Infer miRNA-mRNA interactions using paired expression data from a single sample

Yue Li
yueli@cs.toronto.edu

April 12, 2014

1 Introduction

MicroRNAs (miRNAs) are small ($\sim$ 22 nucleotides) RNA molecules that base-pair with mRNA primarily at the 3′ untranslated region (UTR) to cause mRNA degradation or translational repression (Bartel [2009]). The expression levels of miRNA and mRNA are usually measured by microarray or RNA-seq. Paried expression profiling of both miRNA and mRNA enables identifying miRNA-mRNA interactions within an individual and calls for a new computational method. We develop *Roleswitch* to infer Probabilities of MiRNA-mRNA Interaction Signature (ProMISe) using paired expression data from a single sample (paper in preparation). Roleswitch takes as inputs two expression vectors of N mRNAs and M miRNAs and a $N \times M$ seed match matrix containing the number of target sites for each mRNA i and miRNA k. The program then outputs a probability matrix that mRNA i (row) being a target of miRNA k (column). User can provide `roleswitch()` with a seed-match matrix. Otherwise, `getSeedMatrix` will be invoked to retrieve seed-match matrix from existing (online) database.

Briefly, Roleswitch operates in two phases by inferring the probabilities of mRNA (miRNA) being the targets ("targets") of miRNA (mRNA), taking into account the expression of all of the mRNAs (miRNAs) due to their potential competition for the same miRNA (mRNA). Due to mRNA transcription and miRNA repression events simutaneously happening in the cell, Roleswitch assumes that the total transcribed mRNA levels are higher than the observed (equilibrium) mRNA levels and iteratively updates the total transcription of each mRNA targets based on the above inference. Based on our extensive tests on cancer data from TCGA, Roleswitch identifies more validated targets comparing with existing methods (paper in preparation). Additionally, the inferred ProMISe rivals expression profiles in cancer diagnosis yet provides unique opportunities to explore oncogenic mRNA-miRNA interactions (paper in preparation). The algorithm is outlined as follow:

1. Infer mRNA i targeted by miRNA k taking into account the hidden total expression of 1 . . . N mRNA and miRNA k

2. Estimate total transcription level of mRNA i

3. Infer miRNA k "targeted" by mRNA i taking into account 1 ... M miRNA and mRNA i expression

4. Repeat 1-3 until convergence

2 Simulation

To help appreciate the model, this section demonstrates a toy example using simulated data of 10 mRNAs and 4 miRNAs. Specifically, we generated expression of 10 mRNAs and 4 miRNAs from Gaussian distribution using `rnorm` with mean and standard deviation set to 3 and 1, respectively. The 10×4 seed matrix were generated (using `rpois`) from a Poisson distribution with $\lambda = 0.2$.

```
> library(Roleswitch)
> # simulated example
> N <- 10
> M <- 4
> x.o <- matrix(abs(rnorm(N, mean=3)))
> rownames(x.o) <- paste("mRNA", 1:nrow(x.o))
> colnames(x.o) <- "mRNA expression"
> # miRNA expression
> z.o <- matrix(abs(rnorm(M, mean=3)))
> rownames(z.o) <- paste("miRNA", 1:nrow(z.o))
> colnames(z.o) <- "miRNA expression"
> # simulate target sites
> c <- matrix(rpois(nrow(z.o)*nrow(x.o), 0.2), nrow=nrow(x.o))
> # ensure each miRNA (mRNA) has at least one
> # seed (seed match) to a mRNA (miRNA)
> c[apply(c,1,sum)==0, sample(1:ncol(c),1)] <- 1
> c[sample(1:nrow(c),1),apply(c,2,sum)==0] <- 1
> dimnames(c) <- list(rownames(x.o), rownames(z.o))
> # simulate true labels
> rs.pred <- roleswitch(x.o, z.o, c)
```

As shown in Fig 1, the top panels (left to right) display the observed mRNA and miRNA expression, seed-match matrix and inferred total mRNA expression (A-D); the bottom panels display the probability matrix of miRNA-mRNA (i.e. miRNA targeting mRNA), mRNA-miRNA (i.e. mRNA "targeting" miRNA), the dot product of the above two matrices, and the convergence rate (E-H). Such simple example is sufficient to highlight several important features of the proposed model. First and most obviously, mRNA that does not carry a seed match for miRNA has zero probability of being a target of that miRNA, regardless the expression levels, and vice versa (Fig 1C,E-G). Second, $p(t_{i,k}^{(x)}|\mathbf{x}^{(t)}, z_k, \mathbf{c}_{.,k})$ (miRNA-mRNA) (Fig 1E) and $p(t_{i,k}^{(z)}|x_i^{(t)}, \mathbf{z}, \mathbf{c}_{i,.})$ (mRNA-miRNA) (Fig 1F) differ in many cases for the same pair of miRNA and mRNA. Third, the joint probabilities (ProMISe, Fig 1G) reflect both aspects of the targeting mechanisms and can differentiate many cases, where the probabilities are equal in either $p(t_{i,k}^{(x)}|.)$ (Fig 1E) or $p(t_{i,k}^{(z)}|.)$ (Fig 1F). Finally, $p(t^{(x)}|.)$ (or $p(t^{(z)}|.)$) converges quickly in only a few iterations (Fig 1H). The same

```
> diagnosticPlot(rs.pred)
```

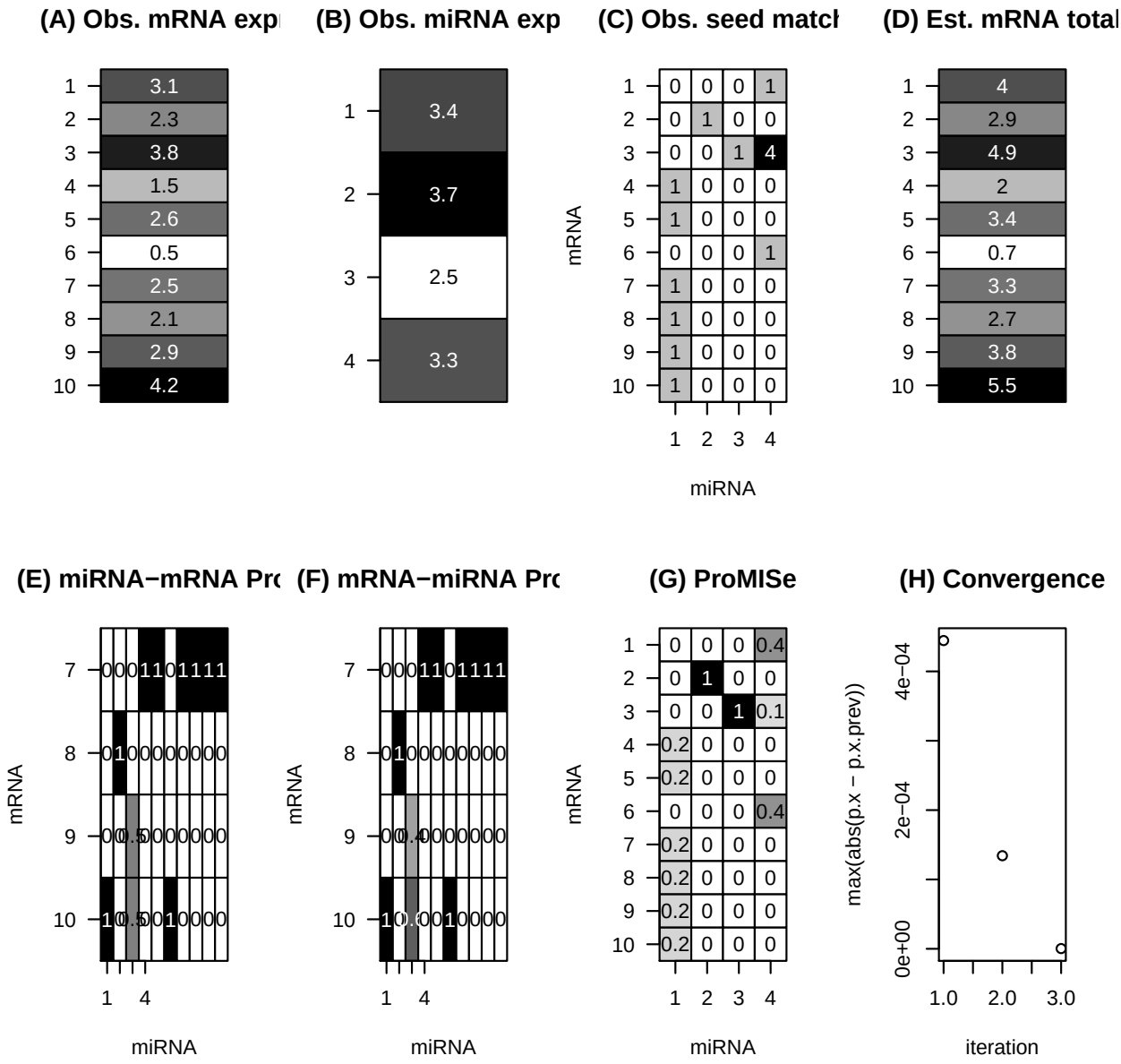

Figure 1: From left to right, the top panel displays (A) the 10 simulated mRNA expression, (B) 4 miRNA expression, (C) the 10×4 seed-match matrix, and (D) the inferred total mRNA expression by the proposed model; the bottom panel displays (E) the inferred probabilities of the 4 miRNAs targeting the 10 mRNA (miRNA-mRNA), (F) the probabilities of the 10 mRNA "targeting" the 4 miRNAs (mRNA-miRNA), (G) the dot product of the above two matrices (which is defined as ProMISe), and (H) the convergence rate. The coloured rectangles on the particular interaction scores help explain the properties of model in the main text.

holds true for practically large number of mRNAs and miRNAs: the model converges within 10 iterations at $tol = 10^{-5}$.

3 Real test

In this section, we demonstrate the real utility of *Roleswitch* in construct ProMISe from a single sample. The test data of miRNA and mRNA expression for the same individual (barcode ID: TCGA-02-0001-01) were downloaded from TCGA GBM. We linear transformed the data to to the non-negative scale since negative expression will produce unexpected results.

```
> data(tcga_gbm_testdata)
> # rescale to non-negative values (if any)
> if(any(x<0)) x <- rescale(as.matrix(x), to=c(0, max(x)))
> if(any(z<0)) z <- rescale(as.matrix(z), to=c(0, max(z)))
```

Next, we obtain seed-match matrix using `getSeedMatrix` from pre-compiled and processed human target site information `hsTargets` saved in the *microRNA* package, originally downloaded from Microcosm (`http://www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5/`) Griffiths-Jones et al. [2008]. For each mRNA-miRNA pair, we calculated the number of corresponding target sites. For multiple transcripts of the same gene, we used transcripts with the longest 3′UTR. The end result is a $N \times M$ seed-match matrix of N distinct mRNAs each corresponding to a distinct gene and M distinct miRNAs. The execution of the follwing code requires *microRNA* package.

```
> seedMatrix <- getSeedMatrix(species="human")
> seedMatrix <- seedMatrix[match(rownames(x),
+            rownames(seedMatrix), nomatch=F),
+                  match(rownames(z), colnames(seedMatrix), nomatch=F)]
> x <- x[match(rownames(seedMatrix),rownames(x), nomatch=F),,drop=F]
> z <- z[match(colnames(seedMatrix),rownames(z), nomatch=F),,drop=F]
```

We now apply Roleswitch to the test data:

```
> rs.pred <- roleswitch(x,z,seedMatrix)
```

To demonstrate the quality of the prediction, we compare the Roleswitch predicted ProMISe with using seed-match matrix alone. Among the 100-1000 rank with 100-interval from each method, we counted validated targets downloaded from mirTarBase (Hsu et al. [2011]) (http://mirtarbase.mbc.nctu For Roleswitch, we use the joint probability matrix as its final inference of ProMISe. For seed-match matrix, which does not consider expression data, the mRNA-miRNA interaction was simply ranked by the corresponding total number of target sites - the more target sites a mRNA i has for miRNA k, the more likely it is the miRNA target. To ascertain functional interaction, we restrict the valdiated miRNAs to validated miRNAs that has nonzero expression values in the sample since miRNAs that do not express at all will not have any target in that sample (evan though they may be validated in some other cell-line or tissues).

We first need to process the validated targets to make it having the same order of `dimnames` as the `seedMatrix`:

```
> # reorder valdiated taregts to match with seedMatrix
> validated <- lapply(colnames(seedMatrix), function(j) {
+
+         as.matrix(rownames(seedMatrix)) %in%
+                 as.matrix(subset(mirtarbase, miRNA==j)$`Target Gene`)
+ })
> validated <- do.call("cbind", validated)
> dimnames(validated) <- dimnames(seedMatrix)
```

We now count the validated targets from the top rank targets from Roleswitch and Seed-match matrix and plotted the results in barplot. As shown in Fig 2, Roleswitch predicted more validated targets of the expression miRNAs than using Seed-match alone. For more extensive comparison with other exisitng methods, please refer to our paper (once it is published).

```
> toprank <- seq(from=100,to=1000,by=100)
> toprank_eval <- function(pred, decreasing=T, mirna.expr) {
+
+         expressed.miRNA <- rownames(mirna.expr)[mirna.expr > 0]
+
+         tmp <- validated
+
+         tmp[, !colnames(tmp) %in% expressed.miRNA] <- FALSE
+
+         valid <- which(as.numeric(tmp)==1)
+
+         tp <- sapply(toprank, function(n)
+                 sum(head(order(pred, decreasing=decreasing), n) %in% vali
+
+         data.frame(rank=toprank, validated=tp)
+ }
> rs.toprank <- data.frame(toprank_eval(as.numeric(rs.pred$p.xz),
+                 mirna.expr=z), type="GBM", method="Roleswitch")
> seed.toprank <- data.frame(toprank_eval(as.numeric(seedMatrix),
+                 mirna.expr=z), type="GBM", method="Seed Matrix")

> require(ggplot2)
> df <- rbind(rs.toprank, seed.toprank)
> gg <- ggplot(data=df, aes(x=factor(rank), y=validated, fill=method)) +
+
+         theme_bw() + geom_bar(stat="identity", position="dodge") +
+
+         scale_x_discrete("Top rank") +
+
+         scale_y_continuous("Validated targets of expressed miRNAs")
```

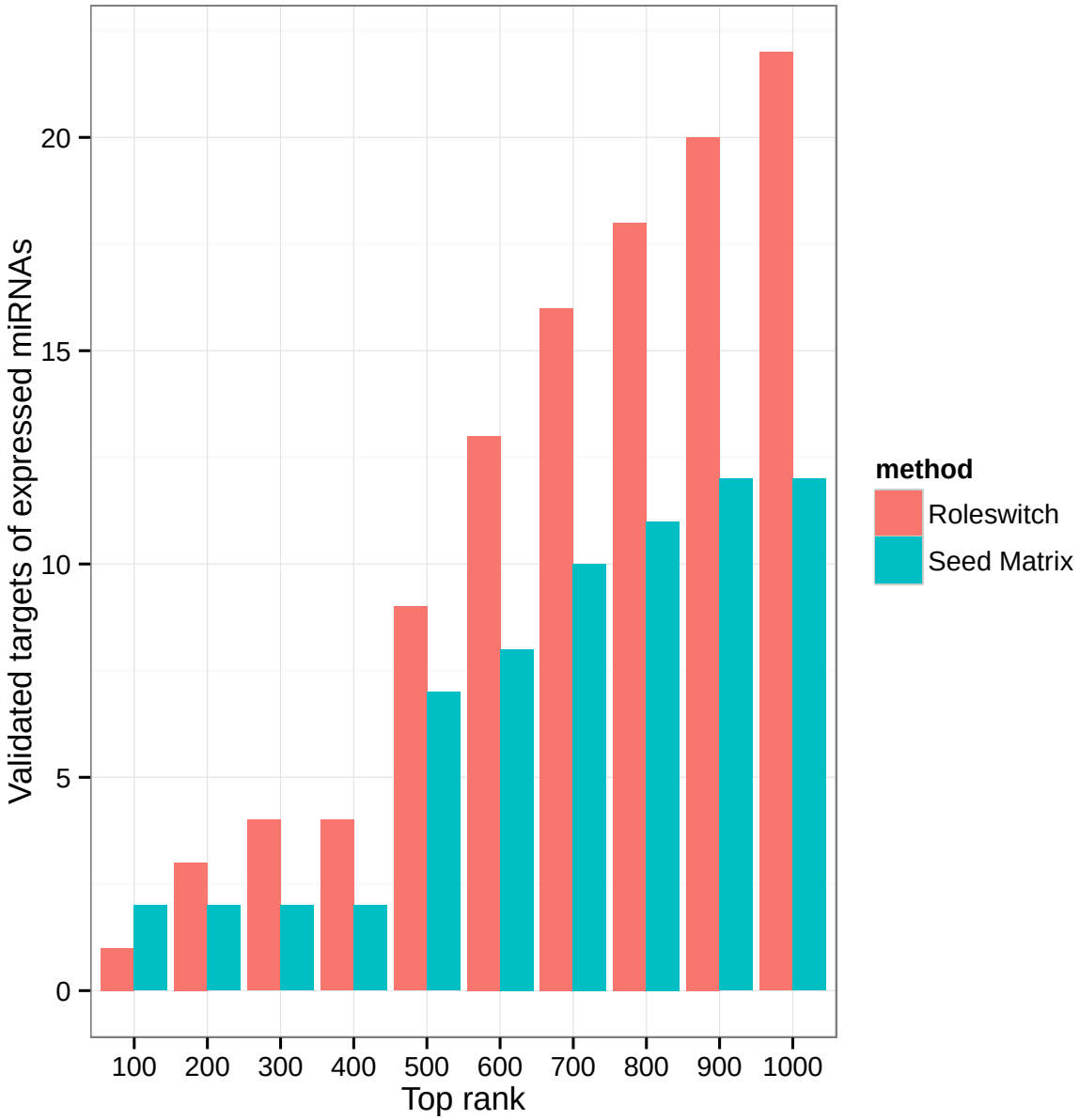

Figure 2: The number of validated targets selected by Roleswitch and seed match matrix among their top rankings.

4 Working with `eSet/ExpressionSet`

Roleswitch supports `eSet` or `ExpressionSet` from *Biobase* as input for mRNA expression.[1] For an eSet containing multiple samples, Roleswitch will only take the first sample. Thus, user needs to run Roleswitch multiple times on distinct samples or average the probe values across multiple samples (replicates). In the following example, Roleswitch converts the probe ID to gene symbol using packages dedicated for the chip (i.e., hgu95av2 in the example below), avearges multiple probe values for the same gene or miRNA, and constructs seed match matrix for human, automatically. Fig **??** depicts the 36×7 probabilities matrix generated by Roleswitch for the distinct 36 mRNAs being the targets of 7 distinct miRNAs.

```
> # mRNA expression from eSet
> dataDirectory <- system.file("extdata", package="Biobase")
> exprsFile <- file.path(dataDirectory, "exprsData.txt")
> exprs <- as.matrix(read.table(exprsFile, header=TRUE, sep="\t",
+                   row.names=1, as.is=TRUE))
> eset <- ExpressionSet(assayData=exprs[,1,drop=F], annotation="hgu95av2")
> annotation.db <- sprintf("%s.db", annotation(eset))
> # miRNA expression
> mirna.expr <- matrix(
+         c(1.23, 3.52, 2.42, 5.2, 2.2, 1.42, 1.23, 1.20, 1.37),
+         dimnames=list(
+         c("hsa-miR-148b", "hsa-miR-27b", "hsa-miR-25",
+       "hsa-miR-181a", "hsa-miR-27a", "hsa-miR-7",
+       "hsa-miR-32", "hsa-miR-32", "hsa-miR-7"), "miRNA Expression")
+ )
> rs <- roleswitch(eset, mirna.expr)
> promise <- rs$p.xz[apply(rs$p.xz,1,sum)>0, apply(rs$p.xz,2,sum)>0]
```

5 Session Info

```
> sessionInfo()

R version 3.1.0 (2014-04-10)
Platform: x86_64-unknown-linux-gnu (64-bit)

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=C
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
```

[1] For miRNA expression, user still need to provide a $1 \times M$ matrix containing expression of M miRNAs

```
> color2D.matplot(promise, extremes=c("white", "red"),
+   main=sprintf("ProMISe"), axes=FALSE, xlab="", ylab="", show.values=T)
> axis(1,at=0.5:(ncol(promise)-0.5),las=3,labels=sub("hsa-","", colnames(pr
> axis(2,at=0.5:(nrow(promise)-0.5),las=2,labels=rownames(promise))
```

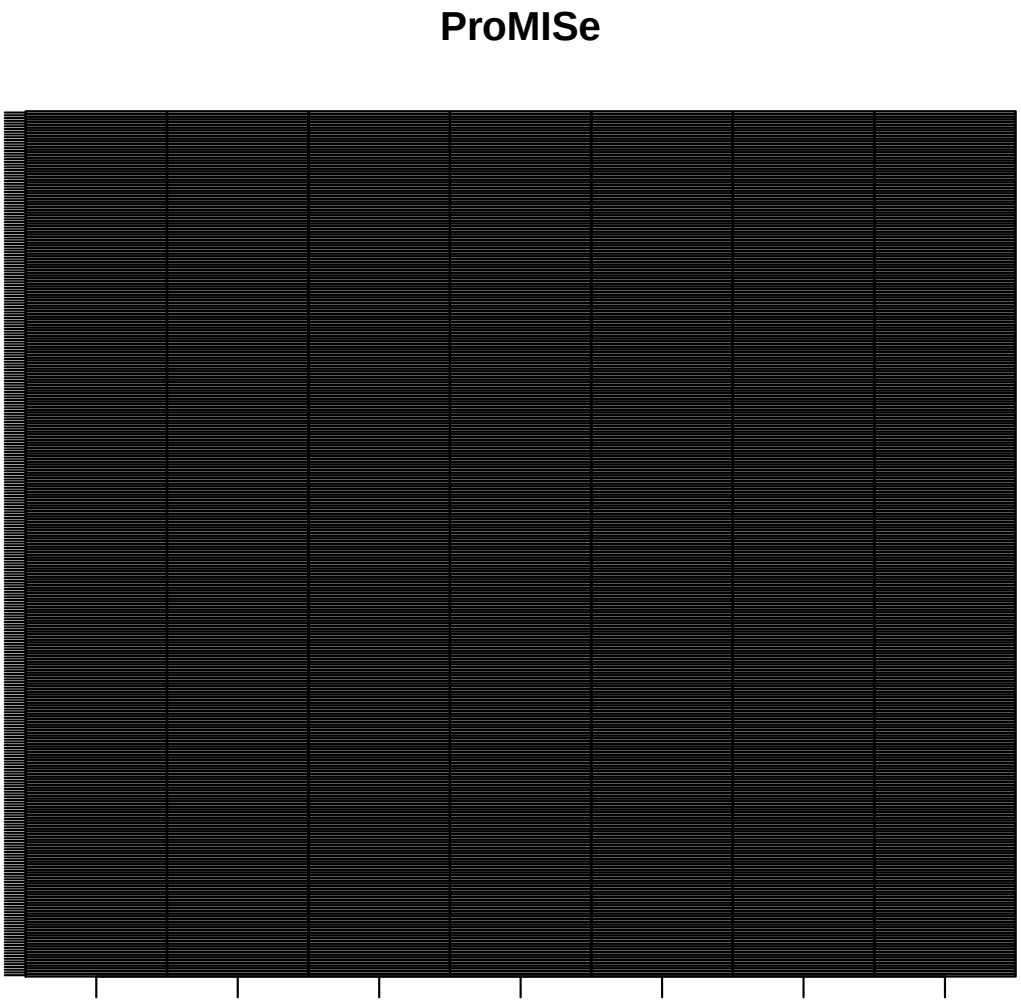

Figure 3: ProMISe generated from Section 4. Probabitlies are displayed in the cells. The intensity of the red color corresponds to the magnitude of the probabilisties (i.e. the higher the probability the more red).

```
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] parallel  stats     graphics  grDevices utils     datasets  methods
[8] base

other attached packages:
 [1] hgu95av2.db_2.14.0   org.Hs.eg.db_2.14.0  RSQLite_0.11.4
 [4] AnnotationDbi_1.26.0 GenomeInfoDb_1.0.0   ggplot2_0.9.3.1
 [7] Roleswitch_1.2.0     DBI_0.2-7            Biobase_2.24.0
[10] Biostrings_2.32.0    XVector_0.4.0        IRanges_1.21.45
[13] BiocGenerics_0.10.0  biomaRt_2.20.0       microRNA_1.22.0
[16] plotrix_3.5-5        reshape_0.8.4        plyr_1.8.1
[19] pracma_1.6.4

loaded via a namespace (and not attached):
 [1] MASS_7.3-31        RColorBrewer_1.0-5 RCurl_1.95-4.1     Rcpp_0.11.1
 [5] XML_3.98-1.1       colorspace_1.2-4   dichromat_2.0-0    digest_0.6.4
 [9] grid_3.1.0         gtable_0.1.2       labeling_0.2       munsell_0.4.2
[13] proto_0.3-10       reshape2_1.2.2     scales_0.2.3       stats4_3.1.0
[17] stringr_0.6.2      tools_3.1.0        zlibbioc_1.10.0
```

References

David P Bartel. MicroRNAs: Target Recognition and Regulatory Functions. *Cell*, 136(2):215–233, January 2009.

Sam Griffiths-Jones, Harpreet Kaur Saini, Stijn van Dongen, and Anton J Enright. miRBase: tools for microRNA genomics. *Nucleic acids research*, 36(Database issue):D154–8, January 2008.

Sheng-Da Hsu, Feng-Mao Lin, Wei-Yun Wu, Chao Liang, Wei-Chih Huang, Wen-Ling Chan, Wen-Ting Tsai, Goun-Zhou Chen, Chia-Jung Lee, Chih-Min Chiu, Chia-Hung Chien, Ming-Chia Wu, Chi-Ying Huang, Ann-Ping Tsou, and Hsien-Da Huang. miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic acids research*, 39 (Database issue):D163–9, January 2011.