MiRNA-RIP 1.0

User Manual

October 2018

MiRNA-RIP is an R (www.r-project.org) package to predict miRNA regulators of gene expression using matched gene- and miRNA expression profiles. In this manual we will briefly describe the workflow and present a Rectum cancer data set from The Cancer Genome Atlas (TCGA) as an example study. MiRNA-RIP has been tested on a Linux OS.

MiRNA-RIP author:	Volker Ast
Availability:	http://www.leibniz-hki.de/en/mirnarip.html
Manual author:	Volker Ast & Rainer König
Contact:	Rainer.Koenig@uni-jena.de

1. Installation

1.1. Required software

MiRNA-RIP requires the following software:
a) R (version 3.2. or later), available from https://cran.r-project.org/mirrors.html
b) RStudio (RStudio Desktop Open Source License), available from https://www.rstudio.com/products/rstudio/download/
c) Gurobi (version 6.52 or later), available from http://www.gurobi.com/downloads/download-center/

For using Gurobi, a licence is necessary. For academic usage this licence is free and can be obtained after registration if your IP-adress is within a university network or other academic institution. Otherwise, you can obtain an online course license which is limited to 2000 constraints per model. For more details, see “Quick Start Guide” on the Gurobi website (http://www.gurobi.com/documentation/).
After installation and activation (grbgetkey) of Gurobi, the Gurobi R-API has to be installed. This works best using Rstudio:

→ **Tools** → **Install Packages** → **Install from Package Archive File**

In case of a win64 installation, the package archive file can be found at:
"<GUROBI_INSTALLATION_PATH/gurobi<VERSION>/win64/R/gurobi_<VERSION>.zip"

For a Linux system, the package archive file can be found at:
"<GUROBI_INSTALLATION_PATH>/gurobi<VERSION>/linux64/R/gurobi_<VERSION>_R_x86_64_<GNU_VERSION>.tar.gz"

=> after installing the Gurobi R-API, make sure you can load the library on the R console:

```
> library("gurobi")
```

1.2 Required R-packages

All R packages can be installed either using the “install.packages()” command or via BioConductor (https://www.bioconductor.org/).

Besides Gurobi, MiRNA-RIP depends on the following R-packages:
a) slam
b) doParallel
c) doMC
d) foreach
e) topGO
f) org.Hs.eg.db

1.3 Installation of MIRNA-RIP

To install MIRNA-RIP, download the package from
http://www.leibniz-hki.de/en/mirnarip.html

On a Unix/Linux system, execute the following command from a shell:

```
R CMD INSTALL MIRNARIP_1.0.tar.gz
```

or from the R command line:

```
install.packages("MIRNARIP_1.0.tar.gz")
```

=> after installing the MIRNA-RIP package, make sure you can load the library on the R console:

```
> library("MIRNARIP")
```

2. Workflow

2.1 What is MIRNA-RIP?

MIRNA-RIP is a Mixed-Integer-Linear-Programing approch to model miRNA - target gene interactions based on expression data. It combines two modeling approaches: a simple linear regression model and a piecewise linear regression model. In either case, for each pair of miRNA - target gene, the expression of the target gene is estimated using the expression of the miRNA. After running each model type individually, the predictions of both models are combined. Optionally, using all predicted genes for single miRNAs, a gene set enrichment analysis using Gene Ontology terms will be performed.

2.2 What data do you need?

1) Matched miRNA- and gene expression data sets. We recommend a minimum number of 50 overlapping samples. Both files must be tab-delimited, gene and miRNA identifiers as rownames.

2) A table of miRNA - target gene interactions. The table must be tab-delimited. The first column contains mature miRNA identifiers (usually with a -3p or -5p suffix), the second column contains the target gene identifiers:

miRNA	**target_gene**
hsa-miR-20a-5p	HIF1A
hsa-miR-146a-5p	CXCR4
hsa-miR-222-3p	STAT5A
...	...

3) A table mapping mature miRNA IDs to pre-miRNA IDs. The table must be tab-delimited. We used miRBase (http://www.mirbase.org/) to identify pairs of mature and pre-miRNA identifiers. The first column contains mature miRNA identifiers, the second column contains the pre-miRNA identifiers:

mature_id	**pre_mirna_id**
hsa-miR-20a-5p	hsa-mir-20a
hsa-miR-146a-5p	hsa-mir-146a
hsa-miR-122-5p	hsa-mir-122
...	...

4) A table mapping experimental miRNA IDs to pre-miRNA IDs. The table must be tab-delimited. The experimental miRNA IDs must match the miRNA identifiers used in the miRNA expression data set defined in **1)**. The first column contains experimental miRNA identifiers, the second column contains the pre-miRNA identifiers:

exp_mirna	**pre_mirna**
hsa-mir-20a	hsa-mir-20a
hsa-mir-146a	hsa-mir-146a
hsa-mir-365-2	hsa-mir-365b
hsa-mir-365-1	hsa-mir-365a
...	...

5) A list of miRNAs you want to model. Make sure that the miRNAs have expression values with a minimum of variance and at least one target gene with existing gene expression values mapped. The files must contain one column without header:

hsa-mir-20a
hsa-mir-146a
hsa-mir-365-2
...

2.3 How does it work?

A schematic workflow is depicted in Figure 1. In the following, we will briefly explain the single modules.

1a) Starting with miRNA- and gene expression data, the intersection samples of both data sets are determined and stored in a separate file.

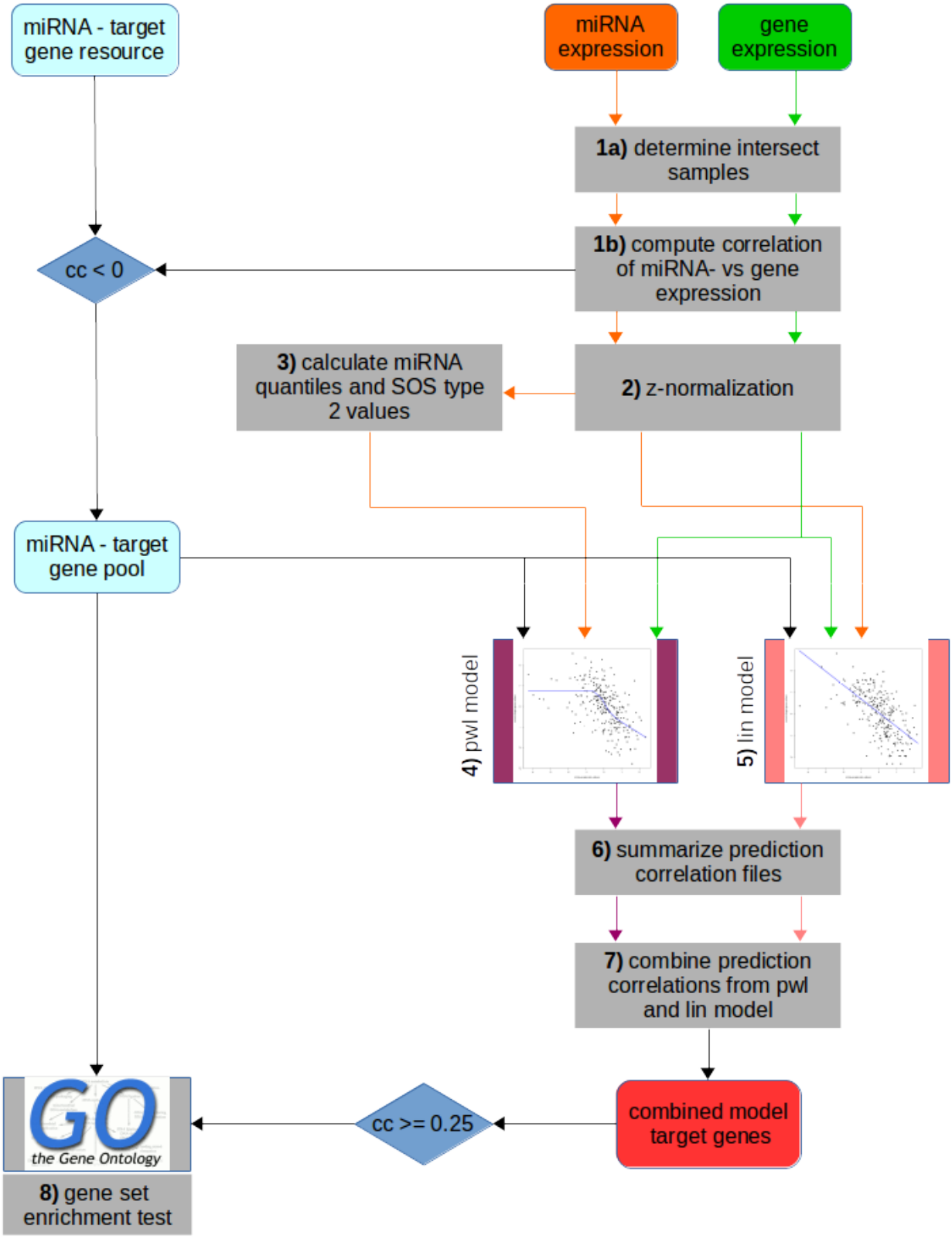
1b) For all miRNAs and target genes available in the expression files, Pearson's correlation of miRNA- and gene expression is computed. Only overlapping samples determined in the previous step are included. This step will generate a (#Genes x #miRNAs) - matrix of Pearson's correlation coefficients. This procedure is running in parallel jobs but still could take some time depending on the size of the expression data.

2) Both the miRNA- and the gene expression data matrix are z-normalized.

3) For each miRNA, the 0%, 20%, 40%, 60%, 80% and 100% quantiles are determined using the z-normalized data generated in step 2. MiRNAs with NA values or insufficient variance (i.e. minimum value = median values) will be removed. SOS type 2 values are computed per miRNA and sample. For each miRNA, a single (#Samples x #quantiles) matrix containing SOS type 2 values will be generated and stored as a single file.

4) The piecewise linear model is set up for every miRNA - target gene pair individually. MiRNA- target gene pairs are retrieved from an external resource (i.e. TarBase); only target genes with a correlation coefficient below a cutoff (default: 0) will be considered. Depending on the number of cross validations (default: 5), the sample set is split into (#cross-validation) multiple subsections. Excluding one subset at a time, the model is solved with the remaining subsets by determining the β- and the x parameter using the SOS type 2 values calculated in step 3. The β- and the x parameters are used to calculate the predicted gene expression of the respective target gene for the samples in the left out partition. This whole process is repeated (#cross-validation) times. Finally, the correlation of predicted and measured gene expression is calculated and stored in a file.

Figure 1: The MIRNA-RIP workflow



5) The setup of the linear model is analogous to the piecewise linear model except that only one β - parameter is fitted for each miRNA - target gene pair and the z-transformed miRNA values are used instead of the SOS type 2 values.
6) For each model type, the single prediction correlation files are combined in one file.
7) For each miRNA - targert gene pair, the prediction correlation coefficients from the linear and the piecewise linear model are compared. In the combined model, the better correlation coefficient is used as the final prediction value for the respective miRNA - target gene pair.
8) (Optional) A gene set (Gene Ontology BP) enrichment analysis using TopGO is performed. The predictions from the combined model established in step 7 are filtered for miRNA - target gene pairs with a prediction performance above a cutoff (default: 0.25). These target genes are considered to be "well predicted" by the specific miRNA and are used for the enrichment analysis. As a background, all potential target genes of the specific miRNAs are taken into account. The user can define a minimum number of target genes (default: 5) and TopGO specific parameters as GO node size (default: 5) and the significance cutoff (default:0.05).

3. Example

On the basis of a Rectum cancer data set from The Cancer Genome Atlas (TCGA), we will demonstrate how to run MIRNA-RIP.

3.1. Data files

1) The file “RNAseqV2_expression_matrix_update_symbols.csv” contains gene expression data with gene symbol identifiers. The matched miRNA expression data can be found in “miRNA_seq_expression_matrix.csv”. Alternatively, you can download the gene- and miRNA expression data files is via the UCSC Xena Cancer Browser. Go to https://xenabrowser.net/hub/, click on "TCGA hub", choose "TCGA Rectal Cancer" and download gene expression (IlluminaHiSeq) and miRNA expression (IlluminaHiseq) data.
2) The file "miRNA_target_gene_TarBase.csv" contains experimentally validated miRNA - target gene interactions extracted from TarBase (http://carolina.imis.athena-innovation.gr/diana_tools/web/index.php?r=tarbasev8%2Findex). We did not consider the applied detections methods. Only interactions observed in human cells were included.
3) The files "total_mature_pre_mirna_mapping.csv" and "total_experimental_pre_mirna_mapping.csv" provide mappings of mature miRNA IDs to

pre-miRNA IDs and experimental miRNAs to pre-miRNA IDs for the TCGA Rectal Cancer data set as described in section 2.2.

4) Create a file "mirna_candidates.txt" containing all your miRNA IDs of interest. In this example, we will analyse only one miRNA, "hsa-mir-192".

5) Download the file "start_MIRNARIP.R".

3.2. Define your data

1) Copy all files listed in 3.1 in a directory, for instance "/home/TCGA_rectum/data/".

2) Open the file "start_MIRNARIP.R"

3) Modify the output and the data path:

```
basic_output_path = "/home/TCGA_rectum/MIRNARIP_result/"
data_path = "/home/TCGA_rectum/data/"
```

4) Modify the name of the data files if they are named differently.

5) Modify the number of cores that will be used for parallel computing. Default to 40 cores for step 1, 4, 5 and 8.

6) Optional: modify the paramters neg_cor_cutoff, num_fold_cross_validation, quantiles and beta_m if needed.

3.3 Start the workflow

```
source("/path/to/start_MIRNARIP.R")
```

Please note: depending of the data set, the number of miRNAs and the number of target genes, this workflow can take a long time to finish!

Once the workflow has finished, the result files can be found in the basic_output_path

- model/mirna/linear/results/ contains the predictions performances of the linear model
- model/mirna/piecewise_linear/results/ contains the predictions performances of the piecewise linear model
- model_combination/mirna_combined_model_cor_summary.csv contains the combined predictions from both models for all modeled miRNAs and target genes:

mirna	**gene**	**cor**
hsa-mir-192	ABCA8	0.22
hsa-mir-192	ABCB5	0.53
hsa-mir-192	ABI2	0.23
...	...	...

Per default, you would use the model combination results and genes with a correlation of predicted and measured gene expression of at least 0.25 for further analysis. Theses genes are more likely to be targets of the specific miRNA according to the used data set(s) and condition(s).

3.4 Gene set enrichment analysis

Optionally, you could perform a gene set enrichment analysis with the predicted target genes. To define "well predicted" target genes, we set the cutoff good_prediction_cutoff to 0.25. You can modify this cutoff if you want to include target genes with a lower prediction correlation. In addition, you can change the TopGO specific parameters min_target_gene_num, go_node_size and top_go_sign_cutoff. All potential target genes listed in the miRNA - target gene interactions table (described in **2.2.2**) are used as background gene set.

If you want to include the gene set enrichment analysis in the workflow, uncomment the R-code at the very end of "start_MIRNARIP.R" and re-start the workflow.

Using the predictions from the combined model, the folder model_combination/enrichment/GO/good_prediction_genes_background_all_target_genes contains enrichment files for the "well predicted" target genes of all modeled miRNAs.

The following table shows enriched GO terms for "well predicted" target genes of hsa-mir-192:

GO.ID	Term	Annotated	Significant	Expected	classic	adj_p_value	sign_genes
GO:0007420	brain development	5	4	0.76	0.0019	0.43725	ACTB,CADM1,EMX2,GRIN2A
GO:0032940	secretion by cell	11	6	1.66	0.0019	0.43725	CADM1,FGF7,NRXN3,SLC5A7,S
GO:0050877	neurological system p	11	6	1.66	0.0019	0.43725	EIF4EBP2,GRIN2A,NRXN3,SIX1
...	...	...	...	...	...	...	...

4 References

[1] The Cancer Genome Atlas Network. Comprehensive Molecular Characterization of Human Colon and Rectal Cancer. *Nature*. 2012;487(7407):330-337. doi:10.1038/nature11252.

[2] Dimitra Karagkouni, Maria D. Paraskevopoulou, Serafeim Chatzopoulos, Ioannis S. Vlachos, Spyros Tastsoglou, Ilias Kanellos, Dimitris Papadimitriou, Ioannis Kavakiotis, Sofia Maniou, Giorgos Skoufos, Thanasis Vergoulis, Theodore Dalamagas, Artemis G. Hatzigeorgiou; DIANA-TarBase v8: a decade-long collection of experimentally supported

miRNA–gene interactions, Nucleic Acids Research, gkx1141, https://doi.org/10.1093/nar/gkx1141

[3] Ana Kozomara, Sam Griffiths-Jones; miRBase: annotating high confidence microRNAs using deep sequencing data, Nucleic Acids Research, Volume 42, Issue D1, 1 January 2014, Pages D68–D73, https://doi.org/10.1093/nar/gkt1181