Vignette for "Transcriptomic Meta-analysis Reveals Up-regulation of Gene Expression Functional In Osteoclast Differentiation in Human Septic Shock"

Samanwoy Mukhopadhyay and Saroj Kant Mohapatra*

January 28, 2017

Contents

1 Loading the necessary libraries and the data

This is a vignette for analysing the data of the manuscript titled "Up-regulation of Osteoclast Differentiation is Associated with Septic Shock". Meta-analysis of publicly available gene expression data sets reveals up-regulation of osteoclast differentiation pathway associated with septic shock. For easy reproducibility of the analysis described in that manuscript, an R data package **nibmgss** has been created. The basic steps for data analysis on this package are described below. Electronic search was performed on medical literature and gene expression databases. Selection of studies was based on the organism (human subjects), tissue of origin (circulating leukocytes) and the platform technology (gene expression microarray) [Fig.1]. Gene-level meta-analysis was conducted on the six selected studies to identify the genes consistently differentially expressed in septic shock. These genes were then subjected to pathway analysis. Fig.2 depicts the flowchart of the analysis plan.

Some preliminaries before we start. Go to the directory with data and set it as a working directory.

```
> datapath <- "./"
```

*skm1@nibmg.ac.in

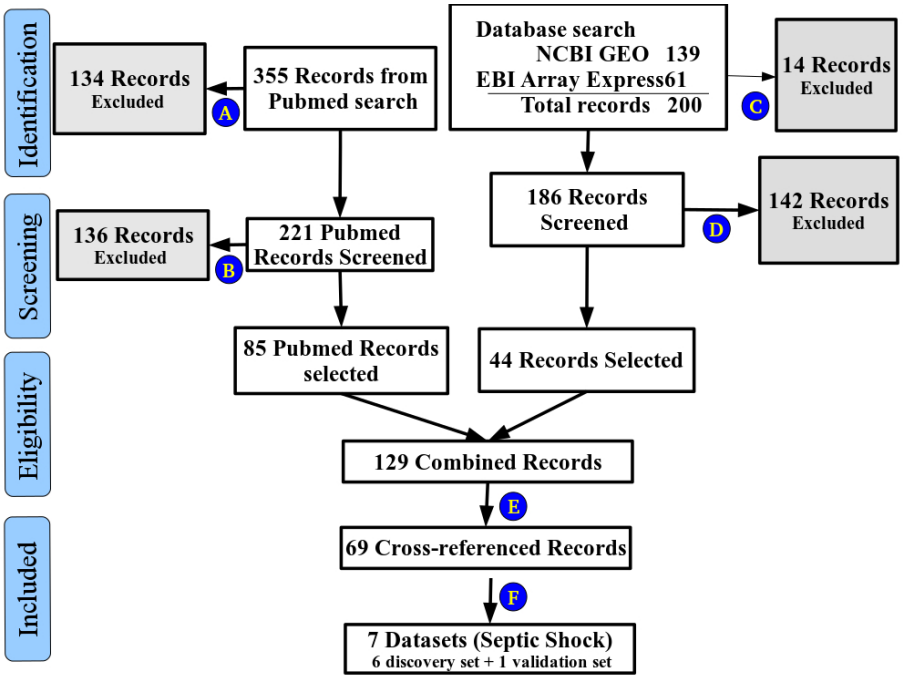

Figure 1: Selection of studies

*Source*ing the "prelim.R" file loads libraries and some useful function definitions. You may look in for the details.

```
> source("Rcode/prelim.R")
```

Now, load the package with the *library* command and get the expression set object *ss.eset*. *allegs* includes all entrez ids for which there are expression data in septic shock.

```
> library("ssnibmg")
> data("ss.eset")
> allegs <- featureNames(ss.eset[[1]])
```

Now the required libraries and the processed transcriptome data are loaded in to the R environment. Let us first see the structure of the expression data matrix (first five rows and first five columns).

```
> head(exprs(ss.eset[[1]])[1:5,1:5])
```

```
       GSM350139 GSM350140 GSM350141 GSM350142 GSM350143
1          0.869     0.986     0.918     0.890     0.987
10         1.034     1.335     1.059     1.000     0.698
100        0.992     0.695     0.993     1.123     1.436
1000       0.903     1.148     0.860     1.245     0.849
10000      0.980     0.672     1.159     0.852     0.924
```


```
       GSM350139 GSM350140 GSM350141 GSM350142 GSM350143
1          0.869     0.986     0.918     0.890     0.987
10         1.034     1.335     1.059     1.000     0.698
100        0.992     0.695     0.993     1.123     1.436
1000       0.903     1.148     0.860     1.245     0.849
10000      0.980     0.672     1.159     0.852     0.924
```


The rownames of the expression matrix contains Enterez Gene IDs and the columns contain the gene expression data across different samples.

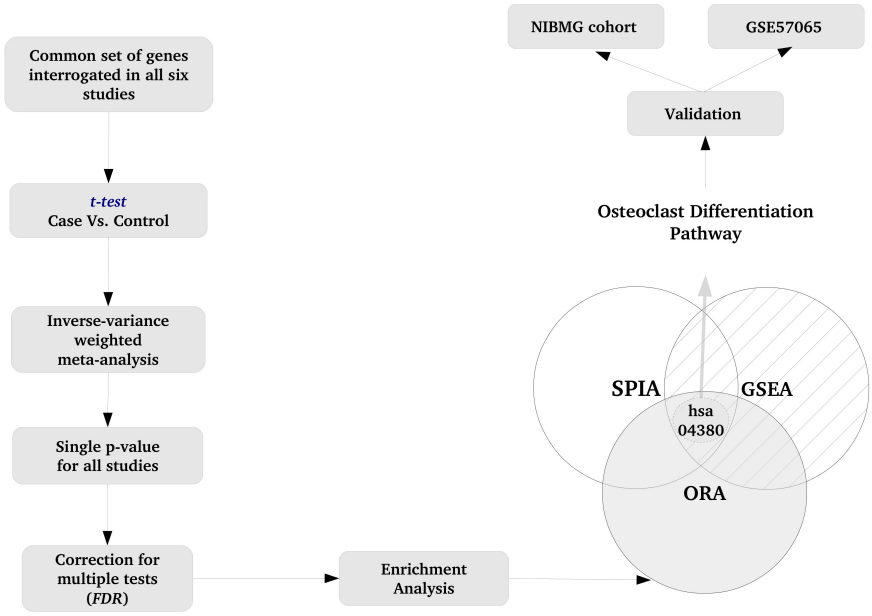

Figure 2: Flowchart of Analysis plan

2 Gene-level Meta-analysis

Differential expression was measured in terms of both log-fold change and p-value. For each gene, SS was compared with control and the six p-values were combined and adjusted for multiple testing (refer to the code below for details on meta-analysis) to generate a single p-value per gene. For each gene, the six log-fold changes were averaged to produce a single log-fold change. Using stringent criteria (adjusted p-value < 0.01, fold change of 2 or more), we discovered 200 genes that were consistently up-regulated in SS. As noted in Fig. 3, there are more up-regulated (than down-regulated) genes in these data sets.

```
> source("Rcode/runGenelevelMetaanalysis.R")
```

```
[1] "Now analysing study  GSE13904"
[1] "Now analysing study  GSE26378"
[1] "Now analysing study  GSE26440"
[1] "Now analysing study  GSE4607"
[1] "Now analysing study  GSE8121"
[1] "Now analysing study  GSE9692"
```

3 Analysing the data: Pathway enrichment analysis

Load the pathway annotation data from KEGG.

```
> data("genes.by.pathway")
> data("pathways.list")
```

3.1 Over Representation Analysis

First we perform **ORA** (Over Representation Analysis) on 200 up-regulated genes obtained from the previous step by applying hypergeometric test. The basic idea is as follows. Let us consider two lists of genes: the first list being the set of up-regulated genes, and the second being the member genes of a given KEGG pathway. The task is to find out if genes belonging to this pathway are also likely to be part of the list of up-regulated genes. This is captured in a 2×2 contingency table as shown in Table 1. Further details may be found in the code file "runORA.R".

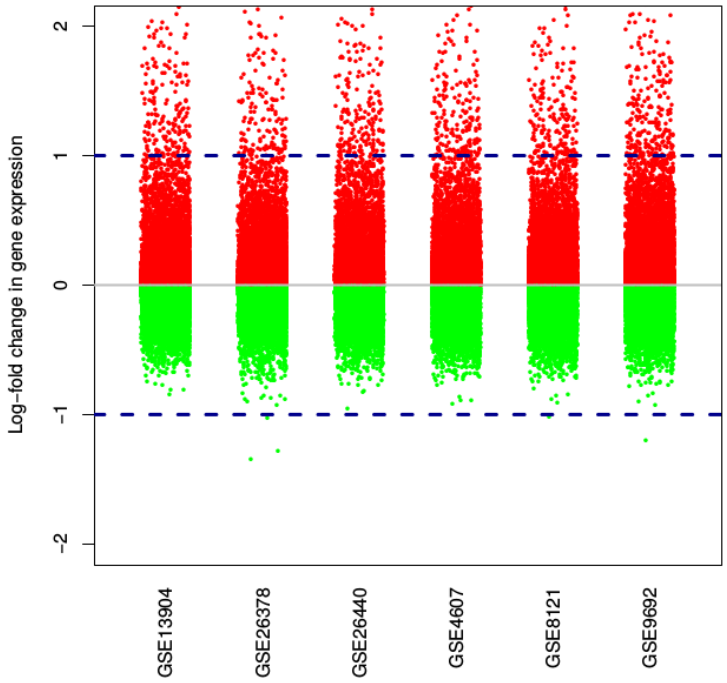

Figure 3: Simple Dotplot of logfold change in each of the six Studies. The dotted horizontal lines refer to two-fold change in either direction.

	Number of genes in the given KEGG pathway	Number of genes NOT in the given KEGG pathway
Number of genes up-regulated in SS	n_{11}	n_{12}
Number of genes NOT up-regulated in SS	n_{21}	n_{22}

Table 1: **A** 2×2 **contingency table.** The table shows the four quantities of interest while estimating if a pathway is over-represented (enriched) among the set of differentially expressed genes.

```
> source("Rcode/runORA.R")

[1] Over-representation analysis: using KEGGREST
                p odds expected
hsa04610 5.87e-06 11.5    0.692
hsa04380 5.38e-05 6.64     1.31
hsa05202 0.000461 4.77      1.8
                                                                    Name
hsa04610     Complement and coagulation cascades - Homo sapiens (human)
hsa04380              Osteoclast differentiation - Homo sapiens (human)
hsa05202 Transcriptional misregulation in cancer - Homo sapiens (human)
```

The pathway ids are saved to an object named "keggids.ora".

```
> keggids.ora <- rownames(oraGene2Kegg.up[oraGene2Kegg.up[,1]<0.001,])
```

3.2 Gene Set Enrichment Analysis

Next we performed **GSEA** (Gene Set Enrichment Analysis) GSEA calls upon a global (i.e., genome-wide, not limited to any pre-selected list) search strategy to detect the KEGG pathway(s) with significant

up-regulation in SS compared to control. It takes about 10 minutes on a reasonably powered computer, and you may want to save the result to a file, and read it from there in subsequent sessions (read the code in the file "runGSEA.R" for more details).

```
> source("Rcode/runGSEA.R")
```

The pathways returned significant by GSEA are saved to an object named "keggids.gsea".

```
> keggids.gsea <- names(which(gseaFp.up==0.0))
```

3.3 Signaling Pathway Impact Analysis

SPIA (Signaling Pathway Impact Analysis) on the data. SPIA combines elements of ORA and GSEA, with attention to gene-gene interactions and pathway topology. It takes about 30 minutes on a reasonably powered computer, and you may want to save the result to a file, and read it from there in subsequent sessions (read the code in the file "runSPIA.R" for more details).

```
> source("Rcode/runSPIA.R")
```

The pathways returned significant by SPIA are saved to an object named "keggids.spia".

```
> keggids.spia <- paste("hsa",top10ids, sep="")
```

The three keggid lists are now intersected to identify the common pathway returned by all three methods (refer to the Venn diagram in Fig. 2).

```
> intersect(intersect(keggids.ora, keggids.gsea), keggids.spia)
```

```
[1] "hsa04380"
```

Intersection of three results gives us the single common significantly upregulated KEGG pathway *hsa04380* (Osteoclast Differentiation Pathway).

4 Validation Cohort

Let us load data from the validation cohort of SS patients.

```
> data("esetn.b")
> data("esetn")
```

Correction for batch effect has been performed (Fig. 4).

```
> par(mfrow=c(1,2))
> boxplot(exprs(esetn.b), main="Before correction", las=2)
> boxplot(exprs(esetn), main="After correction", las=2)
```

4.1 Permutation test for enrichment

Now we will perform a permutation-based enrichment test to provide evidence for over-all up-regulation of the pathway *hsa04380* in SS (validation cohort). For this, we are using the function *permutationTest* of the package **resample** to calculate the permutation-based p-value accounting for correlation among the pathway genes.

```
> source("Rcode/getPvalResample.R")
```

First, we calculate the proportion of significantly up-regulated ($p < 0.05$) genes in the pathway by using a two-sample t-test to test for up-regulation of each pathway gene. This is observed to be 0.447. Next, we reshuffle the sample groups (i.e. case/control status) 100000 times and similarly calculate the proportion of up-regulated genes for each permutation replicate. Finally, the permutation-based p-value is obtained as the proportion of replicates where the simulated proportion is greater than the observed value.

4.2 Generating case vs control scatterplot

We calculate the mean expression values (for the two groups: control and SS) and generate the scatterplot as shown in Fig.5.

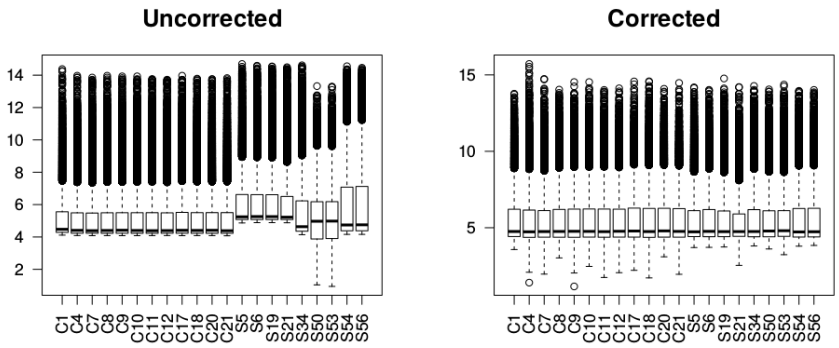

Figure 4: Plots showing correction for batch effect. In the right panel, the samples are all aligned at the median gene expression.

```
> source("Rcode/drawScatterplot.R")
```

In this plot, each point corresponds to a single gene. The points near the identity line (the diagonal in Fig. 5), correspond to genes with similar expression level in control and SS groups. The genes that are up-regulated in SS are expected to be significantly deviated from the diagonal toward the SS axis. Indeed, for most of the genes, there is much higher expression in SS, as shown in this scatterplot. The up-regulation of these genes in SS is statistically significant ($p = 0.00028$, permutation test for enrichment). Additional testing (for each gene; unpaired t-test between control and SS) reveals individual genes significantly up-regulated after multiple-testing correction at an FDR level of 0.05. An expression filter is applied to identify the genes that show a high fold-change (2 or more) and are expressed in significant amounts (intensity of 100 or more). These genes are shown in red on the plot.

4.3 Generating Boxplot of key 25 genes of the Osteoclast Dfferentiation pathway

On looking at individual genes, we find that 60 genes were FDR significant.

```
> which.lowp <- which(padjn<0.05)
> print(length(which.lowp))
```

```
[1] 60
```

There are 25 genes that are FDR significant and pass the expression filters.

```
> which.lowp.lowexp <- which(padjn<0.05 & xcon>log2(100) & xss-xcon>1)
> egs.toshow <- names(which.lowp.lowexp)
> print(length(egs.toshow))
```

```
[1] 25
```

List the 25 gene symbols.

```
> print(as.character(unlist(mget(egs.toshow, org.Hs.egSYMBOL))))
```

```
 [1] "MAPK1"   "MAPK3"   "PIK3CG" "IFNGR1" "IFNGR2" "IL1B"   "NFKBIA" "JUNB"
 [9] "NCF2"    "NCF4"    "OSCAR"  "LILRB2" "LILRA3" "LILRA2" "LILRA6" "FCGR1A"
[17] "FCGR2A" "SIRPA"   "TYROBP" "SYK"    "PLCG2"  "SPI1"   "IFNAR1" "IFNAR2"
[25] "GAB2"
```

```
> source("Rcode/drawBoxplot25genes.R")
```

Fig.6 shows the boxplots of the highly significant 25 genes of the pathway *hsa04380* up-regulated in SS. Green color corresponds to the control subjects while the red color corresponds to the cases of SS. Gene symbols are shown at the bottom. For each gene, log-intensity of gene expression has been normalized to the median expression of the control group.

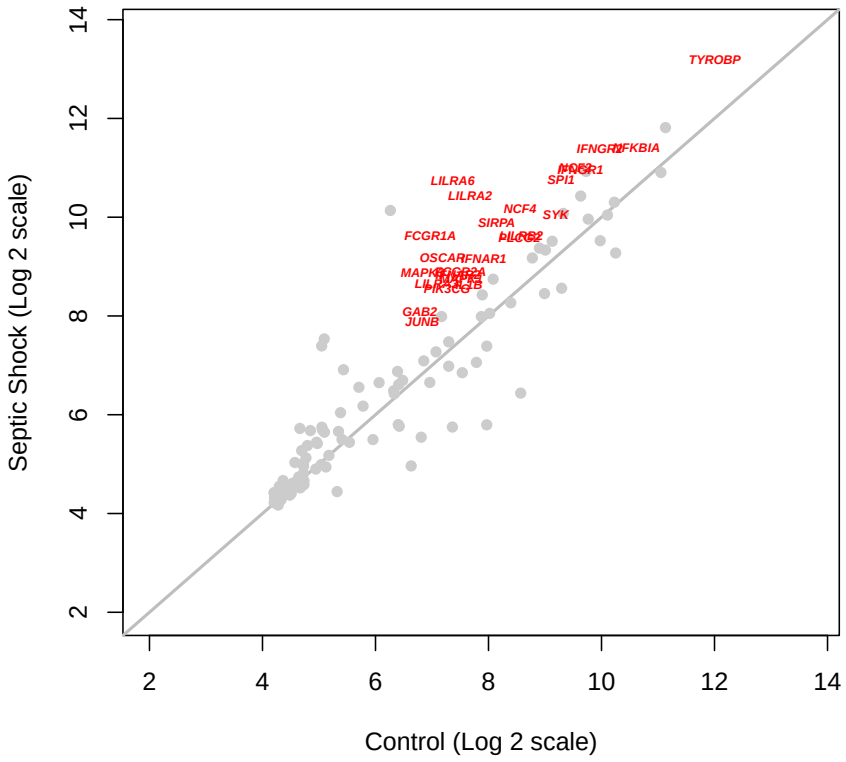

Figure 5: Scatterplot of the *hsa04380* pathway gene expression in NIBMG validation cohort

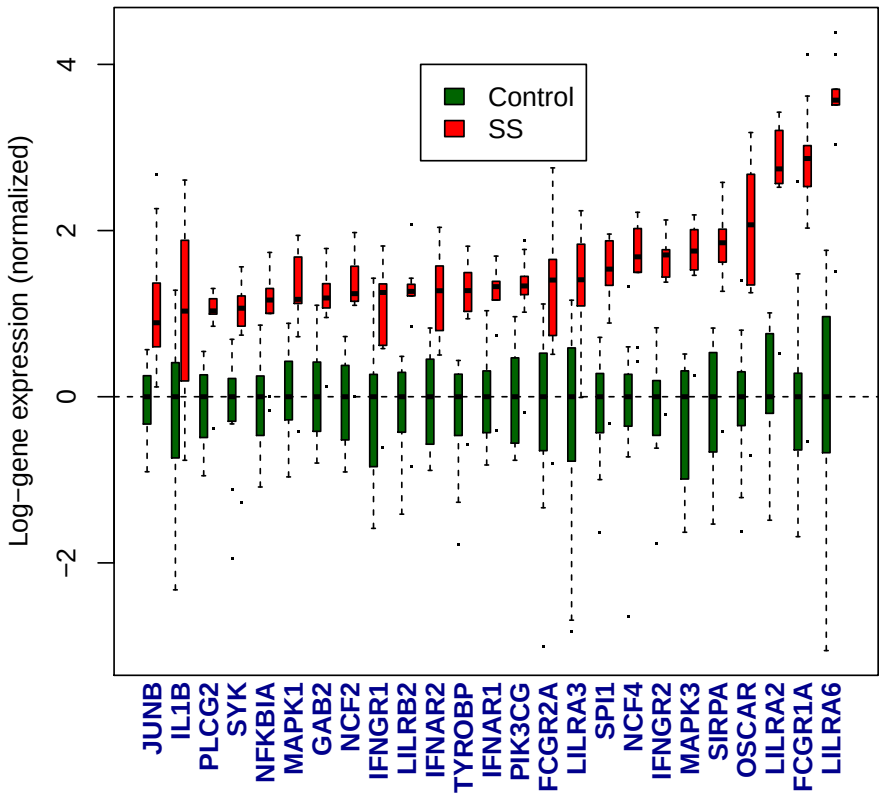

Figure 6: boxplot of key 25 the pathway genes

5 Validation of 25 genes in an Independent Validation Cohort

we start by loading the validation data and preprocessing the data before analysing it.

```
> data(gset)
> source("Rcode/analyse_validation_cohort.R")
```

We performed Principal Component Analysis of SS vs healthy controls in a new validation dataset, for the 25 significant genes from osteoclast differentiation pathway. Following is the code chunk that produces the $3D$ PCA plot [Fig.7]of control vs SS cases.

```
> pc<-prcomp(data.frame(t(dat)),scale=TRUE)
> mycol<-rep(c("red"),length(dx))
> mycol[which(dx=="Control")]<-c("green")
> pca3d(pc, components = 1:3, radius=3, col = mycol,title = NULL, new = FALSE,axes.color = "grey", b
```

```
[1] 0.432 0.264 0.184
Creating new device
```

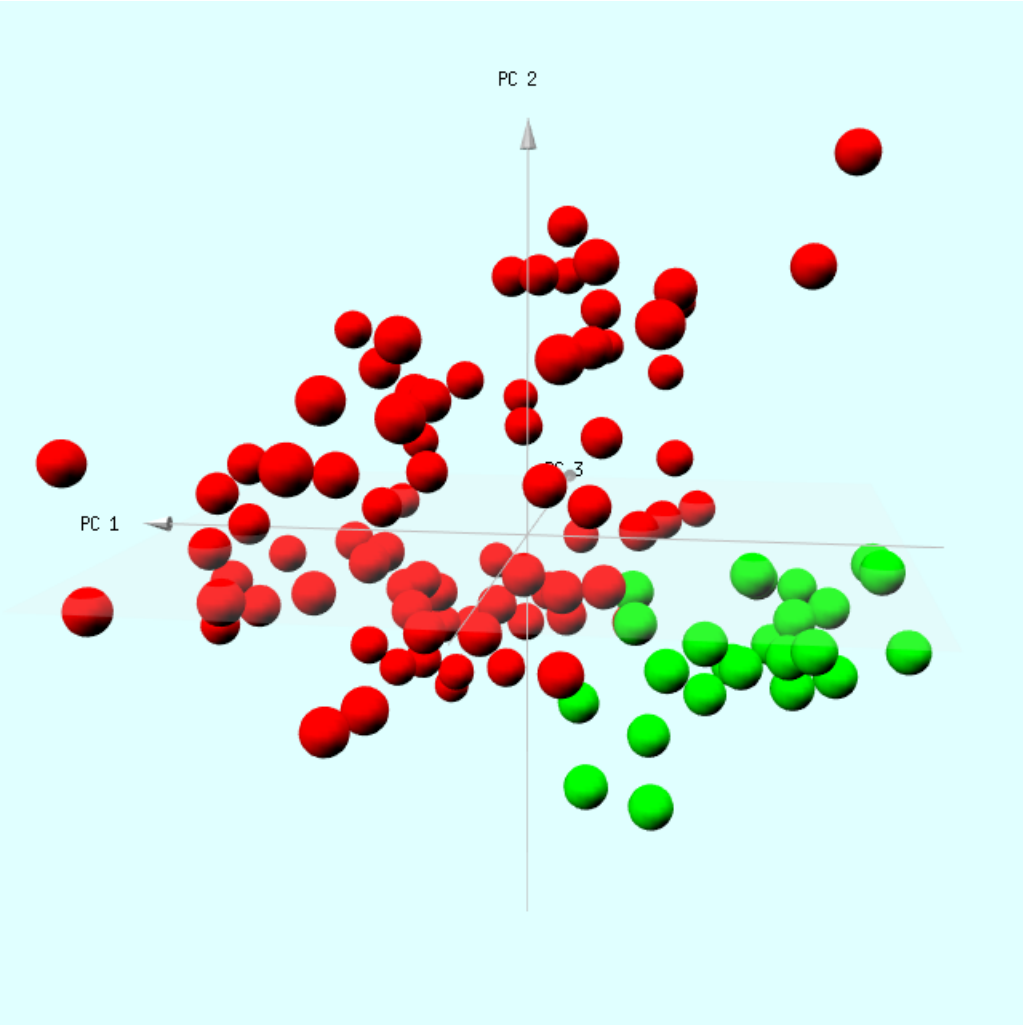

Figure 7: **3D PCA plot of control and cases in validation dataset GSE57065:** The SS cases are shown in color red and healthy controls in green

6 Survival Analysis with 25 genes

```
> source("Rcode/runSurvivalAnalysis.R")
```

The outcome information provided in study GSE4607, were subject to analysis for the expression of 25 selected genes from osteoclast differentiation pathway. Asterisk(*) suggests statistically significant ($p < 0.1$) difference in expression between the survival and non-survival SS cases.

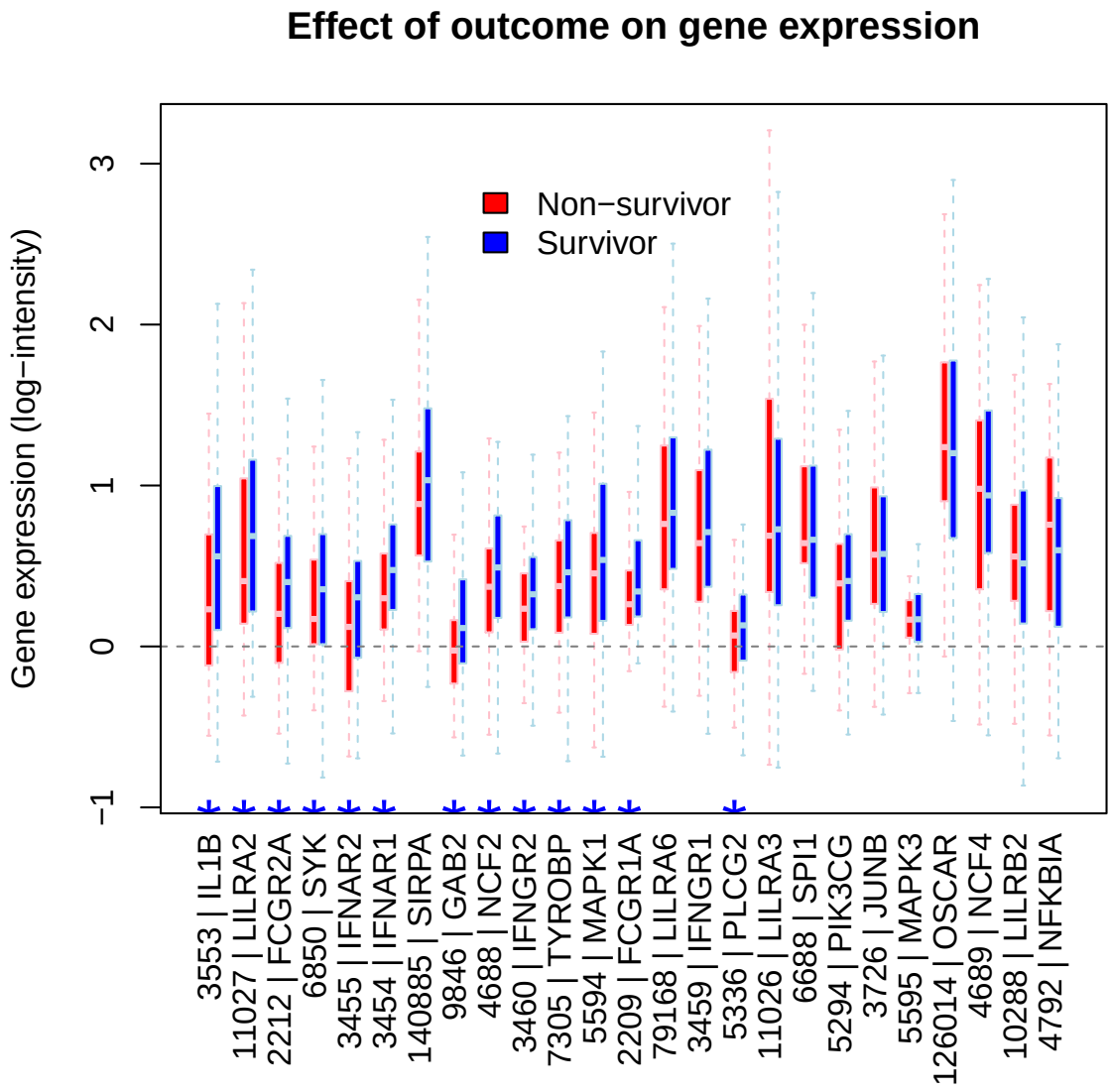

Figure 8: **Outcome-wise expression of 25 selected genes:**

7 Topology analysis of the selected genes

The 25 key genes were divided in to three groups based on location within the KEGG pathway: membrane, nucleus or intermediate. For each group, log-fold changes in gene expression were displayed as box plot Fig9

```
> prox.score <- read.delim(file="Metadata/egs25proximalityscore.txt", header=T, sep="\t")
> o <- order(prox.score[,4])
> egs25 <- as.character(prox.score[o,1])
> lfc1 <- lfc[egs25]
> keggsym25 <- as.character(prox.score[o,2])
> mycol <- as.character(prox.score[o,5])
> prox <- as.character(prox.score[o,"Position"])
> par(mar=c(12,5.5,2,2))
> plotdat <- split(lfc1, prox)
> b <- boxplot(plotdat[c("Membrane","Intermediate","Nucleus")],
 outline=F, col=c("red","pink","brown"), ylim=c(0,1.3),
 axes=F, ylab="Relative Gene Expression (log2 scale)")
>  box(col="gray")
>  axis(2)
>  mtext(text=c("Membrane","Intermediate","Nucleus"),
 side=1, line=0.5, at=1:3, col="darkblue",
 font=2, las=2, cex=2)
```

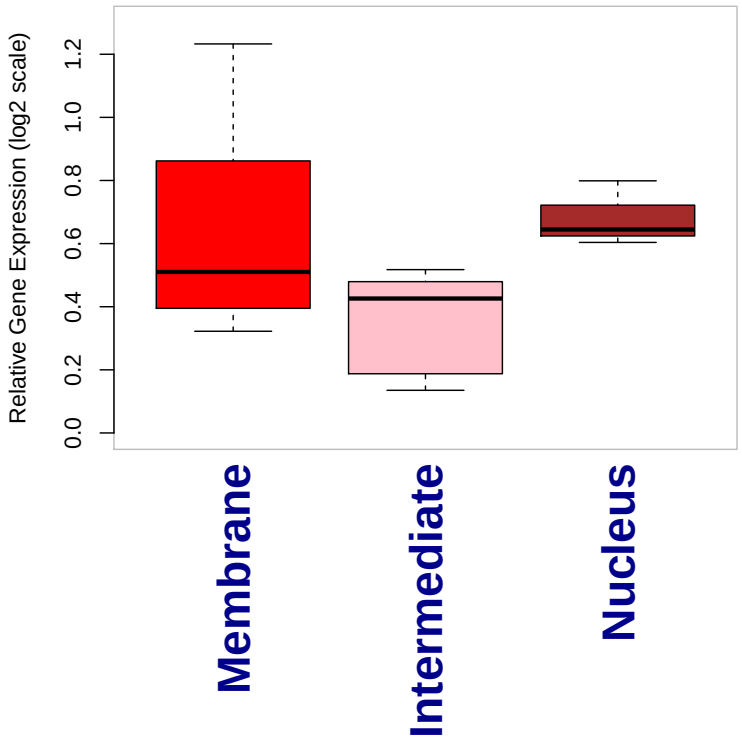

Figure 9: **Topology analysis of key 25 the pathway genes**

8 Acknowledgement

We acknowledge help from our laboratory colleagues at the National Institute of Biomedical Genomics, Kalyani for generation of data from subjects of septic shock, and Dr. Samsiddhi Bhattacharjee for generation of permutation-based p-value.

9 Session Information

```
> sessionInfo()

R version 3.2.3 (2015-12-10)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 16.04.1 LTS

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

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

other attached packages:
 [1] resample_0.4              ssnibmg_1.0
 [3] impute_1.42.0             sva_3.18.0
 [5] mgcv_1.8-8                nlme_3.1-124
 [7] stringr_1.0.0             gplots_3.0.1
 [9] pca3d_0.8                 rgl_0.95.1441
[11] KEGGprofile_1.12.0        KEGGREST_1.8.1
[13] SPIA_2.22.0               KEGGgraph_1.28.0
[15] Category_2.34.2           GO.db_3.1.2
[17] Matrix_1.2-3              GSEABase_1.30.2
[19] graph_1.46.0              annotate_1.46.1
[21] XML_3.98-1.3              illuminaHumanv2.db_1.26.0
[23] hgu133plus2.db_3.2.2      org.Hs.eg.db_3.2.3
[25] RSQLite_1.0.0             DBI_0.5-1
[27] AnnotationDbi_1.32.3      IRanges_2.4.8
[29] S4Vectors_0.8.11          genefilter_1.50.0
[31] limma_3.26.9              GEOquery_2.36.0
[33] Biobase_2.30.0            BiocGenerics_0.16.1

loaded via a namespace (and not attached):
 [1] gtools_3.5.0       splines_3.2.3      lattice_0.20-33    RBGL_1.44.0
 [5] survival_2.38-3    KEGG.db_3.1.2      zlibbioc_1.14.0    Biostrings_2.36.4
 [9] caTools_1.17.1     biomaRt_2.24.1     GenomeInfoDb_1.4.3 KernSmooth_2.23-15
[13] xtable_1.7-4       gdata_2.17.0       XVector_0.8.0      ellipse_0.3-8
[17] TeachingDemos_2.10 png_0.1-7          stringi_1.0-1      grid_3.2.3
[21] tools_3.2.3        bitops_1.0-6       magrittr_1.5       RCurl_1.95-4.7
[25] httr_1.0.0         R6_2.1.1
```