How To Use MiRSEA

Junwei Han

May 31, 2015

Contents

1 Overview

This vignette demonstrates how to easily use the MiRSEA package. The package can identify dysregulated pathways by a novel method of microRNA(miRNA) set enrichment analysis(MiRSEA). Regulation of dysregulated pathway of microRNAs concentrated at the top or bottom of the miRNA List.Our system constructs the gene sets of pathways from three database(Kyoto Encyclopedia of Genes and Genomes(KEGG); Reactome; Biocarta;) and target gene sets of human microRNAs(miRNAs) from four database(TarBaseV6.0; mir2Disease; miRecords; miRTarBase;).The MiRSEA can quantify the strength of the pathway regulated by miRNAs.It gets a weighting matrix of strength of the pathway regulated by each miRNA(see the section 2).The MiRSEA uses the weighted Kolmogorov-Smirnov statistic to calculate a miRNA set enrichment score(ES),which is in order to assess if the pathway is associated the specific phenotype(see the section 3).When users input interesting pathway name,the MiRSEA package also can create a running enrichment plot and a Heat Map of the miRNAs(see the section 4)

2 Get the pathway-miRNA correlation profile(pmSET) and a weighting matrix

The section introduces how to obtain the pathway-miRNA correlation file and a p.value weighting matrix.The miRNA and target gene data are collected from the four popular public microRNA databases(TarBase

V6.0; mir2Disease; miRecords and miRTarBase). Each rows of the dataframe represents a target gene set of miRNA. The human pathways data are collected from the three popular public databases (KEGG, Reactome,Biocarta).Each rows of the downloaded dataframe represents a gene set of pathway,whose first and second column are the pathway name and source.

MiRSEA calculate a weighting matrix by hypergeometric test,which represent strength of the pathway regulated by each miRNA.Each row of the weighting matrix represents a pathway,whose columns represent miRNA . The weighting value(w) of the matrix is 1- p value of hypergeometric test,which can quantify the strength of the pathway regulated by the miRNAs.The smaller p value is represent the bigger strength of regulate.For each human pathway,MiRSEA get a regulated miRNA set of the pathway(pmSET,w>0).

The following commands can obtain the pathway-miRNA correlation file(pmSET) and a weighting matrix of p value.

```
> #getting KEGG pathway and human miRNAs Correlation profile(pmSET)
> #and getting a weighting matrix of human miRNAs
> p22m<-Corrp2miRfile(pathway="kegg",species="example")
> #getting a weighting matrix of human miRNAs
> p_value<-p22m$p
> p_value[1,1:15]
```

```
  hsa-miR-221   hsa-miR-222  hsa-miR-124a   hsa-miR-127  hsa-miR-122a
            1             1             1             1             1
 hsa-miR-199b  hsa-miR-106a hsa-miR-17-5p   hsa-miR-101   hsa-miR-29b
            1             1             1             1             1
  hsa-miR-504   hsa-miR-19a    hsa-miR-21   hsa-miR-15b    hsa-let-7a
            1             1             1             1             1
```

```
> #getting the column names of matrix(miRNA names)
> miRnames<-colnames(p_value)
> miRnames[1:10]
```

```
 [1] "hsa-miR-221"   "hsa-miR-222"   "hsa-miR-124a"  "hsa-miR-127"
 [5] "hsa-miR-122a"  "hsa-miR-199b"  "hsa-miR-106a"  "hsa-miR-17-5p"
 [9] "hsa-miR-101"   "hsa-miR-29b"
```

```
> #getting the row names of matrix(pathway names)
> pathway.names<-rownames(p_value)
> pathway.names[1:2]
```

```
[1] "KEGG_GLYCOLYSIS_GLUCONEOGENESIS" "KEGG_CITRATE_CYCLE_TCA_CYCLE"
```

```
> #getting the set of regulating miRNAs of each pathway(pmSET)
> p2miR<-p22m$p2miR
> p2miR[1,1:5]
```

```
[1] "KEGG_GLYCOLYSIS_GLUCONEOGENESIS"
[2] "http://www.broadinstitute.org/gsea/msigdb/cards/KEGG_GLYCOLYSIS_GLUCONEOGENESIS"
[3] "hsa-miR-133a"
[4] "hsa-miR-133b"
[5] ""
```

```
>##write the results to tab delimited file.
>write.table(p_value,file="p_value.txt",sep="\t")
>##write the results to tab delimited file.
>write.table(p2miR,file="p2miR.gmt",sep="\t",row.names=FALSE,col.names=FALSE)
```

3 Discovering the dysregulated pathways based on miRNA set enrichment analysis

The section introduces the miRNA Set Enrichment Analysis (MiRSEA) method for identifying canonical biological pathways associated with a specific phenotype.MiRSEA identify dysregulated pathways by calculating the weighted Kolmogorov-Smirnov statistic of the microRNA set(pmSET),which regulate genes in the pathway(see the section 2). MiRSEA integrates pathway structure that is regulated by miRNAs and differential expression of miRNA among two phenotypes(e.g.tw-score).MirSEA operates on all miRNAs from an experiment and get a miRNA list ranking ordered by the weighted signal to noise ratio(tw-score).Finally, the weighted Kolmogorov-Smirnov statistic is used to prioritize the pathways by mapping the miRNAs in the pmSET to the miRNA list (see the section 3.2)

3.1 calculate signal to noise ratio for miRNA

For each miRNA,The function `S2N`can calculate the differential expression scores(signal to noise ratio) of cancer samples and control samples.The following commands can calculate the differential expression scores(signal to noise ratio) of miRNAs in a given miRNA expression dataset.

```
> #input example expression dataset
> A<-matrix(runif(200),10,20)
> ##input a class.labels("0" or "1") of the expression dataset
> a1<-rep(0,20)
> a1[sample(1:20,5)]=1
> a1<-sort(a1,decreasing=FALSE)
> #Calculate the differential expression score for miRNAs
> M1<-S2N(A, class.labels=a1, miR.labels=seq(1,10), nperm=100)
> #print the top five observed results to screen
> M1$obs.s2n.matrix[1:5,1]

[1]  0.04167061 -0.29039910  0.13547973 -0.30856024  0.13844810

> #print the top five permutations results to screen
> M1$s2n.matrix[1:5,1:5]

             [,1]        [,2]        [,3]        [,4]        [,5]
[1,] -0.472871063  0.23847060 -0.24612075  0.07729647  0.17630743
[2,] -0.007452044 -0.13954986 -0.11470976 -0.12161677  0.05785403
[3,] -0.070787544  0.20387597 -0.07144645 -0.14115533  0.03571008
[4,]  0.481277431 -0.26600462  0.14517391 -0.27576924  0.02335267
[5,] -0.051028263 -0.03579814 -0.02065167 -0.21974615 -0.17667335
```

3.2 Discovering the dysregulated pathways

MiRSEA identify dysregulated pathways by calculating the weighted Kolmogorov-Smirnov statistic of the microRNA set(pmSET),which regulate genes in the pathway.The weighted Kolmogorov-Smirnov statistic is used to evaluate each pathway and the permutation is used to calculate the statistical significance of pathways.

The function `MirSEA` can identify the dysregulated pathways. The following commands can identify the dysregulated pathways in a given miRNA expression dataset with default parameters.

```
> #input example expression dataset
> input.ds <- GetExampleData("dataset")
```

```
> ##input a class.labels of the expression dataset
> input.cls <- GetExampleData("class.labels")
> #get example of p value matrix
> #p_value<-p22m$p
> p_value <- GetExampleData("p_value")
> #get example of correlation profile
> #p2miR<-p22m$p2miR
> p2miR <- GetExampleData("p2miR")
> #identify dysregulated pathways by using the function MirSEA
> MirSEAresult <- MirSEA(input.ds,input.cls,p_value,p2miR,
+ reshuffling.type = "miR.labels", nperm = 100,
+ weighted.score.type = 1, ms.size.threshold.min = 10,
+ ms.size.threshold.max = 500)
> #print the summary results of up-regulated pathways to screen
> summaryResult1<-MirSEAresult$report.phen1
> summaryResult1[1:5,]

                         Pathway SIZE
1     KEGG_ACUTE_MYELOID_LEUKEMIA   75
2             KEGG_BLADDER_CANCER  101
3   KEGG_CHRONIC_MYELOID_LEUKEMIA  114
4 KEGG_NON_SMALL_CELL_LUNG_CANCER   89
5          KEGG_PANCREATIC_CANCER  109
                                                                      SOURCE
1     http://www.broadinstitute.org/gsea/msigdb/cards/KEGG_ACUTE_MYELOID_LEUKEMIA
2             http://www.broadinstitute.org/gsea/msigdb/cards/KEGG_BLADDER_CANCER
3   http://www.broadinstitute.org/gsea/msigdb/cards/KEGG_CHRONIC_MYELOID_LEUKEMIA
4 http://www.broadinstitute.org/gsea/msigdb/cards/KEGG_NON_SMALL_CELL_LUNG_CANCER
5          http://www.broadinstitute.org/gsea/msigdb/cards/KEGG_PANCREATIC_CANCER
       ES    NES NOM p-val FDR q-val Tag \\% Mir \\% Signal
1 0.50616 1.9352         0         0   0.333  0.0658  0.341
2 0.43799 1.8058         0         0   0.317   0.094  0.326
3 0.43771 1.7558         0         0   0.307  0.0987   0.32
4 0.45229 1.7496         0         0   0.337  0.0952  0.341
5 0.42895 1.7183         0         0   0.303  0.0846  0.318

> #print the summary results of down-regulated pathways to screen
> summaryResult2<-MirSEAresult$report.phen2
> summaryResult2[1:5,]

                                          Pathway SIZE
1                     KEGG_ECM_RECEPTOR_INTERACTION   35
2                        KEGG_N_GLYCAN_BIOSYNTHESIS   21
3                      KEGG_SPHINGOLIPID_METABOLISM   10
4 KEGG_METABOLISM_OF_XENOBIOTICS_BY_CYTOCHROME_P450   10
5                       KEGG_MTOR_SIGNALING_PATHWAY   46
                                                                                SOURCE
1                     http://www.broadinstitute.org/gsea/msigdb/cards/KEGG_ECM_RECEPTOR_INTERACTION
2                        http://www.broadinstitute.org/gsea/msigdb/cards/KEGG_N_GLYCAN_BIOSYNTHESIS
3                      http://www.broadinstitute.org/gsea/msigdb/cards/KEGG_SPHINGOLIPID_METABOLISM
4 http://www.broadinstitute.org/gsea/msigdb/cards/KEGG_METABOLISM_OF_XENOBIOTICS_BY_CYTOCHROME_P450
5                       http://www.broadinstitute.org/gsea/msigdb/cards/KEGG_MTOR_SIGNALING_PATHWAY
```

```
        ES     NES NOM p-val FDR q-val Tag \\% Mir \\% Signal
1 -0.55125 -1.7195         0         0     0.4  0.0764  0.385
2 -0.62464 -1.6736         0         0   0.429   0.067   0.41
3 -0.72341 -1.6507         0         0     0.4  0.0247  0.395
4 -0.73447 -1.6474         0         0     0.4  0.0235  0.395
5 -0.50983 -1.6425         0         0   0.283  0.0494  0.284
```

The each row of the summaryResult (data.frame) is a pathway. Columns include 'Pathway name', 'SIZE', 'Pathway Source', 'ES', 'NES', 'NOM p-val','FDR q-val','Tag percentage'(Percent of miRNA set before running enrichment peak),'MiR percentage'(Percent of miRNA list before running enrichment peak),'Signal strength'(enrichment signal strength).

```
>##write the results to tab delimited file.
>write.table(summaryResult1,file="summaryResult1.txt",sep="\t",col.names=FALSE,row.names=FALSE)
>##write the results to tab delimited file.
>write.table(summaryResult2,file="summaryResult2.txt",sep="\t",row.names=FALSE,col.names=FALSE)
```

4 Get running result of a pathway

4.1 Produce report for a pmSET(a miRNA set for pathway)

When users input a interesting pathway,the function `MsReport` can create a report for miRNA set that coordinated regulate this pathway.Msreport is matrix of input pathway which present the detail results. Its columns include "miRNA name", "location of the miRNA in the sorted miRNA list", "tw-scoe of miRNA", "Running enrichment score", "Property of contribution".miRList is a list of drawing parameters for function `PlotHeatMap`,`PlotCorrelation` and `PlotRunEnrichment`.

```
> #get example data
> input.ds <- GetExampleData("dataset")
> input.cls <- GetExampleData("class.labels")
> #get example of p value matrix
> p_value <- GetExampleData("p_value")
> #get example of correlation profile
> p2miR <- GetExampleData("p2miR")
> #get a report of miRNA set for KEGG ERBB pathway
> Results<-MsReport(MsNAME = "KEGG_ERBB_SIGNALING_PATHWAY", input.ds, input.cls,p_value,p2miR)
> # show the report of top five miRNA in the pathway
> Results[[1]][1:5,]
```

```
  #             MiR LIST LOC TW-SCORE    RES CORE_ENRICHMENT
1 1     hsa-miR-424        1     3.09 0.0488             YES
2 2       hsa-miR-7        2     2.34 0.0858             YES
3 3    hsa-miR-34b*        3     1.99  0.117             YES
4 4  hsa-miR-34c-5p        4      1.8  0.146             YES
5 5 hsa-miR-146b-5p        5      1.6  0.171             YES
```

```
> miR.report<-Results[[1]]
> ##write the results to tab delimited file.
> write.table(miR.report,file="miR.report.txt",sep="\t",col.names=FALSE,row.names=FALSE)
> #write the detail results of miRNAs for drawing results
> for(i in 1:length(Results[[2]])){
+ miRList<-Results[[2]][[i]]
```

```
+ filename <- paste("miRPlots",".txt", sep="", collapse="")
+ write.table(miRList, file = filename, quote=F, row.names=F,col.names=F, sep = "\t",append=T)
+ }
```

4.2 Plot global miRNAs correlation(tw-score) profile

The function `PlotCorrelation` can plot global miR correlation profile for weighted differential expression scores(tw-score) of miRs

```
> #get example data
> input.ds <- GetExampleData("dataset")
> input.cls <- GetExampleData("class.labels")
> #get a list of miRNA list result
> #Results<-MsReport(MsNAME="KEGG_ERBB_SIGNALING_PATHWAY", input.ds, input.cls,
> #weighted.score.type = 1)
> #miRlist<-Results[[2]]
> miRlist<-GetExampleData("miRList")

> #plot global miRNA correlation profile
> PlotCorrelation(miRlist)
```

Figure 1 shows the global miRNA correlation profile for weighted differential expression scores(tw-score) of miRNAs.

4.3 Plot running miRNAs enrichment score

The function `PlotRunEnrichment` can plot running miRNAs enrichment score for the pathway result.

```
> #get example data
> input.ds <- GetExampleData("dataset")
> input.cls <- GetExampleData("class.labels")
> #get a list of miRNA list result
> #Results<-MsReport(MsNAME="KEGG_ERBB_SIGNALING_PATHWAY", input.ds, input.cls,
> #weighted.score.type = 1)
> #miRlist<-Results[[2]]
> miRlist<-GetExampleData("miRList")

> #Plot running miRNAs enrichment score for the pathway result
> PlotRunEnrichment(miRlist)
```

Figure 2 shows the running miRNAs enrichment score for the pathway result

4.4 Plot a heat map

The function `PlotHeatMap` can plot a heat map for a miR set which co-regulate pathway

```
> #get example data
> input.ds <- GetExampleData("dataset")
> input.cls <- GetExampleData("class.labels")
> #get a list of miRNA list result
> #Results<-MsReport(MsNAME="KEGG_ERBB_SIGNALING_PATHWAY", input.ds, input.cls,
> # weighted.score.type = 1)
> #miRlist<-Results[[2]]
> miRlist<-GetExampleData("miRList")
```

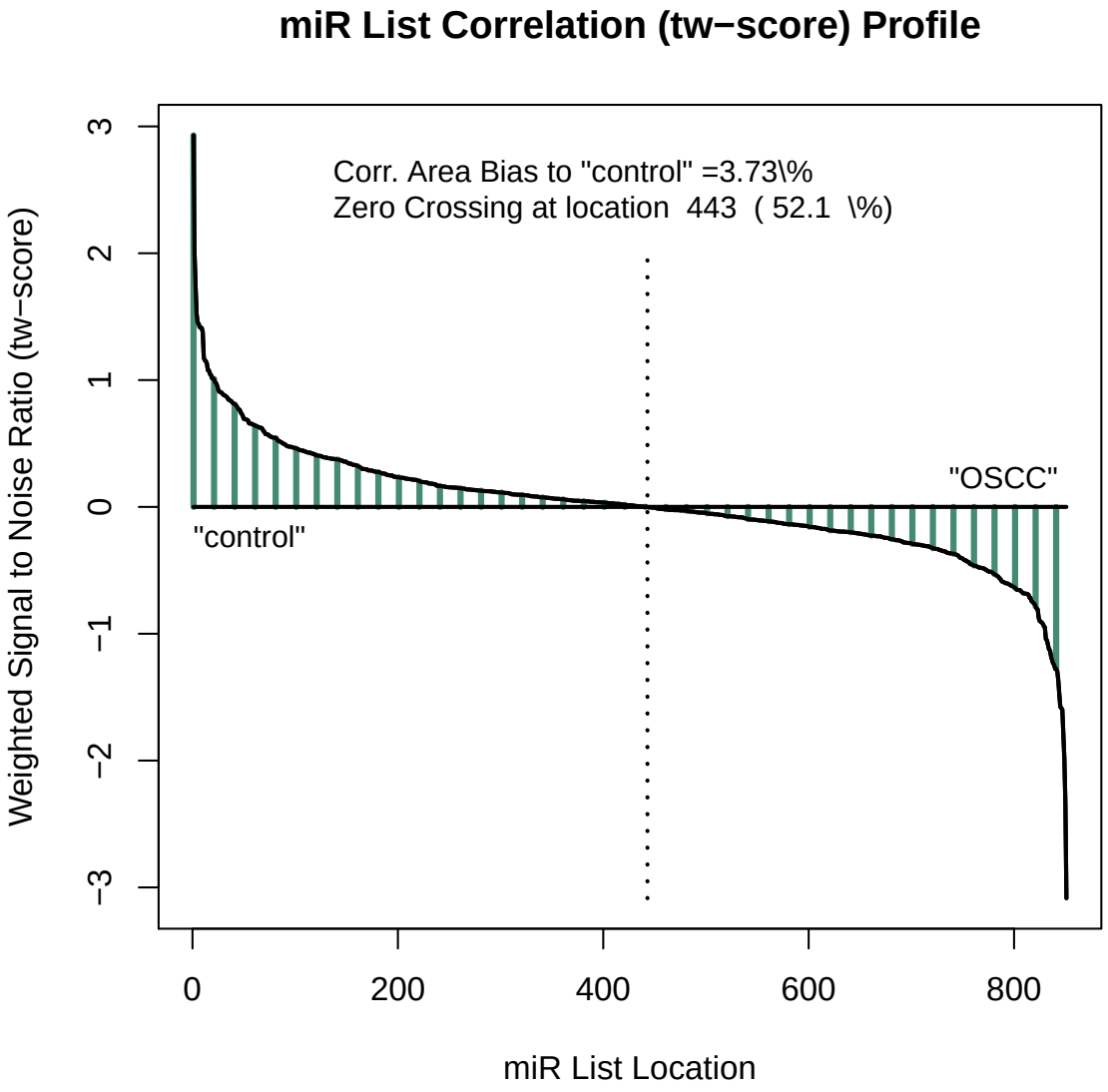

Figure 1: The visualization of global miRNAs correlation profile for weighted differential expression scores(tw-score) of miRNAs.

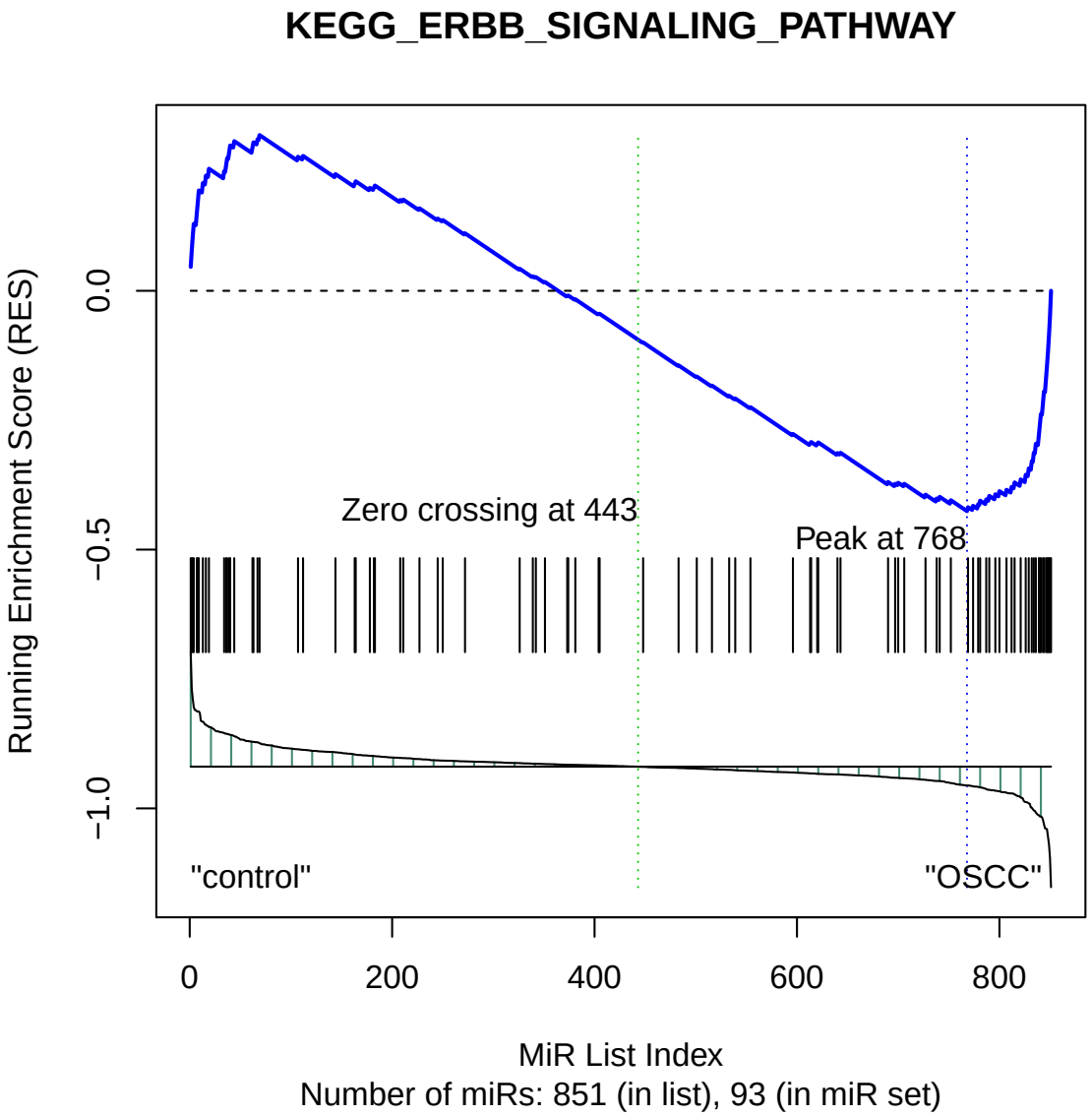

Figure 2: The visualization of the running miRNAs enrichment score for the pathway result

```
> #Plot a heat map for a miRNA set which co-regulate pathway
> PlotHeatMap(miRlist,input.ds,input.cls)
```

Figure 3 shows a heat map for a miRNA set which co-regulate the pathway

5 Data management

The environment variable `envData`, which is used as the database of the system, stores many data relative to pathway analyses. We can use the function `ls` to see the environment variable and use `ls(envData)` to see data in it. These data include `pathway`, `miRTarget`,`example.CLS`,`example.GCT`,`miRList`.For example, the variable `pathway` show some gene sets. The variable `mfile` include some miRs and their target genes ,which we combined from some databases.The variable `example.GCT` is an interesting miRNAs expression data and the variable `example.CLS`is the vector of binary labels(class.labels).The variable `miRList` provides drawing parameters of miRNA set.

```
> ##data in environment variable envData
> ls(envData)

[1] "biocarta"    "example.CLS" "example.GCT" "expMir2Tar"  "kegg"
[6] "miRList"     "p"           "p2miR"       "reactome"
```

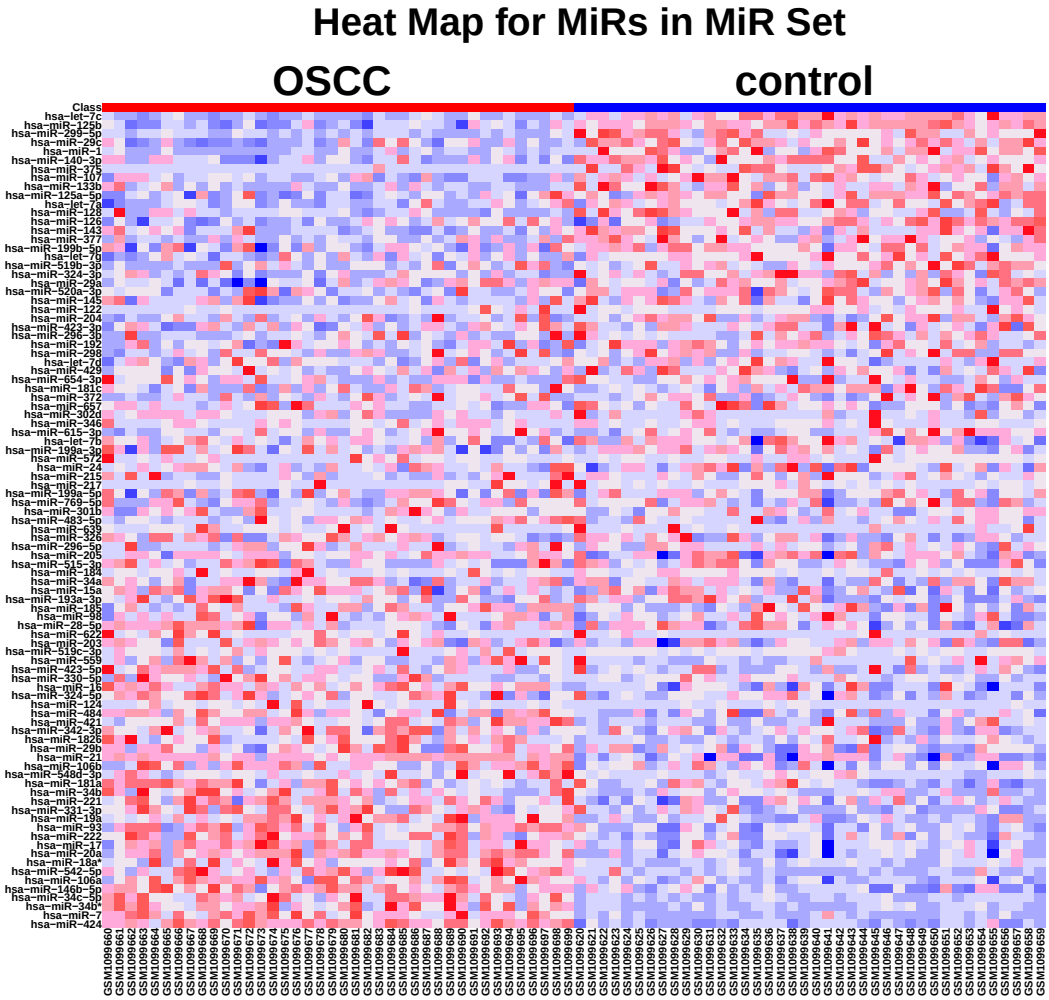

Figure 3: The visualization of heat map for a miRNA set

6 Session Info

The script runs within the following session:

```
R version 2.15.1 (2012-06-22)
Platform: i386-pc-mingw32/i386 (32-bit)

locale:
[1] LC_COLLATE=C
[2] LC_CTYPE=Chinese (Simplified)_People's Republic of China.936
[3] LC_MONETARY=Chinese (Simplified)_People's Republic of China.936
[4] LC_NUMERIC=C
[5] LC_TIME=Chinese (Simplified)_People's Republic of China.936

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

other attached packages:
[1] MiRSEA_1.0

loaded via a namespace (and not attached):
[1] tools_2.15.1
```

References

[Subramanian *et al.*, 2005] Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. et al. (2005) Gene set enrichment analysis: a knowledgebased approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A, 102, 15545-15550.