Current Progress of High-Throughput MicroRNA Differential Expression Analysis and Random Forest Gene Selection for Model and Non-Model Systems: an R Implementation

Jing Zhang[1], Hanane Hadj-Moussa[1], Kenneth B. Storey[1,*]

[1]Institute of Biochemistry and Department of Biology, Carleton University, 1125 Colonel By Drive, K1S 5B6, Ottawa, Ontario, Canada, http://carleton.ca/

Summary

MicroRNAs are short non-coding RNA transcripts that act as master cellular regulators with roles in orchestrating virtually all biological functions. The recent affordability and widespread use of high-throughput microRNA profiling technologies has grown along with the advancement of bioinformatics tools available for analysis of the mounting data flow. While there are many computational resources available for the management of data from genome-sequenced animals, researchers are often faced with the challenge of identifying the biological implications of the daunting amount of data generated from these high-throughput technologies. In this article, we review the current state of high-throughput microRNA expression profiling platforms, data analysis processes, and computational tools in the context of comparative molecular physiology. We also present RBioMIR and RBioFS, our R package implementations for differential expression analysis and random forest-based gene selection. Detailed installation guides are available at kenstoreylab.com.

1 Introduction

Recently, microRNAs (miRNAs) have come to the forefront as dynamic gene regulators that are proving to be master regulators of most biological functions [1]. This group of highly-conserved, short (17-22 nt) non-coding RNAs is known to regulate over 60% of protein-coding genes in humans. The broad controls that miRNAs exert are in part due to the ability of an individual miRNA to target multiple mRNAs and the fact that a single mRNA transcript may be subject to regulation by various miRNAs [2]. MiRNAs regulate genes mainly at the post-transcriptional level through binding with partial or perfect complementarity to the 3' untranslated region of their mRNA targets, thereby modulating mRNA stability and translation [3]. Partial complementarity leads to translational repression via mechanisms that are not yet fully elucidated but that include sequestering mRNA transcripts in cytoplasmic loci such as stress granules and P-bodies. However, perfect complementarity, a major silencing mechanism, results in mRNA target degradation by Argonaute endonucleases [3]. Other miRNA silencing mechanisms include cases of gene regulation at the transcriptional level in which miRNAs are able to bind directly to DNA regulatory elements, destabilize mRNAs through cleavage-independent processes, and inhibit mRNA:protein interactions by acting as decoys that directly bind to these RNA-binding proteins [4].

* To whom correspondence should be addressed. Email: kenneth_storey@carleton.ca

doi:10.2390/biecoll-jib-2016-306

On-going studies on miRNA biology have led to evaluation of their uses in diagnostic, prognostic, and therapeutic applications for diseases [5]. Their emerging roles in orchestrating development, cell cycle, metabolism, and the molecular and physiological adaptations required for organisms to respond to various environmental stresses has also attracted tremendous attention [6]. MiRNAs have even been shown to regulate pathogen and host interactions [7]. As such, multiple approaches have been developed for characterizing and assessing the functional roles and biomarker potential of miRNAs. The main technologies currently available for high-throughput miRNA expression profiling include: quantitative reverse transcription polymerase chain reaction (qRT-PCR), microarray analysis, and next generation sequencing (NGS) based methods. These approaches have received immense attention in recent years, largely due to their increased accessibility and the advancement in computational capacity and data analysis tools. In this article, we review the current state of miRNA high-throughput expression profiling techniques, data analysis processes, and computational tools in the context of comparative molecular physiology. We also present RBioMIR and RBioFS, two automated and easy-to-use R packages for the assessment of differential expression (DE) and for machine learning-based gene selection.

2 Current miRNA research approaches

The short length and uniqueness of miRNAs made large-scale parallel analysis a technical challenge that was initially systematically addressed using dot blots and northern blots. Currently, there are three main types of high-throughput miRNA expression profiling approaches: qRT-PCR, hybridization-based miRNA microarray, and NGS based small RNA-Sequencing (RNA-Seq) [8]. The qRT-PCR is appealing to many laboratories due to its simplicity, reproducibility, and low cost [9]. This approach typically features either polyadenylation or artificial stem-loop based target amplification [10]. While the targeted nature of detection enables assessment without the need for a reference genome, it also renders qRT-PCR approaches ineffective for the discovery of novel miRNAs, and makes it better suited as a validation method rather than a discovery tool [8]. However, the sensitivity of qRT-PCR allows for the absolute quantification of miRNA transcript abundance levels. When compared to other strategies, the lack of scalability that is characteristic of qRT-PCR renders it inefficient for mass miRNA profiling.

Hybridization-based miRNA microarrays were among the first high-throughput miRNA profiling methods developed. Such methods use a surface fixed with thousands of DNA-based capture probes, designed to be complementary to a specific fluorescently labelled mature miRNA target [11]. While miRNA microarrays are easily scalable and relatively less expensive than other platforms, they are considered less sensitive or specific, and are also not suited for novel miRNA discovery and absolute quantification of miRNA abundance [8].

The third major approach is the NGS-based massively parallel small RNA-Seq technology [12]. The general principle behind small RNA-Seq is the generation of a small RNA cDNA library from the biological samples, followed by adapter ligation, and sequencing [13]. While RNA-Seq is relatively more expensive than other approaches, the continual introduction of newer models and DNA barcoding multiplexing technology have made it more accessible. Small RNA-Seq is also considered to be the main platform for novel miRNA discovery [14]. Limitations of RNA-Seq include the massive computational support required for data analysis and the increased dependence on genome availability that makes it challenging to use with non-sequenced animal models [8].

3 Computational tools and workflows for RNA-Seq-based miRNA analysis

The advent of high-throughput miRNA profiling technologies has led to the generation of large datasets that have made computational tools indispensable for miRNA studies. While numerous bioinformatics tools, both public and custom-made, are available, the same general miRNA-profiling data analysis approach can be readily applied to any of the platforms. This includes: (1) raw data processing, (2) quality assessments, (3) identification of conserved and novel miRNAs, (4) DE analysis, (5) target prediction and novel miRNA discovery, as well as (6) other higher level analyses such as gene set enrichment [8]. While specialized programs can be used to perform each of the steps summarized above, comprehensive miRNA analysis pipelines such as miRanalyzer [15] and DSAP [16] are also available. For a detailed discussion of these miRNA bioinformatics tools and others, see Akhtar et al. (2015) [4].

Homology based search tools that rely on sequence and structure can be used to identify orthologues of conserved miRNAs in numerous species, including non-sequenced animals [17], [18]. These tools generally utilize miRbase [19], the miRNA repository of all known and annotated precursor and mature miRNAs from a range of species. Since many of these tools require well-annotated genomes to effectively identify miRNAs, their usefulness to researchers that work on non-sequenced animal models is limited. As such, more advanced approaches that involve leveraging machine learning strategies and experimental data driven methods are needed for assessing both conserved and novel miRNAs. Machine learning tools use algorithms to 'learn' the sequence, structural, and thermodynamic characteristics of miRNAs [20], [21], where specialized tools such as SMIRP [22] are able to predict novel miRNAs in sequenced non-model organisms. Examples of computational tools that leverage transcriptomic and small RNA-seq data to characterize conserved and novel miRNAs are miRDeep2 [23] and the machine-learning tool miReader [24]. The level of error that is introduced in these studies has been shown to be inconsequential and the DE results for most abundant miRNAs are acceptable on a per project basis [25].

3.1 General workflow for conserved miRNA expression analysis

Here we describe a general data processing and analysis workflow for small RNA-Seq based miRNA expression profiling (DE analysis) with a focus on conserved miRNAs and applications for comparative studies (Fig. 1). It should be noted that the procedure outlined herein is but one example of a miRNA data analysis pipeline and that the present assembly of computational tools can be customized. Overall, the current pipeline features two stages: data processing (Unix Shell environment), and expression analysis (R environment). Since we are focusing on screening conserved miRNAs, we do not require a complete annotated genome or comprehensive transcriptomes for these analyses.

The data processing stage starts with performing read quality checks using FastQC [26], followed by adapter trimming and size filtering with fasq-mcf [27]. Since there is currently no non-miRNA small RNA species database specific to non-sequenced animals, the negative reference database was built using all sequences from the rfam [28] and piRNA databases [29]. The trimmed and filtered reads are then aligned to this negative reference database using bowtie [30] and the unaligned 'clean' reads are carried on to the next steps. Similarly, as the miRNAs from the species of interest have yet to be annotated, the entire mature miRNA database of all species is obtained from miRbase [31] as the positive reference database, to which the

remaining 'clean' reads are aligned using bowtie. HTSeq [32] is then used to count the aligned reads for each of the identified miRNAs, serving as the basis for the upcoming DE analysis steps.

The data analysis stage uses the R programming language. First, read counts are imported to R, and in cases where a miRNA species is duplicated, the highest count should be used. Such read count discrepancy originates from the conservation between the species of interest and the species to which the reads are aligned. R packages edgeR [33] and limma [34] are then used to normalize the read counts before DE analysis with limma. The DE results are then exported as comma separated values (csv) files and visualized using ggplot2 and gplots [35], [36]. The results can then be used for downstream analysis such as machine learning-based gene selection (see below).

To streamline this complex analysis, we have wrapped all steps and required tools in the R package RBioMIR. Unix Shell commands and the RBioMIR user manual can be found at (http://kenstoreylab.com/?page_id=2540). The source code and the R package can be downloaded through GitHub: (https://github.com/jzhangc/git_RBioMIR.git).

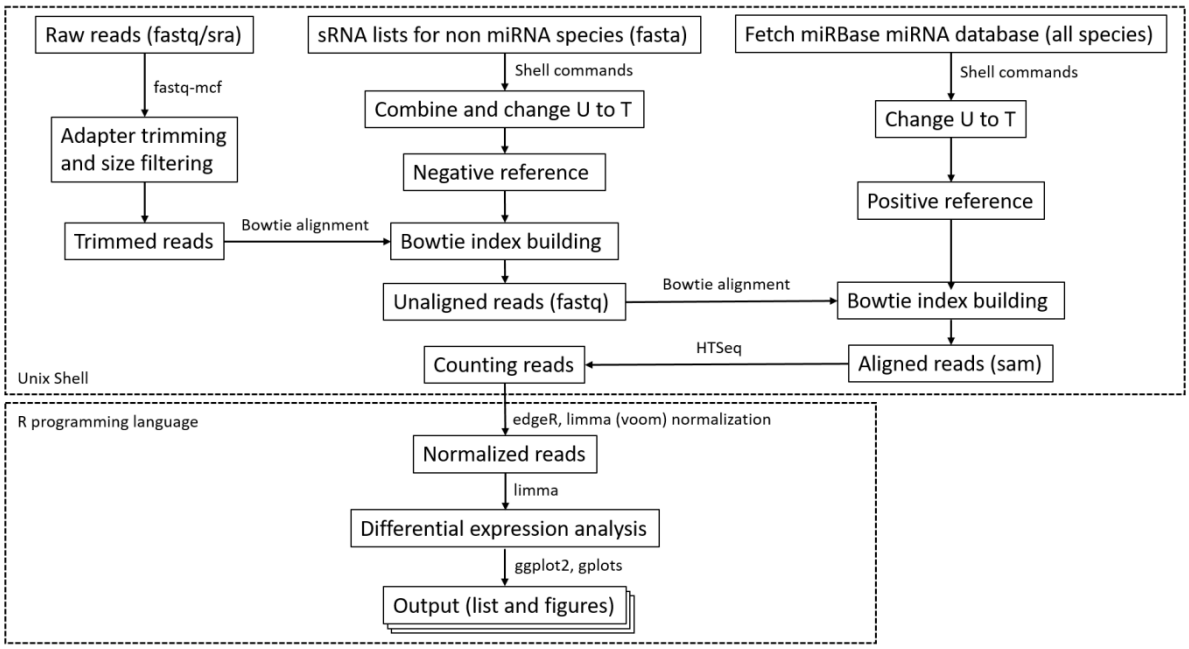

Figure 1: A flowchart depicting the general RNA-Seq data processing and analysis steps necessary for assessing conserved miRNA expression in non-model species.

4 Random forest-based feature selection for identifying important miRNAs

Small RNA-Seq and high-throughput qRT-PCR based miRNA profiling result in high dimensional (i.e., high feature count, also denoted as large *p*) datasets. In the context of comparative molecular physiology, preserving meaningful information from gene expression profiles by eliminating irrelevant features, a process known as feature selection (FS), is critical for mechanistic characterization and biomarker discovery. As such, a myriad of FS methods using, for example, univariate ranking, conventional modelling, or machine learning classifiers have been developed [37], [38].

Random forest (RF) is a machine learning classifier that is well-suited for the processing of high dimensional datasets, due to its robustness, versatility, and high universality. The algorithm handles high dimensionality with low sample size datasets (known as small *n*, large *p*) relatively well as compared to other methods [39]. The core concepts of RF include bootstrap resampling (or bagging), random feature partitioning (i.e., the randomly selected subset of features to model at each tree node), and decision tree modelling [40] (Fig. 2). The bootstrap resampling step enables internal cross-validation by leaving some samples out (known as out-of-bag [OOB] samples), allowing OOB error rates to be used as a measurement for performance (Fig. 2). The random feature partitioning step reduces modelling bias to the dataset (or over-fitting) or to the features. For each subgroup of randomly selected features, the algorithm utilizes either non-parametric methods such as classification and regression tree (CART) or parametric methods to generate a decision tree [40–42]. Therefore, the final RF model represents a collection of decision trees, leading to unbiased feature evaluation and high classification power. It has been demonstrated that RF is capable of handling binary, multi-categorical, as well as regression FS tasks [40]. The algorithm provides variable importance (VI) metrics that act as a starting point for the FS workflow. It is worth noting that, due to the heuristic nature of RF, a bootstrapping repeat of RF-FS run is highly valuable.

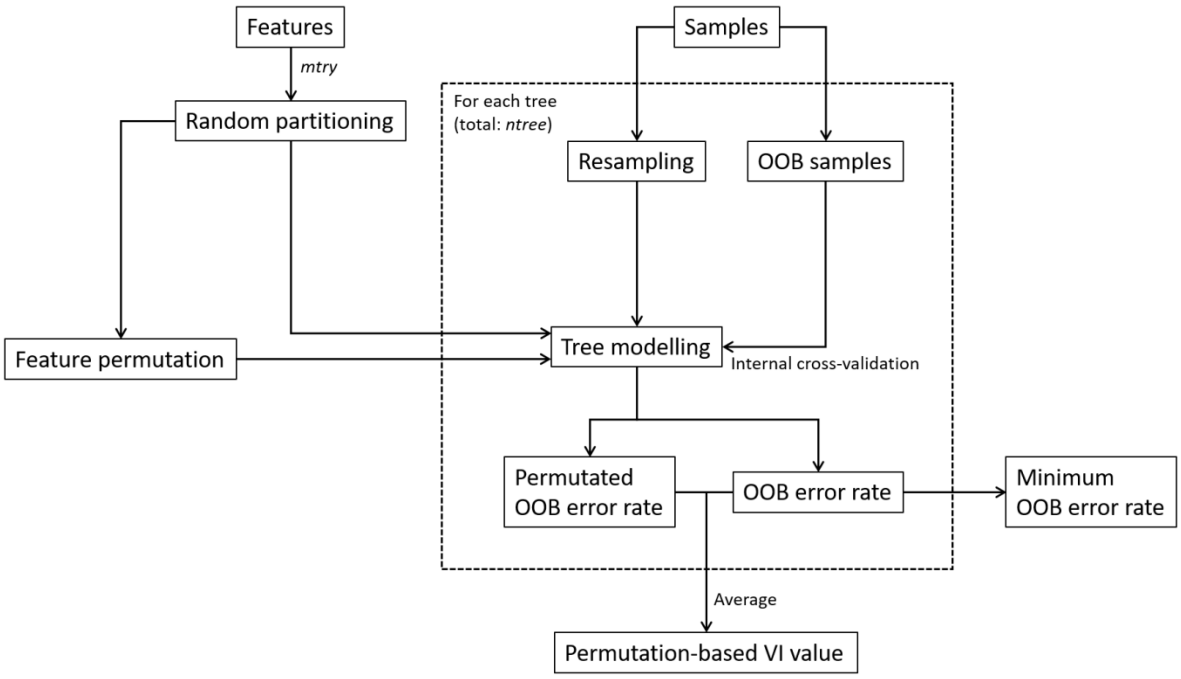

Figure 2: A flowchart of the core algorithm for random forest (RF).

Various FS algorithms such as Boruta and RF-based recursive feature elimination (RF-RFE) have been developed [43], [44], each of which has its distinct specialties. For example, the former focuses on retaining all the relevant features, whereas RF-RFE emphasizes the isolation of the minimal number of features. In the case of investigating and selecting key miRNA targets that facilitate physiological responses under select environmental conditions, identifying all relevant features might be more appropriate. Additionally, since the regulatory connection among miRNAs may be manifested as statistical correlation [45], it is crucial to reflect this consideration when establishing an RF-FS procedure. Since recursive RF approaches have been shown to demonstrate superior performance when handling correlated features [46], they might be preferred for the processing of gene expression datasets. It is also imperative to evaluate the

computational power and time requirements. For example, although being able to identify all features, Boruta's requirements for large sample sizes and computational power can be demanding [44]. Moreover, applying a univariate statistics filter prior to RF FS may help integrate prior knowledge when selecting features, leading to a more focused study around specific research goals. Accordingly, we consider the following when implementing an RF-FS workflow:

a. Selecting all relevant features;
b. Taking feature correlation into consideration;
c. Incorporating a univariate statistics filter;
d. Options for reducing computational time.

4.1 Implementation

Based on the principles above, we present RBioFS, an implementation of an RF-FS workflow for selecting miRNAs relevant to differentiating between physiological phenotypes of interest. The current workflow is based on the strategy first presented by Genuer et al. (2010) [47]. The workflow uses the original RF principle proposed by Breiman (2001) [40] via the R package randomForest [48], and is a combination of univariate statistical filter, recursive RF VI ranking, single CART modeling, and sequential forward selection (SFS)-like recursive RF feature elimination. The original method by Genuer et al. (2010) [47] also features an additional round of SFS, providing a minimal feature list for predictive modelling. However, since our focus is on selecting all relevant features, the list obtained from the first SFS round is considered sufficient.

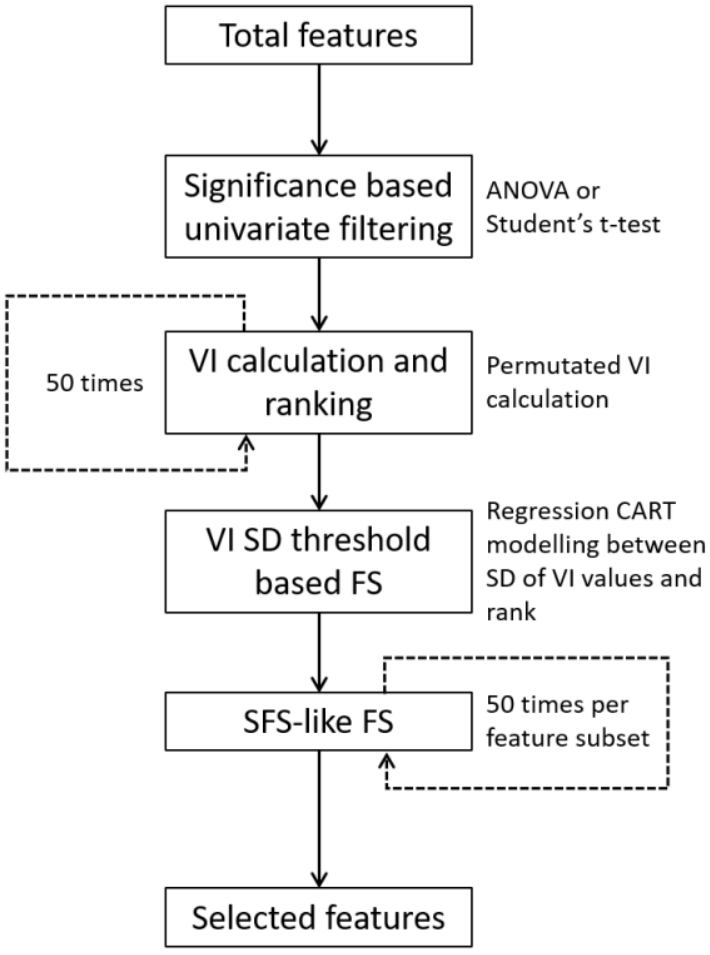

Figure 3: The flowchart of the current RF-FS implementation.

Briefly, an analysis of variance (ANOVA) or Student's t-test, based on the univariate test, is conducted on validated high-throughput miRNA datasets and miRNAs showing no statistically significant changes between phenotypes are discarded. A bootstrapping (50 times) RF run is then performed on the filtered dataset to calculate VI values for each feature. All features are subsequently ranked in descending order based on the mean VI value. It is worth noting that the current implementation uses permutation-based VI values because they are less biased towards correlated features [49]. Specifically, the permutation-based VI value of a given feature is the average difference between OOB error rates from all RF trees with and without the permutation for the feature in the random partition (Fig. 2).

As described in Genuer et al. (2010) [47], the standard deviation (SD) for each feature and corresponding VI mean ranks are subjected to CART regression modelling using ANOVA to estimate the minimum SD using the R package, rpart [50]. This estimate is then used as a threshold for the initial VI rank-based feature elimination step, based on the observation that relevant features tend to exhibit a larger variance in VI values [47]. The remaining features are then subjected to a bootstrapping (50 times) SFS-like process, where features are recursively introduced to the RF runs starting from the feature ranked at the top of the VI ranking. The first subset of features resulting in a mean OOB error rate less than the minimum OOB error rate plus one standard deviation is reported as the final result. A schematic representation of this workflow is depicted in Fig. 3. We wrapped the RF-FS pipeline described herein in the automated R package RBioFS; a package that takes advantage of parallel computing. The RBioFS detailed user manual, sample dataset, and sample results can be found at (http://kenstoreylab.com/?page_id=2542). The package and source code are available on GitHub: (https://github.com/jzhangc/git_RBioFS.git).

4.2 Case study and discussion

Here we demonstrate the use of RBioFS for RF-FS based gene selection on part of the dataset from a study that explored changes in miRNA expression profiles in response to hibernation in a South American marsupial [51]. In the original paper, expression levels of 85 miRNAs were measured in liver and skeletal muscle of euthermic active (control) and torpid animals using a high-throughput qRT-PCR profiling approach.

We applied RBioFS, our in-house R implementation of the RF-FS workflow, to the liver dataset. For univariate statistical filtering, all miRNAs that failed to show significant changes (Student's t-test) were discarded; this reduced the miRNA list size from 85 to 35 targets (Table 1). The permutation based VI values were then calculated and ranked upon 50 times RF runs (Fig. 4A). Based on SD values and ranking, a CART model was established to estimate the threshold for the initial RF feature elimination; this resulted in the selection of 12 miRNAs (Table 1).
An SFS-like selection step was conducted on these 12 miRNAs. Starting with *miR-22-5p*, 12 recursive RF runs (50 times per run) were carried out adding one miRNA each time. Consequently, the first group featuring 6 miRNAs was considered the most important target. The decreasing trend of the OOB error is shown in Fig. 4B. For both RF steps, a tree count, i.e.: times of resampling (denoted as *ntree* in the randomForest package) of 501 was used. A value of $p / 3$ was used as the random feature partitioning scale, i.e.: the number of features for random feature partitioning (denoted *mtry* in the randomForest package), as suggested by Genuer et al. (2010) [47].

Table 1: The miRNAs selected by RBioFS after each FS step. "*dgl*" stands for the species name *Dromiciops gliroides*.

t-test (not ranked)		VI-based selecting (ranked by VI)	SFS-like selection (ranked by VI)
dgl-miR-let-7f-5p	*dgl-miR-99b-5p*	*dgl-miR-22-5p*	*dgl-miR-22-5p*
dgl-miR-1a-5p	*dgl-miR-106b-5p*	*dgl-miR-876-5p*	*dgl-miR-876-5p*
dgl-miR-1b-5p	*dgl-miR-125a-5p*	*dgl-miR-21a-3p*	*dgl-miR-21a-3p*
dgl-miR-10b-5p	*dgl-miR-134-5p*	*dgl-miR-18a-3p*	*dgl-miR-18a-3p*
dgl-miR-16-3p	*dgl-miR-137-5p*	*dgl-miR-16-3p*	*dgl-miR-16-3p*
dgl-miR-18a-3p	*dgl-miR-139-5p*	*dgl-miR-142b-5p*	*dgl-miR-142b-5p*
dgl-miR-20a-5p	*dgl-miR-142b-5p*	*dgl-miR-191-5p*	
dgl-miR-21a-3p	*dgl-miR-145a-5p*	*dgl-miR-34a-5p*	
dgl-miR-22-5p	*dgl-miR-152-5p*	*dgl-miR-23a-5p*	
dgl-miR-23a-5p	*dgl-miR-181a-5p*	*dgl-miR-196a-5p*	
dgl-miR-26a-5p	*dgl-miR-185-5p*	*dgl-miR-219a-5p*	
dgl-miR-27a-5p	*dgl-miR-191-5p*	*dgl-miR-106b-5p*	
dgl-miR-29a-5p	*dgl-miR-193b-5p*		
dgl-miR-34a-5p	*dgl-miR-196a-5p*		
dgl-miR-34c-5p	*dgl-miR-214-5p*		

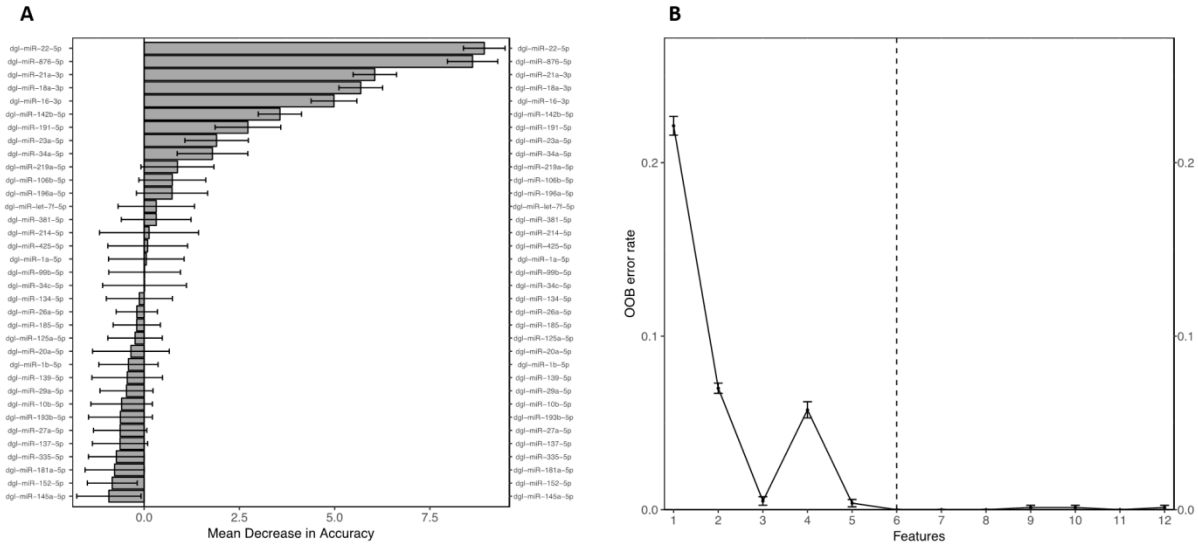

Figure 4: Results from RF-FS conducted on the expression profile of 85 miRNAs from liver of control versus torpid marsupials. (A) Histogram depicting mean VI values (± SD) and ranking of the 35 miRNAs remaining from the univariate filtering; (B) OOB error rate (± SEM) change based on SFS-like selection. The vertical line represents the set of features that resulted in the minimum OOB error rate (+ 1SD). Figures were generated using the R package RBioplot [50]**.**

The group of 6 RF-FS selected miRNAs was further examined using both available literature and miRNA target prediction tools highlighted in this work. It was revealed that these miRNAs significantly downregulated during torpor are likely involved in coordinating a compensatory hepatic thermoregulatory mechanism to facilitate arousal in hibernating marsupials. Indeed, *miR-22-5p* has been implicated as a regulator of lipid metabolism [53] and DIANA-miRPath [54] analysis of the 6 miRNAs revealed that *miR-22-5p*, *miR-21a-3p*, *miR-34a-5p*, and *miR-876-5p* directly target key elements of mitogen-activated protein kinase (MAPK) pathways.

This hypothesis is further supported by the reduced expression level of *miR-142-5p* that is known to be inversely related to MAPK activity [55]. Moreover, recent findings have identified *miR-142-5p* as a temperature-sensitive miRNA, and a known regulator of raised body temperature [56]. Our case study on hibernating marsupial data emphasizes the ability of RBioFS to effectively sort and reveal the differentiating and most crucial set of miRNAs that occur in response to metabolic or environmental perturbations in a particular system. Indeed, the utility of RBioFS is not limited to comparative molecular physiology but is also applicable for miRNA studies of development and disease. Furthermore, the ability of RBioFS to select the most important miRNA targets, capable of differentiating between healthy and patient groups, emphasizes it ability to aid in the discovery of biomarkers in various conditions.

5 Conclusion

High-throughput expression profiling and advanced computational tools have been instrumental to our understanding of miRNAs and the central role that they play in modulating gene expression. In this review, we highlight the functional characteristics of miRNAs, miRNA-profiling methods, and we outline the main data processing steps and computational tools required to undertake these analyses. We also present our implementations of a small RNA-Seq data analysis pipeline and an RF based feature selection workflow through the R packages RBioMIR and RBioFS, respectively.

Author contributions

All authors were involved in the assembly of the manuscript and approved the final version. JZ and HH performed the data analysis. JZ developed Unix Shell codes, RBioMIR, and RBioFS.

Acknowledgments

This work was supported by a Discovery grant (Grant # 6793) to KBS from the Natural Sciences and Engineering Research Council (NSERC) of Canada. KBS holds the Canada Research Chair in Molecular Physiology and HH holds an Ontario Graduate Scholarship.

References

[1] P. P. Amaral, M. E. Dinger, and J. S. Mattick. Non-coding RNAs in homeostasis, disease and stress responses: an evolutionary perspective. Briefings in functional genomics, 12(3):254–78, May 2013.

[2] M. S. Ebert and P. A. Sharp. Roles for microRNAs in conferring robustness to biological processes. Cell, 149(3):515–24, Apr. 2012.

[3] D. Bartel. MicroRNAsGenomics, Biogenesis, Mechanism, and Function. Cell, 116(2):281–297, Jan. 2004.

[4] M. M. Akhtar, L. Micolucci, M. S. Islam, F. Olivieri, and A. D. Procopio. Bioinformatic tools for microRNA dissection. Nucleic acids research, 44(1):24–44, Jan. 2015.

[5] G. Bertoli, C. Cava, and I. Castiglioni. MicroRNAs: New Biomarkers for Diagnosis, Prognosis, Therapy Prediction and Therapeutic Tools for Breast Cancer. Theranostics, 5(10):1122–1143, 2015.

[6] K. B. Storey. Regulation of hypometabolism: insights into epigenetic controls. Journal of Experimental Biology, 218(1):150–159, 2015.

[7] M. D. Saçar, C. Bağcı, and J. Allmer. Computational Prediction of MicroRNAs from Toxoplasma gondii Potentially Regulating the Hosts' Gene Expression. Genomics, proteomics & bioinformatics, 12(5):228–238, Oct. 2014.

[8] C. C. Pritchard, H. H. Cheng, and M. Tewari. MicroRNA profiling: approaches and considerations. Nature Reviews Genetics, 13(5):358–369, 2012.

[9] R. J. Farr, A. S. Januszewski, M. V. Joglekar, H. Liang, A. K. McAulley, A. W. Hewitte, et al. A comparative analysis of high-throughput platforms for validation of a circulating microRNA signature in diabetic retinopathy. Scientific Reports, 5:10375, Jun. 2015.

[10] K. K. Biggar, C.-W. Wu, and K. B. Storey. High-throughput amplification of mature microRNAs in uncharacterized animal models using polyadenylated RNA and stem–loop reverse transcription polymerase chain reaction. 2014.

[11] P. T. Nelson, D. a Baldwin, L. M. Scearce, J. C. Oberholtzer, J. W. Tobias, and Z. Mourelatos. Microarray-based, high-throughput gene expression profiling of microRNAs. Nature methods, 1(2):155–161, 2004.

[12] J. Shendure and H. Ji. Next-generation DNA sequencing. Nature Biotechnology, 26(10):1135–1145, Oct. 2008.

[13] M. L. Metzker. Sequencing technologies — the next generation. Nature Reviews Genetics, 11(1):31–46, Jan. 2010.

[14] C. J. Creighton, J. G. Reid, and P. H. Gunaratne. Expression profiling of microRNAs by deep sequencing. Briefings in Bioinformatics, 10(5):490–497, Sep. 2009.

[15] M. Hackenberg, N. Rodríguez-Ezpeleta, and A. M. Aransay. miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments. Nucleic acids research, 39(Web Server issue):W132-8, Jul. 2011.

[16] P.-J. Huang, Y.-C. Liu, C.-C. Lee, W.-C. Lin, R. R.-C. Gan, P.-C. Lyu, and P. Tang. DSAP: deep-sequencing small RNA analysis pipeline. Nucleic acids research, 38(Web Server issue):W385-91, Jul. 2010.

[17] J. R. Brown and P. Sanseau. A computational view of microRNAs and their targets. Drug Discovery Today, 10(8):595–601, 2005.

[18] L. P. Lim, N. C. Lau, E. G. Weinstein, A. Abdelhakim, S. Yekta, M. W. Rhoades, C. B. Burge, and D. P. Bartel. The microRNAs of Caenorhabditis elegans. Genes & Development, 17(8):991–1008, Apr. 2003.

[19] A. Kozomara and S. Griffiths-Jones. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Research, 42(D1):D68–D73, Jan. 2014.

[20] L. Li, J. Xu, D. Yang, X. Tan, and H. Wang. Computational approaches for microRNA studies: a review. Mammalian genome : official journal of the International Mammalian Genome Society, 21(1–2):1–12, Feb. 2010.

[21] M. D. Saçar and J. Allmer. Machine learning methods for microRNA gene prediction. Methods in molecular biology (Clifton, N.J.), 1107:177–87, Jan. 2014.

[22] R. J. Peace, K. K. Biggar, K. B. Storey, and J. R. Green. A framework for improving microRNA prediction in non-human genomes. Nucleic acids research, 43(20):e138, Nov. 2015.

[23] M. R. Friedländer, S. D. Mackowiak, N. Li, W. Chen, and N. Rajewsky. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic acids research, 40(1):37–52, Jan. 2012.

[24] A. Jha and R. Shankar. miReader: Discovering Novel miRNAs in Species without Sequenced Genome. PLoS ONE, 8(6), 2013.

[25] K. Etebari and S. Asgari. Accuracy of MicroRNA discovery pipelines in non-model organisms using closely related species genomes. PLoS ONE, 9(1):1–10, 2014.
[26] S. Andrews. FastQC A quality control tool for high throughput sequence data. Bioinformatics, Babraham, 2010. [Online]. Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
[27] E. Aronesty. ea-utils : Command-line tools for processing biological sequencing data. 2011.
[28] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, and S. R. Eddy. Rfam: an RNA family database. Nucleic acids research, 31(1):439–41, Jan. 2003.
[29] S. Sai Lakshmi and S. Agrawal. piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. Nucleic Acids Research, 36(Database):D173–D177, Dec. 2007.
[30] B. Langmead, Langmead, and Ben. Aligning Short Sequencing Reads with Bowtie. in Current Protocols in Bioinformatics, :11.7.1-11.7.14, Hoboken, NJ, USA: John Wiley & Sons, Inc., 2010, 11.7.1-11.7.14.
[31] S. Griffiths-Jones, R. J. Grocock, S. van Dongen, A. Bateman, and A. J. Enright. miRBase: microRNA sequences, targets and gene nomenclature. Nucleic acids research, 34(Database issue):D140-4, Jan. 2006.
[32] S. Anders, P. T. Pyl, and W. Huber. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics, 31(2):166–169, Jan. 2015.
[33] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26(1):139–140, Jan. 2010.
[34] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research, 43(7):e47–e47, Apr. 2015.
[35] G. Warnes, B. Bolker, L. Bonebakker, R. Gentleman, W. Liaw, T. Lumley, M. Maechler, A. Magnusson, S. Moeller, M. Schwartz, and B. Venables. gplots: Various R programming tools for plotting data. R package version 3.0.1. 2016.
[36] H. Wickham. Ggplot2 : elegant graphics for data analysis. Springer, 2009.
[37] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning, (46):389–422, 2002.
[38] C. Ambroise and G. J. McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. Proceedings of the National Academy of Sciences of the United States of America, 99(10):6562–6, May 2002.
[39] R. Díaz-Uriarte, S. Alvarez de Andrés, J. Lee, J. Lee, M. Park, et al. Gene selection and classification of microarray data using random forest. BMC Bioinformatics, 7(1):3, 2006.
[40] L. Breiman. Random Forests. Machine Learning, 45(1):5–32, 2001.
[41] L. Breiman, J. Fredman, C. Stone, and O. RA. Classification and regression trees. Chapman & Hall, 1984.
[42] A. Zeileis, T. Hothorn, and K. Hornik. Model-Based Recursive Partitioning. Journal of Computational and Graphical Statistics, 17(2):492–514, 2008.
[43] M. B. Kursa and W. R. Rudnicki. Feature Selection with the Boruta Package. Journal of Statistical Software, 36(11):1–13, 2010.
[44] M. B. Kursa, Y. Saeys, I. Inza, P. Larrañaga, R. Nielsen, J. Peña, et al. Robustness of Random Forest-based gene selection methods. BMC Bioinformatics, 15(1):8, 2014.
[45] S. G. Chaulk, H. A. Ebhardt, and R. P. Fahlman. Correlations of microRNA:microRNA expression patterns reveal insights into microRNA clusters and global microRNA expression patterns. Mol. BioSyst., 12(1):110–119, 2016.
[46] B. Gregorutti, B. Michel, and P. Saint-Pierre. Correlation and variable importance in random forests. Statistics and Computing, :1–20, Mar. 2016.

[47] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot. Variable selection using random forests. 2010.
[48] A. Liaw and M. Wiener. Classification and Regression by randomForest. R news, 2(December):18–22, 2002.
[49] K. K. Nicodemus, J. D. Malley, C. Strobl, A. Ziegler, L. Breiman, T. Hothorn, et al. The behaviour of random forest permutation-based variable importance measures under predictor correlation. BMC Bioinformatics, 11(1):110, 2010.
[50] T. Therneau, B. Atkinson, and B. Ripley. rpart: Recursive partitioning and regression trees. R package version 4.1-10. 2015.
[51] H. Hadj-Moussa, J. A. Moggridge, B. E. Luu, J. F. Quintero-Galvis, J. D. Gaitán-Espitia, R. F. Nespolo,et al. The hibernating South American marsupial, Dromiciops gliroides, displays torpor-sensitive microRNA expression patterns. Scientific Reports, 6:24627, Apr. 2016.
[52] J. Zhang and K. B. Storey. RBioplot: an easy-to-use R pipeline for automated statistical analysis and data visualization in molecular biology and biochemistry. PeerJ, 4:e2436, Sep. 2016.
[53] C. Koufaris, G. N. Valbuena, Y. Pomyen, G. D. Tredwell, E. Nevedomskaya, C.-H. Lau, et al. Systematic integration of molecular profiles identifies miR-22 as a regulator of lipid and folate metabolism in breast cancer cells. Oncogene, 35(21):2766–2776, May 2016.
[54] I. S. Vlachos, K. Zagganas, M. D. Paraskevopoulou, G. Georgakilas, D. Karagkouni, T. Vergoulis, T. Dalamagas, and A. G. Hatzigeorgiou. DIANA-miRPath v3.0: deciphering microRNA function with experimental support. Nucleic acids research, 43(W1):W460-6, Jul. 2015.
[55] S. Sharma, J. Liu, J. Wei, H. Yuan, T. Zhang, and N. H. Bishopric. Repression of miR-142 by p300 and MAPK is required for survival signalling via gp130 during adaptive hypertrophy. EMBO molecular medicine, 4(7):617–32, Jul. 2012.
[56] J. J.-L. Wong, A. Y. M. Au, D. Gao, N. Pinello, C.-T. Kwok, A. Thoeng, et al. RBM3 regulates temperature sensitive miR-142-5p and miR-143 (thermomiRs), which target immune genes and control fever. Nucleic acids research, 44(6):2888–97, Apr. 2016.