing RT-PCR data. Also, it supports efficient data discovery using the gold standard approach of limma analysis and

linking the information with publicly available databases to extract the biological meaning.

Limitations

miRkit focuses on the data from miScript miRNA PCR Array (Qiagen) technology. Other miRNAs technologies which produce count tables i.e. samples on the columns and miRNA of interest on the row, could be possibly analyzed using our tool by following the format of our input files. To this end, a list of conversion scripts will be developed so that a user can use them to convert outputs from other PCR array technologies than Qiagen, as an input to miRkit. This limitation may be addressed in future versions of the miRkit.

Abbreviations

miRNA: MicroRNAs; RT-PCR: Reverse transcription polymerase chain reaction; miRNome: Human miRNA genome; Ct: Threshold cycle; RTC: Reverse transcription control; PPC: Positive PCR controls; NA: Not available value; GO: Gene ontology; FAIR: Findable, Accessible, Interoperable, Reusable.

Acknowledgements

None.

Authors' contributions

MT designed the study and wrote the manuscript. AT, NP and MCM developed the tool, analyzed the data. AB and AM provided the data and reviewed the submitted version. CP and FP designed, supervised the study, and reviewed the manuscript. All authors contributed to the article. All authors read and approved the final manuscript.

Funding

This research is co-financed by Greece and the European Union (European Social Fund-ESF) through the Operational Programme "Human Resources Development, Education and Lifelong Learning 2014–2020" in the context of the project "microRNA from Bone Marrow Mesenchymal Stem Cell-derived exosomes and defective hematopoiesis in Myelodysplastic Syndromes" (MIS 5048511). This research has also been co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation,

under the call RESEARCH–CREATE–INNOVATE (GenOptics, project code: T2E1 K-00407). Finally, this work was also supported by the “Hellenic Network for Precision Medicine” in the framework of the Hellenic Republic—Siemens Settlement Agreement.

Availability of data and materials

The tool is freely available under a MIT license from <https://bio.tools/mirkit>, and offers fast execution times with low memory usage. The miRkit was applied on artificial data provided by Qiagen.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Institute of Applied Biosciences, Centre of Research and Technology Hellas, 57001 Thessaloniki, Greece. ²Haemopoiesis Research Laboratory, School of Medicine, University of Crete, 71003 Heraklion, Greece. ³Department of Hematology, School of Medicine, University of Crete, 71003 Heraklion, Greece. ⁴Department of Genetics, Development and Molecular Biology, School of Biology, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece.

Received: 19 May 2021 Accepted: 15 September 2021

Published online: 26 September 2021

References

- Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*. 2004;116(2):281–97.
- O'Brien J, Hayder H, Zayed Y, Peng C. Overview of microRNA biogenesis, mechanisms of actions, and circulation. *Front Endocrinol*. 2018;9:402.
- Trang P, Weidhaas JB, Slack FJ. MicroRNAs as potential cancer therapeutics. *Oncogene*. 2008;27(Suppl 2):S52–7.
- Hydbring P, Badalian-Very G. Clinical applications of microRNAs. *F1000Res*. 2013;2:136.
- Ha TY. MicroRNAs in human diseases: from cancer to cardiovascular disease. *Immune Netw*. 2011;11(3):135–54.
- Peng Y, Croce CM. The role of MicroRNAs in human cancer. *Signal Transduct Target Ther*. 2016;1:15004.
- Git A, Dvinge H, Salmon-Divon M, Osborne M, Kutter C, Hadfield J, et al. Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression. *RNA*. 2010;16(5):991–1006.
- Rao X, Huang X, Zhou Z, Lin X. An improvement of the 2^{Δ(ΔCT)} method for quantitative real-time polymerase chain reaction data analysis. *Biostat Bioinform Biomath*. 2013;3(3):71–85.
- Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7):e47.
- Hsu SD, Lin FM, Wu WY, Liang C, Huang WC, Chan WL, et al. miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic Acids Res*. 2011;39(Database issue):D163–9.
- Sethupathy P, Corda B, Hatzigeorgiou AG. TarBase: a comprehensive database of experimentally supported animal microRNA targets. *RNA*. 2006;12(2):192–7.
- Paraskevopoulou MD, Georgakilas G, Kostoulas N, Vlachos IS, Vergoulis T, Reczko M, et al. DIANA-microT web server v50: service integration into miRNA functional analysis workflows. *Nucleic Acids Res*. 2013;41(Web Server issue):W169–73.
- Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinform*. 2013;14:128.
- Dvinge H, Bertone P. HTqPCR: high-throughput analysis and visualization of quantitative real-time PCR data in R. *Bioinformatics*. 2009;25(24):3325–6.
- Lamprecht A-L, Garcia L, Kuzak M, Martinez C, Arcila R, Martin Del Pico E, Dominguez Del Angel V, van de Sandt S, Ison J, Martinez PA, McQuilton P, Valencia A, Harrow J, Psomopoulos F, Gelpi JL, Chue Hong N, Goble C. Towards FAIR principles for research software. *Data Sci*. 2020;3:37–59.

Publisher's Note

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

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

