miRNApath Pathway Enrichment for miRNA Expression Data

James M. Ward*, Yunling Shi, John P. Cogswell, Cindy Richards

October 13, 2014

Contents

1 Introduction

Gene set enrichment has been well-studied in the context of differential gene expression, but has not been evaluated to a great extent yet for microRNAs (miRNAs.) The method used here by Cogswell et al uses perhaps the classical gene set enrichment paradigm, hypergeometric enrichment. The fundamental change in the algorithm was to include some provision for the many-to-many situation where multiple miRNAs may be predicted to target a single gene, or one miRNA is predicted to target multiple genes. We took

the basic approach to represent this paradigm shift by simply representing each individual miRNA-gene prediction. Although expression levels of the predicted gene transcripts and absolute expression levels of the differentially expressed miRNAs could potentially be used to fine-tune the algorithm, we found utility in treating each prediction event equally, then surveying the results. An exception to this statement was that multiple miRNA binding events to a single gene were treated once - this aspect of the algorithm may be the first to evolve due to potential synergistic effects of multiple miRNA binding sites.

Pathway effects of miRNAs have not been sufficiently validated on a global scale to support any statements of fact, so we use the phrase "calculated signal" to indicate the mathematical likelihood of a pathway being affected by miRNA changes. Features we observed which built confidence were: observations that in general most pathways were unidirectionally regulated; many of the hits displayed multiple miRNAs predicted to affect the same gene, and multiple genes affected in the pathway (i.e. not just one miRNA and a host of predictions); pathway predictions were most often represented across gene and miRNA families, lessening the potential bias of sequence homology or identity; as the miRNA prediction P-value was made less restrictive, signal was quickly lost altogether, presumably either due to the overwhelming universe size or the noise far outweighing the signal. Nonetheless, validated miRNA targets have been demonstrated at the lenient P-value range, though we contend far fewer than would be seen at the strict P-value range.

In general, the observation was that certain predictions tended to "reinforce" the calculated signal, which may suggest that these predictions are more biologically relevant. We found that using Miranda and RnaHybrid separately, then using just the overlapping predictions, the results tended to be the same with few exceptions. These methods are closely related to one another, so this observation is biased, however this method is made available specifically to enable evaluations and enhancements over time as our scientific understanding of miRNAs advances.

This report has been produced as a reproducible document. It contains the actual computer instructions for the method it describes, and these in turn produce all results, including the figures and tables that are shown here. The computer instructions are given in the language R, thus, in order to reproduce the computations shown here, you will need an installation of R (version 2.3 or greater) together with a recent version of the package *miRNApath* and of some other add-on packages.

First, we load the package.

```
> library("miRNApath")
```

2 Requirements for using miRNApath

This vignette assumes you have available a tab-delimited text file with differential miRNA expression data along with some representation of the miRNAs included in the experiment overall. Another text file should be supplied with miRNA-gene predictions. Finally, a text file containing gene-path links should be supplied. These files are kept separate since combining them quickly yields prohibitively large datasets, well beyond the R and operating system memory limits.

3 Loading miRNA expression data

The first basic step in the workflow is to load data into the "mirnapath" object. The method assumes you have a tab-delimited file with miRNA names in one column, and assay ID values in another column. The assay ID values are used to discriminate multiple miRNA assays for the same miRNA. If this situation does not apply, simply re-use the column for miRNA names in the function argument below.

This vignette begins with data contained within the package, and is written to a file simply to provide an example of the data format, and to model the end-user workflow.

```
> ## Start with miRNA data from this package
> data(mirnaobj);
> ## Write a file as example of required input
> write.table(mirnaobj@mirnaTable, file="mirnaTable.txt",
+     quote=FALSE, row.names=FALSE, col.names=TRUE, na="",
+     sep="\t");
> ## Now essentially load it back, but assign column headers
> mirnaobj <- loadmirnapath( mirnafile="mirnaTable.txt",
+     pvaluecol="P-value", groupcol="GROUP",
+     mirnacol="miRNA Name", assayidcol="ASSAYID" );
> ## Display summary information for the resulting object
> mirnaobj;

mirnapath object:
   Length     Class      Mode
```

```
        1 mirnapath        S4

Columns specified:
   mirnacol = "miRNA Name"
   assayidcol = "ASSAYID"
   groupcol = "GROUP"
   filterflagcol = "FILTERFLAG"

Filters Applied:
   none

Number of miRNAs: 196
Number of sample groups: 18
Number of pathways: NA
State: filtered
```

4 Filtering miRNA data for hits

At this point the data only contains miRNA differential expression results, with no delineation of hits versus non-hits to be used during the enrichment technique. The first step therefore is to describe hits in statistical terms, and in this dataset the relevant statistic is the P-value. We define the P-value column header in the actual dataset:

```
> mirnaobj@columns["pvaluecol"] <- "P-value";
```

The miRNApath package currently recognizes P-value (pvaluecol), fold change (foldchangecol), and expression abundance (expressioncol) as allowable filter criteria. Multiple filter criteria can be applied in one function call, or in a series of calls, both to the same result: additive constraints. Thus, all criteria must be fulfilled in order for an entry to be defined as a hit.

We now filter the data using the P-value cutoff 0.05:

```
> mirnaobj <- filtermirnapath( mirnaobj, pvalue=0.05,
+     expression=NA, foldchange=NA );
```

5 Loading miRNA-gene associations

The next step is to load miRNA-gene associations, typically predicted miRNA gene targets (e.g. from miRBase, RNAHybrid, or T-Scan to name a few.)

Again, this package assumes the availability of a tab-delimited file containing miRNA names and gene names. Both sets of names must match identically to those contained in the miRNA data (former), and subsequent gene-pathway data (latter.)

```
> ## Load the miRNA to gene associations
> mirnaobj <- loadmirnatogene( mirnafile="mirnaGene.txt",
+     mirnaobj=mirnaobj, mirnacol="miRNA Name",
+     genecol="Entrez Gene ID",
+     columns=c(assayidcol="ASSAYID") );
> ## Display summary, noting the miRNA-gene predictions
> mirnaobj;

mirnapath object:
   Length     Class      Mode
        1 mirnapath        S4

Columns specified:
   mirnacol = "miRNA Name"
   assayidcol = "ASSAYID"
   groupcol = "GROUP"
   filterflagcol = "FILTERFLAG"
   mirnagene = "miRNA-Gene"
   genecol = "Entrez Gene ID"
   pathwaycol = "Pathway Name"
   pathwayidcol = "PATHWAY_ID"
   pvaluecol = "P-value"

Filters Applied:
   none

Number of miRNAs: 196
Number of sample groups: 18
Number of pathways: 0
State: filtered
```

The mirnapath object now contains filtered data, and miRNA-gene associations.

6 Loading gene-pathway associations

The next step is to load gene-pathway associations, completely analogous to the miRNA-gene step previously.

```
> ## Load the gene to pathway associations
> mirnaobj <- loadmirnapathways( mirnaobj=mirnaobj,
+     pathwayfile="mirnaPathways.txt",
+     pathwaycol="Pathway Name", genecol="Entrez Gene ID");
> ## Display summary, noting the number of pathways reported
> mirnaobj;

mirnapath object:
   Length     Class      Mode
        1 mirnapath        S4

Columns specified:
   mirnacol = "miRNA Name"
   assayidcol = "ASSAYID"
   groupcol = "GROUP"
   filterflagcol = "FILTERFLAG"
   mirnagene = "miRNA-Gene"
   genecol = "Entrez Gene ID"
   pathwaycol = "Pathway Name"
   pathwayidcol = "PATHWAY_ID"
   pvaluecol = "P-value"

Filters Applied:
   none

Number of miRNAs: 196
Number of sample groups: 18
Number of pathways: 771
State: filtered
```

The data is prepared for pathway enrichment once the miRNA data is loaded and appropriate column headers defined, the data is filtered to describe hits and non-hits, the miRNA-gene and gene-pathway associations are also loaded. We now run pathway enrichment:

7 Running miRNA pathway enrichment

```
> Groups = unique(mirnaobj@mirnaTable[,
+         mirnaobj@columns["groupcol"] ]);
> mirnaobj <- runEnrichment( mirnaobj=mirnaobj, Composite=TRUE,
+    groups=Groups[grep("^AD.+(UP|DOWN)",Groups)], permutations=0 );
```

The composite flag indicates whether to treat the fully expanded miRNA-gene combinations as separate enrichment events (TRUE), or whether to treat all effects on one gene as one collective event. The latter case reverts to the classic un-ordered hypergeometric enrichment technique. However the expansion of combinations is the current method chosen to represent the multiple predicted effects of miRNAs to one gene, and the predicted effect of one miRNA to multiple genes. The algorithm will identify statistically significantly enriched results when the combination of these effects is greater than would be anticipated by random chance.

To provide some added confidence to the P-value predictions, a random permutation algorithm has been provided. This permutation algorithm randomizes miRNA hits, as opposed to randomizing all possible miRNA-gene combinations, since we're making the assumption that the predictions although imperfect at least provide a stable reference frame for the technique. Each miRNA is typically predicted to affect 20 to potentially hundreds of genes. Therefore one hopes that a coordinated effect seen across a suite of changing miRNAs is quite unlikely to occur by random chance. The permutation algorithm tests exactly that theory, and thus far tends to strengthen the statistically significant P-value results while weakening most of the already weak P-values, as compared to the P-values from the hypergeometric test.

The final step is to export a tabular summary of results. There are many possible ways of formatting and visualizing the results, and more will be included in this package over time as they evolve. The simplest approach is to list the results with as much content as can be used to reformat any number of ways outside this package:

8 Exporting miRNA enrichment results

```
> finaltable <- mirnaTable( mirnaobj, groups=NULL, format="Tall",
+    Significance=0.1, pvalueTypes=c("pvalues","permpvalues"),
+    maxStringLength=42 );
```

```
> ## Display only the first few rows of the best P-values...
> finaltable[sort(finaltable[,"pvalues"], index.return=TRUE)$ix,][1:5,];

          pvalues Measured pathway mirnaGenes Enriched pathway mirnaGenes
2651 1.035112e-03                          24                         13
4831 1.035112e-03                          24                         13
210  1.075882e-03                          14                         10
164  1.128459e-02                          44                         15
201  1.129498e-03                          12                          9
     Genes Enriched miRNAs Enriched Total mirnaGenes Total filtered mirnaGenes
2651                 5                5            6347                       1481
4831                 5                5            6347                       1481
210                  3                7            6231                       1779
164                  9               10            6253                       1173
201                  4                6            6231                       1779
                              Group PATHWAY_ID Pathway Name
2651 AD Hippo vs control Hippo, UP        265         <NA>
4831 AD Hippo vs control Hippo, UP        483         <NA>
210  AD Cereb vs control Cereb, UP        210         <NA>
164    AD MFG vs control MFG, DOWN        164         <NA>
201  AD Cereb vs control Cereb, UP        201         <NA>
```

The Significance is the threshold of P-values for hypergeometric enrichment. We find that fairly inclusive cutoffs allows greater comparisons across the data, specifically providing a sufficient backdrop of negative results to contrast the positive results. When multiple sample groups are provided, a heatmap of enrichment across the groups has shown to be a rapid way of comparing and contrasting group effects. The results serve two purposes: comparisons may show that results are non-random, while differences can potentially describe functional differences across the groups. The P-values can help describe significance, especially using the permutation P-value which randomizes the significant miRNAs prior to running the algorithm 'n' times.

Once the enrichment data is available, there are numerous ways to visualize or further analyze the results. Some straightforward examples follow:

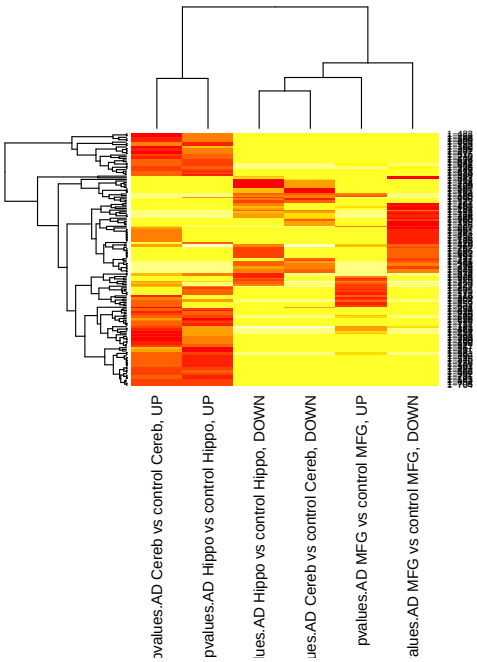
9 Displaying a heatmap summary

The heatmap display seems particularly useful for comparing changes across a fairly large number of sample groups. It is useful to check consistency in

pathways enriched across groups with respect to fold changes, which is potentially one way of discriminating random effects. The code below extracts numeric P-values from the miRNApath object, subsets the data for pathways enriched in 5 or more sample groups, restricts the display to only the UP and DOWN changes, then displays a heatmap. Any of the steps can be altered to suit the experiment.

In this case, the pathway data has been filtered to remove any pathways that cannot be published, and therefore the Significance term is more lenient than normal. However, a value of 0.1 or even higher (e.g. 0.3 as below) is not undesirable mainly because it allows greater ability to view results across sample groups. When interpreting the data below, the recommendation is generally to trust statistically significant P-value as normal (e.g. potentially 0.05 or below) and use the other shadings to help evaluate possible cross-group effects. One could observe that very few pathways appear to be shared across fold change direction. Without interpreting the meaning of UP and DOWN, the data may at least suggest which pathways are activated, and further suggest that the pathways activated are generally affected in one consistent fold change direction.

```
> ## Example which calls heatmap function on the resulting data
> widetable <- mirnaTable( mirnaobj, groups=NULL, format="Wide",
+     Significance=0.3, na.char=NA, pvalueTypes=c("pvalues") );
> ## Assign 1 to NA values, assuming they're all equally
> ## non-significant
> widetable[is.na(widetable)] <- 1;
> ## Display a heatmap of the result across sample groups
> pathwaycol <- mirnaobj@columns["pathwaycol"];
> pathwayidcol <- mirnaobj@columns["pathwayidcol"];
> rownames(widetable) <- apply(widetable[,c(pathwaycol,
+    pathwayidcol)], 1, function(i)paste(i, collapse="-"));
> wt <- as.matrix(widetable[3:dim(widetable)[2]], mode="numeric");
> pathwaySubset = apply(wt, 1, function(i)length(i[i<0.2])>1)
> ## Print out a heatmap
> par(ps="8");
> heatmap(log2(wt[pathwaySubset,]), scale="row",
+   cexRow=0.9, margins=c(15,12));
>
```


pvalues.AD Cereb vs control Cereb, UP
pvalues.AD Hippo vs control Hippo, UP
ues.AD Hippo vs control Hippo, DOWN
ues.AD Cereb vs control Cereb, DOWN
pvalues.AD MFG vs control MFG, UP
alues.AD MFG vs control MFG, DOWN