Gene expression

iMIRAGE: an R package to impute microRNA expression using protein-coding genesAritro Nath ¹, Jeremy Chang^{2,3} and R. Stephanie Huang ^{1,*}¹Department of Experimental and Clinical Pharmacology, University of Minnesota Twin Cities, Minneapolis, MN 55455, USA, ²Biological Sciences Collegiate Division, The University of Chicago, Chicago, IL 60637, USA and ³Weill Cornell Medical College, Weill Cornell Medicine, New York City, NY 10021, USA

*To whom correspondence should be addressed.

Associate Editor: Yann Ponty

Received on March 4, 2019; revised on November 30, 2019; editorial decision on December 16, 2019; accepted on December 17, 2019

Abstract

Summary: MicroRNAs (miRNAs) are critical post-transcriptional regulators of gene expression. Due to challenges in accurate profiling of small RNAs, a vast majority of public transcriptome datasets lack reliable miRNA profiles. However, the biological consequence of miRNA activity in the form of altered protein-coding gene (PCG) expression can be captured using machine-learning algorithms. Here, we present iMIRAGE (imputed miRNA activity from gene expression), a convenient tool to predict miRNA expression using PCG expression of the test datasets. The iMIRAGE package provides an integrated workflow for normalization and transformation of miRNA and PCG expression data, along with the option to utilize predicted miRNA targets to impute miRNA activity from independent test PCG datasets.

Availability and implementation: The iMIRAGE package for R, along with package documentation and vignette, is available at <https://aritrnath.github.io/iMIRAGE/index.html>.

Contact: rshuang@umn.edu

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

MicroRNAs (miRNAs) are short, endogenous non-coding RNAs ranging from 19 to 25 nucleotides in length that silence gene expression by recruiting Argonaute proteins to the 3'-UTR of target mRNAs (Gebert and MacRae, 2019). Despite constituting only about 2% of all genes (Bartel, 2009), miRNAs control the expression of most protein-coding genes (PCGs), thereby influencing virtually all biological processes (Vidigal and Ventura, 2015). The dysregulation of miRNAs is implicated in the pathogenesis of many human diseases, including cardiovascular disease, neurodevelopmental disorders, autoimmune disease and cancer (Bushati and Cohen, 2007; Li and Kowdley, 2012).

Accurate profiling of the miRNA transcriptome requires overcoming several challenges due to the small size of the mature miRNA transcript, the lack of a poly-A tail and overall lower abundance compared to other transcripts (Pritchard *et al.*, 2012). Consequently, it is difficult to infer miRNA activity from the vast majority of RNAseq and microarray expression datasets, which did not utilize specialized sample and library preparation methods to profile small RNAs. To address this problem, we present imputed miRNA activity from gene expression (iMIRAGE), an R package to impute the expression of miRNAs using PCGs.

The expression of PCGs is under the control of various transcription factors, epigenetic regulators and non-coding RNAs, including

miRNAs. Therefore, the activity of a miRNA is expected to contribute to the altered expression patterns of direct or indirect target genes (Gurtan and Sharp, 2013). Based on this assumption, our method utilizes machine-learning models trained on high-quality PCG and miRNA transcriptome data to impute the miRNA activity on an independent PCG dataset.

2 Features and implementation

The iMIRAGE workflow involves training models on user-supplied PCG and miRNA transcriptome datasets, and predicting miRNA expression using the independent PCG dataset of the samples of interest (Fig. 1A). The package facilitates harmonization of the input data followed by cross-validation and imputation with the option of using miRNA target information.

2.1 Data cleanup and pre-processing

To impute the miRNA activity using the independent PCG data of interest, the user provides a training dataset containing matched PCG and miRNA transcriptome data, for example, from GTEx (Jensen *et al.*, 2017) for normal tissues or from GDC for cancer tissues (Grossman *et al.*, 2016). It is recommended that mature miRNA transcriptome profiles from small RNAseq or microarrays

Fig. 1. iMIRAGE workflow and example. (A) Flow chart depicting features and typical workflow. Various iMIRAGE functions that perform the necessary steps are indicated in blue above the arrows. Optional cleanup and pre-processing steps are italicized. (B) Density plot comparing imputed prostate cancer miRNA expression using various options with measured miRNAseq expression. RF and KNN refer to RF and k-NN regression, and Target refers to the use of predicted targets as training feature. The correlation between RNAseq and miRNAseq profiles is shown for comparison. Dashed lines indicate mean prediction accuracy. (Color version of this figure is available at *Bioinformatics* online.)

designed to capture miRNA expression are used as training datasets. Moreover, it is important to select training datasets that cover the same sample conditions as the PCG data of interest (e.g. the same tissue-type or cell lines). A minimum sample size of 50 is recommended.

The training (PCG and miRNA) and test (PCG) datasets are first harmonized using the *match.gex* function (required), which retains gene IDs that are common across the datasets. Next, the *filter.exp* function removes miRNAs or PCGs profiles expressed below a user-defined threshold. By default, all genes with expression > 0 in at least 75% of the samples are retained. The *pre.process* function performs a number of additional transformation and normalization steps. First, this function removes training features with zero variance and then scales the expression of each gene to a mean of zero and unit SD. In addition, users have the option of enabling log transformation ($\log_2(x + 1)$) or upper-quantile normalization for unprocessed datasets. Both the cleanup and pre-processing steps are optional. However, we strongly recommend variance filtering to remove features that do not contribute to prediction models, and scaling to counter the effects of values of large magnitude, as minimal pre-processing. The expression levels of the miRNAs are not required to be scaled before use.

2.2 Imputation and accuracy

After cleanup and pre-processing of the input data, the *imimage.cv* function determines imputation accuracy by performing K-fold cross-validation on the training PCG and miRNA data. This function generates performance metrics, including as Spearman's coefficients, correlation *P*-value and root mean squared error, to identify miRNAs that can be accurately imputed using the supplied training datasets. Subsequently, the miRNA(s) of interest can be imputed using the test PCG data using the *imimage* function. Both cross-validation and imputation function can use either all PCGs or built-in predicted miRNA targets from the TargetScan database (Agarwal *et al.*, 2015) as training features, with the option of using either random forests (RF), support vector machines or k-nearest neighbors

(KNN) algorithms for training and prediction. In addition, the package allows the user to provide their own miRNA-target gene information for use in the imputation workflow.

2.3 Application

We demonstrate the application of iMIRAGE by comparing the RNAseq-measured and imputed miRNA expression of 477 primary prostate cancer tumors (Abeshouse *et al.*, 2015) (Fig. 1B, Supplementary Table S1). In this example, we imputed the expression of miRNAs in samples that were profiled in both the RNAseq and miRNAseq datasets. We then identified miRNAs that were present in the TargetScan database and expressed at non-zero levels in both datasets in at least 75% of the samples, yielding 81 miRNAs. Next, the models were trained with RFs or k-NN using all PCGs or only predicted targets using RNAseq PCG profiles and miRNAseq data of the prostate cancer samples. The density curves in Figure 1B shows the concordance of RNAseq-measured and imputed miRNA profiles with the measured miRNAseq data. This analysis demonstrates the benefit of imputed miRNA over RNAseq data.

We highlight the application of iMIRAGE to impute miRNA profiles of independent datasets using breast cancer miRNA datasets that were generated using different sequencing platforms on mutually exclusive set of samples (Supplementary Table S2). Additional examples focusing on lymphoblastoid cell lines are also discussed in Supplementary Table S3. Using the breast cancer datasets, we evaluated the ability to impute expression of miRNAs that are relevant to the pathogenesis or progression of the disease (Supplementary Table S4). Next, we obtained the average cross-validation accuracy within each dataset and imputation accuracies from each pair of independent training and test dataset (Supplementary Table S5). We observed that most phenotypically relevant miRNAs exhibit good cross-validation and imputation accuracies, even across varying sequencing or array platforms. In addition, we found that the differential expression patterns of the majority of miRNAs were recapitulated when imputed using independent datasets (Supplementary Table S6, Supplementary Figs. S1–S3). To show that the ability to impute miRNAs is not limited to cancer samples alone, we demonstrate the application with lymphoblastoid cell lines (Supplementary Fig. S3). Finally, we provide a summary of the cross-validation and imputation analyses in independent datasets in Supplementary Table S7, demonstrating that imputations are generally accurate when the training and independent PCG datasets are of the same tissue-type, while imputations across datasets of vastly different tissue of origin are not advisable.

2.4 Availability

The iMIRAGE package can be installed in R by downloading the repository from <https://github.com/aritronath/iMIRAGE>. Additional instructions for use along with a detailed discussion of the analyses using the TCGA breast cancer datasets along with the capabilities of iMIRAGE are provided in the package vignette (<https://aritronath.github.io/iMIRAGE/articles/imimage.html>). A small subset of these datasets are also included in the iMIRAGE library and are used as examples in the package documentation. The prostate cancer data used in the manuscript example (Fig. 1B) is available at <https://osf.io/78mjxl/>.

Acknowledgement

The authors would like to thank Dr Hae Kyung Im at the University of Chicago for her guidance and critical evaluation of this work.

Funding

This work was supported by an NIH/NCI grant [1R01CA204856-01A1], and a research grant from the Avon Foundation for Women.

Conflict of Interest: none declared.

References

- Abeshouse, A. *et al.* (2015) The molecular taxonomy of primary prostate cancer. *Cell*, **163**, 1011–1025.
- Agarwal, V. *et al.* (2015) Predicting effective microRNA target sites in mammalian mRNAs. *eLife*, **4**, e05005.
- Bartel, D.P. (2009) MicroRNAs: target recognition and regulatory functions. *Cell*, **136**, 215–233.
- Bushati, N. and Cohen, S.M. (2007) MicroRNA functions. *Annu. Rev. Cell Dev. Biol.*, **23**, 175–205.
- Gebert, L.F.R. and MacRae, I.J. (2019) Regulation of microRNA function in animals. *Nat. Rev. Mol. Cell Biol.*, **20**, 21–37.
- Grossman, R.L. *et al.* (2016) Toward a shared vision for cancer genomic data. *New Engl. J. Med.*, **375**, 1109–1112.
- Gurtan, A.M. and Sharp, P.A. (2013) The role of miRNAs in regulating gene expression networks. *J. Mol. Biol.*, **425**, 3582–3600.
- Jensen, M.A. *et al.* (2017) The NCI genomic data commons as an engine for precision medicine. *Blood*, **130**, 453–459.
- Li, Y. and Kowdley, K.V. (2012) MicroRNAs in common human diseases. *Genom. Proteom. Bioinform.*, **10**, 246–253.
- Pritchard, C.C. *et al.* (2012) MicroRNA profiling: approaches and considerations. *Nat. Rev. Genet.*, **13**, 358–369.
- Vidigal, J.A. and Ventura, A. (2015) The biological functions of miRNAs: lessons from *in vivo* studies. *Trends Cell Biol.*, **25**, 137–147.