Knowledge Based Gene Set analysis (KB-GSA): A novel method for gene expression analysis

Master's Thesis in Bioinformatics-2010

Author:
Trishul Jadhav
a09trija@student.his.se

Supervisor:
Jane Synnergren
jane.synnergren@his.se

School of Life Sciences
University of Skövde
Box 408
SE-541 28 Skövde
Sweden

Abstract

Microarray technology allows measurement of the expression levels of thousand of genes simultaneously. Several gene set analysis (GSA) methods are widely used for extracting useful information from microarrays, for example identifying differentially expressed pathways associated with a particular biological process or disease phenotype. Though GSA methods like Gene Set Enrichment Analysis (GSEA) are widely used for pathway analysis, these methods are solely based on statistics. Such methods can be awkward to use if knowledge of specific pathways involved in particular biological processes are the aim of the study. Here we present a novel method ***(Knowledge Based Gene Set Analysis: KB-GSA***) which integrates knowledge about user-selected pathways that are known to be involved in specific biological processes. The method generates an easy to understand graphical visualization of the changes in expression of the genes, complemented with some common statistics about the pathway of particular interest.

Introduction

The microarray technology has become a vital tool in biological and biomedical research. Global gene expression analysis is now widely used for exploring the differences between samples e.g. normal and diseased tissues. Differentially expressed genes (DEG) from microarray experiments (static and time series experiment [1]) are commonly identified using statistical methods. After determining a list of DEG, one can correlate these genes to known pathways and identify pathways that are induced under the specific biological condition studied [2].

Methods for pathway analysis are mainly based on one of two approaches viz. Individual Gene Analysis (IGA) and Gene Set Analysis (GSA) [3]. ***IGA*** methods aim at identifying individual genes whose changes in expression are associated with phenotype(s) (e.g. normal and diseased tissues). Significance Analysis of Microarray (SAM) [4] is one of the statistical methods that are widely used for identification of DEG [3, 5]. The ***GSA methods*** identify functionally related pathways that are linked to a disease or a specific biological process. GSA methods select differentially expressed pathways by ranking the pre-defined gene sets [3, 6]. The GSA methods identify gene sets with subtle but coordinated expression [3].

Related work

The GSA methods are aimed at identifying affected biological processes or pathways in a microarray experiment. Recently several GSA methods have been proposed for pathway enrichment analysis viz. *Global Test* [7], *Significance Analysis of Function and Expression* (SAFE) [8], *sigPathway* [9] – just to name a few. The ***Global Test*** [7] determines whether predefined gene sets/pathways are differentially expressed and determines if the global gene expression patterns of differentially expressed gene set are related to some clinical outcome of interest, for example, Acute Lymphoic Leukemia or Acute Myeloid Leukemia. The linear models are used in Global Test to calculate '*Q-statistic*' which describes relationships between gene expression profiles in a gene set and the clinical outcomes. The ***SAFE*** assess significant changes in gene expression across experimental conditions [8]. The assessment in SAFE is carried out by using local and global statistics. The SAFE uses local statistics like t-test to measure association between

gene expression profile and phenotypes. Then the SAFE uses global statistics like Wilcoxon rank sum to evaluate difference between local statistics of the given gene set and that of the reference set. The ***sigPathway*** [9] was developed by Tian *et al* and it determines whether a pathway is coordinately related with a phenotype (wild type and mutant).

However, the ***Gene Set Enrichment Analysis*** **(GSEA)** [10-11] is one of the first proposed GSA methods and is still commonly used for pathway analysis. The ***GSEA*** is based on the hypothesis that none of the predefined gene sets are associated with the phenotype. The Kolmogorov-Smirnov statistical test is used as scoring function to conclude which gene sets/pathways contain high scoring DEG [10]. The output from the GSEA is an ordered list of pathways/gene sets that are differentially expressed in the investigated dataset. However, one cannot explore a particular pathway if it has a low ranking. Ranking of the pathway reflects extent of dysregulation of the pathway. Top ranking suggest larger changes in expression and vice versa. This is a limitation in those cases where you have information about which pathway(s) you want to analyze in detail under a certain biological stimuli. Therefore, instead of ranking of pathways, prior knowledge regarding pathways and the biological processes which they regulate can be utilized for the pathway analysis. The use of prior knowledge is likely to produce more focused and relevant output that is easier to interpret and understand than the output received from GSEA.

Problem description and motivation

Microarrays measure the expression levels of thousands of genes simultaneously. The challenge is to extract useful information from all these expression profiles and make a biological interpretation of the results. One approach is to investigate which pathways that are induced in the gene expression dataset. Methods based on gene set analysis (GSA) are widely used for identification of differentially expressed pathways and elucidation of biological processes [12]. One example of such a method is Gene set enrichment analysis (GSEA) [10], which scores gene sets with *slight but harmonized* changes in expression or gene sets with differential expression. But ranking of pathways becomes immaterial if one has prior knowledge about pathways that are known to be involved in a specific biological process. Therefore, there is a need to develop a new approach as a complement to existing methods that can be utilized for pathway analysis when prior knowledge regarding which pathways those are relevant to elucidate is available. The purpose of this approach is to

aid the user with simple statistics and graphical visualization of the changes in expression of the genes in a pathway of particular interest.

Hypothesis

We hypothesize that prior knowledge of which pathways that are involved in a specific biological process can be utilized for pathway analysis.

Aims and objectives

The aim is to develop a supportive tool for pathway analysis that uses prior knowledge of user selected pathways that are important in a specific biological process. *The First objective* of this work is to identify which genes that are involved in a particular pathway. There are databases like Kyoto Encyclopedia of Genes and Genomes (KEGG) [13] that store various kind of information about pathways and their genes which can be utilized for the analysis. *The Second objective* is to develop an algorithm to evaluate gene expression for pathways involved in various biological processes by employing simple statistical measures and graphical representation of the expression values. *The Third objective* is to illustrate the usability of the tool by analyzing a real dataset from hepatocyte differentiation.

The overall aim is to present a new method called ***Knowledge Based Gene Set Analysis (KB-GSA)*** for pathway analysis. *KB-GSA* utilizes prior knowledge of pathways and biological processes and generates a user friendly statistical and graphical visualization of changes in expression levels of genes involved in a pathway.

Material and Methods

Gene expression data

The performance of the GSEA and the proposed novel method was tested on a real data set from Human Embryonic Stem cell lines that are available in NCBI Gene Expression Omnibus database (GEO; http://www.ncbi.nlm.nih.gov/geo/) and which is accessible through the GEO series accession number GSE13460. The dataset consists of gene expression levels from two groups of samples representing duplicate samples of wild type miR-122 (control) and triplicate samples expressing mutant miR-122 (treatment), carrying mutation in 3 nucleotides within the seed sequence.

GSEA analysis

The GSEA method was applied using default values for all parameters except that the metric "ratio of classes" was used to rank the genes according to how their expression levels correlate with phenotypes (wild type and mutant type). The recommended FDR (maximum of 25%) was used to identify significant gene sets. The gene sets were obtained from the Molecular Signature Database (MSigDB) v2.5 (April 2008 release).

KB-GSA analysis

The proposed novel method was implemented in the programming language R (v2.11.1). An analysis starts with uploading a microarray dataset, a predefined dataset (set of genes involved in a pathway to be analyzed) and cut off threshold value (fold change). After uploading the inputs, KB-GSA computes DEG (up and down regulated genes). DEGs are calculated by using fold change (FC) between the expression values of samples from two classes (e.g. wild type and mutant samples). Two-fold or greater differences in expression are generally required for achieving statistically significant results. Thus the differentially expressed genes in a pathway are identified with threshold of FC=2 or more. The R-code for KB-GSA can be found in the Appendix 1.

$$\text{Fold Change } (FC_k) = \log2(E_{ka}) - \log2(E_{kb})$$

Where, E_{ka} and E_{kb} are the expression values for the ***gene k*** in ***state 'a'*** and ***state 'b'***.

Pathway selection

Pathways for the KB-GSA analysis were selected from two pathway databases viz. KEGG [14]: Wingless signaling pathway (WNT), Cytokine-cytokine receptor interaction (CCR), Transforming growth factor beta pathway (TGF), Hedgehog signaling pathway (Hh), Janus kinase/signal transducers and activators of transcription (JAK/STAT) pathway, Mitogen-activated protein kinase (MAPK) pathway, Hepatocyte Growth Factor (HGF), Mammalian target of rapamycin pathway (mTOR), Vascular Endothelial Growth Factor (VEGF), Biocarta: Integrin Signaling Pathway and AKT Signaling Pathway.

Results and discussion

To illustrate the capability of the KB-GSA to give an 'easy to understand' output of the pathway analysis a dataset from hepatocyte differentiation [15] was used as an example. For comparison, the GSEA method was also applied on the same dataset. The dataset consists of 2 controls and 3 treated samples. The aim of the original study [15] (for which this data set was generated) was to determine whether miR-122 over expression in hESCs can direct the differentiation towards 'liver like' cells. The miR-122 is a liver specific microRNA. Increased levels of miR-122 during embryonic development [16] suggest that miR-122 might play implicit role in liver tissue specification and development. Several studies have shown that specific pathways, like WNT [17], HGF [18], Hedgehog [19], TGFβ [20], are associated with signaling processes during tissue specification and development [21]. Signaling pathways were selected for KB-GSA analysis (2 from BioCarta and 9 from KEGG [14]) that are known to be involved in tissue specification and development [21-23].

The GSEA results

The GSEA algorithm analyzes whether a dataset have genes that are related to a phenotype. It ranks the genes with respect to the difference between the expressions of two groups. The GSEA method was applied on the dataset from hepatocyte differentiation [15] to search for enriched gene set(s) from the curated gene sets (C2) of MSigDB. The MSigDB database is a collection of curated gene sets relevant for pathway analysis.[10]. The settings of the GSEA parameters were kept as default, except that the metric "ratio of classes" was used to rank the genes and how their expression levels correlate with phenotypes [10-11]. GSEA evaluates the distribution of genes in the queried gene set using statistics like Kolmogorov-Smirnov test, Enrichment score (ES), normalized enrichment score (NES). Positive (close to 1) ES indicates over-representation (up-regulation) of genes at the top of the ranked gene list while negative ES (close to -1) represents enrichment of the genes at the bottom of the ranked gene list (down-regulation). NES are obtained by normalizing the ES for enriched gene sets with different gene set sizes. Thus, NES is used for comparing results between different gene sets. Enriched gene sets with FDR less than 25% are likely to produce interesting hypothesis and are first choice for additional analysis.

GSEA identified 545 gene sets with positive ES score and 874 gene sets with negative ES score. The fact that none of the gene sets was found to be significant at FDR < 25% underlines the result obtained by *Tzur et al* [15] that over-expression of miR-122 does not have significant effects on hESCs differentiation. The list of gene sets identified as enriched by the GSEA contains several pathways having no obvious connection to signaling processes during cell development, e.g., PGC1A PATHWAY (ES=0.55, NES=1.51) which is coupled to tissue specific co-activator, while pathways known to be related to such signaling processes did not occur in the list of pathways found enriched by GSEA, e.g., Hepatocyte Growth Factor (HGF) pathway and AKT pathway.

The KB-GSA results

The KB-GSA method was employed to test the effect of miR-122 on the genes of pathways known to be involved in hepatocyte differentiation signaling process. Instead of using all gene sets in a database, only those pathways which are known to be relevant to the cell signaling process [21-23] were selected like WNT [17], HGF [18], Hedgehog [19], TGFβ [20] and Notch [24].

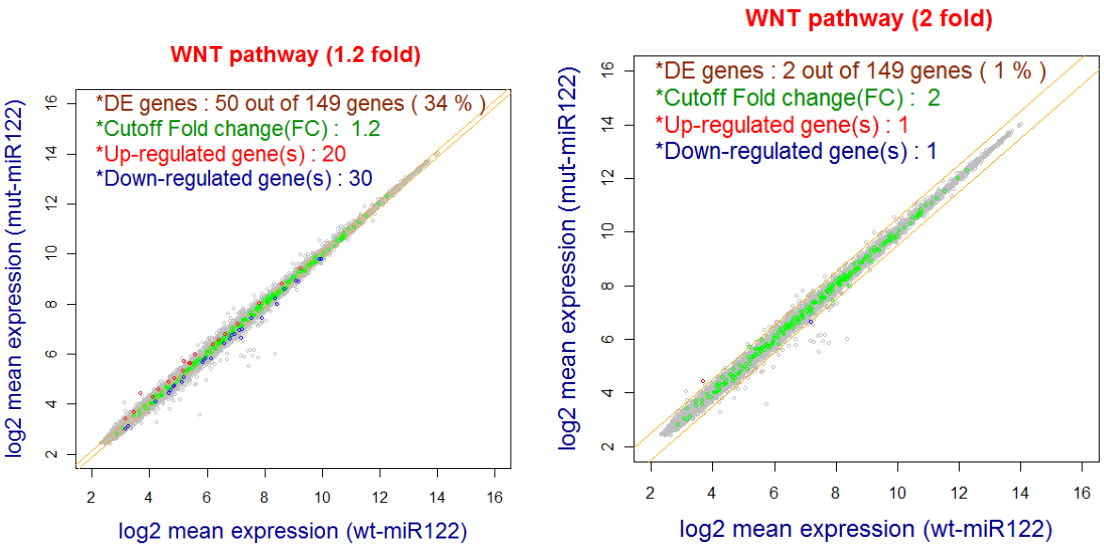
Results from the analysis are summarized in the Table 1. Two fold or greater differences in up-regulation and down-regulation in expression are, in general considered as statistically significant. At FC=2, the KB-GSA identified four pathways viz. Wingless signaling pathway (WNT), Cytokine-cytokine receptor interaction (CCR), Transforming growth factor beta pathway (TGF), Hedgehog signaling pathway (Hh). But the numbers of DEG in each pathway are less than 1%, suggesting that none of the selected pathways are enriched in the hepatocyte differentiation dataset. Most of the selected pathways contain 30% or more DE genes at FC=1.2. But the FC=1.2 is very less to be considered as statistically significant. Thus, the overall pathways analysis shows that the mir-122 had no or very little effect on hepatocyte differentiation.

Table 1. Summary of KB-GSA for hepatocyte differentiation data. The up and down regulated genes in the selected pathways at FC=1.2 and FC=2 are enlisted.

Gene set/data set	# Genes	FC (1.2)			FC (2.0)		
		UP	DN	%	UP	DN	%
Hepatocyte differentiation dataset	22215	1863	2596	20	22	46	0.3
WNT	149	20	30	34	1	1	1
CCR	259	48	44	36	1	-	0.4
TGF	85	23	19	49	1	-	1
Hh	38	5	14	33	-	1	2
MAPK	257	43	56	39	-	-	
JAK_STAT	157	18	27	29	-	-	
INT	25	5	2	28	-	-	
HGF	37	11	5	42	-	-	
mTOR	51	6	9	29	-	-	
AKT	24	2	7	38	-	-	
VEGF	73	8	15	32	-	-	

WNT: Wingless signaling pathway; ***CCR***: Cytokine-cytokine receptor interaction; ***TGF***: Transforming growth factor beta pathway; ***Hh :*** Hedgehog signaling pathway; ***JAK_STAT*** : Janus kinase/signal transducers and activators of transcription (JAK/STAT); ***MAPK:*** Mitogen-activated protein kinase; ***INT:*** Integrin Pathway; ***HGF:*** Hepatocyte Growth Factor; ***mTOR:*** Mammalian target of rapamycin pathway; ***VEGF***: Vascular endothelial growth factor; **UP:** up-regulated, **DN:** down-regulated; **%:** percentage of affected genes; **# genes:** number of genes in the pathway; **FC:** fold change.

Figure 1 illustrates a graphical representation of the output from KB-GSA, taking WNT signaling pathway as an example. The graphical output shows the number of genes that identified differentially expressed in a pathway at specified threshold by KB-GSA. The graphical representation for the other tested pathways can be found in Appendix 2.



1a 1b

Figure 1. Illustrating effect of miR-122 over-expression on WNT signaling pathways at FC=1.2 (1a) and 2(1b). *Red dots: up-regulated genes; *Blue dots: down-regulated genes, *green dots: pathways genes, *gray dots: total dataset from hepatocyte differentiation

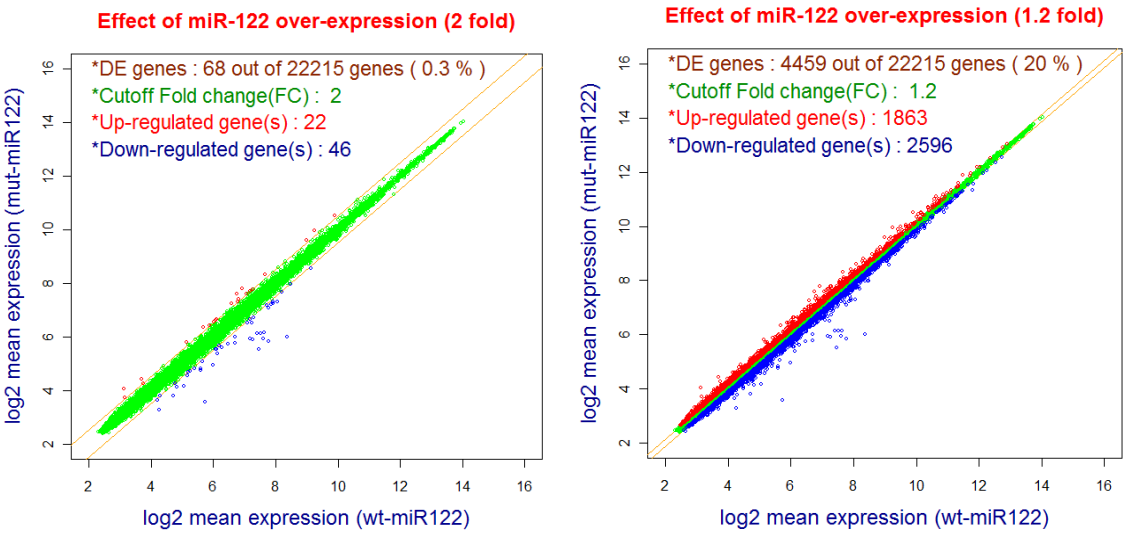
Even though the group gene analysis (GSA) is the main focus, it is often interesting to look at individual genes within the group. Such single genes within the selected pathways can be identified using KB-GSA. DEG in the selected pathways identified by KB-GSA are manually compiled in Table 2.

Table 2. Differentially expressed genes in different pathways in hepatocyte differentiation

Gene set/data set	FC 1.5		FC 2.0	
	UP	DN	UP	DN
WNT	*C1D*	*OAZ1*	*CSRP1, C1D, PPP2R1A*	*RPL35, TAF10 , C19orf50, OAZ1 , HSPA9*
CCR	*RPL11 , NARS ,RPL17, ARF1, , HSP90B1, YWHAZ, DNAJB1 , UBE2L3 , SF3B3 , CALU , SARS*	*GDI2 , RPL12 , GNAS , FAM120A , PSMB7*	*SF3B3*	-
TGF	*YY1 , NONO , GUK1, WDR1 , ATP6V0B , KARS, RPS25*	*TARDBP RPL6 BAT1*	*KARS*	-
Hh	-	*CAPNS1, RPL35*		*RPL35*
MAPK	*RPS24 , P4HB , DNAJB1 XBP1 , KDELR2 , SKP1 GNB1 , MGEA5*	*TAF10, NDRG1, GSN, DNAJA1, RHOC.*	-	-
JAK_STAT	*PRKAR1A , SF3B3*	*GDI2, RPL17 RPL32*	-	-
INT	*PARK7*	-	-	-
HGF	*RPL18 DAD1*	-	-	-
mTOR	*ERH*	-	-	-
AKT	-	*GDI2*	-	-
VEGF	-	-	-	-

WNT: Wingless signaling pathway; ***CCR***: Cytokine-cytokine receptor interaction; ***TGF***: Transforming growth factor beta pathway; ***Hh:*** Hedgehog signaling pathway; ***JAK_STAT:*** Janus kinase/signal transducers and activators of transcription (JAK/STAT); ***MAPK:*** Mitogen-activated protein kinase; ***INT:*** Integrin Pathway; ***HGF:*** Hepatocyte Growth Factor; ***mTOR:*** Mammalian target of rapamycin pathway; ***VEGF***: vascular endothelial growth factor; **FC:** fold change; **UP:** up-regulated genes, **DN:** down-regulated genes.

The overall behavior of the dataset from hepatocyte differentiation [15] due to miR-122 over-expression can be visualized by the scatter plots (Figure 2) at different fold change thresholds viz. FC= 1.2 and FC=2. Figure 2a shows that the miR-122 had very little effect on the hepatocyte differentiation (0.3% DE genes) at FC=2 underlying the results of Tzur *et al*. [15].



2a. 2b.

Figure 2. A scatter plot illustrating the overall effect of miR-122 on hESCs differentiation at fold change (FC) 2, and 1.2. The distance between boundaries, linear to the diagonal is equal to the threshold fold change *Red dots: up-regulated genes; *Blue dots: down-regulated genes, *green dots: unchanged genes at threshold fold change.

The total numbers of DEG (at FC=1.2) in the hepatocyte differentiation data are more than the number of DEG in the selected pathways. The difference between DEG suggests the possibility of other pathways regulating hepatocyte differentiation and other biological process. On the other hand, very few genes are differentially expressed at FC=2 suggesting that miR-122 had no effect on hepatocyte differentiation.

Conclusion

Although several GSA methods exist for pathway analysis, very few of them consider prior biological knowledge of biological processes, diseases and related pathways. The challenge for a GSA method is to filter out pathways that are relevant to the expression dataset to be analyzed. Often biologists performing microarray analysis have a few specific pathways in mind in which they have particular interest in. Thus, occasionally it is sensible to investigate only few such pathways where previous knowledge is available.

This study describes a pathway analysis method (KB-GSA) for analysis of microarray data. The method is based on combining prior knowledge of pathways related to a specific biological process with simple statistical measure. KB-GSA is useful for visualizing microarray gene expression data with simple statistics and easy to understand pathway

analysis results. Special attention is given to visualization of the results to make the interpretation easy. The graphical output consists of a scatter plot showing the number of DEG at specified cutoff threshold for each investigated pathway. Overall, the proposed method provides biologists with an efficient novel strategy that uses prior knowledge of pathways for analysis of microarray data.

Future work

Future work will mainly cover the development the KB-GSA algorithm to analyze time series microarray data. Currently the KB-GSA is designed to analyze microarray dataset using fold change as a statistical measure. Results from different statistical measures, viz. t-like tests, can be compared. The tool is aimed at users with a biological background. However, currently, the algorithm requires knowledge of R-programming. The algorithm can be simplified and made interactive by providing a graphical user interface.

References

1. Bar-Joseph Z: **Analyzing time series gene expression data.** *Bioinformatics* 2004, **20:**2493-2503.

2. Curtis RK, Oresic M, Vidal-Puig A: **Pathways to the analysis of microarray data.** *Trends Biotechnol* 2005, **23:**429-435.

3. Nam D, Kim SY: **Gene-set approach for expression pattern analysis.** *Brief Bioinform* 2008, **9:**189-197.

4. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proc Natl Acad Sci U S A* 2001, **98:**5116-5121.

5. Rivals I, Personnaz L, Taing L, Potier MC: **Enrichment or depletion of a GO category within a class of genes: which test?** *Bioinformatics* 2007, **23:**401-407.

6. Abatangelo L, Maglietta R, Distaso A, D'Addabbo A, Creanza TM, Mukherjee S, Ancona N: **Comparative study of gene set enrichment methods.** *BMC Bioinformatics* 2009, **10:**275.

7. Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC: **A global test for groups of genes: testing association with a clinical outcome.** *Bioinformatics* 2004, **20:**93-99.

8. Barry WT, Nobel AB, Wright FA: **Significance analysis of functional categories in gene expression studies: a structured permutation approach.** *Bioinformatics* 2005, **21:**1943-1949.

9. Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ: **Discovering statistically significant pathways in expression profiling studies.** *Proc Natl Acad Sci U S A* 2005, **102:**13544-13549.

10. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci U S A* 2005, **102:**15545-15550.

11. Subramanian A, Kuehn H, Gould J, Tamayo P, Mesirov JP: **GSEA-P: a desktop application for Gene Set Enrichment Analysis.** *Bioinformatics* 2007, **23:**3251-3253.

12. *Turkish journal of electrical engineering & computer sciences : TJEECS*. Ankara, Turkey : Scientific and Technological Research Council of Turkey (TÜB\0130TAK), 2009-.

13. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes.** *Nucleic Acids Res* 2000, **28:**27-30.

14. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y: **KEGG for linking genomes to life and the environment.** *Nucleic Acids Res* 2008, **36:**D480-484.

15. Tzur G, Levy A, Meiri E, Barad O, Spector Y, Bentwich Z, Mizrahi L, Katzenellenbogen M, Ben-Shushan E, Reubinoff BE, Galun E: **MicroRNA expression patterns and function in endodermal differentiation of human embryonic stem cells.** *PLoS One* 2008, **3:**e3726.

16. Chang J, Nicolas E, Marks D, Sander C, Lerro A, Buendia MA, Xu C, Mason WS, Moloshok T, Bort R, et al: **miR-122, a mammalian liver-specific microRNA, is processed from hcr mRNA and may downregulate the high affinity cationic amino acid transporter CAT-1.** *RNA Biol* 2004, **1:**106-113.

17. McLin VA, Rankin SA, Zorn AM: **Repression of Wnt/beta-catenin signaling in the anterior endoderm is essential for liver and pancreas development.** *Development* 2007, **134:**2207-2217.

18. Schmidt C, Bladt F, Goedecke S, Brinkmann V, Zschiesche W, Sharpe M, Gherardi E, Birchmeier C: **Scatter factor/hepatocyte growth factor is essential for liver development.** *Nature* 1995, **373:**699-702.

19. Sicklick JK, Li YX, Jayaraman A, Kannangai R, Qi Y, Vivekanandan P, Ludlow JW, Owzar K, Chen W, Torbenson MS, Diehl AM: **Dysregulation of the Hedgehog pathway in human hepatocarcinogenesis.** *Carcinogenesis* 2006, **27:**748-757.

20. Shen MM: **Nodal signaling: developmental roles and regulation.** *Development* 2007, **134:**1023-1034.

21. Kung JW, Currie IS, Forbes SJ, Ross JA: **Liver development, regeneration, and carcinogenesis.** *J Biomed Biotechnol* 2010, **2010:**984248.

22. Pires-daSilva A, Sommer RJ: **The evolution of signalling pathways in animal development.** *Nat Rev Genet* 2003, **4:**39-49.

23. Gerhart J: **1998 Warkany lecture: signaling pathways in development.** *Teratology* 1999, **60:**226-239.

24. Tanimizu N, Miyajima A: **Notch signaling controls hepatoblast differentiation by altering the expression of liver-enriched transcription factors.** *J Cell Sci* 2004, **117:**3165-3174.

Appendix 1.

R-code and tutorial for KB-GSA

The appendix describes how the KB-GSA algorithm works. The functionality of the KB-GSA is demonstrated using hepatocyte differentiation dataset available through GEO (GSE13460). The experiment contains 2 normal samples while 3 sample expression mutant miR-122 as described in Tzur *et al.* [15]. The script and input material can be downloaded from **http://tinyurl.com/38mmdrt .**

Input: expression data set, pathway genes and fold change (cutoff threshold)

Output: differentially expressed genes (percentage and individual genes)

- The analysis starts with uploading expression dataset.

```
#loading expression data
        > data_exp <- read.table("processed_data.txt",
        sep = "\t", header = T, as.is = T, dec = ".")
```

- The next step is to calculate mean of expression values for both samples. Since the data available at GEO is already log2 transformed, for calculating mean it was transformed to normal scale.

```
#calculating mean for MUT and WT

  #for MUT samples
        >xMUT<-grep("^MUT", colnames(data_exp))
        >xxMUT<-2^data_exp[,xMUT]

        >meanMUT<-apply(xxMUT,1,mean)

        # adding mean column to data_exp
        >data_exp["meanMUT"]<-meanMUT


  #for WT samples
        >xWT<-grep("^WT", colnames(data_exp))
        >xxWT<-2^data_exp[,xWT]
        >meanWT<-apply(xxWT,1,mean)

  # adding mean column to data_exp
        >data_exp["meanWT"]<-meanWT
```

- Log2 transformation of mean values

```
  #log transformation
        >logMUT<-log2(meanMUT)
        >data_exp["logMUT"]<-logMUT
```

```
>logWT<-log2(meanWT)
>data_exp["logWT"]<-logWT
```

- Second input to the algorithm is the genes of pathway to be analyzed.

```
>x11()

#2. laoding pathway
>data_pathway <- read.table("gene_WNT.txt",
sep = "\t", header = T, as.is = T, dec = ".")
```

- The algorithm then plots identifies common genes between expression dataset and pathway genes

```
#3.WHICH: printing mean for common genes by WHICH
>index_nor_GS<-
which(data_exp[,"gene_symbol"]%in%data_pathway[,1])
```

- Algorithm then plots all genes and common genes in different colors

```
#4 PLOTTING graphs
```

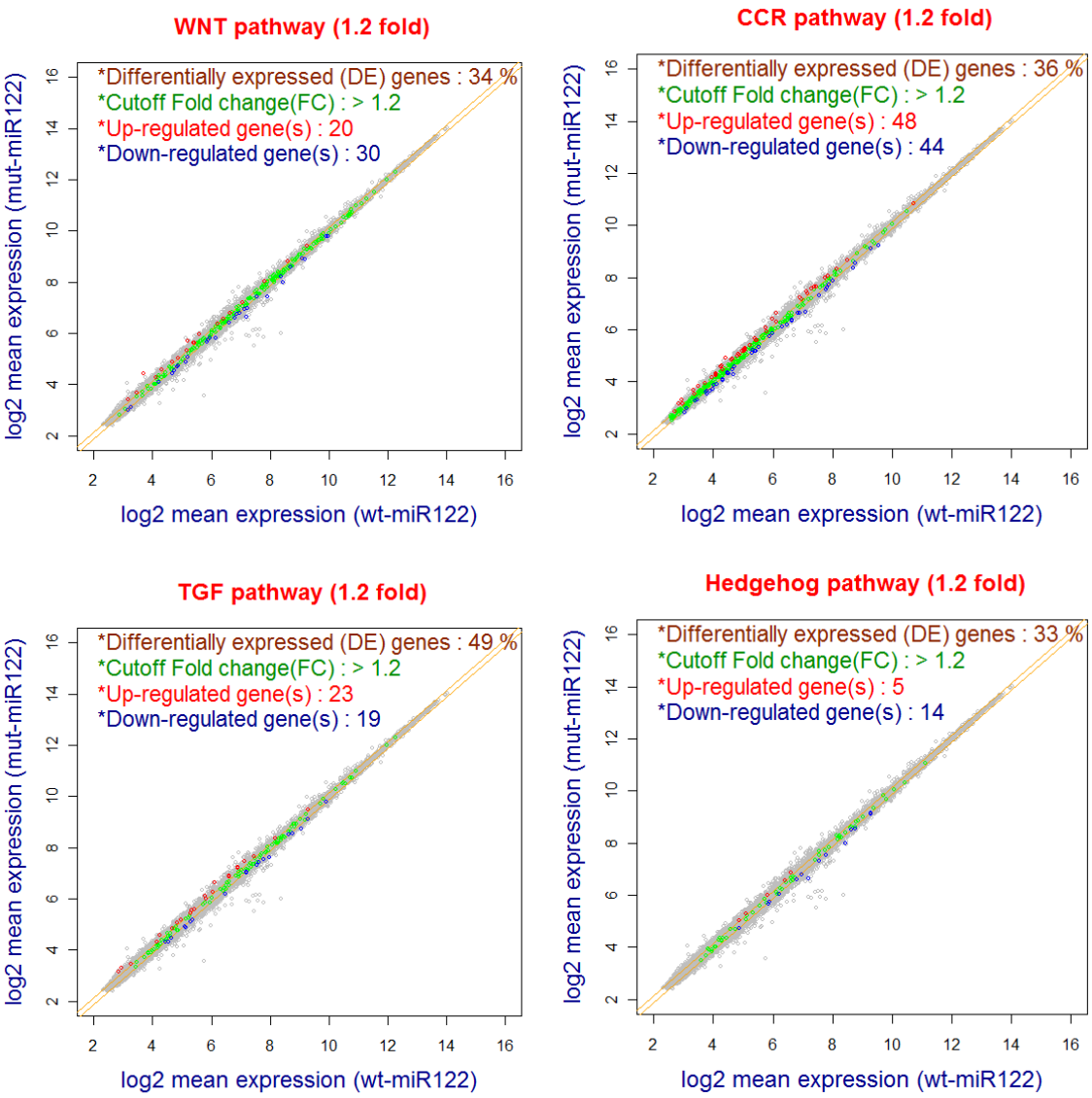
- Here we set cutoff threshold

```
#setting cutoff threshold
   >threshold<-log2(2)

#plotting general GRAY PLOT

> plot(logWT,logMUT, col = grey",cex=0.6,xlim=c(2,16),ylim=c(2,16),
  main=" WNT pathway (2 fold)", col.main="red",
col.sub="blue",xlab="log2 mean expression (wt-miR122)",
  ylab="log2 mean expression (mut-miR122)", col.lab="blue4",
cex.lab=1.5, cex.main=1.5)


>abline(a=threshold/2 , b = 1, col = "orange",lwd=1.8) # UP LINE

>abline(a=-threshold/2 , b = 1, col = "orange",lwd=1.8)#DOWN LINE


#PLOTING mut and wt values in gray plot
> x1 <- data_exp[index_nor_GS,"logWT"]
> y1 <- data_exp[index_nor_GS,"logMUT"]

> points(x1,y1,col="green",cex=0.6)

#ploting up reg
> nor_GS<-data_exp[index_nor_GS,]

> nor_GS_FC_WT_MUT<-nor_GS[,"logMUT"]-nor_GS[,"logWT"]

> upReg<-which(nor_GS_FC_WT_MUT>threshold/2) #setting threshold condt

> x2 <- nor_GS[upReg,"logWT"]          #x coordinate for up
> y2 <- nor_GS[upReg,"logMUT"] #y coordinates

> points(x2,y2,col="red",cex=0.6) #plotting up in red

#plotting down reg
```

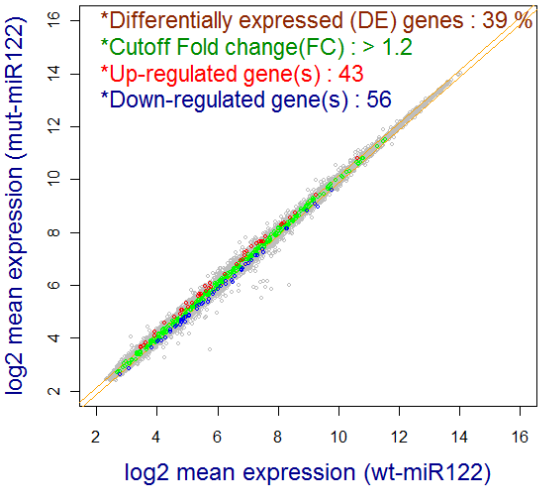
```
> downReg<-which(nor_GS_FC_WT_MUT<(threshold/2*-1))
> x2 <- nor_GS[downReg,"logWT"]
> y2 <- nor_GS[downReg,"logMUT"]

> points(x2,y2,col="blue",cex=0.6)

#printing info

> percent<-((length(upReg)+length(downReg))/nrow(data_pathway))*100

> text(1.8,16,paste("*DE genes :",length(upReg)+length(downReg),"out
of",nrow(data_pathway),"genes (", format(percent, digit=1),"%",")"),pos=4,cex=1.4,
col="orangered4")

> text(1.8,15,paste( "*Cutoff Fold change(FC) : ",2^threshold),pos=4,cex=1.4,
col="green4")

> text(1.8,14,paste("*Up-regulated gene(s) :",length(upReg)),pos=4,cex=1.4,
col="red")

> text(1.8,13,paste("*Down-regulated gene(s) :",length(downReg)),pos=4,cex=1.4,
col="blue4")

#printing differentially expressed genes.
print("UP")

data_exp[upReg,"gene_symbol"]

print("Down")

data_exp[downReg,"gene_symbol"]
```

Appendix 2

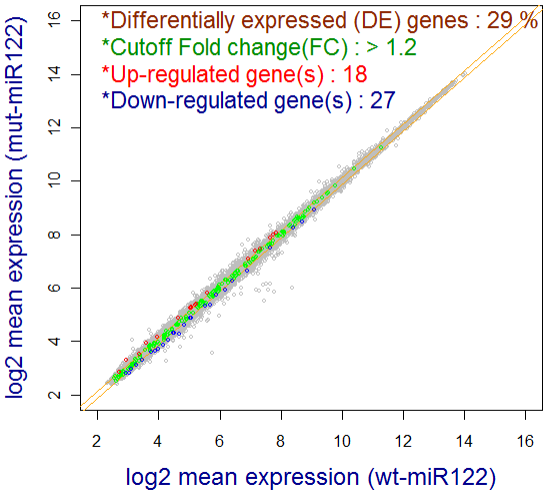
Graphical representation of KB-GSA output for the tested pathways



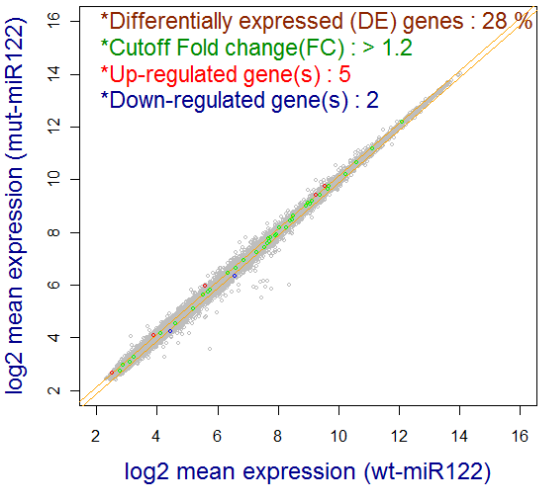
MAPK pathway (1.2 fold)



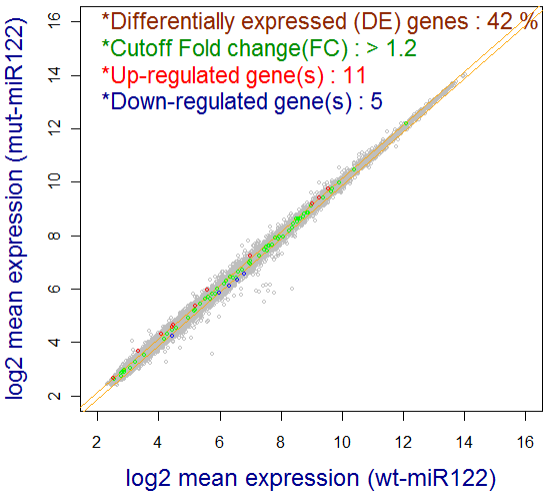
JAK_STAT pathway (1.2 fold)



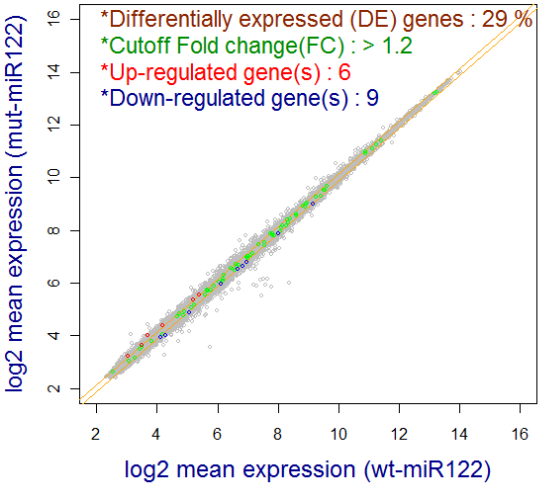
INT pathway (1.2 fold)



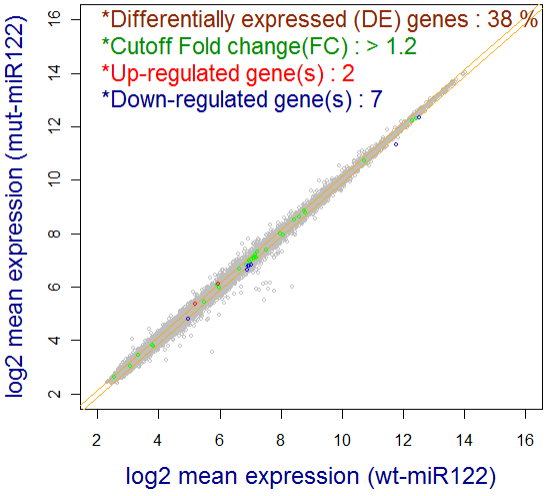
HGF pathway (1.2 fold)

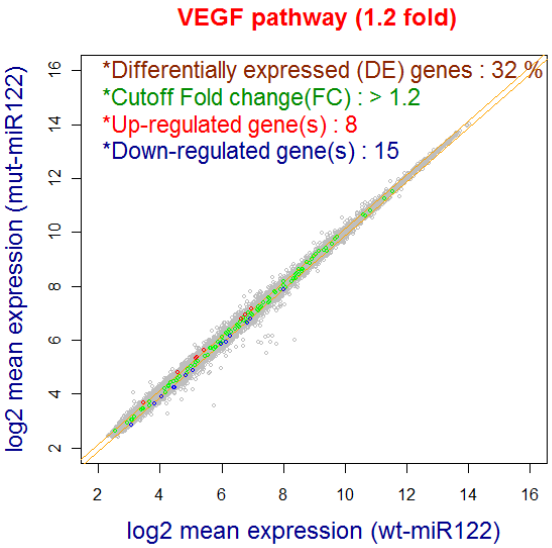


mTOR pathway (1.2 fold)



AKT pathway (1.2 fold)




VEGF pathway (1.2 fold)
*Differentially expressed (DE) genes : 32 %
*Cutoff Fold change(FC) : > 1.2
*Up-regulated gene(s) : 8
*Down-regulated gene(s) : 15
log2 mean expression (mut-miR122)
log2 mean expression (wt-miR122)
2
4
6
8
10
12
14
16