Journal of Biomedical Informatics 70 (2017) 77–84

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Analysis and modelling of septic shock microarray data using Singular Value Decomposition

Srinivas Allanki [a,*], Madhulika Dixit [a], Paul Thangaraj [b], Nandan Kumar Sinha [c]

[a] Laboratory of Vascular Biology, Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences Building, Indian Institute of Technology Madras, Chennai 600 036, India
[b] Department of Cardiothoracic Surgery, Apollo Hospital, Chennai 600 006, India
[c] Department of Aerospace Engineering, Indian Institute of Technology Madras, Chennai 600 036, India

ARTICLE INFO

Article history:
Received 13 January 2017
Revised 28 April 2017
Accepted 8 May 2017
Available online 10 May 2017

Keywords:
Septic shock
Microarray
Singular value decomposition
Microarray modelling
Predictive modelling

ABSTRACT

Being a high throughput technique, enormous amounts of microarray data has been generated and there arises a need for more efficient techniques of analysis, in terms of speed and accuracy. Finding the differentially expressed genes based on just fold change and p-value might not extract all the vital biological signals that occur at a lower gene expression level. Besides this, numerous mathematical models have been generated to predict the clinical outcome from microarray data, while very few, if not none, aim at predicting the vital genes that are important in a disease progression. Such models help a basic researcher narrow down and concentrate on a promising set of genes which leads to the discovery of gene-based therapies. In this article, as a first objective, we have used the lesser known and used Singular Value Decomposition (SVD) technique to build a microarray data analysis tool that works with gene expression patterns and intrinsic structure of the data in an unsupervised manner. We have re-analysed a microarray data over the clinical course of Septic shock from Cazalis et al. (2014) and have shown that our proposed analysis provides additional information compared to the conventional method. As a second objective, we developed a novel mathematical model that predicts a set of vital genes in the disease progression that works by generating samples in the continuum between health and disease, using a simple normal-distribution-based random number generator. We also verify that most of the predicted genes are indeed related to septic shock.

1. Introduction

With the advent of the microarray technology, it has become quite trivial to get a bird's-eye-view on the entire expression profile of a given sample at a genomic level. Being a high throughput technique, enormous amounts of data has been generated across the globe and there arises a need for more efficient techniques of analysis, both in terms of speed and accuracy. Computational analysis of microarray data is majorly divided into two types: Supervised and Unsupervised analysis methods [1,2]. Supervised methods are used when we have prior knowledge of the system. Unsupervised methods are exploratory in nature and are employed to cluster genes/samples into groups with similar expression profiles (e.g. Hierarchical clustering). However, these clustering methods have got their own drawbacks. They are usually not applied over the whole set of microarray genes, but are limited to a small subset of genes that are filtered using certain variables (fold change & p-value). This approach doesn't seem to be optimal, given that we are ruling out some of the vital biological signals that may work at rather lower amplitudes [3]. On the other hand, one cannot call it as a complete unsupervised analysis, unless the genes on the entire array are used for clustering, which requires better computational facility, which is not accessible for all. This opens the need for better unsupervised methods that analyse all the microarray genes to bring out the intrinsic structure of the data with minimal data loss and computational resources.

Singular Value Decomposition (SVD) is one such linear algebra technique. The use of SVD in microarray data analysis was first illustrated in Alter et al., 2000 by introducing the concept of eigengene and eigenarray, which are unique, independent and uncorrelated orthonormal superpositions of the genes and arrays, respectively. SVD is a linear transformation of the data matrix ($\hat{e}$) from a genes × arrays space to a reduced eigengenes × eigenarrays space, such that an eigengene is expressed only in the corresponding eigenarray with a corresponding eigenexpression level that

* Corresponding author at: Department III – Developmental Genetics, Max Planck Institute for Heart and Lung Research, Ludwigstrasse 43, Bad Nauheim 61231, Germany.
E-mail address: srinivas.allanki@mpi-bn.mpg.de (S. Allanki).

http://dx.doi.org/10.1016/j.jbi.2017.05.005
1532-0464/

gives their relative significance. To give a biological sense, each eigengene and eigenarray can be correlated with a regulatory process and a cellular state, respectively [4]. This algorithm decomposes/divides the data matrix into three different matrices as shown in Eq. (1). $\hat{u}$ represents the 'pattern' of genes in the eigenarrays, $\hat{\varepsilon}$ is a diagonal matrix that has the 'weights' or 'eigenexpression levels' $(\varepsilon_1, \varepsilon_2, \ldots \varepsilon_n)$ of eigengenes in their respective eigenarrays in a descending order and $\hat{v}^T$ represents the 'contribution' of each eigengene in the actual experimental arrays. An in-detail explanation of the SVD algorithm is given in Alter et al., 2000.

$$\hat{e}_{n\times m} = \hat{u}_{n\times L}\hat{\varepsilon}_{L\times L}\hat{v}_{L\times m} \quad (1)$$

where n = number of genes, m = number of arrays and $L = \min(n, m)$.

Till date, SVD was used to analyse several microarray datasets [5,6]. These can be split into two major kinds, namely: Time series and Static. Time series is where each sample corresponds to a discrete time point and Static is where they correspond to a different tissue/blood sample [2]. Alter et al., 2000 used SVD in a time series setup to 'remove the basal noise' in data by eliminating the eigenarrays and eigengenes that correspond to experimental noise and then to identify the temporal gene expression patterns. On the other hand, Simek et al., 2003 used SVD in a static setup as a pre-analysis step to 'filter out the most variable genes' based on the expression patterns given by the most significant eigengenes and eigenarrays, across the samples [7]. In this article, the primary objective is to device a novel method of clubbing the two approaches mentioned above to 'remove the basal noise' and then 'filter out the differentially expressed genes', both using SVD. We propose that, this makes it a better unsupervised approach.

The second objective is to develop a simple and novel mathematical model that compares healthy vs disease samples' microarray data and predicts a set of the most probable genes that play a vital role in the disease progression, using the SVD method proposed. Considering only the gene expression values and comparing health and disease, a particular gene's expression can increase (upregulated), decrease (downregulated) or can stand still. Now, comparing the complete array in both healthy and disease state, all the genes can possibly have one of the above mentioned three fates. Using this statement, we hypothesize that a disease sample's gene expression values can be extrapolated from a healthy/low disease samples'. We initially feed the clinically obtained low and high disease samples' gene expression data into the model. It works by generating samples that hypothetically represent intermediate states of disease progression between low and high disease state by extrapolating (increasing or decreasing) the gene expression values of low disease samples step-by-step until they are similar to high disease samples. We use a random number generator to generate random extrapolated gene expression values that are governed by the normal distribution curve obtained from the mean and standard deviation of each individual probe in the array, across low disease samples. We make use of the time course in septic shock microarray data from Cazalis et al., 2014 to validate our proposed method of microarray analysis and to predict the set of the most vital genes that are involved in the progression of septic shock using our model [8].

2. Methods

2.1. Dataset and conventional gene expression analysis

Cazalis et al., 2014 monitored the genome-wide mRNA levels from the Peripheral Blood Mononuclear Cells (PBMCs) of 28 septic shock subjects (median age 62) at three time points (0 h, 24 h and 48 h) from the time of getting admitted in the ICU and the beginning of vasopressor therapy. Blood samples from 25 healthy volunteers (median age 48) were also collected as control samples. The severity of the disease was analysed based on the median of SAPSII value across all the 28 subjects (>45 were regarded as SAPSII-high; <45 as SAPSII-low). Microarray was performed by them using GeneChip Human Genome U133 Plus 2.0 arrays (Affymetrix, Sta. Clara, CA, USA) and we collected the data from the public database, GEO Data sets under the accession number GSE57065 [9].

In order to validate our method, we preliminarily checked if SVD could pick similar set of differentially expressed genes as in the conventional method. We define the efficiency (E) of our method to be the percentage of common genes in both the analyses, as shown in Eq. (2). Since the list of all the differentially expressed gene sets weren't publicly available, we did the conventional microarray analysis using GEO2R tool on GEO database following the same cut-offs and parameters as in Cazalis et al., 2014.

$$\text{Efficiency (E)} = \frac{\text{No. of genes common in both analyses}}{\text{No. of genes conventional in both analyses}} \times 100 \quad (2)$$

In the original article, gene expression comparisons were done in two ways, with the progression of time (H0, H24 & H48) and the severity of disease (SAPSII-high & SAPSII-low). Considering the comparisons with progression of time, samples corresponding to each time point were first normalized with the same set of 25 control samples, as shown in Fig. 1B and the genes that are differentially expressed in all the three comparisons were noted. We think that this 'control-based normalization' approach masks some of the vital temporal changes. In this article, we propose to first normalize H0 with the controls (C vs H0), thereby eliminating all the basal differences between healthy and initial stage of septic shock and then compare the other two time points (H24 & H48) with H0. So, the gene set CG1 as shown in Fig. 1A, will have only the genes that are temporally regulated (H24 vs H0 and H48 vs H0) and are relevant to sepsis compared to controls (C vs H0). Since all the three samples (H0, H24 & H48) are from the same subject, we think that normalizing H24 and H48 with the zero time point sample (H0) would give out a better temporal comparison. We further refer to this as 'H0-based normalization' approach.

2.2. Microarray analysis algorithm using SVD

Method of solving the SVD equation is adopted from Alter et al., 2000 and is used as a tool in the proposed algorithm [4]. Four main steps are involved in the algorithm, as shown in Fig. 2. First, the raw data is normalized using the standard RMA method (Robust Multi-array Average) and baseline transformed to the median of the all samples. In the second step, we apply SVD on the data matrix with the groups of samples placed horizontally in the columns. We then calculate the fraction of eigenexpression level (f) of each eigengene in its own eigenarray using Eq. (3) and check for the value of Shannon entropy ($0 < d < 1$) using Eq. (4). $D = 0$

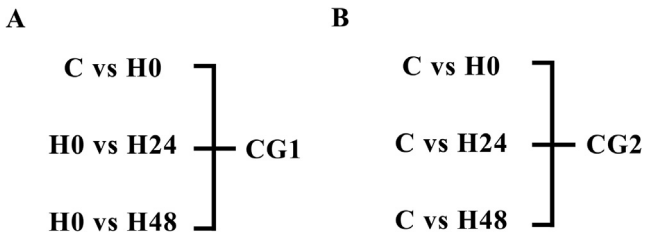

Fig. 1. Types of comparisons. Schematic representation of A. H0-based normalization B. Control-based normalization. CG1 and CG2 are the common genes that are differentially expressed in the respective types of comparisons.

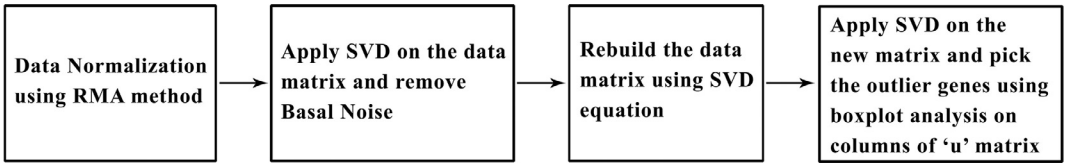

Fig. 2. Flowchart of the algorithm for the proposed SVD method.

represents a redundant dataset in which all the eigenexpression is represented by one single eigengene and $d = 1$ corresponds to a dataset where all the eigenexpression levels are equal [4]. We assume that the percentage of differentially expressed genes is significantly less than the genes that didn't show any change in expression. Since SVD clusters the genes based on the expression patterns across the samples, all the genes that didn't show a significant change in expression will more or less have a similar expression pattern. Therefore, most of the basal noise is represented by the first eigengene. To remove the basal noise, we substitute 'zero' in place of the first eigenexpression level in the $(\hat{\varepsilon})$ matrix, by which we are essentially subtracting the basal noise from all the genes. In the third step, we rebuild the new data matrix $(\widehat{e'})$ by using the SVD equation from Eq. (1).

Now each element of this newly rebuilt data matrix is the standard deviation corresponding to their respective genes. We calculate the log of each element's variance i.e. $\widehat{e''} = \log(\widehat{e'^2})$. We then re-decompose this new data matrix $(\widehat{e''})$ using SVD. We check for f and d again. Now, the *d-value* should be closer to one and therefore the first eigengene has the most significant differentially expressed genes. To filter these genes, the fourth step employs boxplot analysis on the first column of the new $\widehat{u''}$ matrix, a similar approach to Simek's cut-off method [7]. One limitation of this method is that it cannot distinguish between an upregulated and a downregulated gene, since the sign of the samples is lost in one of the data-transformation steps above.

$$f_1 = \varepsilon_l^2 \Big/ \sum_{k=1}^{n} \varepsilon_k^2 \tag{3}$$

$$d = \frac{-1}{\log(L)} \sum_{k=1}^{L} f_k \cdot \log(f_k) \tag{4}$$

where L = min(n, m).

2.3. Prediction model algorithm

Consider two data sets, SAPSII low (low disease – LD) and SAPSII high (high disease – HD), where HD is considered as a progressed state of LD. As shown in the model algorithm in Fig. 3, the first step involves generating the hypothetical intermediate stages' samples (IS) between low disease (LD) and high disease (HD). We use a normal-distribution-based random number generator (NR) for generating these derived samples. First, we calculate the mean (μ_0) and standard deviation (σ_0) of each probe across all the LD samples. Then the NR generates a normal distribution curve for each (μ_0, σ_0) pair and gives out a random number from the curve in the $(\mu_0 - \sigma_0,\ \mu_0 + \sigma_0)$ range. Such generated expression values together for all the probes in an array are thought to represent a 'derived sample'. A derived sample can be called a 'hypothetical intermediate stage' (IS) if we increase σ_0, keeping μ_0 constant for each probe, thereby increasing the range step-by-step. As per our hypothesis, HD samples can be generated by extrapolating the LD samples. Increase in σ_0 is governed by Eq. (5) given below.

$$\sigma_f = \sigma_0(1 + \delta) \tag{5}$$

where σ_f is the varied standard deviation of a particular gene, σ_0 is the original standard deviation from the clinical samples of the same gene and δ is the fraction of increase in standard deviation. We keep increasing the δ-*value* until a particular IS comes closer to HD at the level of gene expression. Each of these ISs with increasing δ-*value* can be thought as the samples obtained from the same individual at different stages of disease progression between LD and HD.

We generate ISs at equal intervals between 0 & D (an estimated δ). Using the same procedure, we generate derived samples from clinical controls (using their (μ_0, σ_0)), LDs, ISs and HDs in quintuplicates and arrange all of them in a data matrix in the same order column-by-column and cluster the samples column-wise using Hierarchical clustering. From the dendrogram generated, we can visualize the closeness/similarity of each IS to HD. We pick the closest or the most similar IS to HD and call its δ-*value*, the thresh-

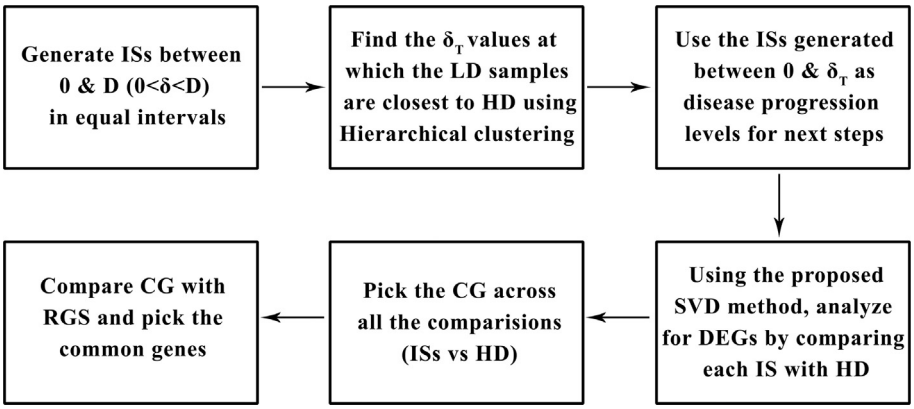

Fig. 3. Flowchart of the proposed Model. IS: Intermediate States, LD: Low Disease (SAPSII-low), HD: High Disease (SAPSII-high), DEG: Differentially Expressed Genes, CG: Common Genes, RGS: Reference Gene Set (DEGs when we compare the actual clinical samples).

old (δ_T). So, if a particular set of genes is being differentially expressed at all these ISs when compared to HD, they probably have an important role to play in the disease progression. Therefore, we employ our proposed SVD algorithm to find the differentially expressed genes sets (DEGSs) between each derived IS and derived HD. We have generated 500 derived samples per group. We find the common genes across all these DEGSs. On the other hand, we also analyse for Reference Gene Sets (RGSs), i.e., comparing the clinical samples (LD vs HD), using the proposed SVD algorithm. Finally, we consider the genes that are common across all the DEGSs and RGSs as the predicted genes. We have used MATLAB for all the programming, Cluster 3.0 and Java Treeview for Hierarchical clustering and generating dendrograms. Euclidian distance method was used for Hierarchical clustering after normalizing both columns and rows by their medians and scale to standard deviation of 1.

2.4. Gene ontology and pathway enrichment analysis

Gene ontology analysis for Biological Processes was done using the software plugin BiNGO in Cytoscape 3.3.0. A Binomial exact *p-value* cut-off of 0.05 was employed. Pathway enrichment analysis was done using the Reactome Database plugin in Cytoscape 3.3.0. The *p-value* was calculated using Benjamini-Hochberg multiple testing algorithm and cut-off was set to be 0.05.

3. Results

3.1. Validating the proposed SVD analysis method

We analysed the microarray data from Cazalis et al., 2014, using both the conventional method and proposed SVD following a control-based normalization approach. In Sections 3.1.1–3.1.3, we try to validate our proposed method by comparing its efficiency (E), observing if similar pathways and biological functions are showing up in Gene Ontology and pathway enrichment analyses and if we are able to replicate the key findings of Cazalis et al., 2014. In Section 3.1.4, we also report that H0-based normalization approach is a better choice by comparing the ontologies and pathways highlighted by it with the control-based normalization.

3.1.1. Efficiency calculation

Each group of samples, either time-wise (H0, H24 & H48) or severity-wise (SAPSII-high & SAPSII-low) was first analysed using the control-based approach (Fig. 1B). The efficiency when we compared C vs H0, C vs H24 and C vs H48 was 58.1%, 59.3% and 64.4% respectively, and it was 53.6%, 48.4%, 54.4%, 49.5%, 63.6% and 68.3% when we compared C vs HighH0, C vs HighH24, C vs HighH48, C vs LowH0, C vs LowH24 and C vs LowH48 respectively, giving an average efficiency of 57.7%. The efficiency is just a preliminary measure of judging the proposed algorithm. It should be noted that we consider only the first eigengene (first column of $(\widehat{u^n})$) to filter the outlier genes.

3.1.2. Gene ontology analysis

We also performed a Gene Ontology analysis using the 1638 common genes (CG2 in Fig. 1) which were modulated at all the three time points when analysed with a control-based normalization approach (numbers shown in the Venn diagram in Fig. 4A). We have compared these ontologies with the original article [8] and also with the genes obtained from our conventional analysis with GEO2R. The first two columns in Table 1 show the statistically significant and non-redundant ontology terms from GEO2R analysis and our proposed SVD analysis (control-based normalization).

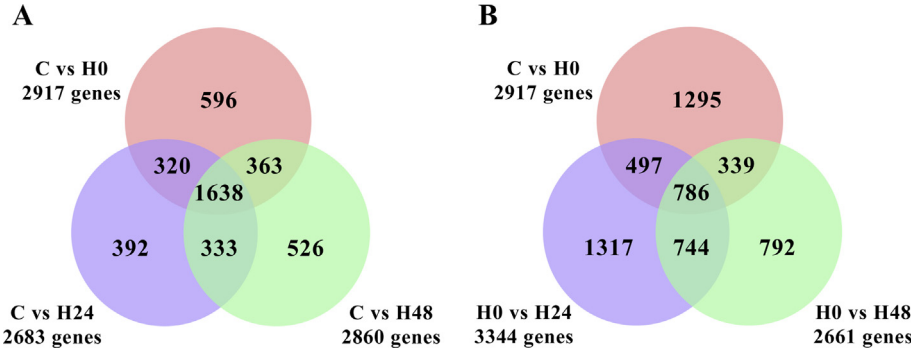

Fig. 4. Venn diagrams showing the number of common outlier genes for each comparison using SVD method. (A) Control-based normalization. (B) H0-based normalization.

Table 1
Gene ontology analysis.

Conventional analysis (control-based normalization)	SVD analysis (control-based normalization)	SVD analysis (H0-based normalization)
Regulation of Apoptosis	Metabolic process[a]	Metabolic process
Cell proliferation	Regulation of apoptosis	Cell proliferation
Leukocyte activation	Cell proliferation	Regulation of apoptosis
Lymphocyte activation	Inflammatory response	Inflammatory response
Cytokine response	Innate immune response	Vesicle mediated transport[b]
Inflammatory response	Leukocyte activation	Response to bacterium
Response to bacterium	Response to bacterium	T-cell activation
T-cell activation	Endocytosis[a]	Innate immune response
Innate immune activation	T-cell activation	Response to hypoxia[b]
MAPK activity regulation	MAPK activity regulation	Blood coagulation[b]

All the ontology terms in the respective columns are in descending order of the percentage of genes they represent in the total number of outlier genes and with p-value < 0.05. First column represents ontology analysis for the common outlier genes in all the three time points using GEO2R. Second and third represent ontology analysis for the common outlier genes in all the three time points using SVD method following Control-based and H0-based normalization approach, respectively.

[a] Ontologies in SVD control-based but not in Conventional.

[b] Ontologies exclusively in SVD H0-based normalization.

Table 2
Pathway enrichment analysis.

Conventional analysis (control-based normalization)	SVD analysis (control-based normalization)	SVD analysis (H0-based normalization)
TCR signaling	Interferon signaling	Interferon signaling
Costimulation by CD28 family	Platelet signaling, activation & aggregation	Platelet signaling, activation & aggregation
Platelet signaling, activation & aggregation	TCA cycle & electron transport chain[a]	Cell surface interaction at vascular wall
Interferon signaling	Cell surface interaction at vascular wall	Response to elevated cytosolic Ca^{2+}
Cell surface interaction at vascular wall	L13a-mediated translational silencing of ceruloplasmin expression	Integrin interaction
MHC class II antigen presentation	TLR signaling	Clotting cascade[b]
Interleukin signaling	Response to elevated cytosolic Ca^{2+}	Degradation of ECM[b]
TLR signaling	Selenoamino acid metabolism[a]	ROS, RNS production in response to bacteria

All the pathways in the respective columns are in descending order of the percentage of genes they represent in the total number of common outlier genes and with p-value < 0.05. First column represents pathway analysis for the common outlier genes in all the three time points using GEO2R. Second and third represent pathway analysis for the common outlier genes in all the three time points using SVD method following Control-based and H0-based normalization approach, respectively.

[a] Pathways in SVD control-based but not in Conventional.

[b] Pathways exclusively in SVD H0-based normalization.

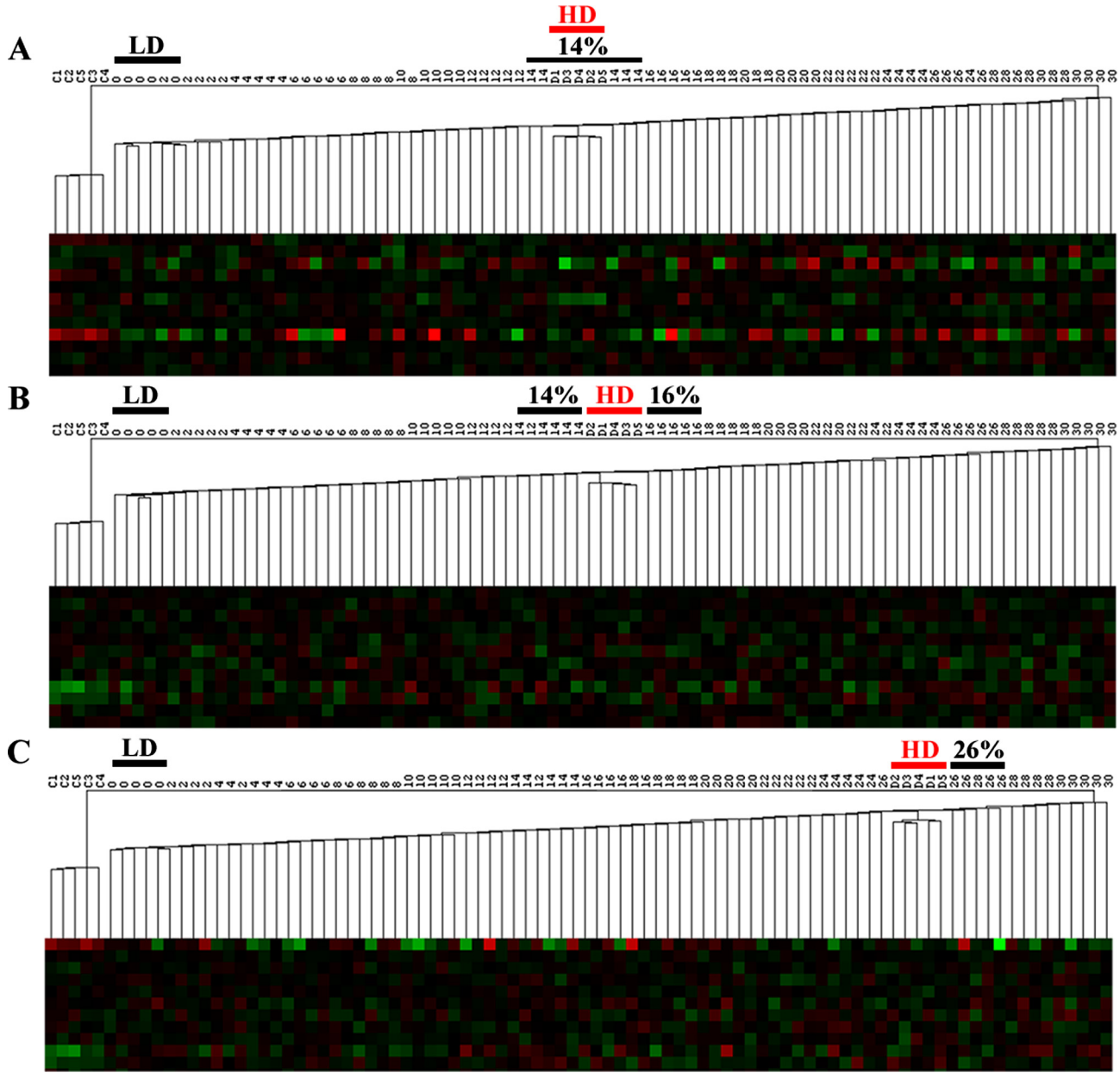

Fig. 5. Dendrograms showing Hierarchical clustering. Hierarchical clustering of the generated intermediate states (0–30% increase in σ_0) with Low Disease (LD), High Disease (HD, D1-D5) and control samples (C1-C5). (A) Showing that the IS generated after 14% increase in σ_0 is the closest to HD at H0. (B) Showing that the IS generated after 14–16% increase in σ_0 is the closest to HD at H24. (C) Showing that the IS generated after 26% increase in σ_0 is the closest to HD at H48.

Pathogen recognition, cytokine/cytokine receptor expression which lead to innate immune response, inflammatory response and cell death (apoptosis) were shown to be deregulated in Cazalis et al., 2014. Both the GEO2R conventional and the proposed SVD analyses were on similar lines showing immune cell activation (T-cell, lymphocyte & leukocyte), regulation of apoptosis, innate immune response, inflammatory response and response to pathogens (bacterium). Besides reproducing results of the conventional method and the original article, our proposed SVD method could also detect 'Metabolic processes' (topping the list) and 'Endocytosis'.

3.1.3. Pathway enrichment analysis

We have also performed a Pathway enrichment analysis using the same set of genes and compared it with the original article and our conventional analysis in the same fashion as explained above. The first two columns of Table 2 show the significantly enriched pathways in the conventional GEO2R and the proposed SVD analyses. The original article reported several immune system related pathways like T-cell receptor signaling, Interferon signaling, Interleukin signaling to be deregulated. Both our conventional GEO2R and SVD analyses yielded similar pathways. Besides, in synchrony with the Gene Ontology analysis, our SVD analysis showed a pronounced deregulation of several metabolic pathways like TCA cycle, Electron transport chain and Seleno-amino acid metabolism, and several platelet-specific pathways. Thus, the proposed SVD method could not only reproduce the conventional method's results, but also could fish out several important ontologies and pathways that weren't highlighted by the conventional method. This was possible due to the fact that SVD works with the pattern of expression across conditions/samples, unlike the conventional method, which only works with the magnitude of expression ($\log_2$ FC). Hence, using the above gene ontology, pathway enrichment analyses and efficiency calculation, we can confidently validate our proposed method.

3.1.4. H0-based normalization shows vital pathways of septic shock

As explained in Section 2.1, we tried a new way of normalizing the samples with controls (Fig. 1A). The efficiency of the proposed SVD method in this case increased to 76.6% and 73.1% for H0 vs H24 and H0 vs H48 respectively, due to the fact that we are normalizing H24 and H48 to H0 from the same set of subjects and not to a control group. The third column of Table 1 shows the Gene ontology analysis for this set of 786 common genes (CG1 in Fig. 1) obtained by using H0-based normalization (numbers shown in Venn diagram in Fig. 4B). It can be clearly observed that this kind of normalization, apart from the other two analyses, additionally picks up 'Response to Hypoxia', 'Blood Coagulation' and 'Vesicle mediated transport'. Similarly, the Pathway enrichment analysis (third column in Table 2) with the same set of genes shows 'Clotting cascade' and 'Degradation of ECM', apart from the common immune and inflammatory pathways as shown by the other two analyses. It is also well known that acute hypoxia induces endothelial inflammation, which leads to the release of platelet growth factors and increased adherence interactions [10]. Given this, 'Cell surface interactions at the vascular wall' and 'Integrin interactions' shown by the pathway enrichment analysis supports the ontologies and pathways highlighted exclusively by the H0-based normalization.

3.2. Gene prediction model

Our second objective was to predict a set of genes that might play a vital role in the disease progression of sepsis by making use of a simple mathematical model as described in Section 2.3. First, we increased the δ-*value* from 0 to 0.3 (0% to 30% increase in σ_0) at equal intervals of 0.02 using Eq. (5) and generated ISs using NR. As shown in Fig. 5A, from the dendrogram we can see that 14% of increase in σ_0 takes the generated IS's closest to HD samples (D1-D5 in Fig. 5A) i.e. $\delta_T = 0.14$ for H0. We can also observe that the control samples (C1-C5 in Fig. 5) are quite distant from any other samples and almost all the quintuplicates for every δ-*value* are clustered together. This increases the confidence of our hypothesis and makes sure that there are no/minimal errors in Hierarchical clustering and NR. Similarly, Fig. 5B and C shows that 16% and 26% increase in σ_0 is required to take IS samples of H24 and H48 closer to their corresponding HD samples (i.e., $\delta_T = 0.16$ and 0.26 respectively). As the disease progresses from H0 to H24 to H48, we can clearly observe an increase in δ_T from 0.14 to 0.26. This forms another proof of our hypothesis that a disease sample is nothing but an extrapolated version of the healthy state, when we consider gene expression values alone.

Now, we use the proposed SVD method to filter out the DEGSs between each IS sample set and its corresponding HD sample, until δ_T is reached. We then look for the common genes among all the DEGSs, in all the three time points and fish out the final set of genes that also show up in all the three RGSs (H0, H24 & H48). We assume that all the generated gene expression values follow

Table 3
Gene ontologies and pathways of the 19 genes predicted by the model.

Gene symbol	Gene name	Ontology & pathways
RORA	RAR-related Orphan Receptor A	HIF1alpha activation, Macrophage activation
XIST	X Inactive Specific Transcript	Non-coding RNA gene
CD24	CD24 molecule	Response to bacterium, LICAM interactions
CEACAM1	Carcinoembryonic Antigen related Cell Adhesion Molecule 1	ECM degradation, Platelet Activation
GBP1	Guanylate Binding Protein 1	Interferon & Cytokine signaling
ASPH	Aspartate beta Hydroxylase	Calcium homeostasis, Cardiac Conduction
GYPA & GYPB	Glycophorin A & B	Platelet membrane proteins, Malarial Pathway
HGF	Hepatocyte Growth Factor	Leukocyte & platelet activation, IL7 Signaling
HPGD	Hydroxyprostaglandin Dehydrogenase	Prostaglandin synthesis & metabolism
IGK	Immunoglobulin kappa locus	Angiotensin activation of ERK
MALAT1	Metastasis Associated Lung Adenocarcinoma Transcript 1	Cancer metastasis
OCLN	Occludin	Cytokine induced regulation of tight junctions
SMCHD1	Structural Maintenance Of Chromosomes Flexible Hinge Domain1	Inactivation of X chromosome by DNA Methylation
SNCA	Synuclein alpha	Neurotransmitter uptake, EGFR signaling
STAT1	Signal Transducer & Activator of Transcription 1	Activated by IL6, EGF, PDGF, IFNA, IFNG
ITGB3	Integrin beta 3	Platelet activation, LICAM, PECAM & Syndecan signaling
IGL	Immunoglobulin Lambda locus	Fc epsilon R1 pathway
TNS1	Tensin 1	ECM interactions, Actin crosslinking to integrins in Focal adhesions

Gene ontologies and pathways that summarize the functions of 19 predicted genes upon a literature survey.

normal distribution. In order to arrive at significant results, 500 arrays/samples were generated for each IS, each IS vs HD comparison was iterated for 5 times and this whole algorithm (from generating ISs to filtering out the common genes) for 3 iterations. The final set of 19 genes as shown in Table 3 appeared in all the three iterations and thus we propose them to be the most probable candidates in Septic shock progression. As a further step to check if the predicted genes are actually involved in septic shock progression, we performed a Gene ontology and pathway enrichment analysis (Table 3) of these 19 genes, which gives out a uniform distribution of all the vital ontologies and pathways observed in septic shock as shown in Tables 1 and 2.

4. Discussion

As stated earlier, several vital biological phenomena occur at a rather lower magnitude of gene expression. Finding the differentially expressed genes based on just the fold change might not extract all the relevant information for the huge amounts of data generated. By setting a cut-off on fold change, we usually pick the top 5–10% of genes. We think that the rest of 90–95% genes have an equal level of say in explaining the phenomenon being studied. This reckons the need for a better way of narrowing down the list. Since SVD works with the 'gene expression patterns' across conditions/samples, using the proposed SVD method, we were able to trace out various additional gene ontologies and pathways related to Septic shock (Tables 1 and 2). SVD teased out certain fundamental metabolic process that are involved in ATP generation such as TCA cycle and ETC, endocytosis, and several platelet-specific pathways in addition to the GEO2R conventional analysis. More than half of the genes involved in the above ontologies and pathways have their $\log_2$ FC less than 1 or even closer to 0.5 (data not shown). Although the average efficiency of SVD turned out to be 57.7% and the proposed method has its own drawbacks (explained further in discussion), one should keep in mind that a significant part of the remaining genes picked by SVD represent biologically-relevant data that conventional analysis would most likely miss out.

A brief look at the pathophysiology of sepsis, would help us understand if the ontologies and pathways picked up exclusively by the proposed SVD method are indeed vital in the disease progression. The onset of sepsis involves an initial systemic immune activation and inflammation as a result of the host immune system recognising the pathogen, followed by an anti-inflammatory response that may help to bring balance in case of a runaway pro-inflammatory first response [11]. This supports the repeated occurrence of immune, inflammatory and apoptosis related ontologies and pathways in both conventional and the proposed SVD method and helps validate the proposed method. In parallel with the immuno-inflammatory response, the pro-inflammatory cytokines activate pro-coagulatory pathways which leads to thrombocytopenia and coagulatory abnormalities like disseminated intravascular coagulation (DIC) and widespread microvascular thrombosis [12]. Having looked at the significant role of coagulatory abnormalities in septic shock progression, 'Blood coagulation' and 'clotting cascade' were only highlighted by the proposed SVD method where the conventional analysis was lacking.

Hypoxic conditions arise due to this microvascular thrombosis and hypotension in septic shock, which leads to oxidative stress and mitochondrial dysfunction. This is followed by an increased apoptosis that paves the path for multiple organ dysfunction and finally death [11]. On the other hand, as reported by Pravda 2014, there arises a hypermetabolic state to meet the increased energy requirements of the system to fight the initial infection [13]. This increased metabolism in turn releases more ROS like Hydrogen peroxide, which may lead to a more oxidized environment within the cell, particularly disrupting mitochondrial function with a sudden change in redox status inside the mitochondria. This in turn impairs its function leading to decreased ATP production – even in the presence of normal oxygen availability – a condition referred to as cytopathic hypoxia [14,15]. Either way, the proposed SVD method could highlight the deregulation of 'Response to hypoxia', 'TCA cycle & Electron transport chain', while the conventional methods failed to do so, thereby reinforcing the validity of the proposed method. In support to this, 7 out of the 18 genes responsible for 'Response to Hypoxia' in the ontology analysis (highlighted in Gene ontology Supplementary file) were found to be localized in the mitochondria (ACSL6, BCL2L1, ALAS2, BNIP3, ABAT, SOD2, and HSP90B1). Hence, with an increased level of confidence, we can say that H0-based normalization of samples has yielded novel and vital information that was previously masked by control-based normalization.

Interestingly, 'Metabolic process' was highlighted with the topmost priority in the proposed SVD method, both in control-based and H0-based normalization approaches (Table 1), which was masked in the conventional method. The deregulation of TCA cycle and ETC shown in pathway enrichment analysis (Table 2) also back the involvement of a hypermetabolic state. Apart from this, Selenoamino acid metabolism was also highlighted in Table 2. There were several reports that show Selenocystine induced ROS-mediated apoptosis in various human cancer cell lines [16,17].

As a second objective of this article, we also proposed a simple mathematical model to predict the vital genes in the progression of a disease. From a set of 54,000 (appx.) probes, we were able to finally come down to 19 individual genes that could possibly have a pivotal role in Septic shock progression. As shown in Table 3, a basic literature survey of these 19 genes showed that they are actually a part of several important septic shock related pathways like cytokine signaling, activation and aggregation of platelets, response to hypoxia and cellular metabolism. ITGB3 surface expression was shown to be upregulated in PI3K-$\gamma^{-/-}$ mice when challenged with intraperitoneal *E. coli* sepsis [18]. CD24 was reported to be associated with Sepsis in two different studies, one showing that it is upregulated in sepsis patients with ARDS (Acute respiratory distress syndrome) when compared with patients only with sepsis and other study showing that CD24-mediated apoptosis via mitochondrial membrane depolarization and ROS in human neutrophils is absent in sepsis conditions [19,20]. STAT1 deficient mice were shown to be resistant to endotoxin-induced and CLP-induced septic shock (Cecal ligation puncture) in two individual reports [21,22] and a detailed explanation of the role of STATs in sepsis was given in two other reviews [23,24].

Also, adding more confidence to our proposed model, we found 4 out of the 19 genes (RORA, IGK, STAT1 and SNCA) were shown to clearly distinguish between SAPSII-high and SAPSII-low subjects in the original article [8]. Another 4 out of the 19 predicted genes (HGF, ITGB3, SNCA, and CD24) were observed to be involved in 'Response to Hypoxia', 'Blood coagulation' and 'ETC' in our own gene ontology analysis (Table 1 and Supplementary information). CEACAM1, HGF, HPGD were shown to be upregulated in a meta-analysis of septic shock microarray data [25]. With these evidences proving that most of the 19 predicted genes are related to sepsis, we think that our hypothesis is valid in predicting important genes in a disease progression. The genes in the 19 predicted genes that haven't been shown to be related to septic shock till date, might have a probable role yet to be discovered.

There are various limitations and assumptions in our proposed SVD method and model algorithm. We assumed that all the genes and arrays work independent of each other, which is not true in case of an actual organism. The proposed SVD method can't

distinguish between up/down-regulated genes, it can just pick the outlier genes. Since we consider only the first eigengene for picking the most significantly deregulated genes, not all the outlier genes can be picked up. There is a lack of proper negative control data set available in current analysis, as the samples were affected by vasopressor therapy, we couldn't conduct a False Discovery Rate (FDR) measurement. The readers are also requested to bear in mind that this method is not ideal for smaller datasets (eg. pathway-specific arrays). With larger datasets (eg. Whole transcriptomic profiles and a good number of biological replicates), the method has enough data points to build and compare the expression profiles for each gene. In the model, we also assume that all the gene expression values follow a normal distribution obtained from the (μ_0, σ_0) of its corresponding clinical data. We also use a random number generator to generate gene expression values based on the normal distribution. We try to overcome this randomness in the data by generating quintuplicates of datasets with each dataset having 500 samples and repeating the whole model in triplicates.

5. Conclusion

In this article we have used the lesser known SVD technique to build a microarray data analysis tool that works with gene expression patterns and intrinsic structure of the data, rather than just fold change and *p-value*. Further elements can be added to this method to distinguish between up/down regulated genes. As a second objective, we have developed a novel and simple mathematical model to generate samples of hypothetical intermediate stages in the progression of a disease and to predict the genes that might play an important role in the disease progression. To our knowledge, this model is the first of its kind in microarray analysis, which deals with random number generator and normal distribution.

Funding

This work was supported by Department of Biotechnology (DBT), Government of India under the project number, [BT/PR12547/MED/30/1456/2014].

Conflict of interest

None declared.

Acknowledgements

We thank the Department Computer Facility (DCF), Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences Building, IIT Madras, for providing space and infrastructure to work in. We also thank our colleagues, Rathnakumar K and Abhiram Charan Tej M, who provided several insights that greatly assisted the work.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2017.05.005.

References

[1] J. Quackenbush, Computational analysis of microarray data, Nat. Rev. Genet. 2 (6) (2001) 418–427.
[2] M. Ringnér, C. Peterson, J. Khan, Analyzing array data using supervised methods, Pharmacogenomics 3 (3) (2002) 403–415.
[3] A. Deutsch, L. Brusch, H. Byrne, G. De Vries, H. Herzel (Eds.), Mathematical Modeling of Biological Systems, vol. I, Berlin, Springer, 2007. ISBN 978-0-8176-4557-1.
[4] O. Alter, P.O. Brown, D. Botstein, Singular value decomposition for genome-wide expression data processing and modeling, Proc. Natl. Acad. Sci. 97 (18) (2000) 10101–10106.
[5] M. Olbryt, M. Jarząb, J. Jazowiecka-Rakus, K. Simek, S. Szala, A. Sochanik, Gene expression profile of B16 (F10) murine melanoma cells exposed to hypoxic conditions in vitro, Gene Expr. 13 (3) (2006) 191–203.
[6] B. Jarząb, M. Wiench, K. Fujarewicz, K. Simek, M. Jarząb, M. Oczko-Wojciechowska, J. Włoch, A. Czarniecka, E. Chmielik, D. Lange, A. Pawlaczek, Gene expression profile of papillary thyroid cancer: sources of variability and diagnostic implications, Can. Res. 65 (4) (2005) 1587–1597.
[7] K. Simek, M. Kimmel, A note on estimation of dynamics of multiple gene expression based on singular value decomposition, Math. Biosci. 182 (2) (2003) 183–199.
[8] M.A. Cazalis, A. Lepape, F. Venet, F. Frager, B. Mougin, H. Vallin, M. Paye, A. Pachot, G. Monneret, Early and dynamic changes in gene expression in septic shock patients: a genome-wide approach, Intens. Care Med. Exp. 2 (1) (2014) 20.
[9] T. Barrett, S.E. Wilhite, P. Ledoux, C. Evangelista, I.F. Kim, M. Tomashevsky, K.A. Marshall, K.H. Phillippy, P.M. Sherman, M. Holko, A. Yefanov, NCBI GEO: archive for functional genomics data sets—update, Nucleic Acids Res. 41 (D1) (2013) D991–D995.
[10] C. Michiels, T. Arnould, J. Remacle, Endothelial cell responses to hypoxia: initiation of a cascade of cellular interactions, Biochim. Biophys. Acta (BBA) – Mol. Cell Res. 1497 (1) (2000) 1.
[11] D.C. Angus, T. Van der Poll, Severe sepsis and septic shock, N. Engl. J. Med. 369 (9) (2013) 840–851.
[12] J.E. Carré, M. Singer, Cellular energetic metabolism in sepsis: the need for a systems approach, Biochim. Biophys. Acta (BBA) – Bioenerg. 1777(7) (2008) 763–771.
[13] J. Pravda, Metabolic theory of septic shock, World J. Crit. Care Med. 3 (2) (2014) 45.
[14] M.P. Fink, Cytopathic hypoxia, Crit. Care Clin. 17 (1) (2001) 219–237.
[15] M. Fink, Cytopathic hypoxia in sepsis, Acta Anaesthesiol. Scand. 41 (S110) (1997) 87–95.
[16] T. Chen, Y.S. Wong, Selenocystine induces caspase-independent apoptosis in MCF-7 human breast carcinoma cells with involvement of p53 phosphorylation and reactive oxygen species generation, Int. J. Biochem. Cell Biol. 41 (3) (2009) 666–676.
[17] C. Fan, W. Zheng, X. Fu, X. Li, Y.S. Wong, T. Chen, Strategy to enhance the therapeutic effect of doxorubicin in human hepatocellular carcinoma by selenocystine, a synergistic agent that regulates the ROS-mediated signaling, Oncotarget 5 (9) (2014) 2853–2863.
[18] E. Ong, X.P. Gao, D. Predescu, M. Broman, A.B. Malik, Role of phosphatidylinositol 3-kinase-γ in mediating lung neutrophil sequestration and vascular injury induced by *E. coli* sepsis, Am. J. Physiol. – Lung Cell. Mol. Physiol. 289 (6) (2005) L1094–L1103.
[19] M. Parlato, F. Souza-Fonseca-Guimaraes, F. Philippart, B. Misset, M. Adib-Conquy, J.M. Cavaillon, S. Jacqmin, D. Journois, A. Lagrange, G.P. de Villechenon, N. Aissaoui, CD24-triggered caspase-dependent apoptosis via mitochondrial membrane depolarization and reactive oxygen species production of human neutrophils is impaired in sepsis, J. Immunol. 192 (5) (2014) 2449–2459.
[20] K.N. Kangelaris, A. Prakash, K.D. Liu, B. Aouizerat, P.G. Woodruff, D.J. Erle, A.J. Rogers, E.J. Seeley, J. Chu, T. Liu, T. Osterberg-Deiss, Increased expression of neutrophil-related genes in patients with early sepsis-induced ARDS, Am. J. Physiol. – Lung Cell. Mol. Physiol. 2015, ajplung-00380.
[21] K. Kamezaki, K. Shimoda, A. Numata, T. Matsuda, K.I. Nakayama, M. Harada, The role of Tyk2, Stat1 and Stat4 in LPS-induced endotoxin signals, Int. Immunol. 16 (8) (2004) 1173–1179.
[22] D. Herzig, G. Fang, T.E. Toliver-Kinsky, Y. Guo, J. Bohannon, E.R. Sherwood, STAT1-deficient mice are resistant to CLP-induced septic shock, Shock (Augusta, GA) 38(4) (2012) 395.
[23] M.J. Scott, C.J. Godshall, W.G. Cheadle, Jaks, STATs, cytokines, and sepsis, Clin. Diagn. Lab. Immunol. 9 (6) (2002) 1153–1159.
[24] A. Matsukawa, STAT proteins in innate immunity during sepsis: lessons from gene knockout mice, Acta Med. Okayama 61 (5) (2007) 239–245.
[25] S. Mukhopadhyay, A. Pandey, P. Thatoi, B.K. Das, B. Ravindran, S. Bhattacharjee, S.K. Mohapatra, Meta-analysis reveals pathway signature of septic shock. bioRxiv. (2016) 051706.