The First Draft

2022-12-10

Contents

1 Materials and Methods **1**
- 1.1 Microarray data collection . . . 1
- 1.2 Pre-processing . . . 2
- 1.3 Differential expression analysis . . . 2
- 1.4 Functional and pathway enrichment analyses . . . 2
- 1.5 Identification of hub genes . . . 2
- 1.6 Machine Learning . . . 3

2 Results and Discussion **3**
- 2.1 Pre-processing . . . 3
- 2.2 Differential expression analysis . . . 5
- 2.3 Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses of the DEGs. . . . 5
- 2.4 Machine Learning . . . 5

Refrences **5**

1 Materials and Methods

1.1 Microarray data collection

Microarray datasets were obtained from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/). In order to come by sufficient classification power for MI samples and others, the sample size of the dataset should be relatively large. Therefore, GSE59867 and GSE56609 were selected as train set and GSE62646 and GSE54475 as validation set for both DEGs and pathway analyse as well as machine learning section. All datasets have been conducted using Affymetrix Human Gene 1.0 ST Array (GPL6244) platform. Only healthy, stable CAD and early stage MI samples were selected from these datasets for further analysis. The basic information for the four GEO datasets evaluated in the current study is provided in Table 1.

Table 1: Basic information of the 4 GEO microarray datasets.

	Platform	Healthy Control	CAD Control	MI	Reference
Training Set					
GSE59867	GPL6244	-	46	111	(Maciejak et al. 2015)
GSE56609	GPL6244	46	-	-	(Matone et al. 2015)
Test Set					
GSE62646	GPL6244	-	14	28	(Kiliszek et al. 2012)

	Platform	Healthy Control	CAD Control	MI	Reference
GSE54475	GPL6244	5	-	-	(Canali et al. 2014)

1.2 Pre-processing

Raw data (CEL files) of the four datasets were downloaded from the GEO and preprocessed using the fRMA package (M. N. McCall, Bolstad, and Irizarry 2010). fRMA allows to preprocess individual microarray samples and combine them consistently for analysis. For each dataset, background correction is performed and then it is quantile normalized based on the reference distribution. During summarization, batch effects are removed and variances of the gene expressions are estimated by taking into account these probe-specific effects. For those multiple probe sets matched to the identical gene, the mean log fold change was retained. This way fRMA can be seen as a batch effect removal technique for different datasets produced using identical microarray platform. Thus, In order to ensure about batch effect removal, the principal component analysis and the relative log expression of all samples were plotted before and after fRMA (Lazar et al. 2013).

1.3 Differential expression analysis

The barcode algorithm proposed by McCall et al. (Matthew N. McCall et al. 2011) transforms the actual expression values into binary barcode values. Huge sets of samples were collected and normalized using fRMA for several platforms. The distribution of the expressed and non-expressed observed intensities for each gene is estimated using these normalized sets. Genes are deemed expressed (and their value coded to 1) or unexpressed (and their value coded to 0) according to the following equation:

$$\hat{x}_{ij} = \begin{cases} 1 & \text{if } x_{ij} >= \mu^{ne} + C \times \sigma^{ne} \\ 0 & \text{otherwise} \end{cases}$$

where x_{ij} is the normalized intensity of gene i in sample j, C is a user-defined parameter, σ^{ne} is the standard deviation of the non-expressed distribution and μ^{ne} is the mean of the non-expressed distribution. The barcode representation of a sample is a vector of ones and zeros denoting which genes are estimated to be expressed (ones) and unexpressed (zeros). The barcode algorithm was implemented by the barcode function in the R fRMA package, and the default value of C was used.

To determine if the expressed ratios differed in the MI group versus the healthy control group, Fisher's exact test for individual genes was carried out upon the barcode values. Genes with a false discovery rate (FDR) of < 0.05, which was calculated through the Benjamini-Hochberg (BH) procedure to adjust for multiple testing issue, were considered as differentially expressed genes.

1.4 Functional and pathway enrichment analyses

Using the R clusterProfilter package, the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis and Gene Ontology (GO) functional annotation were carried out on the differentially expressed genes. The GO analysis included biological process (BP), cellular component (CC) and molecular function (MF) categories. An adjusted $P < 0.05$ was considered to indicate a statistically significant difference. In these analyses, all default parameters were used.

1.5 Identification of hub genes

The online Search Tool for the Retrieval of Interacting Genes (STRING) database (http://string-db.org/) [50] was used to obtain the predicted interactions for the DEGs. The protein–protein interaction (PPI) network of the DEGs was visualized with Cytoscape software (Version 3.9.1, http://www.cytoscape.org/). The CytoHubba plugin in Cytoscape features 12 different algorithms to analyse PPI network topology: ?? Maximal Clique Centrality (MCC), Density of Maximum Neighborhood Component (DMNC), Maximum

Neighborhood Component (MNC), Degree, Component (EPC), BottleNeck, EcCentricity, Closeness, Radiality, Betweenness, Stress and ClusteringCoefficient [51]. The outputs of these algorithms can be integrated to identify hub genes.??

1.6 Machine Learning

The machine learning analysis was performed using Python software, ver. 3.9, numpy (Harris et al. 2020), pandas (McKinney 2010) and Scikit-Learn packages (Pedregosa et al. 2011). In all ML analysis, the train datasets where divided into a train and a development set by 0.7:0.3 ratio.

A two layer architecture has been deployed to the data in order to maximize the prediction values. The first layer will predict whether a sample is healthy or not, and the second layer will separates MI from CAD in the samples which were predicted as not healthy in the first layer. To this end, a distinct ML model was trained for each layer.

1.6.1 First layer: seperating healthy and not healthy samples

Different SVM models using linear, polynomial, and RBF kernels were trained using single miRNAs and different combinations of them. The receiver operating characteristic (ROC) curve were generated for them and area under curves (AUC) were calculated. The best combination were selected based on AUC, accuracy and ??.

1.6.2 Second layer: seperating MI and CAD samples

In order to investigate, whether a single miRNAs could predict the presence of MI with good sensitivity and specificity or not, a simple SVM with linear kernel model has been trained using each differentially expressed miRNA and the ROC curve were generated. Moreover, we tested combination of miRNAs to evaluate their ability to improve the models predictive values. The miRNA combination with the highest AUC has been used to train different algorithms for the sake of finding the model with the best diagnosing capability. To do so, SVM (with linear, polynomial, and RBF kernels), Logistic Regression (LR), Random Forests (RF), k-Nearest Neighbor (kNN), Multi-layer Perceptron (MLP), Gradient Boosting (GB), XGBoost (XGB) and Decision Tree (DT) has been trained using the expression profile of the best combination miRNAs. All models were trained with their pre-set parameters with 10-fold cross-validation.

Criteria for choosing the best model was the highest accuracy on development set between the models with train accuracy of > 0.95. The best model was hypertuned with scikit-opt package (Head et al. 2021) to get the best predictive performance. After finding hypertunning the best model based on the train and development sets, the model performance has been evaluated on the two unseen test datasets (GSE62646 and GSE54475).

2 Results and Discussion

2.1 Pre-processing

The PCA plot of the samples were shown in fig1. As it is clear, there is a complete separation between healthy samples and CAD and MI samples in primary data. Moreover, there is a relative separation between samples with the same disease status in different datasets, which could be considered as a sign of batch effect (Lazar et al. 2013). After pre-processing data with fRMA algorithm, the separation between healthy samples and CAD and MI samples still remain, but the separation between samples with the same disease status has been removed, as a result of batch effect removal.

The RLE plot presented in fig1 also validates batch effect removal. For an efficient batch effect removal method, the individual boxplots will be all distributed around 0 in RLE plot, and inter-quantile distances would be grater than 0.1 (Lazar et al. 2013). The mentioned criteria is not met in primary data, but has been met after conducting fRMA algorithm.

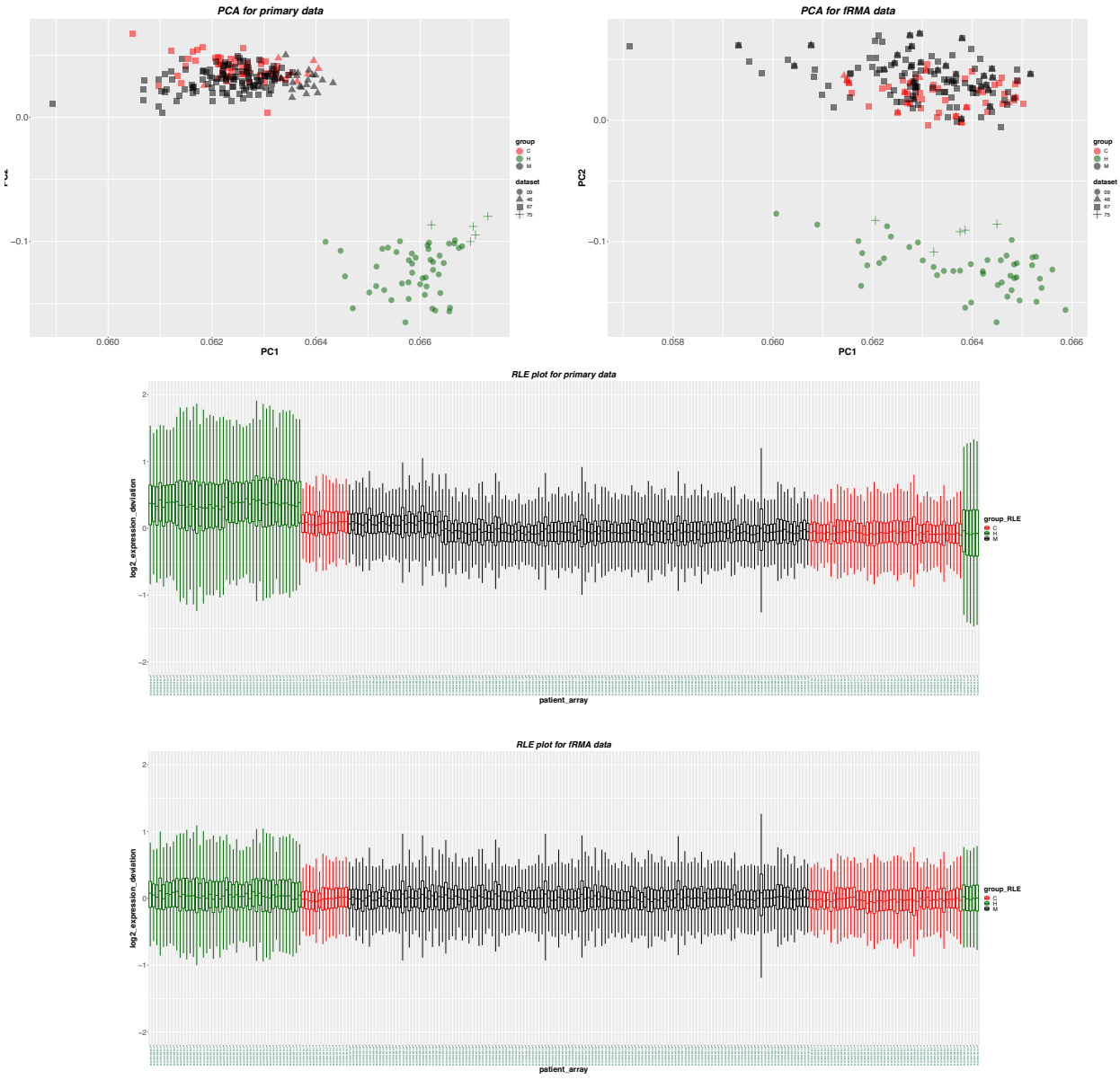

Figure 1: PCA and RLE plot for all samples before and after fRMA.

2.2 Differential expression analysis

According to the cutoff criterion of $FDR < 0.05$, there are 871 DEGs between the MI patients and the healthy controls. Among them, 307 are over-expressed in MI, and 564 are down-expressed in MI compared to the healthy controls.

2.3 Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses of the DEGs.

To explore the biological classification of the DEGs, we performed GO and KEGG pathway enrichment analyses. Many biological functions enriched with the DEGs were associated with the immune cells, as expected. GO enrichment analysis in the cellular component (CC) category suggested that the robust DEGs were enriched in "secretory granule membrane" and "sazurophil granule" (fig2A). In the biological process (BP) category, the robust DEGs were enriched in "mononuclear cell differentiation", "lymphocyte differentiation", "leukocyte activation involved in immune response", and "lymphocyte activation involved in immune response" (fig2B). In the molecular function (MF) category, the robust DEGs were involved in "cadherin binding" and "MHC class I protein binding" (fig2C). KEGG pathway analysis indicated that the robust DEGs were related to the following pathways: "Chemokine signaling pathway", "Lipid and atherosclerosis", "Hematopoietic cell lineage" and "Chronic myeloid leukemia" (fig2D). The above results suggested that the abnormal expression of the DEGs may ??.

2.4 Machine Learning

Among all DEGs, just hsa-miR-186, hsa-miR-32, and hsa-miR-21 are differntially expressed miRNAs. The expression profile of these three miRNAs are presented in fig3. The ROC curves of each miRNA is presented in

Refrences

Canali, Raffaella, Lucia Natarelli, Guido Leoni, Elena Azzini, Raffaella Comitato, Oezgur Sancak, Luca Barella, and Fabio Virgili. 2014. "Vitamin C Supplementation Modulates Gene Expression in Peripheral Blood Mononuclear Cells Specifically Upon an Inflammatory Stimulus: A Pilot Study in Healthy Subjects." *Genes & Nutrition* 9 (3): 390. https://doi.org/10.1007/s12263-014-0390-x.

Harris, Charles R., K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, et al. 2020. "Array Programming with NumPy." *Nature* 585 (7825): 357–62. https://doi.org/10.1038/s41586-020-2649-2.

Head, Tim, Manoj Kumar, Holger Nahrstaedt, Gilles Louppe, and Iaroslav Shcherbatyi. 2021. *Scikit-Optimize/Scikit-Optimize* (version v0.9.0). Zenodo. https://doi.org/10.5281/zenodo.5565057.

Kiliszek, Marek, Beata Burzynska, Marcin Michalak, Monika Gora, Aleksandra Winkler, Agata Maciejak, Agata Leszczynska, Ewa Gajda, Janusz Kochanowski, and Grzegorz Opolski. 2012. "Altered Gene Expression Pattern in Peripheral Blood Mononuclear Cells in Patients with Acute Myocardial Infarction." *PLoS ONE* 7 (11): e50054. https://doi.org/10.1371/journal.pone.0050054.

Lazar, C., S. Meganck, J. Taminau, D. Steenhoff, A. Coletta, C. Molter, D. Y. Weiss-Solis, R. Duque, H. Bersini, and A. Nowe. 2013. "Batch Effect Removal Methods for Microarray Gene Expression Data Integration: A Survey." *Briefings in Bioinformatics* 14 (4): 469–90. https://doi.org/10.1093/bib/bbs037.

Maciejak, Agata, Marek Kiliszek, Marcin Michalak, Dorota Tulacz, Grzegorz Opolski, Krzysztof Matlak, Slawomir Dobrzycki, Agnieszka Segiet, Monika Gora, and Beata Burzynska. 2015. "Gene Expression Profiling Reveals Potential Prognostic Biomarkers Associated with the Progression of Heart Failure." *Genome Medicine* 7 (1): 26. https://doi.org/10.1186/s13073-015-0149-z.

Matone, Alice, Colm M. O'Grada, Eugene T. Dillon, Ciara Morris, Miriam F. Ryan, Marianne Walsh, Eileen R. Gibney, et al. 2015. "Body Mass Index Mediates Inflammatory Response to Acute Dietary Challenges." *Molecular Nutrition & Food Research* 59 (11): 2279–92. https://doi.org/10.1002/mnfr.201500184.

McCall, M. N., B. M. Bolstad, and R. A. Irizarry. 2010. "Frozen Robust Multiarray Analysis (fRMA)." *Biostatistics* 11 (2): 242–53. https://doi.org/10.1093/biostatistics/kxp059.

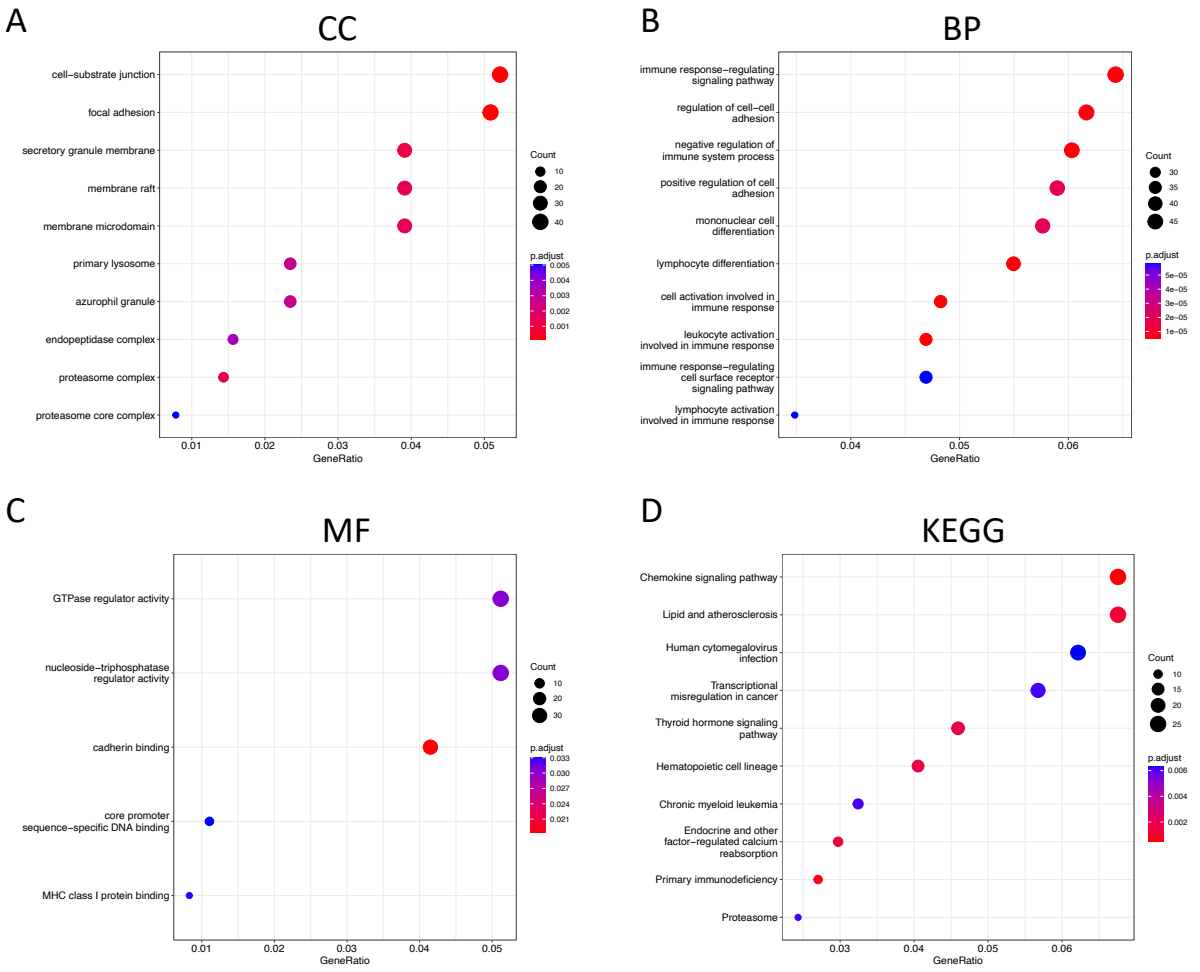

Figure 2: Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways enriched with the DEGs. (A) Biological process terms enriched with the DEGs. (B) Cellular component terms enriched with the DEGs. (C) Molecular function terms enriched with the DEGs. (D) KEGG analysis of the DEGs. The respective pathway involved in MI was identified by using the KEGG pathway database.

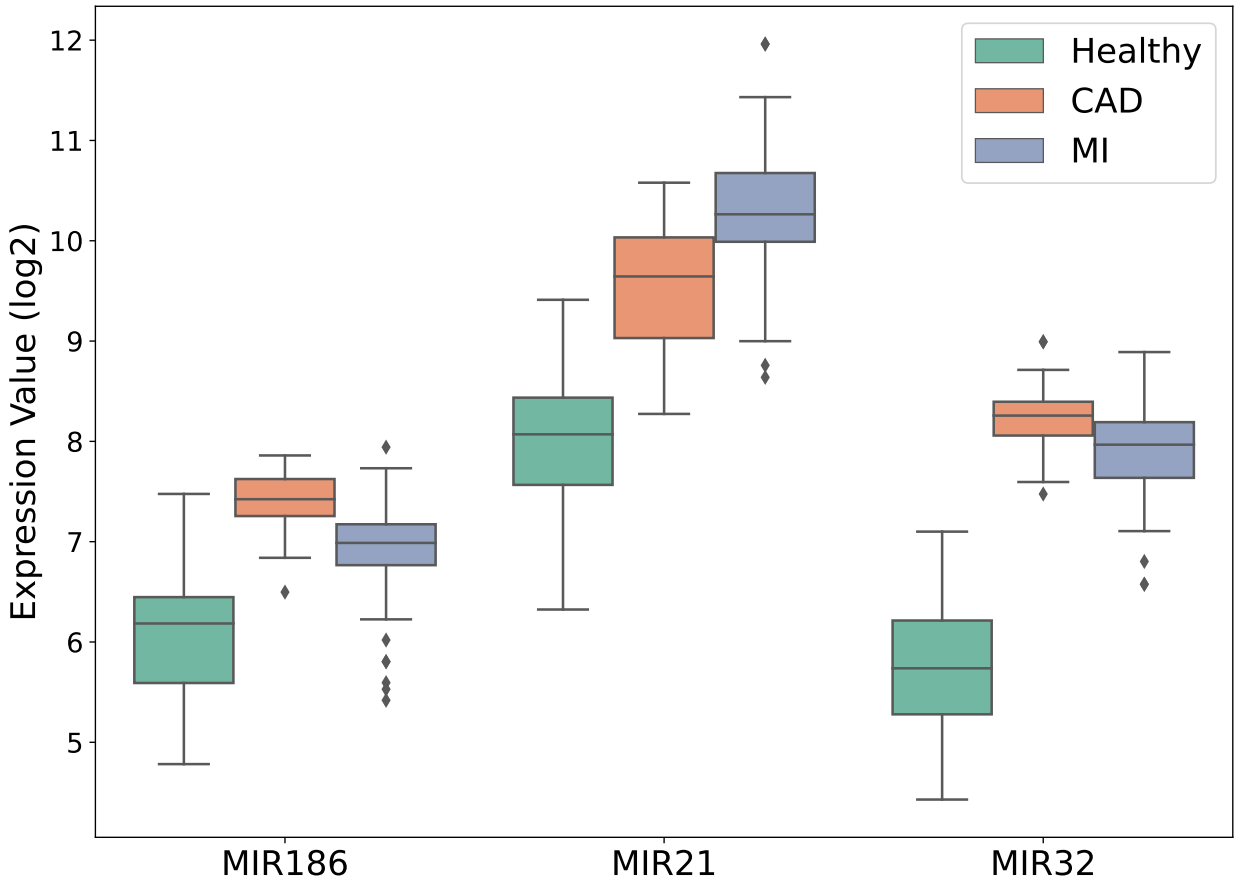

Figure 3: Barplot of differentially expressed miRNAs expression values.

McCall, Matthew N., Karan Uppal, Harris A. Jaffee, Michael J. Zilliox, and Rafael A. Irizarry. 2011. "The Gene Expression Barcode: Leveraging Public Data Repositories to Begin Cataloging the Human and Murine Transcriptomes." *Nucleic Acids Research* 39 (suppl_1): D1011–15. https://doi.org/10.1093/nar/gkq1259.

McKinney, Wes. 2010. "Data Structures for Statistical Computing in Python." In *Proceedings of the 9th Python in Science Conference*, edited by Stéfan van der Walt and Jarrod Millman, 56–61. https://doi.org/10.25080/Majora-92bf1922-00a .

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–30.