Bioinformatics, 31(17), 2015, 2822–2828
doi: 10.1093/bioinformatics/btv220
Advance Access Publication Date: 24 April 2015
Original Paper

Gene expression

Inferring data-specific micro-RNA function through the joint ranking of micro-RNA and pathways from matched micro-RNA and gene expression data

Ellis Patrick[1], Michael Buckley[2], Samuel Müller[1], David M. Lin[3] and Jean Y. H. Yang[1,*]

[1]School of Mathematics and Statistics, University of Sydney, Sydney, NSW 2006, Australia, [2]CSIRO Mathematical & Information Sciences, Clayton South, VIC 3168, Australia and [3]Department of Biomedical Sciences, Cornell University, Ithaca, NY, USA

*To whom correspondence should be addressed.

Associate Editor: Ivo Hofacker

Received on August 24, 2014; revised on April 1, 2015; accepted on April 19, 2015

Abstract

Motivation: In practice, identifying and interpreting the functional impacts of the regulatory relationships between micro-RNA and messenger-RNA is non-trivial. The sheer scale of possible micro-RNA and messenger-RNA interactions can make the interpretation of results difficult.
Results: We propose a supervised framework, pMim, built upon concepts of significance combination, for jointly ranking regulatory micro-RNA and their potential functional impacts with respect to a condition of interest. Here, pMim directly tests if a micro-RNA is differentially expressed and if its predicted targets, which lie in a common biological pathway, have changed in the opposite direction. We leverage the information within existing micro-RNA target and pathway databases to stabilize the estimation and annotation of micro-RNA regulation making our approach suitable for datasets with small sample sizes. In addition to outputting meaningful and interpretable results, we demonstrate in a variety of datasets that the micro-RNA identified by pMim, in comparison to simpler existing approaches, are also more concordant with what is described in the literature.
Availability and implementation: This framework is implemented as an **R** function, *pMim*, in the package *sydSeq* available from http://www.ellispatrick.com/r-packages.
Contact: jean.yang@sydney.edu.au
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Micro-RNA (miRNA) are a class of small non-coding RNA molecules which down-regulate gene expression. Through basepairing, miRNA down-regulate the expression of genes by inhibiting their translation or promoting the degradation of their target messenger-RNA (mRNA). Dysregulation of miRNA can lead to a variety of human diseases with miRNA being shown to play a critical regulatory role in many cellular pathways and functions such as developmental timing, cell death, cell proliferation, immunity and patterning of the nervous system (Mo, 2012). As such, investigating the functional impacts of miRNA may further the understanding of many diseases.

One of the first steps to understand how a miRNA could regulate a biological process is predicting which mRNA can directly interact with a miRNA. Irrespective of a particular dataset or condition, there exist many computational algorithms for predicting the target genes of miRNA, such as TargetScan (Lewis *et al.*, 2005), miRBase (Griffiths-Jones *et al.*, 2008) and PicTar (Krek *et al.*, 2005).

These algorithms essentially attempt to identify whether a mRNA contains the binding motif of a miRNA. This remains an active area of research as there is generally not a large consensus between algorithms (Jayaswal *et al.*, 2009).

Attempting to identify a miRNA that is regulating a set of genes in response to a particular condition or treatment is also an established problem. A simple approach for identifying candidate miRNA-mRNA regulatory relationships is to identify a differentially expressed (DE) miRNA and look for large negative pair-wise correlations between that miRNA and the genes that it is predicted to bind to (Havelange *et al.*, 2011; Li *et al.*, 2011) Another common approach is to identify a DE miRNA and then use predicted binding target information to perform an enrichment analysis on the DE genes or miRNA-mRNA correlations (Nam *et al.*, 2009; Xu and Wong, 2013). An important but often overlooked factor to both of these approaches is the combination of the significance from the test on the miRNA with the test on its targets. Not combining these tests may produce results that are overly conservative as two *p*-value cutoffs are performed instead of one (Yang *et al.*, 2014). This could result in biologically significant signal being missed.

As miRNA can potentially target hundreds or even thousands of genes, if a miRNA and its targets are identified as DE then further pathway enrichment analysis is often performed to make any results biologically interpretable. This adds another level of statistical complication and yet another *P*-value cutoff. Practically this style of analysis often produces a list of DE miRNA, a list of DE miRNA targets and a list of pathway analyses for each miRNA. These nested lists can quickly become unmanageable, in some sense broadening an exploratory analysis instead of focusing it.

Instead of performing functional pathway analysis as an after thought, directly considering how miRNA interact with annotated pathways may help the interpretation of their function. It is well documented that the predicted gene targets of miRNA are often over-represented for annotated functional pathways (Maragkakis *et al.*, 2009; Wang, 2008). Likewise, annotated pathways are over-represented for predicted gene targets of miRNA (see Supplementary Section S3 for further explanation). This observation suggests that pathway databases contain stable information about miRNA-mRNA regulatory networks. If leveraged effectively, this information could improve both the power and interpretability of results.

To this end, we propose an integrative analysis for **p**athways, **Mi**cro-RNA and **m**RNA (pMim) incorporating publicly available biological network information and miRNA target predictions together with experimental RNA-seq and miRNA-seq data. This novel framework is the first to allow for the joint ranking of interesting miRNA and pathways by directly testing for miRNA that target a group of genes from a specific biological pathway in response to a treatment or condition. This approach produces one list of biologically interpretable results instead of the multiple lists generated by the previously described methods. Also, due to its simplicity it is free of the statistically demanding task of estimating complex miRNA-mRNA interactions making it suitable for datasets with small sample sizes.

2 Materials and methods

2.1 pMiM: Pathway, miRNA and mRNA integration

A framework for integrating various data sources to identify regulatory miRNA and their targets, pMim, is described below. This framework will be utilized to form three specific methods cMimDE, pMimDE and pMimCor. The method cMimDE is proposed as a method to identify regulatory miRNA and their targets in an experiment. The concept of mir-pathways is then proposed and used to develop pMimDE and pMimCor as methods for identifying regulatory miRNA that may be targeting a group of genes that share a common biological function.

We first describe in detail the input data sources and notation before describing the methods in detail. Let $\mathbf{Y}$ be a p_g by n matrix corresponding to gene expression estimates for p_g genes in n samples, where the n samples are divided into two conditions where for sample j, $\gamma(j) = 1$ or 2. Similarly let $\mathbf{Z}$ be a matrix of expression estimates for p_{mi} miRNA and the same matched n samples. Assume there also exists some p_{mi} by p_g matrix $\mathbf{M}$, where M_{ij} is equal to one if the ith miRNA is predicted to bind to the jth gene and is zero otherwise. Furthermore, assume there exists some functional or pathway annotation of the genes and represent this as a p_f by p_g matrix $\mathbf{F}$, where F_{ij} is equal to one if the jth gene is in the ith group and zero otherwise.

Prior to integration, the information from matrices $\mathbf{Y}$ and $\mathbf{Z}$ are reduced to summary statistics. The vectors of test statistics $\mathbf{t}^Y$ and $\mathbf{t}^Z$ are calculated for whether each of the genes and miRNA, respectively, are DE between conditions. The p_{mi} by p_g matrix $\mathbf{K}$ contains the probabilities of observing the negative correlations between each of the miRNA and genes. When combining information from tests, Stouffer's method (Stouffer *et al.*, 1949) or a one-sided Pearson's method (OSP) (Pearson, 1934) will be used for testing if *any* or *all* of tests have shown change respectively. Further details regarding the summary statistics and combination methods are described in Supplementary Section S1.

2.1.1 cMimDE: Classic miRNA and mRNA integration using DE

A common approach for highlighting interesting miRNA-mRNA relationships is to first identify a DE miRNA and then use predicted binding target information to perform an enrichment analysis on the DE genes or miRNA-mRNA correlations (Xu and Wong, 2013). This is similar to taking the maximum of the DE miRNA *P*-value and the enrichment analysis *P*-value. Taking the maximum of two *P*-values will result in dramatic reduction of power for the combined *P*-value. In the following, cMimDE is proposed as both a formalization and improvement of this approach.

The method cMimDE tests whether a miRNA is DE and its target genes are DE in the opposite direction. For simplicity, consider finding miRNA with negative log fold change and genes with positive log fold change. A gene set test can then be performed to test whether the target genes of the i^{th} miRNA are up-regulated using Stouffer's method

$$s_i = \Phi\left(\frac{\sum_{j:M_{ij}=1} \Phi^{-1}(P(t_j^Y > 0))}{\sqrt{\sum_{j:M_{ij}=1} 1}}\right), \qquad (1)$$

where Φ is the cumulative distribution function of the standard normal distribution. This gene set test could also have been performed using any method that is favoured, such as by using Fisher Method, a Wilcoxon rank-sum test, over-representation test or gene set enrichment analysis (Subramanian *et al.*, 2005).

To identify those miRNA that are negatively DE and whose target genes are positively DE we can then combine the vectors $\mathbf{t}^Z$ and $\mathbf{s}$ using OSP, where

$$P_i^{cMimDE} = P\Big(U < -2\log\Big(\Big(1 - P(t_i^Z < 0)\Big)(1 - s_i)\Big)\Big) \qquad (2)$$

and U is distributed as a chi-squared distribution with $2n$ degrees of freedom. This test will rank similarly to taking the maximum of the two *P*-values from the miRNA DE and its corresponding gene set test.

2.1.2 pMimDE: Pathway, miRNA and mRNA integration using DE

In addition to finding miRNA that are potentially regulating their target genes it may be of interest to find miRNA that are regulating a set of genes that also share some common biological function or outcome. This test can be implemented by performing gene set tests on the genes that lie in the intersection of the binding target predictions of a miRNA and a biological pathway. These intersections will be referred to as mir-pathways. Stouffers method can again be used to perform this gene set test and find evidence of whether a set of genes in a particular mir-pathway (i.e. genes that are both targeted by the *i*th miRNA and also belong to the *k*th functional pathway) are up-regulated. Let the matrix **S** have entries that correspond to these tests

$$S_{ik} = \Phi\left(\frac{\sum_{j:M_{ij}=1 F_{kj}=1} \Phi^{-1}(P(t_j^Y > 0))}{\sqrt{\sum_{j:M_{ij}=1 F_{kj}=1} 1}}\right). \quad (3)$$

S_{ik} whose corresponding mir-pathways contain less than two intersecting genes will be defined as empty.

These gene set tests, **S**, can then be combined with the miRNA test statistics t^Z using OSP. However, as using OSP is similar to taking the maximum of two *P*-values, then as multiple gene set tests are being performed on the target genes of each miRNA, it can be expected that one of the *P*-values of these gene set tests will be smaller than the miRNA *P*-value due to multiple testing. This will result in the ranking of the minimum combined *P*-values for each miRNA being very close to the ranking of the miRNA *P*-values. To account for these multiple testing issues a multiple testing correction will be applied to each row of **S** before combining using OSP. Many of the mir-pathways may contain similar genes and hence their expression may be correlated. A conservative approach will be taken, ignoring this correlation and performing a row-wise Benjamini-Hochberg (Benjamini and Hochberg, 1995) correction to **S** and calling this $\mathbf{S}^{fdr}$.

To identify those miRNA that are negatively DE and whose target genes are both positively DE and share some common biological function the elements of the matrix $\mathbf{S}^{fdr}$ can now be combined with the elements of the vector $\mathbf{t}^Z$ using OSP, where

$$P_{ik}^{pMimDE} = P\Big(U < -2\log\Big(\Big(1 - P(t_i^Z < 0)\Big)(1 - S_{ik}^{fdr})\Big)\Big), \quad (4)$$

U is chi-squared distributed with $2n$ degrees of freedom and $\mathbf{P}^{pMimDE}$ is a p_{mi} by p_f matrix.

2.1.3 pMimCor: Pathway, miRNA and mRNA integration using correlation

Finding sets of genes that are DE in the opposite direction to a miRNA is very similar to requiring the expression of genes to be negatively correlated with the expression of their corresponding miRNA between conditions if a true relationship exists. In the following, the method pMimDE is extended to simply require that the expression of a miRNA and gene be negatively correlated. As the correlation between the expression of a miRNA and its targets is expected to be negative, the test described for pMimDE can be modified by replacing the definition of S_{ik} with

$$S_{ik} = \Phi\left(\frac{\sum_{j:M_{ij}=1 F_{kj}=1} K_{ij}}{\sqrt{\sum_{j:M_{ij}=1 F_{kj}=1} 1}}\right). \quad (5)$$

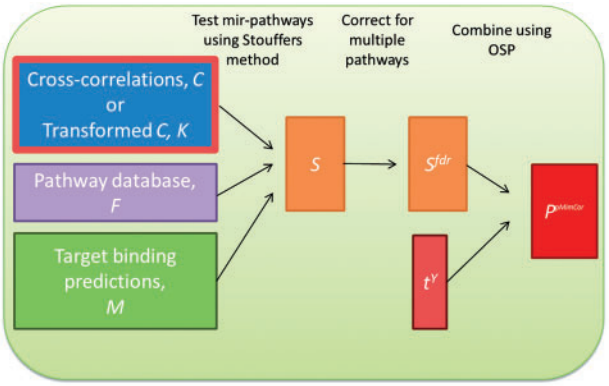

Fig. 1. A visual representation of pMimCor

Here, K_{ij} is a monotone transformation of the cross-correlation between the *i*th miRNA and *j*th gene given by either Equation (S4) or (S5) in Supplementary file S1. This approach will be refered to as pMimCor. A visual representation is given in Figure 1

2.2 Data

Four datasets with matched miRNA-Seq and RNA-Seq samples will be used to evaluate the proposed methods. The first, Notch, consists of matched mRNA and miRNA enriched cortical samples from three wild type mice and three conditional *Notch2* knockout mice. In this dataset, *Notch2* has been conditionally deleted in the brain with a nestin-cre driver. Cortical samples were isolated from matched mutant and control adult mice The other three are from The Cancer Genome Atlas (TCGA) and were chosen to represent a broad spectrum of cancers. These datasets were analysed from a prognostic perspective and were hence split into groups to represent good prognosis (GP) and poor prognosis (PP). The ovarian serous cystadenocarcinoma data (Ovarian) has 26 matched mRNA and miRNA enriched samples in GP and 23 in PP. The skin cutaneous melanoma data (Melanoma) has 19 matched mRNA and miRNA enriched samples in GP and 21 in PP. While the lung adenocarcinoma data (Lung) has 16 matched mRNA and miRNA enriched samples in GP and 17 in PP. Further information on how these datasets were processed and split into prognosis groups can be found in Supplementary File S1.

2.3 Evaluation

Three approaches will be used to evaluate the performance and viability of the proposed methods cMimDE, pMimDE and pMimCor. For these evaluations TargetScan (Lewis *et al.*, 2005) and KEGG (Kanehisa and Goto, 2000) are used as the predicted binding target and pathway databases respectively. Both databases are accessed via the R packages *multiMiR* (Ru *et al.*, 2014) and *KEGG.db* (Carlson, 2014).

2.3.1 Evaluating signal in the data

One of the key limitations of working with small sample sizes is the inability to use sampling methods, such as cross-validation, to improve confidence about results. However, the experimental data is not the only input into our proposed methods. As pMimCor and pMimDE rely on large pathway and target matrices, we propose that these are resampled instead.

A method for evaluating the output of pMimCor when run on the four evaluation datasets is described as follows:

Resampling scheme:

> pMimCor is implemented using KEGG and TargetScan as inputs. The number of significant mir-pathways are calculated at an arbitrary *P*-value cut-off. This is then repeated one hundred times, finding the average number of significant mir-pathways, after randomly reassigning the genes for each pathway in KEGG, reassigning the target genes for each miRNA in TargetScan, and reassigning genes in both KEGG and TargetScan.

The *P*-values have not been adjusted for multiple testing across miRNA. If there is signal related to miRNA regulation in the data, the binding target predictions are accurate and there exists annotated phenotypes that are associated with miRNA regulation then resampling the pathway and/or target matrices should reduce the number significant mir-pathways found. The probit rank transformation is used when estimating the probability of a miRNA-gene correlation in the Notch dataset while the Fisher transformation is used in the TCGA datasets.

2.3.2 Evaluation via literature search

The performance of the proposed methods will be assessed by their concordance with results from a literature search. For example, the Notch experiment was designed to study the loss of *Notch2* function in the brain in the context of neurodegeneration. With this in mind, PubMed can be used to identify miRNA that have been associated with neurodegeneration. This information will allow the concurrent verification of the relationships between miRNA regulation, the loss of *Notch2* function and any effect on neurodegeneration and the effectiveness of our proposed analysis framework.

The strategy for performing and evaluating the literature search is outlined as follows:

Literature search strategy:

> A PubMed search was performed for each miRNA that was observed in the data. Each miRNA was included in the following analysis if it had at least one publication referring to it in the abstract. A second batch of searches was performed searching for each miRNA and a key word, e.g. 'neurodegeneration'. A miRNA was then classified as being associated with neurodegeneration if it had at least one hit on this search. Treating this information as an approximation of the truth, it was then used to calculate the number of true positives (TP) and false positives (FP) for four methods; miRNA DE, a moderated t-test on just the miRNA data; cMimDE, a classic miRNA and mRNA integration using just miRNA and gene DE; pMimDE and pMimCor, pathway, miRNA and mRNA integration using DE correlation respectively.
>
> As pMimDE and pMimCor have many *P*-values for each miRNA, the minimum of these for each miRNA was taken to provide a rank for the miRNA.

A miRNA is labelled as TP for a method if that method identifies it as interesting and it had a hit in the search. Alternatively, a miRNA is labelled as FP if the method identifies it as interesting and it did not have a hit in the search. This approach will be biased towards previous research and/or methodology. The search terms 'Melanoma', 'Ovarian' and 'Adenocarcinoma' were used for the Melanoma, Ovarian and Lung datasets respectively.

2.3.3 Evaluation of prognostic performance

To further assess the concordance of the miRNA identified by each of the methods and their corresponding target genes a prognostic strategy will be used. If a miRNA is regulatory with respect to a condition of interest, its target genes should be viable prognostic markers.

Prognostic strategy:

> The three most significant miRNA from each method are selected after being applied to a complete dataset. The samples of the dataset are then split into five folds. Four of these folds are used to calculate differential gene expression with respect to prognosis. The 10 most significantly DE genes that are also targeted by the three significant miRNA are then used to build a classifier using diagonalized linear discriminant analysis (DLDA) trained on the four folds. This classifier is then tested on the remaining fold. The process is repeated for each fold and replicated 100 times. A classification error is then calculated.

This prognostic strategy is applied to the Melanoma, Ovarian and Lung datasets. The four methods for selecting the three most significant miRNA are the same as those outlined in the literature search strategy. Classification errors are also calculated using the top i most differentially expressed genes within the training set, for $i = 2, 3, \ldots, 50$, to give perspective on the impact of the number of genes used.

3 Results

A resampling scheme was used to evaluate if there is miRNA regulation in any of the datasets and to illustrate the importance of having accurate predicted binding target and pathway matrices when implementing pMimDE and pMimCor. Table 1 shows the results from the resampling scheme applied to four different datasets. Randomising either the pathway (KEGG) or predicted binding target (TargetScan) matrices reduces the number of mir-pathways that are called significant in all four datasets. Randomising both matrices reduces the number of mir-pathways further. This suggests that both KEGG and TargetScan matrices contain information pertinent to miRNA regulation. This is also suggestive that miRNA may be regulating gene expression in each of the datasets. While observed in all datasets this pattern is observed most strongly in the Melanoma data, with the number of identified mir-pathways dropping 80% from 92 to 18 mir-pathways after randomization.

In order to compare the miRNA rankings of four methods; miRNA DE, cMimDE, pMimDE and pMimCor; a strategy based on broad literature searches is used. Methods which rank highly miRNA that have been associated with the conditions in each of the datasets in the literature would be considered favourable. Figure 2 plots the TP against FP for the region of small FP for each of the four datasets. Full ROC curves can be found in Supplementary File S1. For this region of small FP, pMimCor performs better than or comparably to all other approaches in every dataset. This is because

Table 1. The number of significant mir-pathways

Randomization	Notch	Melanoma	Ovarian	Lung
None	46.0	92.0	19.0	39.0
miRNA	28.5	23.8	11.0	28.5
KEGG	17.8	41.5	9.3	31.4
miRNA & KEGG	15.6	18.0	5.7	20.9

A table of the average number of significant mir-pathways calculated at an arbitrary *P*-value cut-off of 0.01 for four datasets using pMimCor. Significance is calculated for mir-pathways estimated using TargetScan and KEGG, randomized TargetScan and KEGG, TargetScan and randomized KEGG and both randomized TargetScan and randomized KEGG.

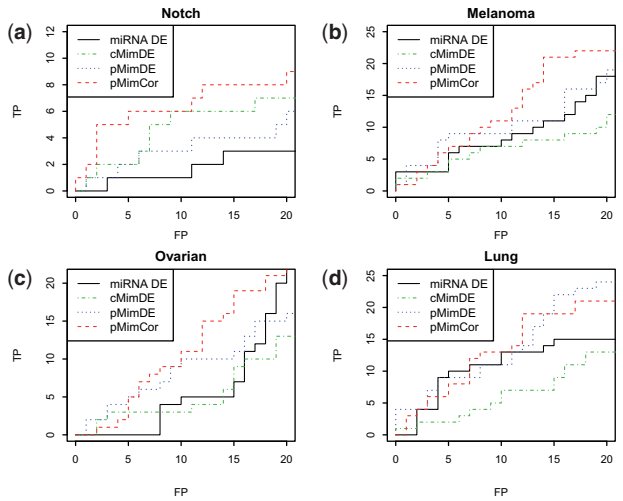

Fig. 2. TP versus FP from PubMed search. True Positives (TP) are plotted against False Positives (FP) for the small FPR regions of the ROC curves in Supplementary Figure 4. The plotted lines are for four methods, miRNA DE (black), cMimDE (green), pMimDE (blue) and pMimCor (red). This is performed on four datasets (**a**) Notch, (**b**) Melanoma, (**c**) Ovarian and (**d**) Lung

for any FP it generally has a higher TP, i.e. its curve is higher than the others. In the Melanoma dataset all methods appear to perform comparably. In the remaining three datasets pMimCor performs most favourably with a simple miRNA DE test generally performing worst. The method pMimCor appears to be identifying more miRNA that have been associated each of the datasets in the literature than the other methods.

The prognostic strategy is used to compare the concordance of the miRNA identified by each of the methods and their corresponding target genes. If a miRNA has been identified as a prognostic marker its target genes should also be able to predict prognosis. Figure 3 shows boxplots of the classification error rates of these target genes for each of the cancer datasets. These error rates were calculated after training a DLDA on the 10 most DE targeted genes of each method. The choice of this number is shown to be not too influential on the results in Supplementary Figure S5. In all three datasets cMimDE, pMimDE and pMimCor deliver error rates less than 0.5. In the Melanoma dataset all miRNA based methods perform similarly, with simply taking the 10 most DE genes performing dramatically worse. In Lung all the integration based methods; cMimDE, pMimDE and pMimCor; perform favourably to the other two approaches who have error rates of 50%. Both pMimDE and pMimCor appear most preferable in the Ovarian dataset. Supplementary Figure S5 demonstrates that these results are relatively robust to the choice of the number of genes under consideration. Unsurprisingly, methods that integrate both miRNA and gene data show greater concordance in behaviour between their identified miRNA and genes.

Our proposed methods pMiMDE and pMimCor rank both miRNA and pathways concurrently which can improve biological interpretation. To illustrate this Table 2 lists the top 10 mir-pathways identified by pMimCor in the Notch dataset. In this dataset, *Notch2* has been conditionally deleted in the brain of mice using a nestin-cre driver. Cortical samples were isolated from matched mutant and control adult mice. The affected mir-pathways are consistent with known *Notch* function in the nervous system. *Notch* signalling has been associated with ALS (Praline *et al.*, 2010) and Alzheimers disease (Chávez-Gutiérrez *et al.*, 2012), and loss of *Notch2* leads to neurodegeneration in the olfactory system (Rodriguez *et al.*, 2008). Furthermore, *Notch* signalling is known to play important roles in axon guidance, synaptic plasticity and learning and memory (Alberi *et al.*, 2013). *Notch* signalling also possesses significant cross talk with other signalling pathways, including the *Jak-STAT* pathway (Kamakura *et al.*, 2004). Finally, *Notch* signalling regulates growth and differentiation, and as such is implicated in many cancers (Balint *et al.*, 2005; Santagata *et al.*, 2004; Wang *et al.*, 2006). In keeping with this, *Notch* is a key regulator of neurogenesis in the brain (Imayoshi *et al.*, 2010). The identified mir-pathways are therefore consistent with the existing literature, and provide a biological context for understanding how misregulation of miRNA in *Notch2* mutants can affect neuronal function.

4 Discussion

The key strength of pMimDE and pMimCor is to rank both miRNA and pathways concurrently, i.e. rank the top mir-pathways. As such, these may be considered favourable approaches for integration as they facilitate interpretation and directly address the biological question. However, quantitatively assessing the performance of these rankings is not trivial in the absence of the known truth. As such, we have assessed the performance of the proposed methods via three evaluation techniques. While individually these techniques are suggestive at best, concurrently they provide a more complete picture of each methods performance.

Our proposed method has recently been applied to an experiment using microarrays to investigate the associations between miRNA and gene expression with BRAF mutations and patient prognosis in stage III metastatic melanoma (Tembe *et al.*, 2014). In comparison to other state of the art methods used in the analysis, our approach provided insights that could not be easily made using the other approaches. It has been our experience that the joint ranking of miRNA and pathways provided by pMim makes the information from studying miRNA-mRNA interactions more easily accessible and interpretable.

The Notch dataset is previously unpublished and has a small sample size yet still identifies signal concordant with the predicted phenotype. This is promising in two ways. Firstly, it demonstrates the benefits of using a predefined network structure as opposed to one derived empirically. Deriving a large complex network using six samples would generally be ill advised. However, using the predefined mir-pathways, pMimCor is able to identify signal, demonstrating that small samples experiments can still be informative even when calculating correlations on only six samples. Secondly, as this dataset is unpublished and still identifies signal concordant with the literature, it protects against publication bias in the evaluation approach.

In both the literature search strategy and prognostic strategy the integration approaches; cMimDE, pMimDE and pMimCor; generally perform better than using only the gene or miRNA data. This potentially demonstrates the value of integrating data to ask complex questions. Alternatively, there are arguments why both of the evaluation strategies should be biased towards the integration approaches and it is reassuring to observe the results concordant with these expected biases. In the literature search strategy a miRNA may be more likely to have been associated with a condition in the literature if that miRNA and its targets have been observed as changed. Whilst in the prognostic strategy, as the miRNA for each method are chosen outside of cross-validation performance could be biased towards those methods that use gene information.

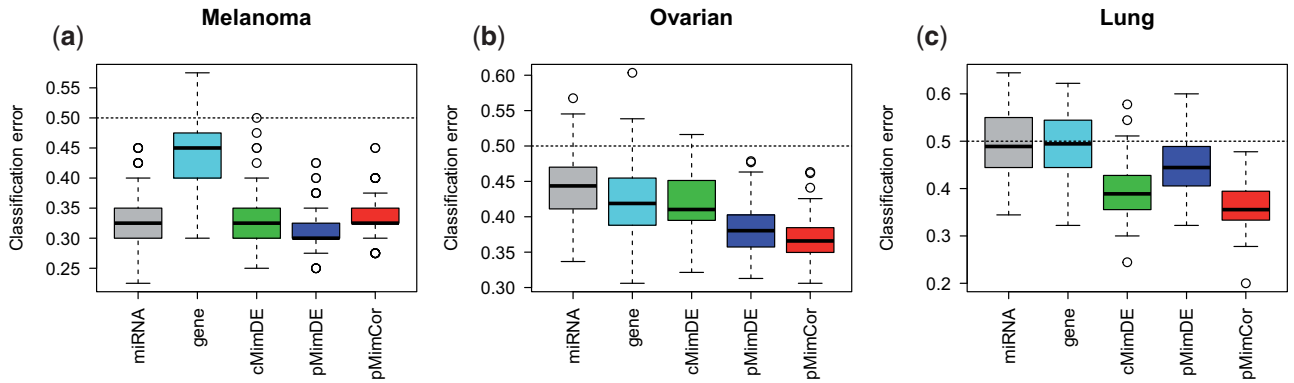

Fig. 3. Classification error rates for prognosis. The classification error rates from 100 replicates of 5-fold cross-validation performed on a LDA classifier built with 20 genes is shown. The 20 genes are select as the targets from the three most significant miRNA from a moderated t-test (miRNA), cMimDE, pMimDE and pMimCor. Also shown are the error rates from simply the 20 most DE genes (gene). These error rates are calculated on three datasets (**a**) Melanoma, (**b**) Ovarian and (**c**) Lung

Table 2. Table of pMimCor output

miRNA	Pathways	Direction	Score
mir-342	Apoptosis, Toxoplasmosis, Small cell lung cancer	Up	0.0006
mir-340	Vascular smooth muscle contraction	Down	0.0013
mir-342	Jak-STAT signalling pathway, Pathways in cancer	Up	0.0017
mir-497	Axon guidance	Up	0.0024
mir-760	Jak-STAT signalling pathway, Amyotrophic lateral sclerosis (ALS), Pancreatic cancer, Chronic myeloid leukemia	Up	0.0032
mir-497	Jak-STAT signalling pathway	Up	0.0035
mir-497	Small cell lung cancer, non-small cell lung cancer	Up	0.0038
mir-96	Long-term potentiation	Down	0.0052
mir-96	Amyotrophic lateral sclerosis (ALS)	Down	0.0054
mir-497	T cell receptor signalling pathway, B cell receptor signalling pathway	Up	0.0063

A table of the top 10 mir-pathways identified by pMimCor in the Notch dataset. The table consists of four columns; the miRNA that was identified as changed, the direction the miRNA expression changed, a significance score of the mir-pathway and the pathway contained the target genes.

The general framework described in this paper has some natural extensions. It is possible to extend pMim to cover experiments with multiple conditions by combining the p-value from testing the equality of group means (using a one-way ANOVA) of the miRNA expression with the correlations between the miRNA and the genes in a mir-pathway. Also, many significance combination approaches can be modified to include weights including Stouffer's method (Liptak, 1958). Weights could be included to increase the impact of genes that are thought to be highly influential on the biology in question.

Our integration approach has raised some statistically interesting questions that should warrant further investigation. To transcend the typical modularized analysis framework, our approaches make use of *P*-value combination methods to combine significance. By combining significance from a DE test of one miRNA with tests on multiple gene sets, pMim suffers from an issue associated with correcting for multiple comparisons. While there has been some research into correction for multiple comparisons in the presence of correlated pathways (Holmans *et al.*, 2009), this methodology could be extended further to account for having correlated summary statistics from multiple miRNA. It would also be interesting to establish whether it is more appropriate to correct for multiple testing before or after utilising a *P*-value combination method like Fisher or OSP.

In addition to considering how individual miRNA regulate a set of gene, miRNA are also known to act in groups to regulate genes. One broad group of statistical methods aims to identify groups of genes that are potentially being regulated by a miRNA or group of miRNA. Such methods include canonical correlation analysis (Witten and Tibshirani, 2009), nonnegative matrix factorization (Zhang *et al.*, 2011), multivariate random forests (Jayaswal *et al.*, 2011) and integrative Bayesian analysis (Wang *et al.*, 2013). Statistically these methods can be thought of as ways of identifying relationships between a high dimensional multivariate response and high dimensional multivariate covariates. Unfortunately, as these methods are attempting to identify large complex networks their results are very dependent on sample size. As with the simpler univariate tests, the interpretation of the outputs from these methods can quickly become intimidating due to the many pathway analyses being performed on multiple lists of results. For datasets with sufficient sample size, our framework may be used as a first pass of analysis before implementing more sophisticated or intricate approaches.

We can easily incorporate other functional databases into the pMim framework. The framework is not restricted to any specific databases such as the TargetScan and KEGG databases that were used as the predicted binding target and pathway databases respectively in this paper. Depending on the question of interest and the desired interpretation, scientists could easily substitute the pathway database with other functional relevant databases such as those on GSEA (Subramanian *et al.*, 2005) or DAVID (Dennis *et al.*, 2003). The resampling scheme outlined in the evaluation, as well as a further strategy outlined in Section 3 of the supplementary file, could both be used to assess the viability of any alternate predicted binding target or functional database. We have included an evaluation of the

information in the protein-protein interaction database iRef (Razick *et al*., 2008) in Section 5 of the supplementary file.

5 Conclusion

We described a general framework, pMim, for integrating miRNA expression, gene expression and annotation information. From this framework three methods were proposed, cMimDE, pMimDE and pMimCor. The method cMimDE is a formalization and mild improvement of previously described approaches for identifying regulatory miRNA. The methods pMimDE and pMimCor are both novel methods for identifying if a miRNA is regulating a set of genes that also share some common biological function. Each of these methods relies on a predefined network structure derived from miRNA binding target predictions and/or biological pathway annotations making them suitable for use on experiments with limited replication. The proposed pMimDE and pMimCor are the first methods to jointly rank miRNA and pathways in relation to a condition of interest which greatly facilitates biological interpretation.

Acknowledgement

We thank the reviewers for their useful contributions.

Funding

This work was supported in part by ARC through grants FT0991918 (Y.Y.) and DP130100488 (S.M.,Y.Y.), Australian Postgraduate Award (E.P.) and NIH 5R21AG033241 (D.L.).

Conflict of Interest: none declared.

References

Alberi,L. *et al*. (2013) Notch signaling in the brain: in good and bad times. *Ageing Res. Rev.,* **12**, 801–814.

Balint,K. *et al*. (2005) Activation of notch1 signaling is required for beta-catenin-mediated human primary melanoma progression. *J. Clin. Invest.,* **115**, 3166–3176.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B,* **57**, 289–300.

Carlson,M. (2014) *KEGG.db: A set of annotation maps for KEGG*. R package version 2.14.0.

Chávez-Gutiérrez,L. *et al*. (2012) The mechanism of γ-secretase dysfunction in familial alzheimer disease. *EMBO J.,* **31**, 2261–2274.

Dennis,G. Jr, *et al*. (2003) David: database for annotation, visualization, and integrated discovery. *Genome Biol.,* **4**, P3.

Griffiths-Jones,S. *et al*. (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Res.,* **36**, D154–D158.

Havelange,V. *et al*. (2011) Functional implications of microRNAs in acute myeloid leukemia by integrating microRNA and messenger RNA expression profiling. *Cancer,* **117**, 4696–4706.

Holmans,P. *et al*. (2009) Gene ontology analysis of gwa study data sets provides insights into the biology of bipolar disorder. *Am. J. Hum. Genet.,* **85**, 13–24.

Imayoshi,I. *et al*. (2010) Essential roles of notch signaling in maintenance of neural stem cells in developing and adult brains. *J. Neurosci.,* **30**, 3489–3498.

Jayaswal,V. *et al*. (2009) Identification of microRNAs with regulatory potential using a matched microRNA-mRNA time-course data. *Nucleic Acids Res.,* **37**, e60.

Jayaswal,V. *et al*. (2011) Identification of microRNA-mRNA modules using microarray data. *BMC Genomics,* **12**, 138.

Kamakura,S. *et al*. (2004) Hes binding to STAT3 mediates crosstalk between Notch and JAK-STAT signalling. *Nat. Cell Biol.,* **6**, 547–554.

Kanehisa,M. and Goto,S. (2000) Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.,* **28**, 27–30.

Krek,A. *et al*. (2005) Combinatorial microRNA target predictions. *Nat. Genet.,* **37**, 495–500.

Lewis,B.P. *et al*. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell,* **120**, 15–20.

Li,X. *et al*. (2011) Comparative mRNA and microRNA expression profiling of three genitourinary cancers reveals common hallmarks and cancer-specific molecular events. *PLoS One,* **6**, e22570.

Liptak,T. (1958) On the combination of independent tests. *Magyar Tud Akad Mat Kutato Int Kozl,* **3**, 171–197.

Maragkakis,M. *et al*. (2009) Diana-microt web server: elucidating microrna functions through target prediction. *Nucleic A,* **37**, W273–W276.

Mo,Y.-Y. (2012) MicroRNA regulatory networks and human disease. *Cell Mol. Life Sci.,* **69**, 3529–3531.

Nam,S. *et al*. (2009) Microrna and mrna integrated analysis (mmia): a web tool for examining biological functions of microrna expression. *Nucleic Acids Res.,* **37**, W356–W362.

Pearson,K. (1934) On a new method of determining" goodness of fit". *Biometrika,* **26**, 425–442.

Praline,J. *et al*. (2010) CADASIL and ALS: a link?. *Amyotroph Lateral Scler,* **11**399–401.

Razick,S. *et al*. (2008) irefindex: a consolidated protein interaction database with provenance. *BMC Bioinformatics,* **9**, 405.

Rodriguez,S. *et al*. (2008) Notch2 is required for maintaining sustentacular cell function in the adult mouse main olfactory epithelium. *Dev. Biol.,* **314**, 40–58.

Ru,Y. *et al*. (2014) The multimir r package and database: integration of microrna-target interactions along with their disease and drug associations. *Nucleic Acids Res.*, **42**, e133.

Santagata,S. *et al*. (2004) JAGGED1 expression is associated with prostate cancer metastasis and recurrence. *Cancer Res.,* **64**, 6854–6857.

Stouffer,S. *et al*. (1949) The American Soldier, Adjustment during Army Life. Vol.1, Princeton University Press, Princeton.

Subramanian,A. *et al*. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA,* **102**, 15545–15550.

Tembe,V. *et al*. (2014) Microrna and mrna expression profiling in metastatic melanoma reveal associations with braf mutation and patient prognosis. *Pigment Cell Melanoma Res.*, **28**, 254–266.

Wang,W. *et al*. (2013) iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics,* **29**, 149–159.

Wang,X. (2008) mirdb: a microrna target prediction and functional annotation database with a wiki interface. *RNA,* **14**, 1012–1017.

Wang,Z. *et al*. (2006) Notch-1 down-regulation by curcumin is associated with the inhibition of cell growth and the induction of apoptosis in pancreatic cancer cells. *Cancer,* **106**, 2503–2513.

Witten,D.M. and Tibshirani,R.J. (2009) Extensions of sparse canonical correlation analysis with applications to genomic data. *Stat. Appl. Genet. Mol. Biol.,* **8**, Article28.

Xu,J. and Wong,C.-W. (2013) Enrichment analysis of miRNA targets. *Methods Mol. Biol.,* **936**, 91–103.

Yang,P. *et al*. (2014) Direction pathway analysis of large-scale proteomics data reveals novel features of the insulin action pathway. *Bioinformatics,* **30**, 808–814.

Zhang,S. *et al*. (2011) A novel computational framework for simultaneous integration of multiple types of genomic data to identify microrna-gene regulatory modules. *Bioinformatics,* **27**, i401–i409.