# Gene Set Data for Pathway Analysis in Mouse

Valerie Bares and Xijin Ge

November 29, 2016

Department of Mathematics and Statistics, South Dakota State University

## 1 Data Introduction

Gene Set Knowledgebase (GSKB) is a comprehensive knowledgebase for pathway analysis in mouse. GSKB is similar to MSigDB (molecular signature database), developed at Broad Institute (http://www.broadinstitute.org/gsea/msigdb). GSKB is intended for mouse, while MSigDB is for human. GSKB is created to support pathway analysis using software like Gene Set Enrichment Analysis (GSEA), etc.

The GSKB contains various gene sets, corresponding to pathways and functional categories. There are seven different types of gene sets:

- mm_GO: gene sets from Gene Ontology for mouse (Mus musculus)
- mm_location: Gene sets based on chromosomal location
- mm_metabolic: metabolic pathways
- mm_miRNA: Target genes of microRNAs, predicted or experimentally verified
- mm_pathway: Currated pathways
- mm_TF: Transcription factor target genes.
- mm_other

In addition, we also compiled a large collection of gene lists representing differentially expressed genes manually collected from literature. This dataset is too big and is only available on our web site. At the end of this vignette, we show how these gene sets can be used.

## 2 Data Description

Interpretation of high-throughput genomics data based on biological pathways constitutes a constant challenge, partly because of the lack of supporting pathway database. We created a functional genomics knowledgebase in mouse, which includes 33,261 pathways and gene sets compiled from 40 sources such as Gene Ontology, KEGG, GeneSetDB, PANTHER, microRNA and transcription factor target genes, etc. Detailed information on these 40 sources and the citations is available http://ge-lab.org/gskb/Table%201-sources.pdf.

In addition, we also manually collected and curated 8,747 lists of differentially expressed genes from 2,526 published gene expression studies to enable the detection of similarity to previously reported gene expression signatures. These two types of data constitute a comprehensive Gene Set Knowledgebase (GSKB), which can be readily used by various pathway analysis software such as gene set enrichment analysis (GSEA).

More information about this data is available here http://ge-lab.org/gskb/. A paper describing these data are currently in revision by Database: The Journal of Biological Databases and Curation.

# 3 Loading the data

The datasets can be loaded using the `data` function. Here we load the specific types of gene sets, microRNA target genes. We also show a portion of the first gene set. Users can change "mm_miRNA" to any of the 7 categories such as "mm_pathway" to load other types of gene sets.

```
> library(gskb)
> data(mm_miRNA)
> mm_miRNA[[1]][1:10]
```

```
 [1] "MIRNA_MM_BETEL_MMU-LET-7A"
 [2] "BETEL_MMU-LET-7A; Good mirSVR score Conserved; The microRNA.org resource: targets and
 [3] "NSUN4"
 [4] "DCX"
 [5] "KCNK6"
 [6] "PBX1"
 [7] "PHF8"
 [8] "RACGAP1"
 [9] "EFHD2"
[10] "DCBLD2"
```

# 4 Using the Data

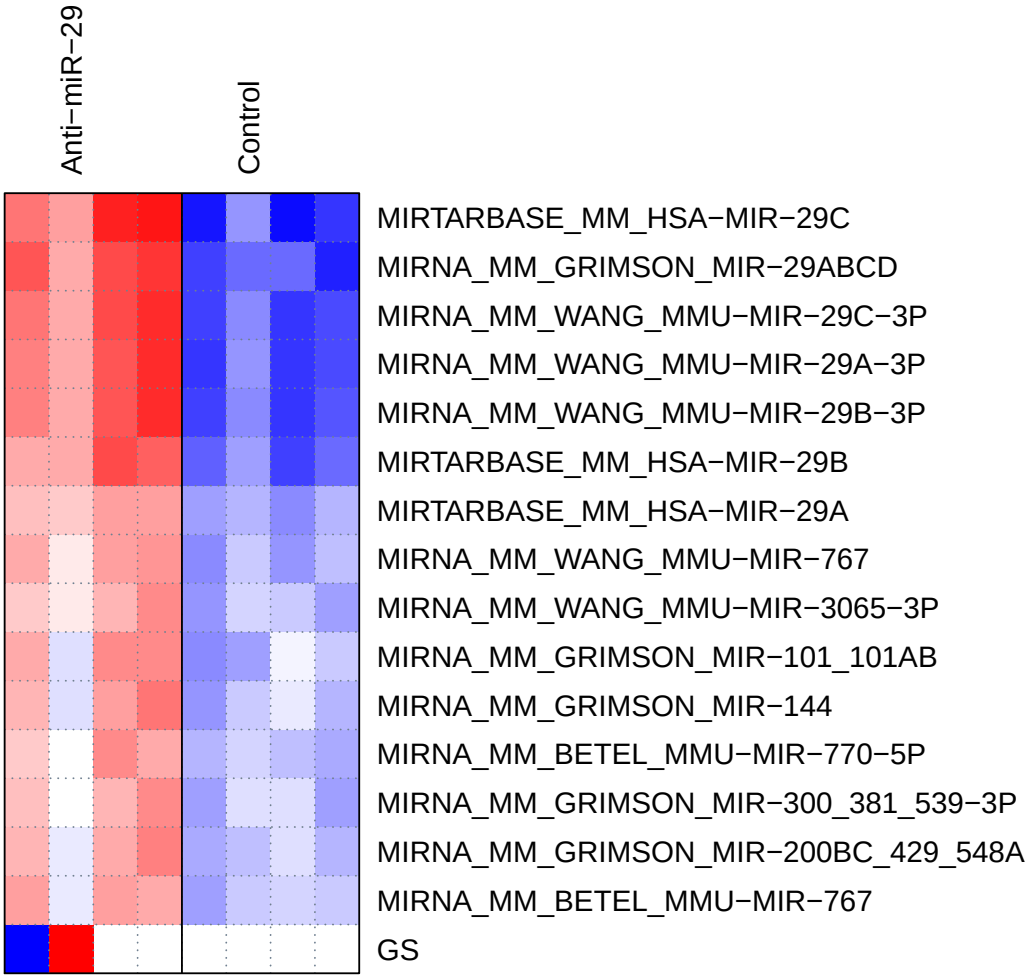
The following examples use the knowledge base gene sets to run pathway analysis on a microarray gene expression data obtained from Gene Expression Ominibus(GSE40261). In this experiment, microRNA-29b was blocked by antisense oligonecleotides. We will first use the PAGE: Parametric Analysis of Gene Set Enrichment method (http://www.biomedcentral.com/1471-2105/6/144), as implemented in the PGSEA bioconductor package. Then we will use GSEA to analyze it.

## 4.1 Pathway analysis using PGSEA

This example is using PGSEA with the microRNA gene sets.

```
> source("http://bioconductor.org/biocLite.R")
> biocLite("PGSEA")
```

```
> library(PGSEA)
> library(gskb)
> data(mm_miRNA)
> gse<-read.csv("http://ge-lab.org/gskb/GSE40261.csv",header=TRUE, row.name=1)
> # Gene are centered by mean expression
> gse <- gse - apply(gse,1,mean)
> pg <- PGSEA(gse, cl=mm_miRNA, range=c(15,2000), p.value=NA)
> # Remove pathways that has all NAs. This could be due to that pathway has
> # too few matching genes.
> pg2 <- pg[rowSums(is.na(pg))!= dim(gse)[2], ]
> # Difference in Average Z score in two groups of samples is calculated and
> # the pathways are ranked by absolute value.
> diff <- abs( apply(pg2[,1:4],1,mean) - apply(pg2[,5:8], 1, mean) )
> pg2 <- pg2[order(-diff), ]
> sub <- factor( c( rep("Control",4),rep("Anti-miR-29",4) ) )
> smcPlot(pg2[1:15,],sub,scale=c(-12,12),show.grid=TRUE,margins=c(1,1,7,19),col=.rwb)
```




This figure shows the top 15 pathways. As expected, miRNA-29 related gene sets are identified as differentially regulated pathway.

## 4.2 GSEA

This example is using GSEA with the miRNA gene set. Gene expression data is reformated to a GCT format. And also a phenotype vector file is created in the CLS format. See GSEA help file for more information on these files. http://www.broadinstitute.org/gsea/. These files are read from our web site. R program for GSEA is also downloaded and saved on our web site.

```
> library(gskb)
> data(mm_miRNA)
> ## GSEA 1.0 -- Gene Set Enrichment Analysis / Broad Institute
>
> GSEA.prog.loc<- "http://ge-lab.org/gskb/GSEA.1.0.R"
> source(GSEA.prog.loc, max.deparse.length=9999)
```

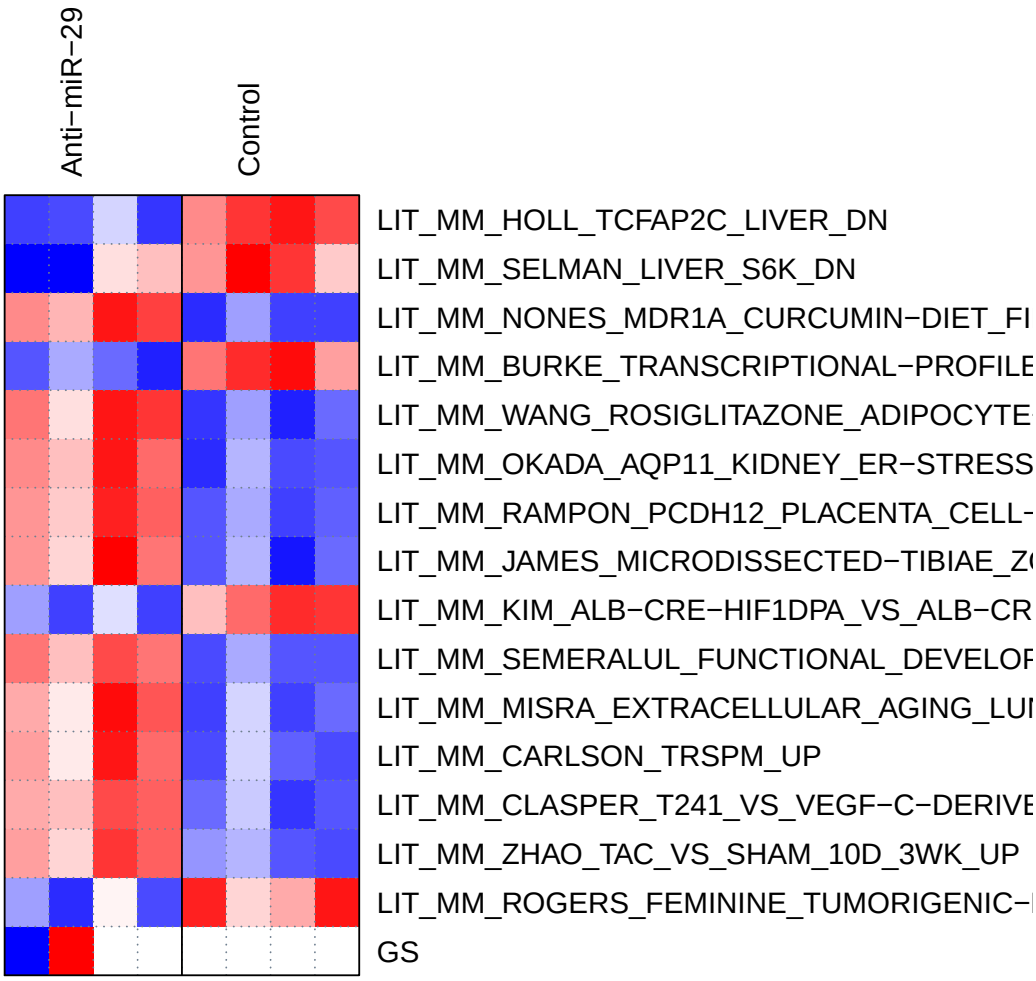
```
> GSEA(
+     # Input/Output Files :-------------------------------------------------
+
+     # Input gene expression Affy dataset file in RES or GCT format
+     input.ds = "http://ge-lab.org/gskb/mouse_data.gct",
+
+     # Input class vector (phenotype) file in CLS format
+     input.cls = "http://ge-lab.org/gskb/mouse.cls",
+
+     # Gene set database in GMT format
+     gs.db = mm_miRNA,
+
+     # Directory where to store output and results (default: "")
+     output.directory = getwd(),
+
+     #  Program parameters :-------------------------------------------------
+     doc.string = "mouse",
+     non.interactive.run = T,
+     reshuffling.type = "sample.labels",
+     nperm = 1000,
+     weighted.score.type =  1,
+     nom.p.val.threshold = -1,
+     fwer.p.val.threshold = -1,
+     fdr.q.val.threshold = 0.25,
+     topgs = 10,
+     adjust.FDR.q.val = F,
+     gs.size.threshold.min = 15,
+     gs.size.threshold.max = 500,
+     reverse.sign = F,
+     preproc.type = 0,
+     random.seed = 3338,
+     perm.type = 0,
+     fraction = 1.0,
+     replace = F,
+     save.intermediate.results = F,
+     OLD.GSEA = F,
+     use.fast.enrichment.routine = T
+ )
```

This will produce many output files in the current folder and the user can examine these figures and tables.

## 4.3 Additional Data

In addition to the 7 types of gene sets listed above, we also compiled a large collection of gene lists representing differentially expressed genes manually collected from literature. This dataset is too big and is only available on our web site. http://ge-lab.org/gskb/2-MousePath/MousePath_Co-expression_gmt.gmt Users can read these files directly from our website and use it in pathway analysis.

```
> library(PGSEA)
> library(gskb)
> d1 <- scan("http://ge-lab.org/gskb/2-MousePath/MousePath_Co-expression_gmt.gmt", what='
> mm_Co_expression <- strsplit(d1, "\t")
> names(mm_Co_expression) <- sapply(mm_Co_expression, '[[', 1)
> pg <- PGSEA(gse, cl=mm_Co_expression, range=c(15,2000), p.value=NA)
> # Remove pathways that has all NAs. This could be due to that pathway has
> # too few matching genes.
> pg2 <- pg[rowSums(is.na(pg))!= dim(gse)[2], ]
> # Difference in Average Z score in two groups of samples is calculated and
> # the pathways are ranked by absolute value.
> diff <- abs( apply(pg2[,1:4],1,mean) - apply(pg2[,5:8], 1, mean) )
> pg2 <- pg2[order(-diff), ]
> sub <- factor( c( rep("Control",4),rep("Anti-miR-29",4) ) )
> smcPlot(pg2[1:15,],sub,scale=c(-12,12),show.grid=TRUE,margins=c(1,1,7,19),col=.rwb)
```


Anti−miR−29
Control
LIT_MM_HOLL_TCFAP2C_LIVER_DN
LIT_MM_SELMAN_LIVER_S6K_DN
LIT_MM_NONES_MDR1A_CURCUMIN−DIET_FI
LIT_MM_BURKE_TRANSCRIPTIONAL−PROFILE
LIT_MM_WANG_ROSIGLITAZONE_ADIPOCYTE
LIT_MM_OKADA_AQP11_KIDNEY_ER−STRESS
LIT_MM_RAMPON_PCDH12_PLACENTA_CELL−
LIT_MM_JAMES_MICRODISSECTED−TIBIAE_Z
LIT_MM_KIM_ALB−CRE−HIF1DPA_VS_ALB−CR
LIT_MM_SEMERALUL_FUNCTIONAL_DEVELOP
LIT_MM_MISRA_EXTRACELLULAR_AGING_LUI
LIT_MM_CARLSON_TRSPM_UP
LIT_MM_CLASPER_T241_VS_VEGF−C−DERIVE
LIT_MM_ZHAO_TAC_VS_SHAM_10D_3WK_UP
LIT_MM_ROGERS_FEMININE_TUMORIGENIC−
GS

# 5 Session Info

The version number of R and packages loaded for generating the vignette were:

```
> sessionInfo()
```

```
R version 3.3.2 (2016-10-31)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 16.04.1 LTS

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

attached base packages:
[1] parallel  stats4    stats     graphics  grDevices utils     datasets
[8] methods   base

other attached packages:
 [1] PGSEA_1.48.0         annaffy_1.46.0       KEGG.db_3.2.3
 [4] GO.db_3.4.0          AnnotationDbi_1.36.0 IRanges_2.8.1
 [7] S4Vectors_0.12.0     Biobase_2.34.0       BiocGenerics_0.20.0
[10] gskb_1.6.1

loaded via a namespace (and not attached):
[1] Rcpp_0.12.8     digest_0.6.10   DBI_0.5-1       RSQLite_1.1
[5] BiocStyle_2.2.1 tools_3.3.2     memoise_1.0.0
```