RBiomirGS: an all-in-one miRNA gene set analysis solution featuring target mRNA mapping and expression profile integration

Jing Zhang[1] and Kenneth B. Storey[2]

[1] Schulich School of Medicine & Dentistry, University of Western Ontario, London, Canada
[2] Institute of Biochemistry, Departments of Biology and Chemistry, Carleton University, Ottawa, Canada

ABSTRACT

Background. With the continuous discovery of microRNA's (miRNA) association with a wide range of biological and cellular processes, expression profile-based functional characterization of such post-transcriptional regulation is crucial for revealing its significance behind particular phenotypes. Profound advancement in bioinformatics has been made to enable in depth investigation of miRNA's role in regulating cellular and molecular events, resulting in a huge quantity of software packages covering different aspects of miRNA functional analysis. Therefore, an all-in-one software solution is in demand for a comprehensive yet highly efficient workflow. Here we present RBiomirGS, an R package for a miRNA gene set (GS) analysis.

Methods. The package utilizes multiple databases for target mRNA mapping, estimates miRNA effect on the target mRNAs through miRNA expression profile and conducts a logistic regression-based GS enrichment. Additionally, human ortholog Entrez ID conversion functionality is included for target mRNAs.

Results. By incorporating all the core steps into one package, RBiomirGS eliminates the need for switching between different software packages. The modular structure of RBiomirGS enables various access points to the analysis, with which users can choose the most relevant functionalities for their workflow.

Conclusions. With RBiomirGS, users are able to assess the functional significance of the miRNA expression profile under the corresponding experimental condition by minimal input and intervention. Accordingly, RBiomirGS encompasses an all-in-one solution for miRNA GS analysis. RBiomirGS is available on GitHub (http://github.com/jzhangc/RBiomirGS). More information including instruction and examples can be found on website (http://kenstoreylab.com/?page_id=2865).

Subjects Biochemistry, Bioinformatics, Computational Biology, Molecular Biology, Data Mining and Machine Learning

Keywords Logistic regression, Pathway analysis, Transcriptome, Gene set enrichment, Molecular biology, Post-transcriptional regulation

Submitted 17 October 2017
Accepted 23 December 2017
Published 12 January 2018

Corresponding author
Kenneth B. Storey,
Kenneth_Storey@carleton.ca

Academic editor
Claus Wilke

Additional Information and Declarations can be found on page 14

DOI **10.7717/peerj.4262**

INTRODUCTION

MicroRNA (or miRNA) is a ~22 nucleotide long small RNA species and is mostly recognized as a negative gene expression regulator on a post-transcriptional level (*He & Hannon, 2004*). miRNAs have been proposed as biomarkers and/or therapeutic targets

for medical disorders such as drug-induced liver injury and cancer (*Mitchell et al., 2008*; *Wang et al., 2009*). Additionally, the primary structure of many miRNAs exhibits high level of conservation across species (*Zhang & Storey, 2013*), enabling smooth transfer of knowledge between different model systems.

Gene expression gene set (GS) analysis associates expression profiles with the functional outcome under specific experimental conditions and phenotypes. miRNA and coding gene expression GS analyses share the same general goal: to identify the significantly affected biological events from a given expression profile. The commonly used GS databases include gene ontology (GO) term (*Ashburner et al., 2000*) and KEGG (*Kanehisa & Goto, 2000*). Several GS techniques have been developed to directly incorporate differential expression (DE) results, such as gene set enrichment analysis (GSEA) (*Subramanian et al., 2005*). Even though it has been reported that these methods hold a more thorough and complete GS evaluation for coding genes (*Mootha et al., 2003*; *Subramanian et al., 2005*), the popular methods for miRNA research still rely on pre-selecting differentially expressed targets. Briefly, the commonly used miRNA GS analysis procedure starts with obtaining the list of the differentially expressed miRNAs, followed by searching for their target mRNAs, and then comparing the mRNA list with the GS databases (*Long et al., 2013*; *Chen et al., 2013*). However, it has been demonstrated that such method and its variations tend to exhibit bias of various origins (*Khatri, Sirota & Butte, 2012*; *Bleazard et al., 2015*). Moreover, the information on directionality from these methods is either indirect or lacking. One strategy to tackle the issue is to directly integrate miRNA DE results and transfer the information to the target mRNAs as a quantifiable metric.

There are a variety of computational analysis tools covering various aspects of miRNA studies, ranging from miRNA prediction, miRNA:mRNA interaction prediction and functional annotation (*Gomes et al., 2013*; *Akhtar et al., 2016*). As a result, multiple standalone tools are typically required to complete a miRNA GS workflow, e.g., mRNA target mapping, GS database preparation, GS enrichment, and results visualization. Practically, researchers usually face the challenge of constructing a pipeline for each project with multiple software packages and web services, which present incoherent connections between steps. Therefore, it is beneficial to establish a bioinformatic solution that searches multiple databases for mRNA target mapping and enables seamless navigation between analysis steps with minimal user intervention. Moreover, it is also critical to provide users with multiple entry points to the pipeline so that it is possible to customize and integrate only the functionalities necessary to their specific workflow. Here we present the R package RBiomirGS, a comprehensive miRNA GS analysis framework capable of performing the following tasks: (i) thorough target mRNA mapping, (ii) calculation of miRNA regulatory effect for target mRNAs, (iii) GS enrichment, and (iv) data visualization.

METHODS

As shown in Fig. 1: users provide the miRNA identity list and associated DE results, as well as GS database file. The RNA mapping module takes the miRNA list and searches multiple databases for miRNA:mRNA interactions, resulting in either a validated or predicted target

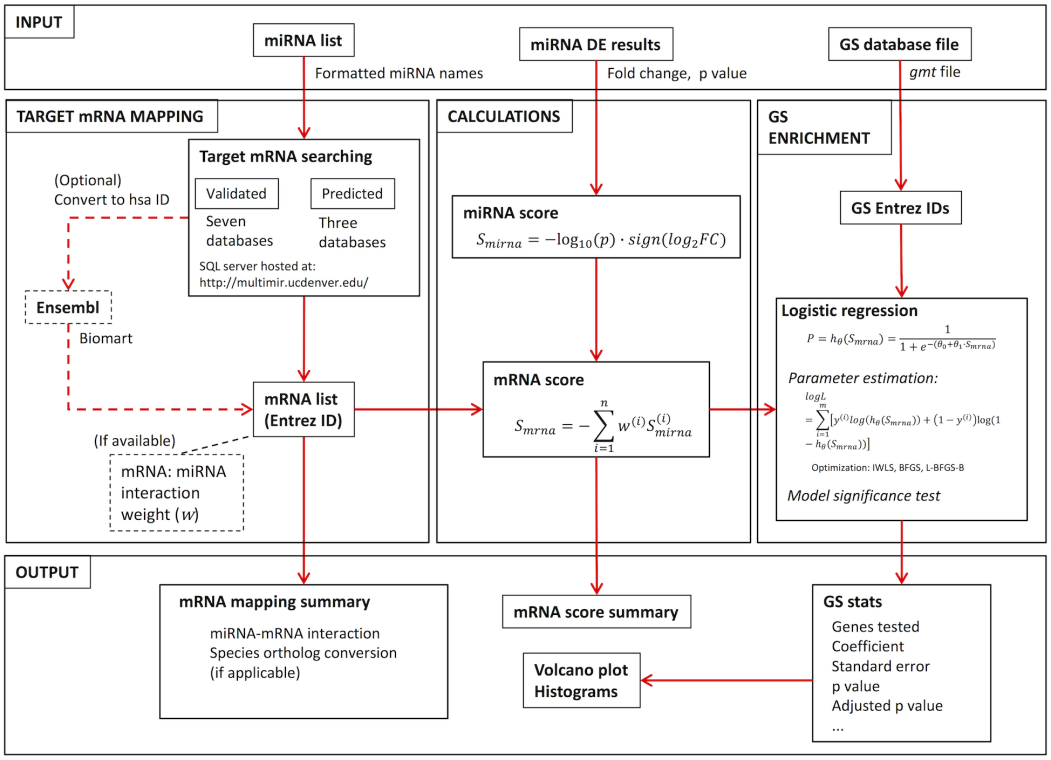

Figure 1 **RBiomirGS workflow, showing input and output configuration, as well as all the functionality modules.**

Full-size DOI: 10.7717/peerj.4262/fig-1

mRNA list. Fold change (FC) and *p* value from the miRNA DE list are then used to calculate a miRNA expression score for each miRNA measured, from which a miRNA impact score for target mRNAs is generated. With the mRNA score and GS database file, GS enrichment is then conducted using logistic regression. The package was built using R version 3.4.0 (*R Core Team, 2017*).

Target mRNA mapping module

RBiomirGS features a target mRNA mapping module that utilizes multiple miRNA:mRNA interaction databases, whose information is hosted on a SQL server at University of Colorado Cancer Centre (http://multimir.ucdenver.edu/). Information for both predicted and validated miRNA:mRNA interactions can be retrieved from the server. Although a disease research-focused R interface was developed by the host institution for data query (*Ru et al., 2014*), we assembled our own module for a more general purpose miRNA:mRNA interaction search with additional code optimizations such as parallel computing. The current module takes advantage of multiple databases for a more complete mapping result. For the experimentally validated miRNA:mRNA interactions, miRecords, mirTarBase and TarBase were used (*Xiao et al., 2009*; *Chou et al., 2016*; *Sethupathy, Corda & Hatzigeorgiou, 2006*); whereas DIANA-microT-CDS, ElMMo, MicroCosm (http://www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5/info.html), miRanda

(http://microrna.org), miRDB, PicTar, PITA, and TargetScan were searched for predicted interactions (*Paraskevopoulou et al., 2013*; *Gaidatzis et al., 2007*; *Betel et al., 2008*; *Wang, 2008*; *Krek et al., 2005*; *Kertesz et al., 2007*; *Lewis, Burge & Bartel, 2005*; *Grimson et al., 2007*; *Friedman et al., 2009*; *Garcia et al., 2011*). It is worth noting that DIANA-microT-CDS, PicTar, PITA and TargetScan are skipped for rat miRNAs. Currently, the mapping module supports human, rat and mouse miRNAs.

The core function of the target mRNA mapping module is *rbiomirGS_mrnascan*. The input file for this function is a list of miRNA names following the standard miRNA naming convention (http://www.mirbase.org/help/nomenclature.shtml). The function submits SQL queries to the server using the input miRNA list. The returned results are then output as both R list objects and as *csv* files to the working directory. By setting the species code (*hsa*, *rno* or *mmu* for human, rat or mouse, respectively), the function will search the databases accordingly. The argument *queryType* governs whether to search for validated or predicted interactions. For the output file, the universal column elements for both validated and predicted queries include Database, Mature miRNA miRBase accession number, Mature miRNA ID (name), Target gene symbol, Target gene Entrez ID, and Target gene Ensembl ID. The output results file will also contain column elements that are unique to the two query types.

miRNA score and mRNA score

The core idea behind the current GS analysis strategy is to quantitatively estimate the miRNA regulatory effect on the target mRNAs, through which the miRNA impact on specific functional gene sets can be evaluated. Based on the initial study by *Garcia-Garcia et al. (2016)*, a miRNA score is first calculated featuring the directionality presented in log FC (or $\log_2$FC), and log transformed p value (or $-\log_{10}(p)$). The equation is as follows:

$$S_{mirna} = -\log_{10} p \cdot sign(\log_2 FC) \tag{1}$$

As shown in Eq. (1), the S_{mirna} is a linear combination of the sign of $\log_2$FC and $-\log_{10}(p)$. Integrating p value and the sign of $\log_2$FC ensures that both significance and directionality of the change are taken into consideration. S_{mirna} can be calculated either with or without prior filtering of miRNAs. Although either approaches are valid, using the whole miRNA list both reduces the influences from thresholding method and enables a GS analysis resembling the core principle of a competitive GS enrichment approach (*De Leeuw et al., 2016*), thereby ensuring high compatibility and statistical power.

Upon obtaining S_{mirna}, the mRNA score (S_{mrna}) can be calculated. The current calculation is a modification of the approach proposed by *Garcia-Garcia et al. (2016)*. Such score is a quantitative representation of the potential regulatory effect on the target mRNAs from miRNAs. The equation is as follows:

$$S_{mrna} = -\sum_{i=1}^{n} w^{(i)} S_{mirna}^{(i)} \tag{2}$$

Equation (2) shows that the S_{mrna} of a mRNA is a sign reversed summation of the S_{mirna} of all the upstream miRNAs. The term n is the number of upstream miRNAs for the mRNA

of interest; and w is the miRNA:mRNA affinity score, with values set as 1 by default, i.e., no difference between interactions. However, users can set such score using a numeric vector if available.

Logistic regression-based GS enrichment

With S_{mrna} calculated with Eq. (2) and the GS database file, RBiomirGS uses logistic regression to enrich gene sets. Such approach is based on the core concept that a specific gene set is affected if its member genes are also regulated, either at the expression level or by influence from other regulatory factors such as miRNA. Practically, the goal is to assess if a gene can be categorized into a gene set solely based on its S_{mrna} value. As such, the enrichment algorithm models the probability of a gene with a specific S_{mrna} value belonging to a gene set. Mathematically, such probability is represented by the logistic regression sigmoid function (or hypothesis function):

$$P = h_\theta(S_{mrna}) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 \cdot S_{mrna})}} \tag{3}$$

As seen in Eq. (3), P is the aforementioned probability, which represents the hypothesis function of logistic regression with parameter vector θ. Transformation of Eq. (3) gives the equation below:

$$\log\left(\frac{P}{1-P}\right) = \theta_0 + \theta_1 \cdot S_{mrna} \tag{4}$$

Equation (4) shows that the function is the log odds ratio of a gene belonging to the gene set of interest, given the associated S_{mrna} value. Coefficient θ_1 stands for the change in the log odds ratio of the gene belonging to the gene set of interest by a unit change in S_{mrna}.

The model parameter is estimated based on the principle of maximum likelihood (*Fu & Li, 1993*). Specifically, the following log likelihood function is maximized:

$$\log L = \sum_{i=1}^{m} \left[y^{(i)} \log(h_\theta(S_{mrna})) + \left(1 - y^{(i)}\right) \log(1 - h_\theta(S_{mrna})) \right]$$

where y is the dummified membership to the gene set of interest for a gene, with 1 representing a member, 0 otherwise; m is the number of genes tested. RBiomirGS uses multiple optimization algorithms for finding the optimal parameter value for the model, including iteratively reweighted least square (IWLS), BFGS, and limited memory BFGS-B (L-BFGS-B) (*Byrd et al., 1994*; *Roger, 1987*; *Wolke & Schwetlick, 1987*). Such approach enables users to choose according to the volume of data and available computational power. RBiomirGS utilizes both generalized linear model (*glm*) function with logit link function natively included in R language, and a manual implementation of the logistic regression sigmoid function and log likelihood function. Specifically, the R native *glm* with logit link function uses IWLS by default; and the other two optimization methods work by applying general optimization function to the manual logistic regression implementation. To demonstrate the difference in performance with a specific dataset, an analysis of variance (ANOVA) test was conducted on the data from the case study using the statistical analysis R package RBioplot (*Zhang & Storey, 2016*).

The model significance test is carried out through a Wald test:

$$z = \frac{\widehat{\theta_1}}{s_{\widehat{\theta_1}}}$$

where $\widehat{\theta_1}$ is the estimated model coefficient by maximum likelihood method; and $s_{\widehat{\theta_1}}$ represents the standard error for the estimated model coefficient. The GS p value is then obtained using the z score. For IWLS, t value is used instead to calculate the GS p value with one degree of freedom. All GS p values are then adjusted using a false discovery rate (FDR) (*Benjamini & Hochberg, 1995*).

The calculation of the scores and logistic regression analysis are achieved through the function *rbiomirgs_logistic*. The scores, along with the GS database file, are then passed to the logistic modelling process. Similar to the target mRNA mapping function, argument *objTitle* sets the file name prefix. The miRNA DE object can be set using the *mirna_DE* argument. The arguments *var_mirnaName*, *var_mirnaFC* and *var_mirnaP* are used to set the column elements for miRNA names, FC and p value, respectively. The target mRNA object can then be set using argument *mrnalist*. The *mrna_Weight* argument is used to incorporate the miRNA:mRNA interaction weight matrix, if available. The *gs_file* argument is used to set the GS database file. The parameter optimization algorithm can be set using argument *optim_method*. By default, FDR is used to adjust the GS p value via argument *p.adj*. The GS enrichment results are exported as a *csv* file. A *txt* file detailing iterations to convergence if either BFGS or L-BFGS-B is used. The function also outputs the result to the R environment so that data visualization can be carried out.

Data visualization module

The current package includes a data visualization module utilizing the R package ggplot2 (*Wickham, 2009*). Specifically, the results can be plotted using bar graph and volcano plot. For bar graphs, two types of plots are featured in the package through function *rbiomirgs_bar*. Specifically, the horizontal bar graph inside the volcano plot depicts the overall distribution of the model coefficient (log odds ratio change per unit S_{mrna}) for all the gene sets tested; whereas the vertical bar graph shows the gene sets with top model coefficient values. The function ranks the absolute coefficient values and plots the top user defined gene sets. The bar graph is model coefficient $\pm$ standard error. Users can choose to only plot the significantly enriched gene sets on the bar graphs, as shown in the case study. The volcano plot is carried out by the *rbiomirgs_volcano* function. Users can set the p value threshold and the number of top gene sets to display on the graph. Additionally, users can freely use other plotting packages to meet their specific data visualization needs.

RESULTS

We demonstrate the usage and performance of RBiomirGS using the liver data from a study assessing the role of miRNAs in facilitating daily torpor in hibernating South American marsupials (*Hadj-Moussa et al., 2016*). The original study assessed 85 miRNAs in the liver and skeletal muscle of aroused and torpid marsupials using a qPCR approach. Given that the miRNome has yet to be fully characterized for the marsupials, the study used mouse

A

	A	B	C
1	miRNA	FC	pvalue
2	mmu-miR-let-7f-5p	0.58	0.034
3	mmu-miR-1a-5p	0.32	0.037
4	mmu-miR-1b-5p	0.5	0.001
5	mmu-miR-7a-5p	0.74	0.142
6	mmu-miR-9-3-5p	0.59	0.07
7	mmu-miR-10b-5p	0.57	0.006
8	mmu-miR-15a-5p	0.977	0.862
9	mmu-miR-15b-5p	0.906	0.575
10	mmu-miR-16-3p	0.79	0.048
11	mmu-miR-17-5p	0.84	0.339
12	mmu-miR-18a-3p	0.39	0.012
13	mmu-miR-20a-5p	0.4	0.009
14	mmu-miR-21a-3p	0.37	0.028
15	mmu-miR-22-5p	0.77	0.032
16	mmu-miR-23a-5p	0.44	0.029

B

	A	B	C	D	E	F	G
1	database	mature_mirna_acc	mature_mirna_id	target_symbol	target_entrez	target_ensembl	score
2	diana_microt	MIMAT0000526	mmu-miR-15a-5p	R3hdm2	71750	ENSMUSG00000025404	1
3	diana_microt	MIMAT0000526	mmu-miR-15a-5p	Kif23	71819	ENSMUSG00000032254	1
4	diana_microt	MIMAT0000526	mmu-miR-15a-5p	Akt3	23797	ENSMUSG00000019699	1
5	diana_microt	MIMAT0000526	mmu-miR-15a-5p	Tspyl2	52808	ENSMUSG00000041096	1
6	diana_microt	MIMAT0000526	mmu-miR-15a-5p	Mgat4a	269181	ENSMUSG00000026110	1
7	diana_microt	MIMAT0000526	mmu-miR-15a-5p	Cttnbp2	30785	ENSMUSG00000000416	1
8	diana_microt	MIMAT0000526	mmu-miR-15a-5p	Cpeb2	231207	ENSMUSG00000039782	1
9	diana_microt	MIMAT0000526	mmu-miR-15a-5p	Zbtb34	241311	ENSMUSG00000068966	1
10	diana_microt	MIMAT0000526	mmu-miR-15a-5p	Pnpla6	50767	ENSMUSG00000004565	1
11	diana_microt	MIMAT0000526	mmu-miR-15a-5p	Cacul1	78832	ENSMUSG00000033417	1
12	diana_microt	MIMAT0000526	mmu-miR-15a-5p	Sall1	58198	ENSMUSG00000031665	1
13	diana_microt	MIMAT0000526	mmu-miR-15a-5p	Ccne1	12447	ENSMUSG00000002068	1
14	diana_microt	MIMAT0000526	mmu-miR-15a-5p	1700025G04Rik	69399	ENSMUSG00000032666	1
15	diana_microt	MIMAT0000526	mmu-miR-15a-5p	Prdm4	72843	ENSMUSG00000035529	1
16	diana_microt	MIMAT0000526	mmu-miR-15a-5p	Rasgef1b	320292	ENSMUSG00000029333	1
17	diana_microt	MIMAT0000526	mmu-miR-15a-5p	Omg	18377	ENSMUSG00000049612	1
18	diana_microt	MIMAT0000526	mmu-miR-15a-5p	9330154J02Rik		ENSMUSG00000056031	1
19	diana_microt	MIMAT0000526	mmu-miR-15a-5p	Ash1l	192195	ENSMUSG00000028053	1
20	diana_microt	MIMAT0000526	mmu-miR-15a-5p	E2f3	13557	ENSMUSG00000016477	0.999
21	diana_microt	MIMAT0000526	mmu-miR-15a-5p	Vegfa	22339	ENSMUSG00000023951	0.999
22	diana_microt	MIMAT0000526	mmu-miR-15a-5p	Cd2ap	12488	ENSMUSG00000061665	0.999
23	diana_microt	MIMAT0000526	mmu-miR-15a-5p	Rab9b	319642	ENSMUSG00000043463	0.999

C

	A	B	C	D	E	F	G
1	database	mature_mirna_id	target_symbol	target_entrez	target_ensembl	experiment	pubmed_id
2	mirecords	mmu-miR-15a-5p	Ccnd1	12443	ENSMUSG00000070348	fluorescence intensity	19723889
3	mirtarbase	mmu-miR-15a-5p	Ccnd1	12443	ENSMUSG00000070348	Luciferase reporter assay	19723889
4	mirtarbase	mmu-miR-15a-5p	Ccnt2	72949	ENSMUSG00000026349	Luciferase reporter assay	21740905
5	mirtarbase	mmu-miR-15a-5p	Eln	13717	ENSMUSG00000029675	Luciferase reporter assay//Microarray//Western blot	21305018
6	mirtarbase	mmu-miR-15a-5p	Bcl2	12043	ENSMUSG00000057329	Luciferase reporter assay//qRT-PCR//Western blot//Reporter assay;Western blot;qRT-PCR	20445066
7	mirtarbase	mmu-miR-15a-5p	Cadm1	54725	ENSMUSG00000032076	Luciferase reporter assay//Reporter assay;Other	18362358
8	mirtarbase	mmu-miR-15a-5p	Bcl2	12043	ENSMUSG00000057329	qRT-PCR//Western blot//Northern blot//Luciferase reporter assay	18931683
9	mirtarbase	mmu-miR-15a-5p	Ccnd1	12443	ENSMUSG00000070348	qRT-PCR//Western blot//Northern blot//Luciferase reporter assay	18931683
10	mirtarbase	mmu-miR-15a-5p	Wnt3a	22416	ENSMUSG00000009900	qRT-PCR//Western blot//Northern blot//Luciferase reporter assay	18931683
11	tarbase	mmu-miR-15a-5p	bcl-2			Reporter assay//qRT-PCR//Western blot	
12	tarbase	mmu-miR-15a-5p	Cadm1	54725	ENSMUSG00000032076	Reporter assay//Sequencing	

Figure 2 Layout for input and target mRNA mapping output files. (A) input file showing required column elements: miRNA name, fold change (FC) and *p* value; (B) output file layout showing results for the predicted target mRNA query; (C) output file layout showing results for the validated target mRNA query.
Full-size DOI: 10.7717/peerj.4262/fig-2

miRNA sequences for primer design. Such approach led to successful amplification of all 85 miRNAs in the marsupial. The case study used the mouse databases for target mRNA mapping. All output files can be downloaded and viewed from supplementary materials. The analysis was carried out on an Apple Macbook Pro computer with Intel Core i5 2.7 GHz dual-core CPU and 8 GB memory.

Figure 2A shows the input file layout. Upon importing the data to the R environment (sample data object name: *liver*), target mRNA mapping is conducted using the *rbiomirgs_mrnascan* function, through the command line: *rbiomirgs_mrnascan(objTitle = "mmu_liver_predicted", mir = liver$miRNA, sp = "mmu", queryType = "predicted", addhsaEntrez = TRUE, parallelComputing = TRUE, clusterType = "FORK")*. Figures 2B and 2C show truncated mapping results for both predicted and validated mapping results for miRNA *mmu-miR-25a-5p*. The mapping results showed that more predicted targets were retrieved than validated targets. The function output R projects as well as one *csv* file per miRNA tested. Since the case study enabled human ortholog Entrez ID conversion functionality, the function exported an R object including the Entrez ID for the human orthologs, with the suffix "*_hsa_entrez_list*" in the name.

Prior to enrichment, GS database files need to be obtained. For the case study, we used *gmt* files for KEGG and GO term databases downloaded from MSigDB

A

	A	B	C	D	E	F	G	H	I
1	GS	converged	loss	gene.tested	coef	std.err	t.value	p.value	adj.p.val
2	KEGG_GLYCOLYSIS_GLUCONEOGENESIS	Y	0.02370989	51	-0.1321968	0.04570857	-2.8921664	0.00383184	0.02639713
3	KEGG_CITRATE_CYCLE_TCA_CYCLE	Y	0.01378175	27	-0.1311647	0.06255063	-2.0969371	0.03601702	0.11901776
4	KEGG_PENTOSE_PHOSPHATE_PATHWAY	Y	0.01206466	23	-0.0981331	0.06362438	-1.5423824	0.12300327	0.27236439
5	KEGG_PENTOSE_AND_GLUCURONATE_INTERCONVERSIONS	Y	0.00531363	9	0.07667825	0.05887293	1.30243647	0.19278863	0.38147538
6	KEGG_FRUCTOSE_AND_MANNOSE_METABOLISM	Y	0.01350173	26	-0.0400732	0.05121837	-0.782399	0.43399335	0.6509071
7	KEGG_GALACTOSE_METABOLISM	Y	0.00886004	16	-0.0451418	0.0660327	-0.6836286	0.49422103	0.70362534
8	KEGG_ASCORBATE_AND_ALDARATE_METABOLISM	Y	0.00536329	9	0.0078778	0.07538492	0.10450104	0.91677324	0.94733235
9	KEGG_FATTY_ACID_METABOLISM	Y	0.01783618	36	-0.0528696	0.04532526	-1.166449	0.24345283	0.4439434
10	KEGG_STEROID_BIOSYNTHESIS	Y	0.00935086	17	-0.0291672	0.06134502	-0.4754619	0.63446519	0.7629928

B

	A	B	C	D	E	F	G	H
1	GS	gene.tested	coef	std.err	loss	z.score	p.value	adj.p.val
2	KEGG_GLYCOLYSIS_GLUCONEOGENESIS	51	-0.1321971	0.04647425	0.02370989	-2.8445237	0.00444779	0.028253926
3	KEGG_CITRATE_CYCLE_TCA_CYCLE	27	-0.1311643	0.06369688	0.01378175	-2.0591955	0.03947551	0.124448223
4	KEGG_PENTOSE_PHOSPHATE_PATHWAY	23	-0.0981331	0.06372651	0.01206466	-1.5399109	0.12358207	0.273646018
5	KEGG_PENTOSE_AND_GLUCURONATE_INTERCONVERSIONS	9	0.07654989	0.05931056	0.00531363	1.2906621	0.19682087	0.385354547
6	KEGG_FRUCTOSE_AND_MANNOSE_METABOLISM	26	-0.0400732	0.05132098	0.01350173	-0.7808347	0.43489973	0.651309523
7	KEGG_GALACTOSE_METABOLISM	16	-0.0451419	0.06635639	0.00886004	-0.6802944	0.49631806	0.703252087
8	KEGG_ASCORBATE_AND_ALDARATE_METABOLISM	9	0.00787781	0.07536779	0.00536329	0.10452491	0.9167528	0.947311227
9	KEGG_FATTY_ACID_METABOLISM	36	-0.0528698	0.04522525	0.01783618	-1.1690327	0.24239045	0.442006114
10	KEGG_STEROID_BIOSYNTHESIS	17	-0.0291671	0.0614788	0.00935086	-0.4744246	0.63519716	0.763217784

Figure 3 Layout for output GS enrichment results file for the case study, using two parameter optimization methods. (A) IWLS method; (B) BFGS method.

Full-size DOI: 10.7717/peerj.4262/fig-3

(http://software.broadinstitute.org/gsea/msigdb). Regarding GO term, separated files were used for biological processes (BP) and molecular function (MF) databases. The case study used the predicted miRNA:mRNA interaction results for enrichment. Furthermore, since all GS database files were based on human genes, we used the human ortholog Entrez ID list. GS enrichment was carried out with the command line (using KEGG database as the example): *rbiomirgs_logistic(objTitle = "mirna_mrna_iwls", mirna_DE = liver, var_mirnaName = "miRNA", var_mirnaFC = "FC", var_mirnaP = "pvalue", mrnalist = mmu_liver_predicted_mrna_hsa_entrez_list, mrna_Weight = NULL, gs_file = "kegg.v5.2.entrez.gmt", optim_method = "IWLS", p.adj = "fdr", parallelComputing = TRUE, clusterType = "PSOCK")*.

We tested all three parameter optimization algorithms on the KEGG analysis to select for the most effective method. The KEGG database included 186 pathways. Firstly, the liver data failed to converge for all the gene sets tested using the L-BFGS-B algorithm. Figure 3 shows a truncated version of the IWLS and BFGS results. The results suggest that both methods led to consistent coefficient values and model significance (Figs. 3 and 4). We found that the IWLS method with parallel computing enabled with the Unix operating system exclusive FORK mode took the least amount of time to converge for KEGG analysis (Fig. 5, based on three repeats). The one-way analysis of variance (ANOVA) test on the computation time suggested the time reduction when using such configuration was significant (Fig. 5).

As such, the following GO term enrichment was also carried out using IWLS and FORK methods. The results showed a similar trend as that of the KEGG analysis (Figs. 4 and 6), where more GO terms with a positive model coefficient value were identified.

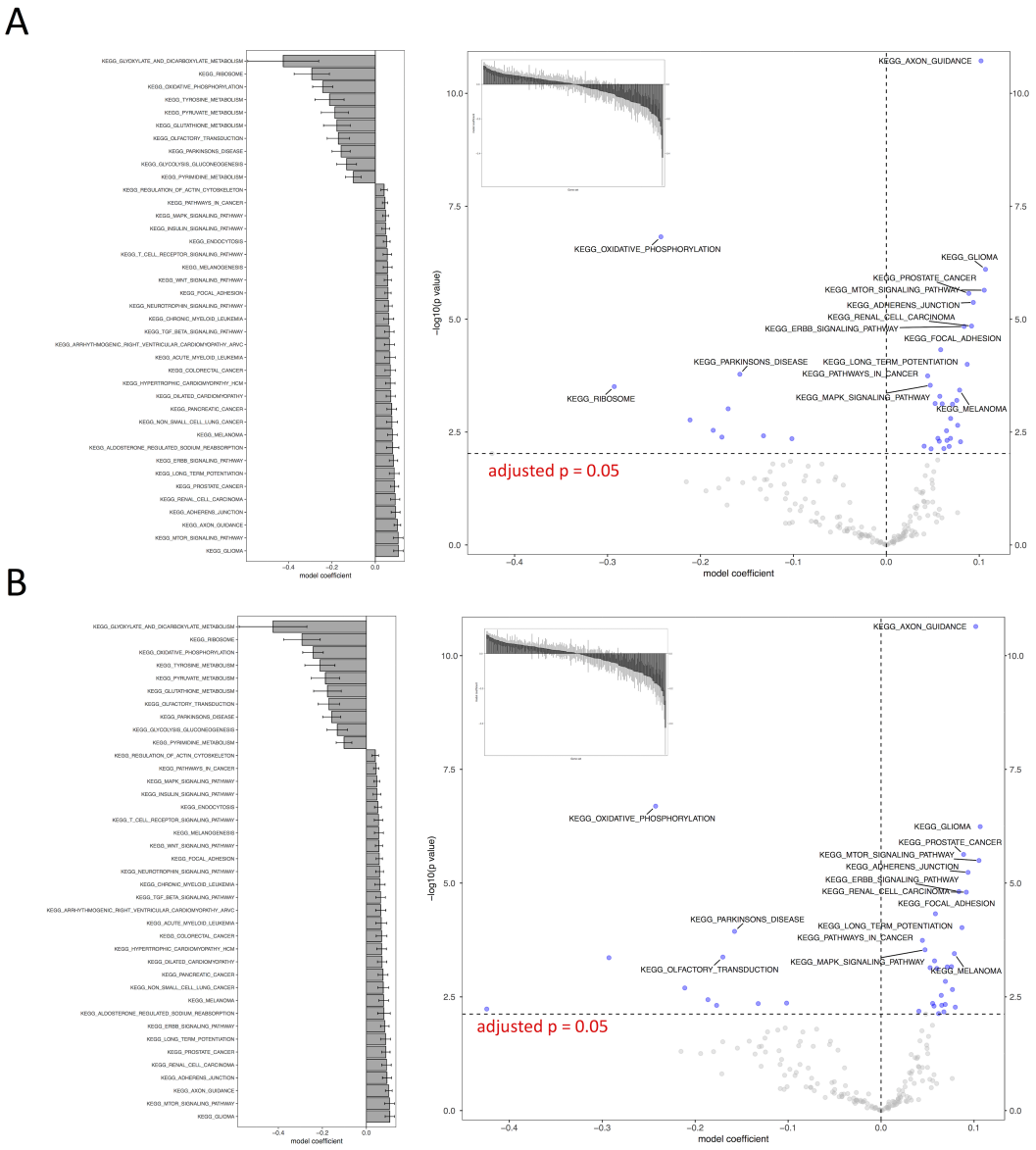

Figure 4 **Visual representation of KEGG pathway analysis for the case study.** Volcano plot depicts the significance and directionality distribution for the KEGG pathway tested (−log p value vs. model coefficient). Blue (upper quadrants) represent the significantly enriched KEGG pathways and the bar graph in the volcano plot shows the overall distribution of model coefficient. Top 15 most significantly enriched KEGG pathways are labeled. Bar graph shows the top 50 enriched gene sets; the bars are model coefficient ± standard error. Only the gene sets with an FDR adjusted $p < 0.05$ are plotted. (A) IWLS method; (B) BFGS method.

Full-size DOI: 10.7717/peerj.4262/fig-4

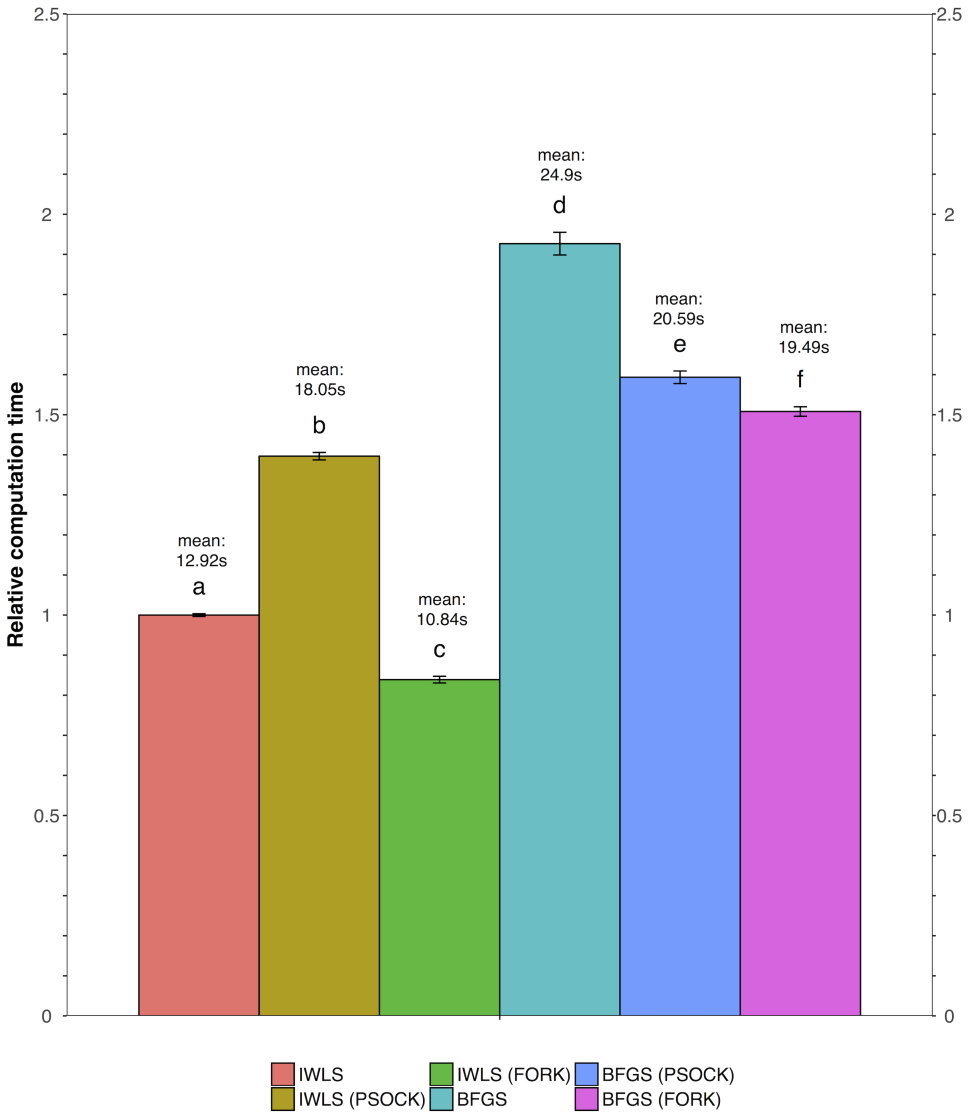

Figure 5 Comparison of relative computation time for parameter optimization between non-parallel and parallel settings using the KEGG database. FORK and PSOCK are parallel computing modes, with the former Unix or Unix-like operating system only. Bars are relative computation time ± SEM; different letters represent statistically significant changes ($p < 0.05$) according to a one-way ANOVA test with a Tukey post-hoc test. The raw mean value for each test is labeled in the graph based on three repeats.
Full-size DOI: 10.7717/peerj.4262/fig-5

DISCUSSION

RBiomirGS requires a miRNA identity list, a DE results list, as well as a GS database file as input (Fig. 1). The package uses fold change (FC) and *p* value to calculate the miRNA score, S_{mirna}. Since the DE results are associated with the miRNAs, both miRNA identity and DE results can be provided in a single *csv* file. The data layout can be viewed in Fig. 2. In addition, due to the modularization of the package functionality, target mRNA mapping can be used as a standalone function, with a list of miRNA names as input. The GS database

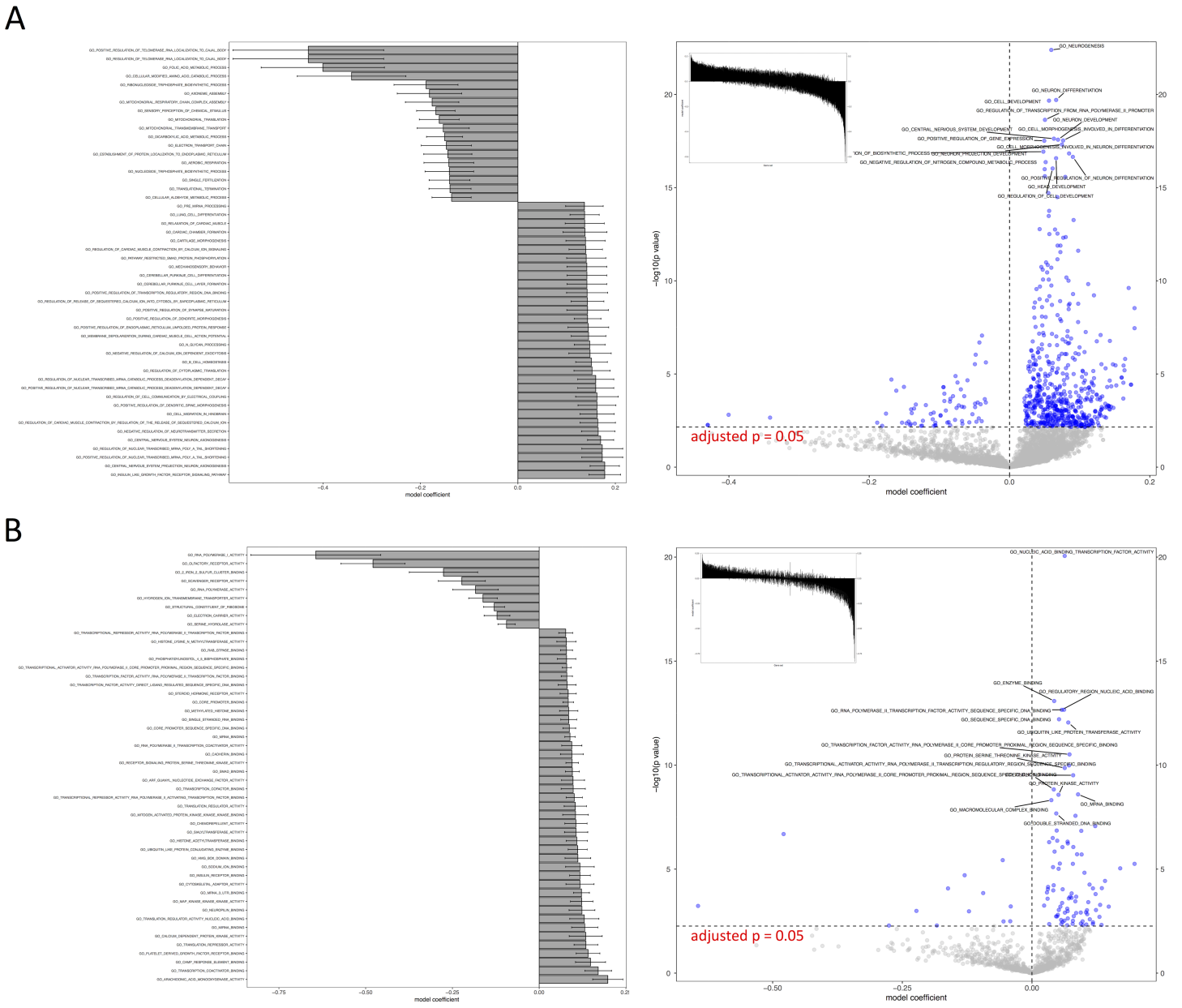

Figure 6 Visual representation of GO term analysis for the case study. (A) biological processes (BP); (B) molecular functions (MF). Additional information is same as Fig. 4.

Full-size DOI: 10.7717/peerj.4262/fig-6

file can be downloaded from various sources. One such source is MSigDB, which indexes popular GS database such as KEGG and GO term. Naturally, databases from other sources can also be used.

To efficiently process high throughput datasets, RBiomirGS implements parallel computing across all major functions. Depending on the user's computer configuration (i.e., number of CPU cores), parallel computing can provide significant speed enhancements. Moreover, both Unix/Unix-like operating system exclusive FORK and

universal PSOCK modes are available for maximizing hardware compatibility. It is worth noting that this feature can be disabled by users. Function *rbiomirgs_logistic* also implements linear algebra for score calculation to reduce computation time.

The target mRNA mapping module also features an optional gene Entrez ID conversion functionality that searches for human gene orthologs on Ensembl databases for rodent models (i.e., mouse or rat). Given the high conservation level in miRNA primary structure across species, such function enables the potential of revealing the miRNA functional implication in human from rodent models. The human Entrez ID conversion function is built upon the open sourced Biomart platform (*Durinck et al., 2005*; *Durinck et al., 2009*). By integrating Biomart software into the package, RBiomirGS connects directly to Ensembl database (http://www.ensembl.org) for human ortholog search using the most up-to-date information. While beneficial, such configuration imposes one limitation of the package wherein an active and functional internet connection is required for the target mRNA mapping function.

RBiomirGS conducts GS analysis through mRNA scores, miRNA scores and logistic regression. The mRNA score S_{mrna} is based on the assumption that, in most cases, miRNAs inhibit target mRNA translation events. Therefore, S_{mrna} represents the inhibitory effect on the mRNA of interest. As the sign reversed summation of S_{mirna}, the biological interpretation of S_{mrna} can be described as the following: In the case of a two-group comparison (i.e., experimental vs control), a positive S_{mrna} means the mRNA of interest might be inhibited more in the control group, whereas a negative value means the mRNA might be under miRNA inhibition upon experimental conditions. In addition, a bigger absolute value represents a stronger miRNA inhibitory effect. Given that S_{mirna} contains directionality information, such approach allows for accumulation and cancelation effects on the mRNA when the mRNA of interest is targeted by multiple miRNAs. Since the strength of the interaction between miRNA and mRNA varies among different miRNAs, it is critical to incorporate such consideration into the S_{mrna} calculation, regardless of the availability of such measurement. Therefore, we added the weight term w to Eq. (2) to accommodate the affinity of the miRNA:mRNA interaction, should such metric be available.

The central goal of the current logistic regression-based classification modelling is to separate the members of a gene set from the rest of the genes using S_{mrna}, which represents the overall miRNA regulatory effect. If a gene can be categorized into a gene set solely based on its S_{mrna}, then said gene set is under miRNA-dependent regulation. As such, based on the model significance test and user customizable GS p value threshold (e.g., FDR adjusted p value < 0.05 by default), a GS model with a significant adjusted p value means that the membership to such gene set for a gene can be determined based on its S_{mrna}, or that the gene set is significantly impacted by miRNA regulation. The biological interpretation of the model coefficient from Eq. (4) can be stated as follows (again, in the context of two-group comparison, i.e., experimental vs control): if the coefficient is positive, miRNA inhibition on target mRNAs might be lifted, thereby leading to less suppression on the gene set of interest in the experimental group. Furthermore, with a positive coefficient, a unit increase in S_{mrna} results in an increased odds ratio of a gene belonging to the gene set of interest. Conversely, a negative value means the opposite. It needs to be clarified

that a positive model coefficient for a gene set means that the gene set of interest might be under more miRNA-dependent inhibition in the control group, as opposed to being activated under the experimental condition. Such observation is closely related to the fact that the miRNA regulation on a pathway is mostly indirect, and represents only one layer of regulation on the mRNAs. As such, another limitation of RBiomirGS is in its limited capacity for evaluating gene set activation when solely relying on miRNA DE results.

The case study demonstrated the usage of RBiomirGS. In general, enrichment on all three GS databases suggested that more gene sets were free from miRNA-dependent inhibition in the livers of torpid marsupials, represented by positive model coefficient values (Figs. 4 and 6). The result is consistent with the observation from the original study where most miRNAs tested showed decreased relative expression levels in liver (*Hadj-Moussa et al., 2016*), leading to less inhibitory effect on their target mRNAs, which in turn resulted in more gene sets independent from miRNA-dependent regulation. For example, such enriched KEGG pathways included mTOR signaling pathway and MAPK signaling pathway, which, when activated, were considered to play critical roles in facilitating torpor (*Hadj-Moussa et al., 2016*). However, the volcano plots in Figs. 4 and 6 suggest that potentially inhibited gene sets in the liver from torpid marsupials exhibited a greater impact by the miRNA, i.e., a wider spread pattern on the x-axis in the negative direction. The KEGG pathways that might be suppressed included Ribosome (KEGG ID: map03010), RNA polymerase (KEGG ID: map03020), Oxidative phosphorylation (KEGG ID: map00190), and Pyruvate metabolism (map00620). Inhibition of those pathways may contribute to suppressing ATP expensive cellular processes such as global gene transcription and protein synthesis, all of which have been reported to be inhibited in other hibernating animals (*Storey, 2010*; *Wu & Storey, 2016*). It is also not a surprise that oxidative phosphorylation and pyruvate metabolism pathways were inhibited under hypometabolic conditions (*Storey, 1997*). Overall, by using RBiomirGS, additional miRNA-dependent regulatory mechanisms that underpin the molecular adaptations facilitating daily torpor in marsupials were revealed.

By incorporating all the core steps into one R package, RBiomirGS eliminates the need for switching between different software packages, or between different software platforms. The package also provides two data visualization functions that can produce three types of plots. Furthermore, the modular structure of RBiomirGS enables various access points to the analysis, with which users can choose the most relevant functionalities for their workflow. With RBiomirGS, users will be able to comprehensively assess the functional implications of the miRNA expression profile under the corresponding experimental condition by minimal input and intervention. Accordingly, RBiomirGS provides an all-in-one and highly accessible miRNA GS analysis solution.

ACKNOWLEDGEMENTS

We thank the Storey lab members for testing the package.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The present study is supported by a Discovery grant from the Natural Sciences and Engineering Research Council of Canada (NSERC) to Kenneth B Storey (grant number: 6793). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Natural Sciences and Engineering Research Council of Canada (NSERC): 6793.

Competing Interests

Kenneth Storey is an Academic Editor for PeerJ.

Author Contributions

- Jing Zhang conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.
- Kenneth B. Storey reviewed drafts of the paper.

Data Availability

The following information was supplied regarding data availability:
GitHub: http://github.com/jzhangc/RBiomirGS.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.4262#supplemental-information.

REFERENCES

Akhtar MM, Micolucci L, Islam MS, Olivieri F, Procopio AD. 2016. Bioinformatic tools for microRNA dissection. *Nucleic Acids Research* **44(1)**:24–44 DOI 10.1093/nar/gkv1221.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nature Genetics* **25**:25–29 DOI 10.1038/75556.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **57**:289–300.

Betel D, Wilson M, Gabow A, Marks DS, Sander C. 2008. The microRNA.org resource: targets and expression. *Nucleic Acids Research* **36(Database issue)**:D149–D153 DOI 10.1093/nar/gkm995.

Bleazard T, Lamb JA, Griffiths-Jones S, Griffiths-Jones S. 2015. Bias in microRNA functional enrichment analysis. *Bioinformatics* **31**:1592–1598 DOI 10.1093/bioinformatics/btv023.

Byrd RH, Lu P, Nocedal J, Zhu C. 1994. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing* **16**:1190–1208 DOI 10.1137/0916069.

Chen M, Zhang X, Liu J, Storey KB. 2013. High-throughput sequencing reveals differential expression of miRNAs in intestine from sea cucumber during aestivation. *PLOS ONE* **8**:e76120 DOI 10.1371/journal.pone.0076120.

Chou CH, Chang NW, Shrestha S, Hsu SD, Lin YL, Lee WH, Yang CD, Hong HC, Wei TY, Tu SJ, Tsai TR, Ho SY, Jian TY, Wu HY, Chen PR, Lin NC, Huang HT, Yang TL, Pai CY, Tai CS, Chen WL, Huang CY, Liu CC, Weng SL, Liao KW, Hsu WL, Huang HD. 2016. miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Research* **44**:D239–D247 DOI 10.1093/nar/gkv1258.

De Leeuw CA, Neale BM, Heskes T, Posthuma D. 2016. The statistical properties of gene-set analysis. *Nature Reviews Genetics* **17**:353–364 DOI 10.1038/nrg.2016.29.

Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W. 2005. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**:3439–3440 DOI 10.1093/bioinformatics/bti525.

Durinck S, Spellman P, Birney E, Huber W. 2009. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols* **4**:1184–1191 DOI 10.1038/nprot.2009.97.

Friedman RC, Farh KK, Burge CB, Bartel DP. 2009. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research* **19**:92–105 DOI 10.1101/gr.082701.108.

Fu YX, Li WH. 1993. Maximum likelihood estimation of population parameters. *Genetics* **134**:1261–1270.

Gaidatzis D, Van Nimwegen E, Hausser J, Zavolan M. 2007. Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics* **8**:69 DOI 10.1186/1471-2105-8-69.

Garcia DM, Baek D, Shin C, Bell GW, Grimson A, Bartel DP. 2011. Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. *Nature Structural & Molecular Biology* **18**:1139–1146 DOI 10.1038/nsmb.2115.

Garcia-Garcia F, Panadero J, Dopazo J, Montaner D. 2016. Integrated gene set analysis for microRNA studies. *Bioinformatics* **32**:2809–2016 DOI 10.1093/bioinformatics/btw334.

Gomes CP, Cho JH, Hood L, Franco OL, Pereira RW, Wang K. 2013. A review of computational tools in microRNA discovery. *Frontiers in Genetics* **4**:Article 81 DOI 10.3389/fgene.2013.00081.

Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP. 2007. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Molecular Cell* **27**:91–105 DOI 10.1016/j.molcel.2007.06.017.

Hadj-Moussa H, Moggridge JA, Luu BE, Quintero-Galvis JF, Gaitán-Espitia JD, Nespolo RF, Storey KB. 2016. The hibernating South American marsupial, *Dromiciops gliroides*, displays torpor-sensitive microRNA expression patterns. *Scientific Reports* **6**:24627 DOI 10.1038/srep24627.

He L, Hannon GJ. 2004. MicroRNAs: small RNAs with a big role in gene regulation. *Nature Reviews Genetics* **5**:522–531 DOI 10.1038/nrg1379.

Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **28**:27–30 DOI 10.1093/nar/28.1.27.

Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E. 2007. The role of site accessibility in microRNA target recognition. *Nature Genetics* **39**:1278–1284 DOI 10.1038/ng2135.

Khatri P, Sirota M, Butte AJ. 2012. Ten years of pathway analysis: current approaches and outstanding challenges. *PLOS Computational Biology* **8**:e1002375 DOI 10.1371/journal.pcbi.1002375.

Krek A, Grün D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, MacMenamin P, Da Piedade I, Gunsalus KC, Stoffel M, Rajewsky N. 2005. Combinatorial microRNA target predictions. *Nature Genetics* **37**:495–500 DOI 10.1038/ng1536.

Lewis BP, Burge CB, Bartel DP. 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**:15–20 DOI 10.1016/j.cell.2004.12.035.

Long C, Jiang L, Wei F, Ma C, Zhou H, Yang S, Liu X, Liu Z. 2013. Integrated miRNA-mRNA analysis revealing the potential roles of miRNAs in chordomas. *PLOS ONE* **8**:e66676 DOI 10.1371/journal.pone.0066676.

Mitchell PS, Parkin RK, Kroh EM, Fritz BR, Wyman SK, Pogosova-Agadjanyan EL, Peterson A, Noteboom J, O'Briant KC, Allen A, Lin DW, Urban N, Drescher CW, Knudsen BS, Stirewalt DL, Gentleman R, Vessella RL, Nelson PS, Martin DB, Tewari M. 2008. Circulating microRNAs as stable blood-based markers for cancer detection. *Proceedings of the National Academy of Sciences of the United States of America* **105**:10513–10518 DOI 10.1073/pnas.0804549105.

Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstråle M, Laurila E, Houstis N, Daly MJ, Patterson N, Mesirov JP, Golub TR, Tamayo P, Spiegelman B, Lander ES, Hirschhorn JN, Altshuler D, Groop LC. 2003. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics* **34**:267–273 DOI 10.1038/ng1180.

Paraskevopoulou MD, Georgakilas G, Kostoulas N, Vlachos IS, Vergoulis T, Reczko M, Filippidis C, Dalamagas T, Hatzigeorgiou AG. 2013. DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. *Nucleic Acids Research* **41**(**Web Server issue**):W169–W173 DOI 10.1093/nar/gkt393.

R Core Team. 2017. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. *Available at http://www.R-project.org/*.

Roger F. 1987. *Practical methods of optimization*. Second Edition. New York: John Wiley & Sons.

Ru Y, Kechris KJ, Tabakoff B, Hoffman P, Radcliffe RA, Bowler R, Mahaffey S, Rossi S, Calin GA, Bemis L, Theodorescu D. 2014. The multiMiR R package and database: integration of microRNA-target interactions along with their disease and drug associations. *Nucleic Acids Research* **42**:e133 DOI 10.1093/nar/gku631.

Sethupathy P, Corda B, Hatzigeorgiou AG. 2006. TarBase: a comprehensive database of experimentally supported animal microRNA targets. *RNA* **12**:192–197 DOI 10.1261/rna.2239606.

Storey KB. 1997. Metabolic regulation in mammalian hibernation: enzyme and protein adaptations. *Comparative Biochemistry and Physiology—Part A: Physiology* **118**:1115–1124 DOI 10.1016/S0300-9629(97)00238-7.

Storey KB. 2010. Out cold: biochemical regulation of mammalian hibernation—a mini-review. *Gerontology* **56**:220–230 DOI 10.1159/000228829.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**:15545–15550 DOI 10.1073/pnas.0506580102.

Wang K, Zhang S, Marzolf B, Troisch P, Brightman A, Hu Z, Hood LE, Galas DJ. 2009. Circulating microRNAs, potential biomarkers for drug-induced liver injury. *Proceedings of the National Academy of Sciences of the United States of America* **106**:4402–4407 DOI 10.1073/pnas.0813371106.

Wang X. 2008. miRDB: a microRNA target prediction and functional annotation database with a wiki interface. *RNA* **14**:1012–1017 DOI 10.1261/rna.965408.

Wickham H. 2009. *ggplot2: elegant graphics for data analysis*. New York: Springer.

Wolke R, Schwetlick H. 1987. Iteratively reweighted least squares: algorithms, convergence analysis, and numerical comparisons. *SIAM Journal on Scientific and Statistical Computing* **9**:907–921 DOI 10.1137/0909062.

Wu CW, Storey KB. 2016. Life in the cold: links between mammalian hibernation and longevity. *Biomolecular Concepts* **7**:41–52 DOI 10.1515/bmc-2015-0032.

Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T. 2009. miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Research* **7(Database issue)**:D105–D110 DOI 10.1093/nar/gkn851.

Zhang J, Storey KB. 2013. Akt signaling and freezing survival in the wood frog, *Rana sylvatica*. *Biochimica et Biophysica Acta/General Subjects* **1830**:4828–4837 DOI 10.1016/j.bbagen.2013.06.020.

Zhang J, Storey KB. 2016. RBioplot: an easy-to-use R pipeline for automated statistical analysis and data visualization in molecular biology and biochemistry. *PeerJ* **4**:e2436 DOI 10.7717/peerj.2436.