"Comprehensive Analysis of Genomic Storm (Transcriptomic) Data, Integrating Clinical Data and Utilizing New and Old Approaches using R and BioConductor Packages"

to be submitted as Major Project in partial fulfilment of the requirement for the degree of

Master of Technology

In

Bioinformatics

Submitted by

Kirti Bhadhadhara

(2K14/Bio/07)

Delhi Technological University, Delhi, India

under the supervision of

Dr. Yasha Hasija

Assistant Professor

Department of Biotechnology

Delhi Technological University, Delhi, India

CERTIFICATE

This is to certify that the dissertation entitled **"Comprehensive Analysis of Genomic Storm (Transcriptomic) Data, Integrating Clinical Data and Utilizing New and Old Approaches using R and BioConductor Packages"** submitted by **Kirti Bhadhadhara (2K14/Bio/07)** in the partial fulfilment of the requirements for the reward of the degree of Master of Technology, Delhi Technological University (Formerly Delhi College of Engineering, University of Delhi), is an authentic record of the candidate's own work carried out by him/her under my guidance. The information and data enclosed in this thesis is original and has not been submitted elsewhere for honoring of any other degree.

Dr. Yasha Hasija

(Project Mentor)

Department of Bio-Technology

Delhi Technological University

(Formerly Delhi College of Engineering, University of Delhi)

DECLARATION

I declare that my major project entitled **"Comprehensive Analysis of Genomic Storm (Transcriptomic) Data, Integrating Clinical Data and Utilizing New and Old Approaches using R and BioConductor Packages"**, submitted to Department of Biotechnology, Delhi Technological University as a result of the work carried out by me at "Genome Informatics Laboratory", Department of Biotechnology, as Major project.

Date: (Kirti Bhadhadhara)

ACKNOWLEDGEMENT

I would like to express my sincere thanks to Delhi Technological University (Formerly Delhi College of Engineering, University of Delhi), Delhi, India for granting permission to carry out my Postgraduate Project.

I would like to take this opportunity to express my gratitude to the Prof. D. Kumar, Professor, HOD, Department of Biotechnology for providing me the opportunity to successfully carry out my project by providing us all the necessary resources and guidance.

I would like to take this opportunity to express my thanks to my project guide Dr. Yasha Hasija, Assistant Professor, Department of Biotechnology for continuously guiding and supporting at each level of the project.

I would also like to thank my fellow colleagues for helping us out when I faced problems in the project.

Kirti Bhadhadhara

2K14/Bio/07

(M.Tech Bioinformatics)

CONTENT:

List of Figures:

List of Tables:

1. ABSTRACT

To determine trauma-specific transcriptomic signatures for septic sub-cohorts.

In retrospective large-scale data analysis, old and new methods were applied, including lagged correlation between transcripts and clinical subtype counts by integrating over 800 samples from trauma patients. Focusing on novel pathways and correlation methods that were revealed (persistently down-regulated) ribosomal genes and changed time profiles of metabolic enzyme precursors /transcripts. Candidates associated to insulin signaling, including HK3, hinted towards "metabolic syndrome". Correlation analysis yielded robust results for LCN2 and LTF ($r>0.9$), but only moderate associations to subtype counts (e.g. top-performing r (Eosinophil, IL5RA)>0.6). Gene Centered Normalization Reduces Ambiguity and Improves Interpretation.

2. INTRODUCTION

Normalization

Normalization is the attempt to compensate for systematic technical differences between chips, to see more clearly the systematic biological differences between samples. Differences in treatment of two samples, especially in labelling and in hybridization, bias the relative measures on any two chips [1].

Systematic non-biological differences between chips are evident in several ways:

- Total brightness differs between chips
- One dye seems stronger than the other (in 2-color systems) on one chip, but not on another
- Typical background is higher in one chip than on another

There are also many non–obvious systematic differences between chips in an experiment, and even between the two channels on a single array. Some causes of systematic measurement variation include:

- Different amounts of RNA
- One dye is more readily incorporated than the other (in 2-color systems)
- The hybridization reaction may proceed more fully to equilibrium in one array than the other

- Hybridization conditions may vary across an array
- Scanner settings are often different, and of course

Murphy's Law predicts even more variation than can be explained simply.

Comparison of two group of samples

The simplest and most common experimental set-up is to compare two groups: for example, Treatment vs. Control, or Mutant vs. Wild type. The issues arising in simple comparisons arise also in more complex settings; it is easier to explain these in the simpler context [2]. The long-time standard test statistic for comparing two groups is the t-statistic:

$$t = (x_{i,1} - x_{i,2}) / s_i,$$

Where xi,1 is the mean value of gene i in group 1, xi,2 is the mean in group 2, and si is the (non-pooled) within-groups standard error (SE) for gene i.

Signal Log Ratio Algorithm

Signal Log Ratio algorithm estimates the measure and the direction of change of a Gene/transcript when two arrays are compared. Each probe pair on the experiment array is compared to the corresponding probe pair in the baseline arrays in the calculation of Signal Log Ratio [3]. This process eliminates differences due to different probe binding coefficients. A One-Step Tukey's Biweight method is used in computing the Signal Log Ratio value by taking a mean of the log ratios of probe pair intensities across the two arrays. The base 2 log scale is used, translating the Signal Log Ratio of 1.0 to a 2-fold increase in the expression level and of -1.0 to a 2-fold decrease. No change in the expression level is thus indicated by a Signal Log Ratio value 0. Tukey's Biweight method also gives estimate of the amount of variation in the data. Confidence intervals are generated from the scale of variation of the data. A 95% confidence interval shows a range of values, which will include the true value 95% of the time. Small confidence interval implies that the expression data is more exact, while large confidence intervals reflect more noise and uncertainty in estimating the true level. Since the confidence intervals attached to Signal Log Ratios are computed from variation between probes, they may not reflect the full width of experimental variation.

Correlation (r)

The correlation of two variables represents the degree to which the variables are related. When two variables are perfectly linearly related, the points in the scatter plot fall on a straight line [3]. Correlation measures only linear relationship. Two summary measures or correlation coefficients, Pearson's correlation and Spearman's rho, are most commonly used. Both of these measure range from perfectly positive linear relationship to perfectly negative linear relationship, or from -1 to 1. It is not wrong to calculate the correlation between variables, which are not linearly related, but it does not make much sense. If the variables are not linearly related, the correlation does not describe their relationships effectively, and no conclusions can be based on the correlation coefficient only. Correlation and scatter plot are a good example of how numerical and graphical tools effectively complement each other.

Log2-transformation

Log2-transformation is often used with DNA microarray experiments. Usually, the intensity ratio is log2-transformed [3]. The resulting new variable is called log ratio. The increase of one in the log ratio means that the actual intensity or expression has doubled.

Intensity ratio

The simplest approach is to divide the intensity of a gene in the sample by the intensity level of the same gene in the control [3].

Hypothesis pair

Before applying the test to the data, a hypothesis pair should be formed. A hypothesis pair consists of a null hypothesis (H0) and an alternative hypothesis (H1) [3]. For the tests, the hypotheses are always formulated as follows:

H0=> There is no difference in means between compared groups

H1=> There is a difference in means between compared groups.

Threshold for p-value

The p-value is usually associated with a statistical test, and it is the risk that we reject the null hypothesis, when it actually is true. Before testing, a threshold for p-value should be decided. This is a cut-off below which the results are statistically significant, and above which the results are not statistically significant. Often a threshold of 0.05 is used. This means that every 20th time

we conclude by chance alone that the difference between groups is statistically signif-icant, when it actually isn't. If the compared groups are large enough, even the tiniest difference can get a significant p-value. In such cases it needs to be carefully weighted whether the statistical significance is just that, statistical significance, or is there real biological phenomenon acting in the background [3].

Fold change

Another means to make the distribution of intensity ratios more symmetrical is to calculate the fold change. The fold change is equal to the intensity ratio, when the expression is higher than one. Below one, the fold change is equal to the inversed intensity ratio [3].

$$\text{For values } >1, \text{ fold change} = \frac{Cy3'}{Cy5'}$$
$$\text{For values } <1, \text{ fold change} = \frac{1}{(Cy3'/Cy5')}$$

The fold change makes the distribution of the expression values more sym-metric, and both under and over-expressed genes can take values between zero and infinity. Note, that the fold change makes the expression values additive in a similar fashion as the log- transformation.

Time series

In a time series experiment expression changes are monitored with samples taken between certain time intervals. Although several replicates can be made per every time point, it should be considered that these replicate chips can possible be made a better use of, if they are added to the time series as sampling points. That is, it should be weighted whether a high precision in every time point is more valuable than the additional information of expression changes new sampling points (time points) produce [3].

Microarray preparation

Microarrays are commonly prepared on a glass, nylon or quartz substrate. Critical steps in this process include the selection and nature of the DNA sequences that will be placed on the array, and the technique of fixing the sequences on the substrate. Affymetrix Company that is a leading manufacturer of gene chips, uses a method adopted from the semiconductor industry with photolithography and combinatorial chemistry [4]. The density of oligonucleotides in their Gene Chips is reported as about half a million sequences per 1.282 cm^2. (Affymetrix web site).

Probe preparation, hybridization and imaging

To prepare RNA probes for reacting with the microarray, the first step is isolation of the RNA population from the experimental and control samples. cDNA copies of the mRNAs are synthesized using reverse transcriptase and then by in vitro transcription cDNA is converted to cRNA and fluorescently labeled. This probe mixture is then cast onto the microarray. RNAs that are complementary to the molecules on the microarray hybridize with the strands on the microarray. After hybridization and probe washing the microarray substrate is visualized using the appropriate method based on the nature of substrate. With high density chips this generally requires very sensitive microscopic scanning of the chip. Oligonucleotide spots that hybridize with the RNA will show a signal based on the level of the labeled RNA that hybridized to the specific sequence. Whereas the dark spots that show little or no signal, mark sequences that are not represented in the population of expressed mRNAs [4].

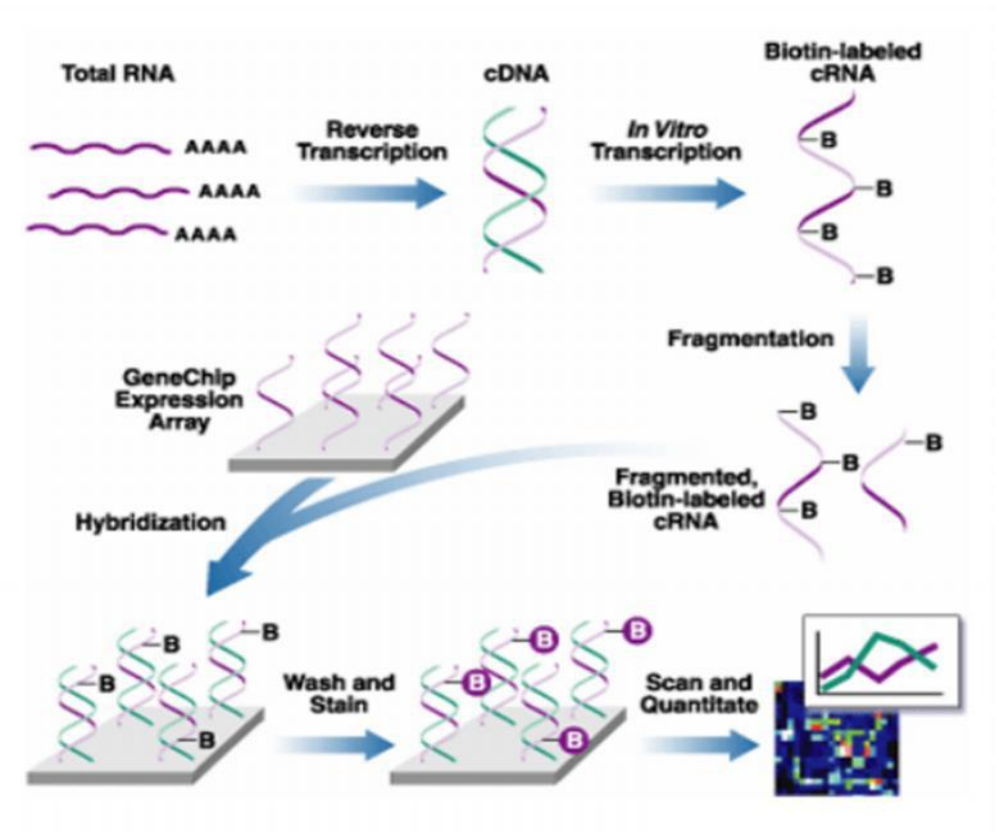


Figure 1: The process of fluorescently labeled RNA probe production.

Low level information analysis

Microarrays measure the target quantity (i.e. relative or absolute mRNA abundance) indirectly by measuring another physical quantity – the intensity of the fluorescence of the spots on the array for each fluorescent dye. These images should be later transformed into the gene expression matrix. This task is not a trivial one because:

1. The spots corresponding to genes should be identified.

2. The boundaries of the spots should be determined.

3. The fluorescence intensity should be determined depending on the background intensity [4].

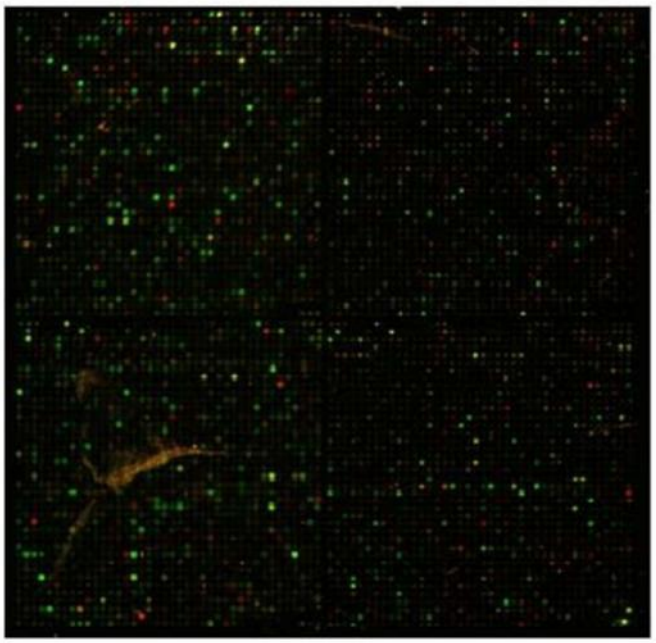


Figure 2: Gene expression data. Each spot represents the expression level of a gene in two different experiments. Yellow or red spots indicate that the gene is expressed in one experiment. Green spots show that the gene is expressed at same levels in both experiments.

In conclusion, microarray-based gene expression measurements are still far from giving estimates of mRNA counts per cell in the sample. The samples are relative by nature. In addition, appropriate normalization should be applied to enable gene or samples comparisons. It is

important to note that even if we had the most precise tools to measure mRNA abundance in the cell; it still wouldn't provide us a full and exact picture about the cell activity because of post-translational changes.

Despite continuing advances in intensive care medicine, severe sepsis and septic shock are currently among the most common causes of morbidity and mortality in intensive care. Moreover, the incidence of severe sepsis and septic shock has increased with ageing of the population over the past decade [5, 6, 7]. According to the University Hospital Jena Website: in Germany, 154,000 new cases of Sepsis occurs every year, killing an average of 150 patients every day. Therefore, Sepsis is regarded as a hidden healthcare disaster. [8]

3. Review of Literature:

Microarray

Microarray technology is a powerful tool for simultaneously evaluating the expression level of thousands of genes in a cell and finally the information that is encoded in the DNA. A microarray is a microscopic slide that contains an ordered series of DNA, RNA proteins or tissues. The DNA microarrays are the most common. A DNA microarray is generally a glass slide or a silicon chip in which thousands of gene sequences are printed. The genes immobilized onto the slide are called the DNA probe. Over this DNA probe, the target DNA or the target RNA (depending on the microarray platform) obtained from the cell under study is hybridized (hydrogen bonded). The amount of hybridization is measured and related to the presence and expression of certain genes in the cell.

SIRS, Sepsis and Septic Shock

For many years' doctors, attending intensive care units used a variety of terms to describe illnesses associated with infection, or illness that looked like infection. These terms included sepsis, septicemia, bacteremia, infection, septic shock, toxic shock etc. Unfortunately there were two problems with these terms: 1. there were no strict definitions for the terms used, and often these words or phrases were used incorrectly. 2, an emerging body of evidence arose which led us to believe that systemic inflammation, rather than infection, was responsible for multi-organ failure [9]. In the early 1990s a consensus conference between the ACCP and the SCCM laid out a new series of definitions [10]:

Infection

A host response to the presence of microorganism or tissue invasion by microorganisms.

Bacteraemia

The presence of viable bacteria in circulating blood.

Systemic Inflammatory Response Syndrome (SIRS)

The systemic inflammatory response to a wide variety of severe clinical insults, manifested by two or more of the following conditions:

- Temperature > 38°C or < 36°C
- Heart rate > 90 beats/min
- Respiratory rate > 20 breaths/min or $PaCO_2$ < 32 mm Hg
- WBC count > 12,000/mm^3, < 4000/mm^3, or > 10% immature (band) forms.

Sepsis

The systemic inflammatory response to infection. In association with infection, manifestations of sepsis are the same as those previously defined for SIRS. It should be determined whether they are a direct systemic response to the presence of an infectious process and represent an acute alteration from baseline in the absence of other known causes for such abnormalities. The clinical manifestations would include two or more of the following conditions as a result of a documented infection:

Severe Sepsis/SIRS

Sepsis (SIRS) associated with organ dysfunction, hypo perfusion, or hypotension. Hypo perfusion and perfusion abnormalities may include, but are not limited to, lactic acidosis, oliguria, or an acute alteration in mental status.

Refractory (Septic) Shock/SIRS Shock

A subset of severe sepsis (SIRS) and defined as sepsis (SIRS) induced hypotension despite adequate fluid resuscitation along with the presence of perfusion abnormalities that may include, but are not limited to, lactic acidosis, oliguria, or an acute alteration in mental status. Patients receiving inotropic or vasopressor agents may no longer be hypotensive by the time they

manifest hypo perfusion abnormalities or organ dysfunction, yet they would still be considered to have septic (SIRS) shock.

Multiple Organ Dysfunction Syndrome (MODS)

Presence of altered organ function in an acutely ill patient such that homeostasis cannot be maintained without intervention.

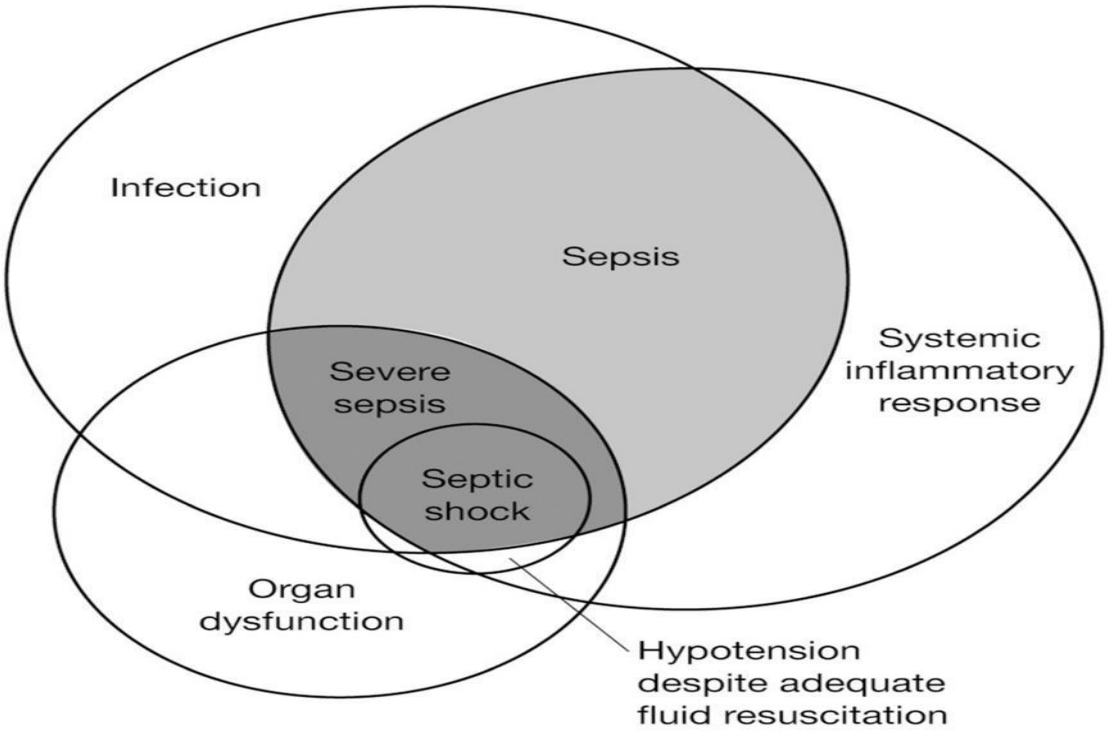


Figure-3: Relationship of Infection, SIRS, Sepsis, Severe Sepsis and Septic Shock [11]

Related Background

Trauma represents a frequent clinical syndrome characterized by the patient's systemic inflammatory response to infection, and carries a very high mortality rate. Trauma injuries frequently lead to infections, sepsis, and multiple organ failure (MOF) [12, 13], which contribute to 51%–61% of late trauma mortality [14]. Traumatic injury with its potential for infection was likely a common cause of death for our human ancestors. Even today, massive injury remains the most common cause of death for those under the age of 45 yr in developed countries [15, 16]. Systematic screening approaches are necessary in order to better diagnose and treat trauma, because it's a complex disease state with time-dependent intra-patient variability [17]. A number

of clinical trials for treating late trauma complications have failed, believed partly due to the inability to identify a proper patient population as well as the limited understanding of the interplay of biological processes underlying post-injury inflammatory complications [18, 19].

Furthermore, potential influential factors in sepsis, including treatment, age, sex and organ failure as well as interactions among these factors are assumed to play a major role in disease progression and are potentially reflected in molecular markers. Only recently has the human injury response been studied systematically at the genomic level and only now is it beginning to become better understood. Prior work has focused on the role of individual mediators [20, 21, 22] or processes such as apoptosis and cellular death in nosocomial infections and organ injury after trauma [23]. Circulating blood leukocytes have the capacity to seek out, recognize, and mount an appropriate inflammatory response at the earliest sign of injury. Blood neutrophils, monocytes, and Natural Killer cells are implicated as primary effectors during the initial inflammation and activation of innate immunity. Severe trauma has also been characterized by immunosuppression, primarily seen on the adaptive immune system with T lymphocyte populations being the most markedly affected cell population [24, 25].

CEL File Description

The CEL file stores the results of the intensity calculations on the pixel values of the DAT file (Contains the pixel intensity values collected from an Affymetrix Scanner). This includes an intensity value, standard deviation of the intensity, the number of pixels used to calculate the intensity value, a flag to indicate an outlier as calculated by the algorithm and a user defined flag indicating the feature should be excluded from future analysis. The file stores the previously stated data for each feature on the probe array [26].

Gene Expression Omnibus (GEO)

GEO is an international public repository that archives and freely distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomics data submitted by the research community [27].

KEGG

Kyoto Encyclopedia of Genes and Genomes; or K.E.G.G., as it is commonly called; is a collection of online databases dealing with genomes, enzymatic pathways, and biological chemicals. The Pathway Database, records networks of molecular interactions in cells and their variants (specific to particular organisms). K.E.G.G. switched to a subscription model,

accessible via FTP in July 2014. KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from genomic and molecular-level information [28].

The Kyoto Encyclopedia of Genes and Genomes was initiated by the Japanese human genome program in 1995 [29]. According to the developers, KEGG is a "computer representation" of the biological system [30]. The KEGG database can be utilized for modelling and simulation, browsing and retrieval of data. It is a part of the systems biology approach.

KEGG is best known for the display of biochemical pathways, but many other functions are now available at KEGG. KEGG is a collection of about 20 databases, which can be divided into three groups covering different biological spaces:

Genes

- KEGG Genes - manually curated from completely sequenced genomes
- DGENES - draft genomes
- EGENES - from EST contigs
- KEGG Orthology - manually defined ortholog groups based on KEGG pathways and BRITE functional hierarchies
- KEGG SSDB - Seq similarity scores

Chemicals and Ligands

- Ligand

Systems

- KEGG Pathway
- KEGG Brite

4. MATERIALS & METHODS

Getting Started with R

R is an implementation of the S language (Becker et al., 1988; Chambers and Hastie, 1992; Chambers, 1998). R is now becoming the most widely used software tools for bioinformatics.

R comes with substantial documentation. There are five manuals: An Introduction to R, The R Language Definition, R Installation and Administration manual, Writing R Extensions and R Data Import and Export. R News is a source of information on R packages and on aspects of the language written at an accessible level. Venables and Ripley (2000) is another reference for programming in the S language.

The R language was primarily designed as a language for data manipulation, modeling and visualization, and many of the algorithms and data structures reflect this. When R is started, a workspace is created and that workspace is where the user creates and manipulates variables. The workspace is an environment, and an environment is a set of binding of names, or symbols, to values. The top-level workspace can be accessed through its name, which is GlobalEnv. The grey background is used for all code examples that were processed by the Sweave system.

Download R

R (www.r-project.org) is a commonly used free Statistics software. R allows to carry out statistical analyses in an interactive mode, as well as allowing simple programming.

Installing R on a Windows System

- Go to http://ftp.heanet.ie/mirrors/cran.r-project.org.
- Under “Download and Install R”, click on the “Windows” option.
- Under “Subdirectories”, click on the “base” option.
- On the next page, a link saying something like “Download R 3.3.1 for Windows” (or R X.Y.Z., where X.Y.Z. gives the version of R, eg. R 3.3.1). Click on this option.
- It may be asked to save or run a file “R-3.3.1-win64.exe”. Choose “Save” and save the file on the Desktop. Then double-click on the icon / link for the file to run / save it.
- Choose English as language.
- The R Setup Wizard will pop up in a window. Click “Next” at the link of the R Setup wizard window.

- The next page says “Information” at the top. Click “Next”.
- The next page says “Select Destination Location” at the top. By default, it will suggest to install R in “C:\Program Files” on computer.
- Click “Next” at the blinking bottom of the R Setup wizard window.
- The next page says “Select components” at the top. Click “Next”.
- The next page says “Startup options” at the top. Click “Next”.
- The next page says “Select start menu folder” at the top. Click “Next”.
- The next page says “Select additional tasks” at the top. Click “Next”.
- R has now been installed. When R has finished, it will appear as “Completing the R for Windows Setup Wizard” appear. Click on “Finish”.
- Check if there is an “R” icon on the desktop of the computer. Double-click on the “R” icon to start R.
- The R console (a rectangle) pops up.

The steps to Install BioConductor Packages:

```
># download the BioConductor installation routines

>source(“http://bioconductor.org/biocLite.R”)

># installing the core packages

>biocLite()
```

Bioconductor

Bioconductor is an open source and open development software project to provide tools for the analysis and comprehension of genomic data. Bioconductor is built on Open Source Platform, R programming language, but does contain contributions in other programming languages. Most Bioconductor components are distributed as R packages and are overseen by the Bioconductor core team, based primarily at the Fred Hutchinson cancer Research Center with other members coming from the various US and international institutions. The main goal of the Bioconductor project is the creation of a durable and flexible software development and deployment environment that meets these new conceptual, computational and inferential challenges.

Data

Publicly available data sets containing gene expression values from blood samples of 167 patients between the ages of 18 and 55 yr, trauma patients (incl. septic and non-septic patients) as well as healthy controls from published studies were retrieved from Gene Expression Omnibus (GEO) [31]. Total blood leukocytes were isolated according to protocols previously published. Total cellular RNA was extracted and hybridized onto an HU133 Plus 2.0 Gene Chip (Affymetrix) according to the manufacturer's recommendations. Sepsis and control samples from the Illumina platform available under GSE36809.

Data Analysis

Briefly, computations were performed using R software (http://www.r-project.org/) v.3.0.2 and Bioconductor [32] packages. To assure comparability due to differences in sample size in Sepsis patients and controls data, a consistent work-flow was applied.

Data was obtained from GEO [33] and were pre-processed using chip definition file from Brain array (v.17, 2013), which aggregates probes into updated gene-centered probe set definitions mapping to Entrez IDs [34]. Further pre-processing was performed using quantile normalization via the RMA method for each sample set and after merging. Differentially expressed genes (DEGs) in the data were filtered according to microarray quality control (MAQC) [35] criteria and standard thresholds as follows: (i) average two-fold difference for pooled groups (sepsis versus non-septic controls); and (ii) false discovery rate (FDR) (Benjamin–Hochberg)-adjusted P-values <0.05 from Wilcoxon test. Illumina data were converted using median averaged signals for technical bead type replicates and re-normalized after merging with Affymetrix data.

Clustering

Clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar (in some sense or another) to each other than to those in other groups. Clustering was performed using the ape (Analyses of Phylogenetic and Evolution) package. **Ape** [36] provides functions for reading and manipulating phylogenetic trees and DNA sequences, computing DNA distances, estimating trees with distance-based methods, and a range of methods for comparative analyses and analysis of diversification. All clustering analyses were performed with agglomerative hierarchical clustering using average linkage. In order to examine the correlation between the cell lines Pearson correlation was used as distance measure.

Enrichment tests

To investigate Sepsis relevant pathways in contrast to the other control samples enrichment tests were used. The test methods either need p-values or t-statistics as an input. Because of multiple hypotheses testing the p-values were Benjamini-Hochberg corrected to control the false discovery rate (*Benjamini & Hochberg, 1995*). The genes were ranked based on their corrected p-value in an ascending order and assigned to pathways using the KEGG database to create gene sets. Mapping of genes to pathways compiled 186 gene sets.

Quantitative Set Analysis for Gene Expression was performed using qusage package [37]. The qusage package accounts for inter-gene correlations using a Variance Inflation Factor technique that extends the method proposed by *Wu et al. (Nucleic Acids Res, 2012)*. In addition, rather than simply evaluating the deviation from a null hypothesis with a single number (a P value), qusage quantifies gene set activity with a complete probability density function (PDF). From this PDF, P values and confidence intervals can be easily extracted [37].

Lagged Correlation

Lagged relationships are characteristic of many natural physical systems. Lagged correlation refers to the correlation between two time series shifted in time relative to one another. Lagged correlation is important in studying the relationship between time series for two reasons. First, one series may have a delayed response to the other series, or perhaps a delayed response to a common stimulus that affects both series. Second, the response of one series to the other series or an outside stimulus may be "smeared" in time, such that a stimulus restricted to one observation elicits a response at multiple observations. For example, because of storage in reservoirs, glaciers, etc., the volume discharge of a river in one year may depend on precipitation in the several preceding years. Or because of changes in crown density and photosynthate storage, the width of a tree-ring in one year may depend on climate of several preceding years. The simple correlation coefficient between the two series properly aligned in time is inadequate to characterize the relationship in such situations. Useful functions we will examine as alternative to the simple correlation coefficient are the cross-correlation function and the impulse response function. The cross-correlation function is the correlation between the series shifted against one another as a function of number of observations of the offset.

If the individual series are auto correlated, the estimated cross-correlation function may be distorted and misleading as a measure of the lagged relationship [38].

The cross-correlation function (ccf) of two time series is the product-moment correlation as a function of lag, or time-offset, between the series. It is helpful to begin defining the ccf with a definition of the cross-covariance function (ccvf). Consider *N* pairs of observations on two time series, u_t and y_t. The sample ccvf is given by [38]:

Lagged correlation was performed between transcripts and clinical subtype counts to check the lagged time effects on LCN2 and HLA-DMB transcripts for Neutrophils and Lymphocytes.

A supplemental web-based portal (Massachusetts General Hospital, 2011) is available to explore in greater detail the largest clinical and genomic database to date from severely injured humans. Data in this study have been deposited in the GEO DataSets site under accession number GSE11375.

5. RESULTS AND DISCUSSION:

Differentially Expressed Genes

After pre-processing the list narrowed down to 18960 gene-centred and annotated features mapped by Entrez IDs. Filtering DEGs by FDR-adjusted P-values <0.05 (Wilcoxon test & T-Test) and average two-fold expression change yielded 1558 features for all pooled samples (septic versus non-septic groups). Because of multiple hypotheses testing the p- values were Benjamini-Hochberg corrected to control the false discovery rate (*Benjamini & Hochberg, 1995*).

TABLE 1: No. of Differentially Expressed Genes

Methods	DEGs	DEGs after FDR Correction
T-Test	13878	13673
Wilcoxon Test	14026	13787
Log2 Fold Change	1598	NA

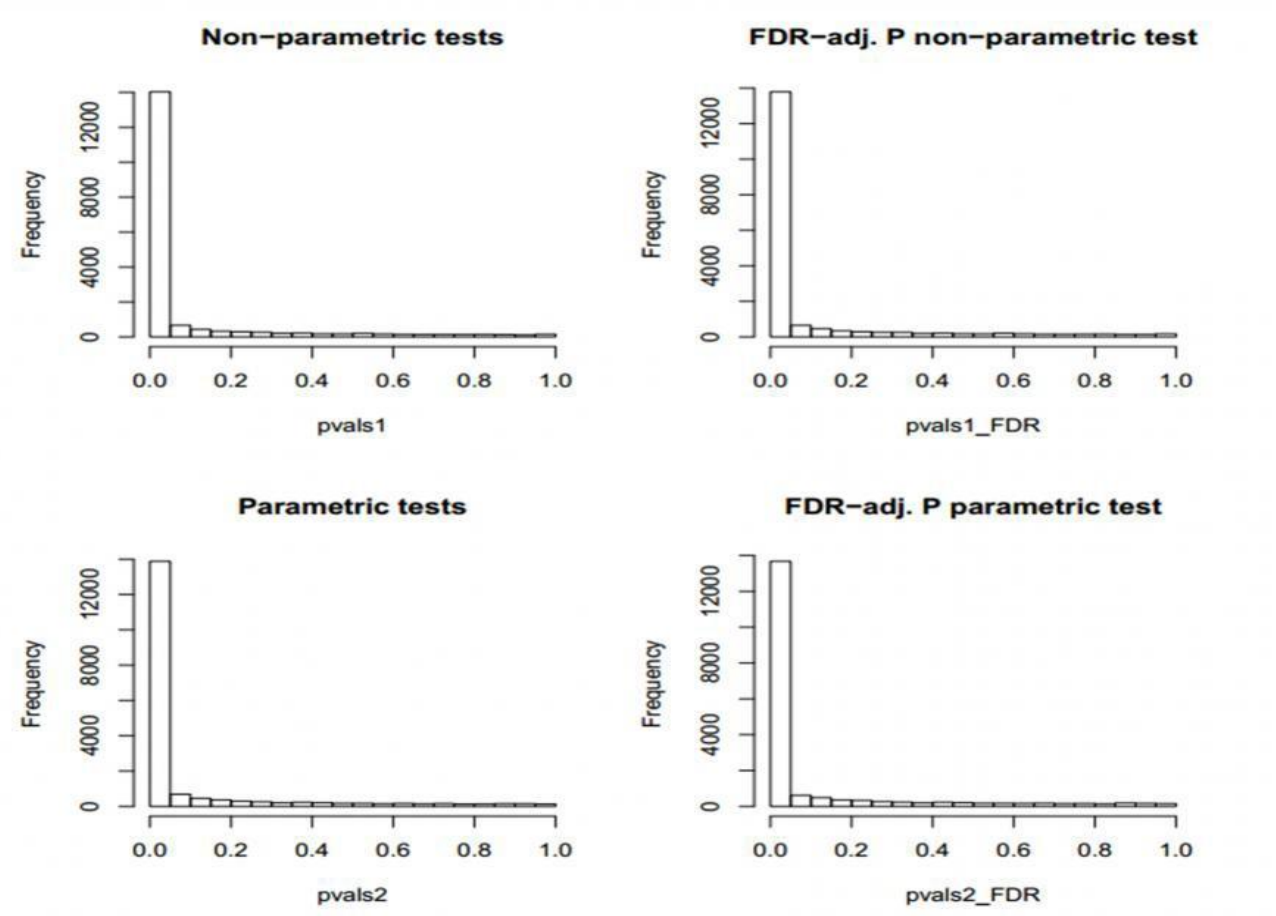


Figure 4: Histograms of p-values with and without multiple tests adj. in parametric and non-parametric version.

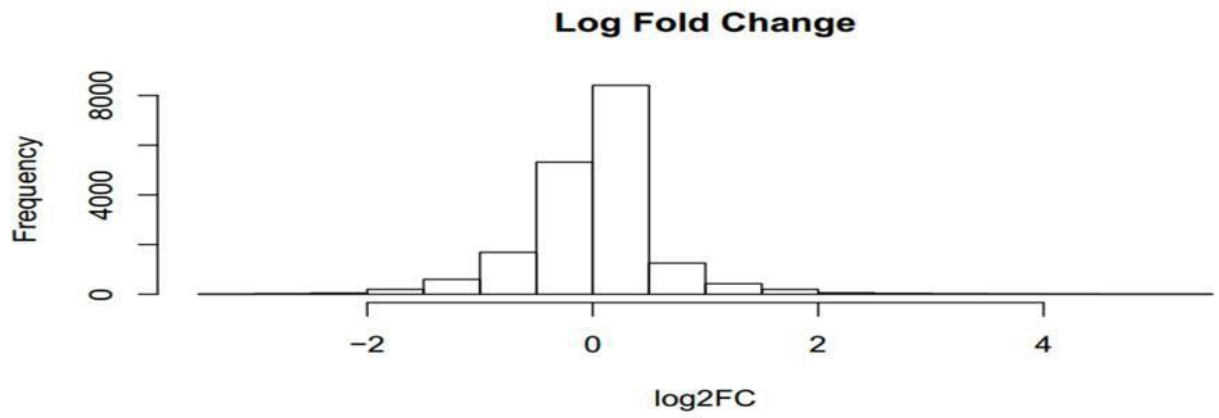


Figure 5: Histogram of Log2 Fold Change

Clustering:

Gender specific clustering was performed taking in consideration the various time points of the patient's data. Red color represents Females and Blue color represents Males. It is depicted from the outcome (figure) that Controls were grouped together with gender specificity and also the samples taken at early time point were grouped together (easily seen from the pinned out part of the dendrogram).

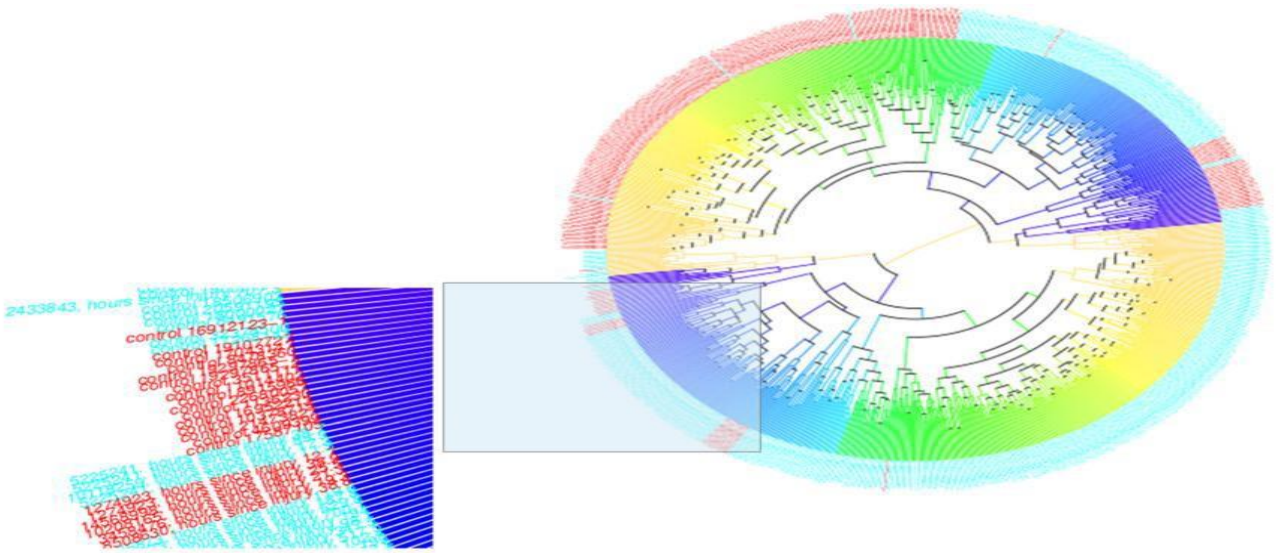


Figure 6: Hierarchical clustering of all samples

Regulation of some important genes:

Some important genes were noticed to be highly up-regulated or down-regulated in the data as mentioned below.

HLA-DMB & LCN2

HLA-DMB belongs to the HLA (Human Leukocyte Antigen) class II beta chain paralogues. This class II molecule is a heterodimer consisting of an alpha (DMA) and a beta (DMB) chain, both anchored in the membrane. DM plays a central role in the peptide loading of MHC class II molecules by helping to release the CLIP (class II-associated invariant chain peptide) molecule from the peptide binding site. Class II molecules are expressed in antigen presenting cells (APC: B lymphocytes, dendritic cells, macrophages). [39]

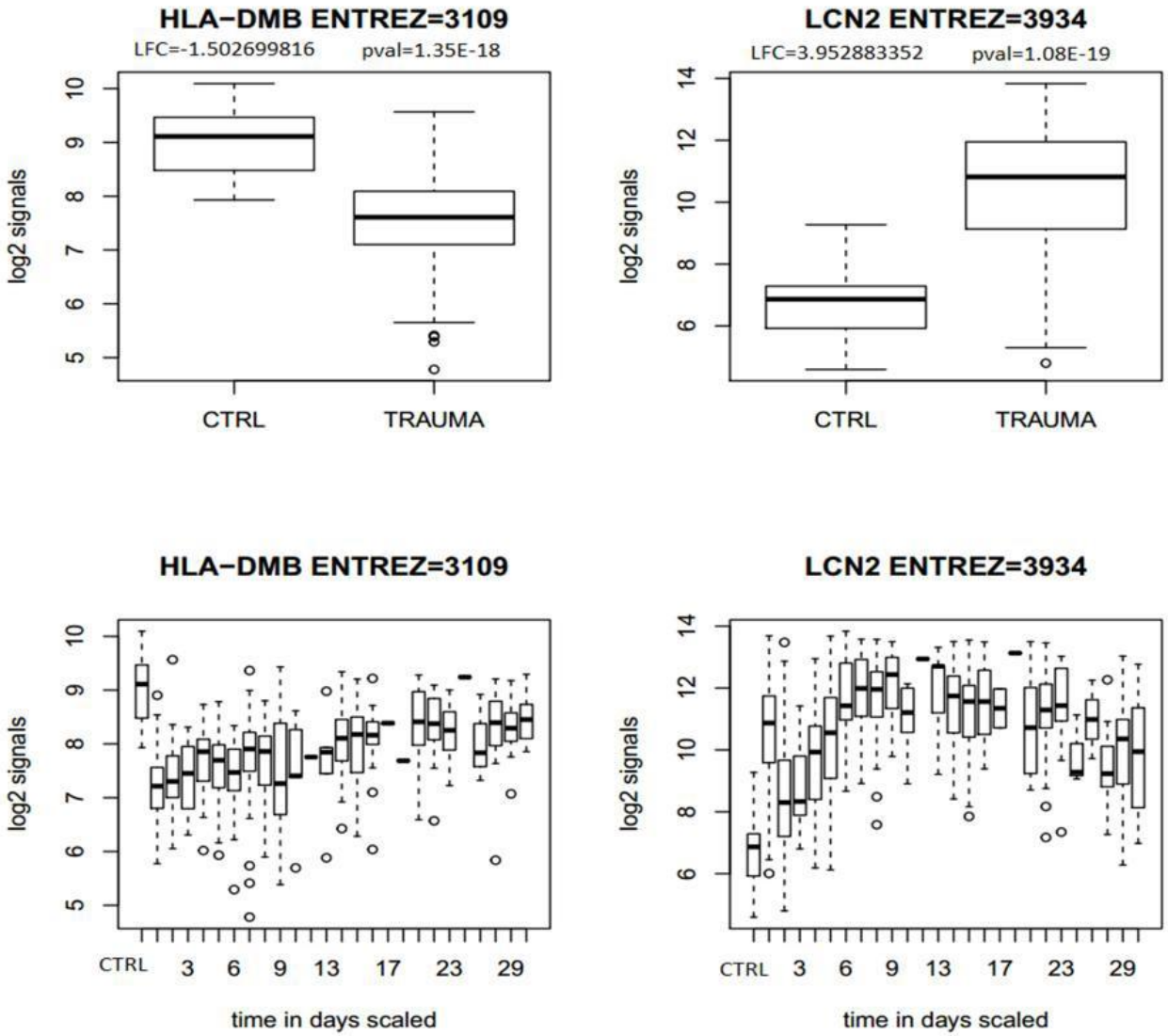


Figure 7: Box Plots of LCN2 & HLA-DMB

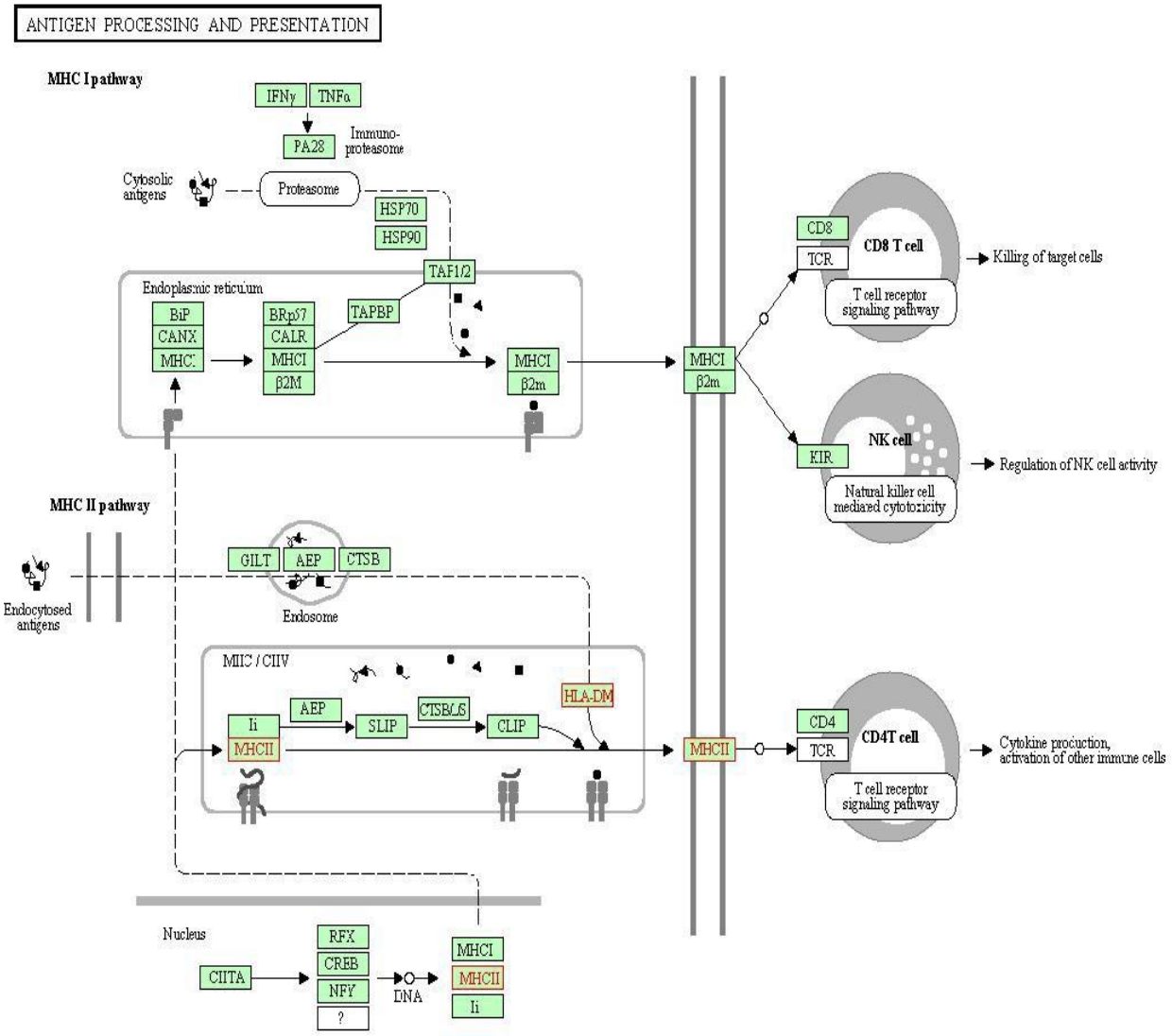


Figure 8: Antigen Processing and Presentation Pathway [41]

T-Cells are a subset of lymphocytes that play a large role in the immune response. The TCR (T-Cell Receptor) is a complex of integral membrane proteins that participates in the activation of T-Cells in response to the presentation of antigen. Stimulation of TCR is triggered by MHC (Major Histocompatibility Complex) molecules on Antigen Presenting Cells that present antigen peptides to TCR complexes and induce a series of intracellular signaling cascades. Engagement of the TCR initiates positive (signal-enhancing) and negative (signal-attenuating) cascades that ultimately result in cellular proliferation, differentiation, Cytokine production, and/or activation-induced cell death. These signaling cascades regulate T-Cell development, homeostasis,

activation, acquisition of effectors' functions and apoptosis. [40]

LCN2 (Lipocalin-2) also known as oncogene 24p3 or Neutrophil Gelatinase-Associated Lipocalin (NGAL). LCN2 is an iron-trafficking protein involved in multiple processes such as apoptosis, innate immunity and renal development. The binding of NGAL to bacterial siderophores is important in the innate immune response to bacterial infection. Upon encountering invading bacteria the toll-like receptors on immune cells stimulate the synthesis and secretion of NGAL. Secreted NGAL then limits bacterial growth by sequestering iron-containing siderophores. LCN2 also functions as growth factor. Originally, NGAL was isolated from a supernatant of activated human neutrophils.[42] Lack of LCN2 expression has been possibly linked to acne could be caused due to lack of gene expression, which possibly can be correct with Isotretinoin.[43,44].

Correlation of LCN 2and LTF

Lipocalin-2 (LCN2) and Lactotransferrin (LTF) found to be highly and positively correlated with the Pearson's product-moment correlation value of 0.944.

Lactoferrin (LF), also known as lactotransferrin (LTF), is a multifunctional protein of the transferrin family. Lactoferrin is a globular glycoprotein with a molecular mass of about 80 kDa that is widely represented in various secretory fluids, such as milk, saliva, tears, and nasal secretions. Lactoferrin is one of the components of the immune system of the body; it has antimicrobial activity (bacteriocide, fungicide) and is part of the innate defense, mainly at mucoses [45]. The important role of lactoferrin in human host defense and especially in lung [46]. Lactotransferrin is a major iron-binding and multifunctional protein found in exocrine fluids such as breast milk and mucosal secretions. Antimicrobial properties include bacteriostasis, which is related to its ability to sequester free iron and thus inhibit microbial growth, as well as direct bactericidal properties leading to the release of lipopolysaccharides from the bacterial outer membrane.

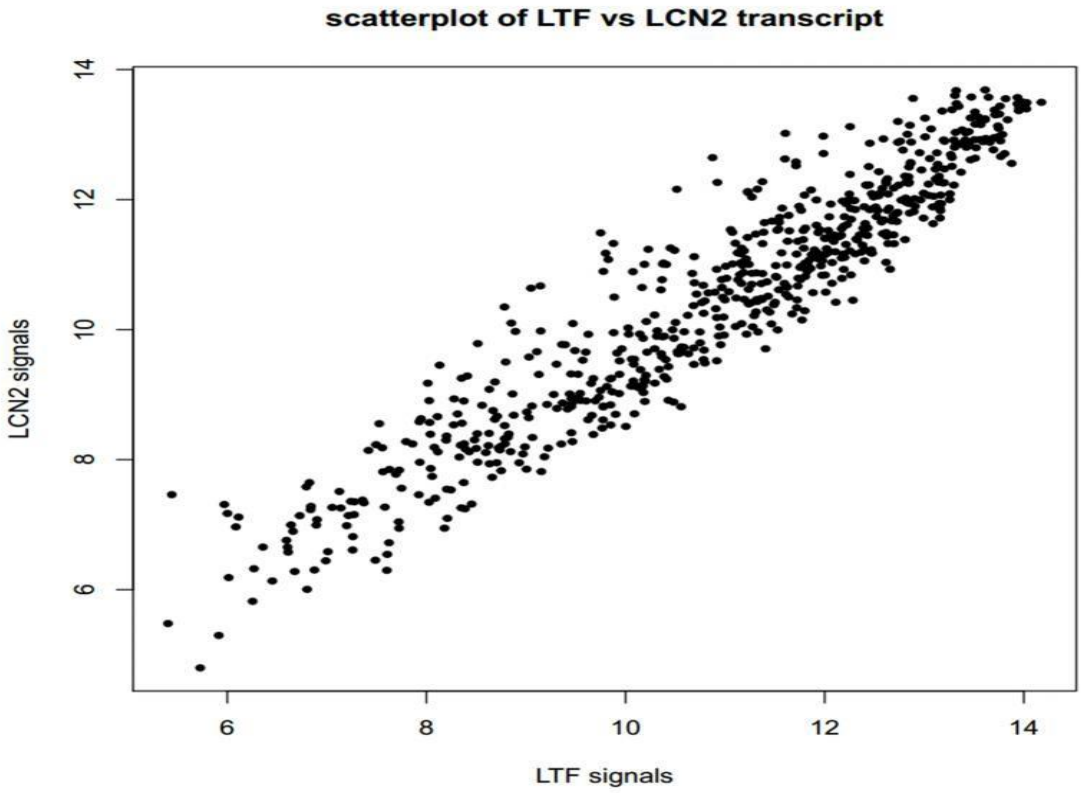


Figure 9: Pearson's product-moment correlation of LCN2 and LTF

(r = 0.9441)

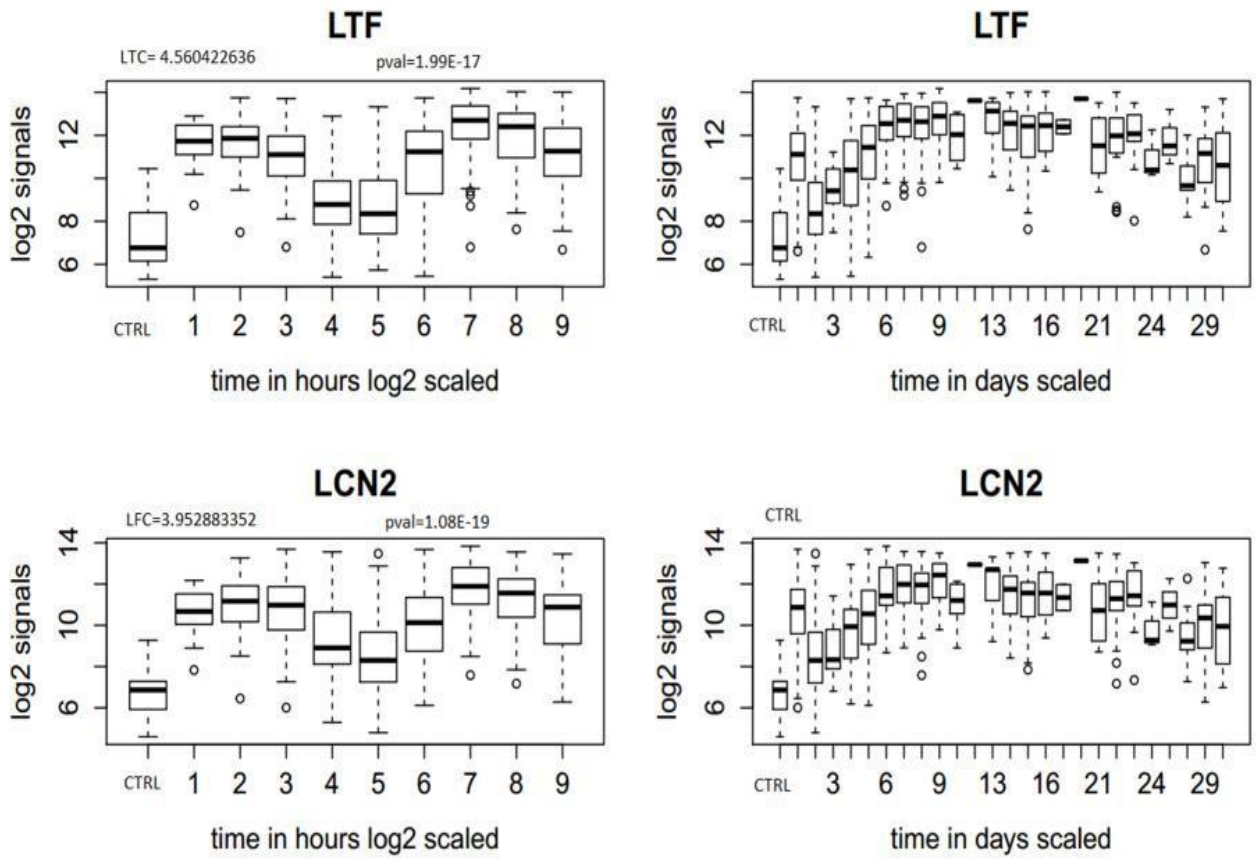


Figure 10: LTF & LCN expression

SLC4A1 & IL5RA

Both the genes have shown up regulation at the later stages in patients. IL5RA was first down regulated in patients, but at later stage, it regulated positively. IL5RA was found to be highly correlating with Eosinophils (r= 0.6136). SLC4A1 gene did not show much effect during initial stages, but was highly up regulated in patients at last stages.

The official name of **SLC4A1** gene is "solute carrier family 4 (anion exchanger), member 1 (Diego blood group)." From NCBI Gene: [47] The protein encoded by this gene is part of the anion exchanger (AE) family and is expressed in the erythrocyte plasma membrane, where it functions as a chloride/bicarbonate exchanger involved in carbon dioxide transport from tissues to lungs. From UniProt:[48] Band 3 is the major integral glycoprotein of the erythrocyte membrane. Band 3 has two functional domains. Its integral domain mediates a 1:1 exchange of inorganic anions across the membrane, whereas its cytoplasmic domain provides binding sites for cytoskeletal proteins,

glycolytic enzymes, and hemoglobin.

IL5RA Gene

The protein encoded by IL5RA (interleukin 5 receptor, alpha) gene is an interleukin 5 specific subunit of a heterodimeric cytokine receptor. Diseases associated with IL5RA include eosinophilic esophagitis (an allergic inflammatory condition of the esophagus, and also called allergic oesophagitis [49]. Symptoms are swallowing difficulty, food impaction, and heartburn) [50] and among its related super-pathways are STAT3 Pathway and Interleukin receptor SHC signaling. GO annotations related to this gene include protein binding and interleukin-5 receptor activity.

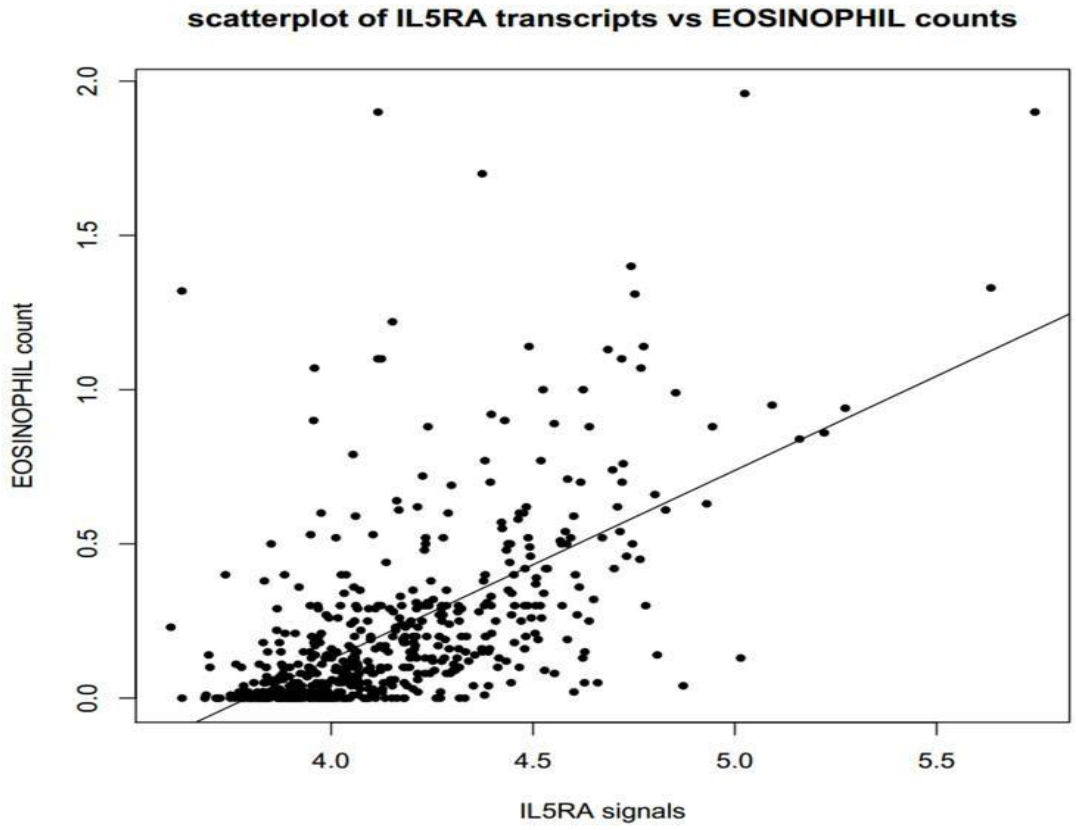


Figure 11: Scatter plot showing Correlation of IL5RA with Eosinophils

(r= 0.6136)

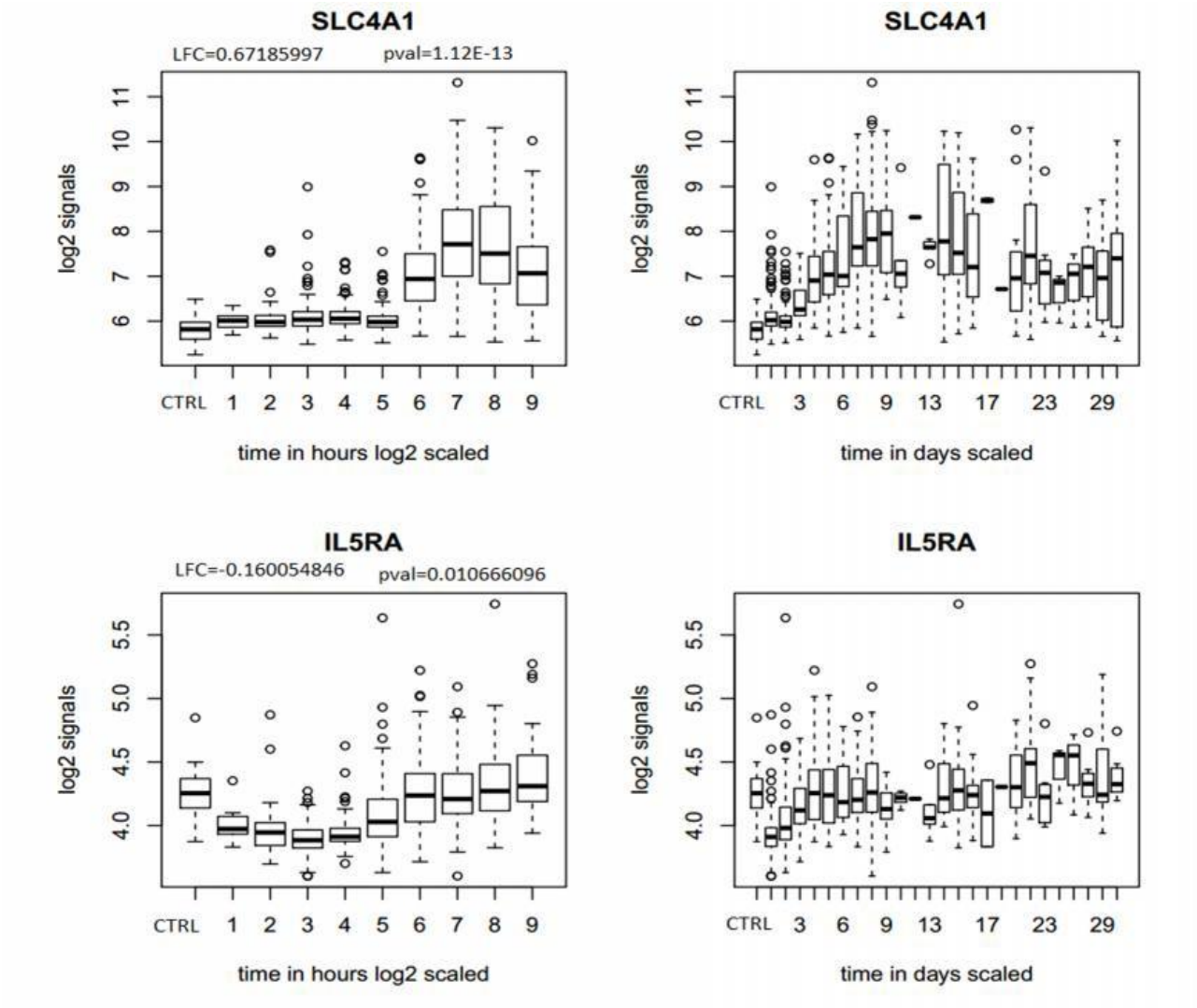


Figure 12: Plots of IL5RA and SLC4A

Gender Linked Genes:

DDX3Y

DX3Y (DEAD (Asp-Glu-Ala-Asp) box helicase 3, Y-linked) is a protein-coding gene and characterized by the conserved motif Asp-Glu-Ala-Asp (DEAD), are putative RNA helicases. This gene has a homolog on the X chromosome (DDX3X). The gene mutation causes male infertility, Sertoli cell only syndrome or severe hypo-spermatogenesis, suggesting that this gene plays a key role in the spermatogenic process [51],[52]. Diseases associated with DDX3Y include spermatocytoma and infertility.

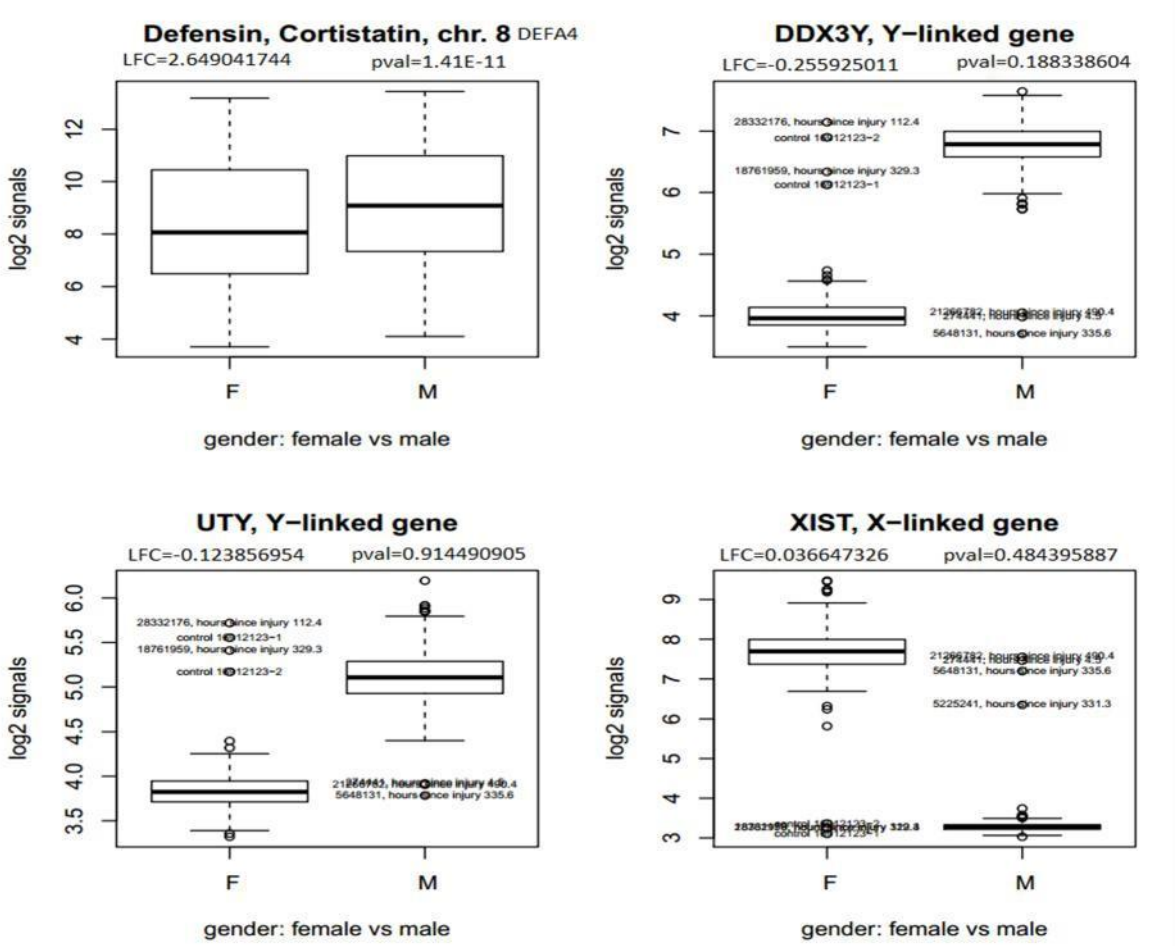


Figure 13: Sex linked genes (outliers identified)

Gene Set Enrichment Analysis (GSEA)

GSEA was performed by the qusage (Quantitative set analysis for gene expression) package. GSEA was performed to determine whether a priori defined set of genes shows statistically significant, concordant differences between two biological states. I obtained following top regulated pathways, on the basis of log fold change:

Table 2: Top KEGG pathways Enriched

Activity	**Log Fold Change**	**Knowledgebase**	**Category**	**FDR-adj.P-value**
Up regulated	0.4514927	KEGG	FolateBiosynthesis (hsa00790)	0.000000e+00
Up regulated	0.4231749	KEGG	Pathogen Escherichia Coli Infection (hsa05130)	0.000000e+00
Up regulated	0.3830689	KEGG	FC Gamma R Mediated Phagocytosis (hsa04666)	0.000000e+00
Up regulated	0.3755925	KEGG	Leukocyte Transendothelial Migration (hsa04670)	0.000000e+00
Up regulated	0.3552252	KEGG	Galactose Metabolism (hsa00052)	0.000000e+00
Up regulated	0.1783564	KEGG	Glycolysis gluconeogenesis (hsa00010)	0.000000e+00
Down	-0.5441087	KEGG	Ribosome (hsa03010)	2.284539e-13
Down regulated	-0.4133383	KEGG	Primary Immunodeficiency (hsa05340)	0.000000e+00
Down regulated	-0.3768175	KEGG	Valine, Leucine and Isoleucine Biosynthesis (hsa00290)	0.000000e+00
Down regulated	-0.3728089	KEGG	Aminocyl tRNA Biosynthesis (hsa00970)	0.000000e+00
Down regulated	-0.3619649	KEGG	Graft versus Host Disease (hsa05332)	3.689376e-12

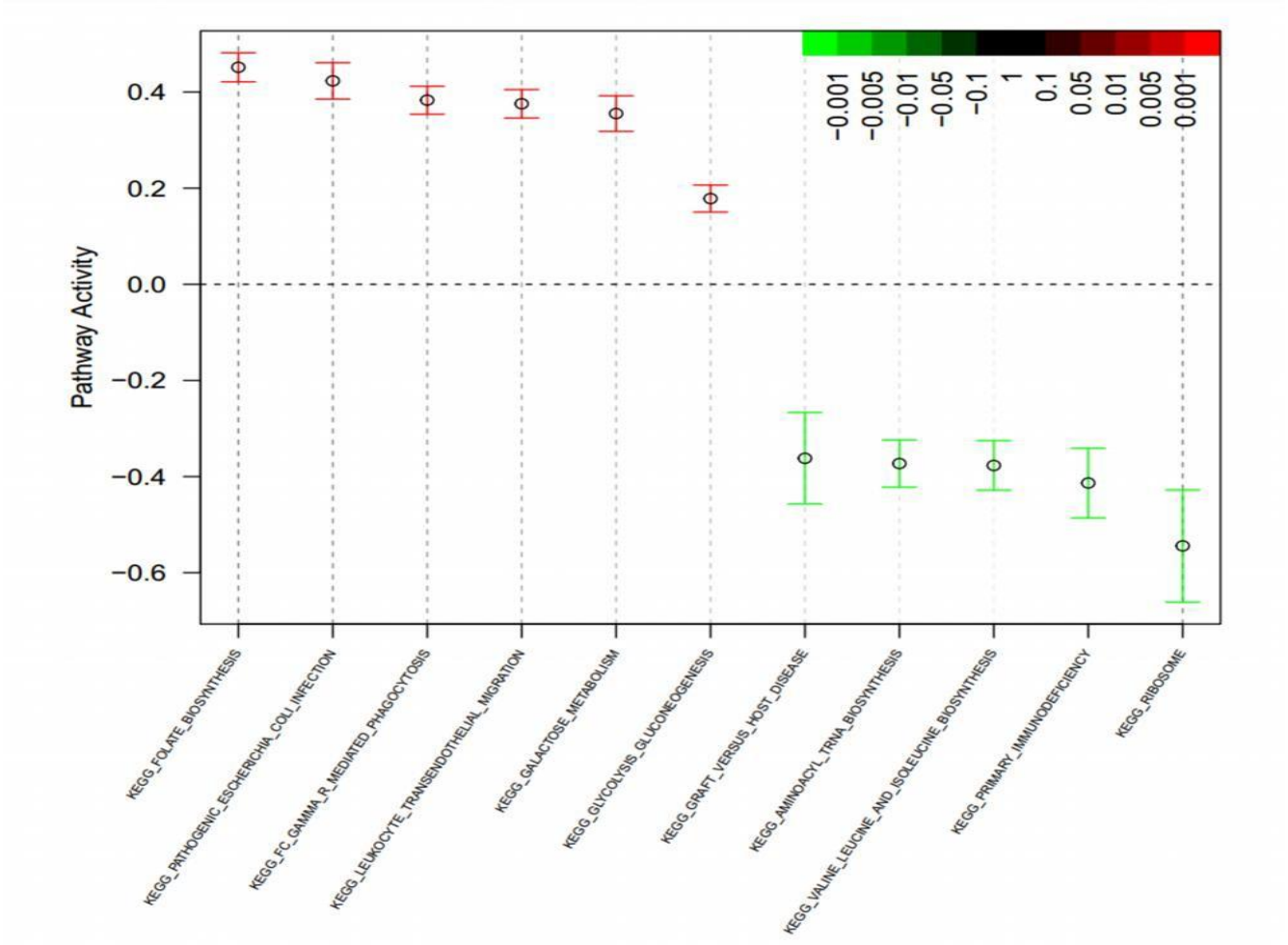


Figure 14: Top Up and Down regulated KEGG pathways

KEGG Mapper

KEGG Mapper was used (http://www.genome.jp/kegg/tool/map_pathway2.html) to construct the pathways with color codes, displaying the genes and their positions in the pathway. I made the color code as follows:

- Up regulated genes- RED
- Not much regulated genes- PINK
- Down regulated genes- BLUE

Glycolysis & Gluconeogenesis

Glycolysis is the process of converting glucose into pyruvate and generating small amounts of

ATP (energy) and NADH (reducing power). It is a central pathway that produces important precursor metabolites: six-carbon compounds of glucose-6P and fructose-6P and three-carbon compounds of glycerone-P, glyceraldehyde-3P, glycerate-3P, phosphoenolpyruvate, and pyruvate. Acetyl-CoA, another important precursor metabolite, is produced by oxidative decarboxylation of pyruvate. Gluconeogenesis is a synthesis pathway of glucose from non-carbohydrate precursors. It is essentially a reversal of glycolysis with minor variations of alternative paths [MD: M00003]. This pathway is the most important pathway, as it produces energy, which is required by the cells, in order to function. Septic patients have this metabolic pathway up-regulated, as the rate of metabolism increases in the patients [53].

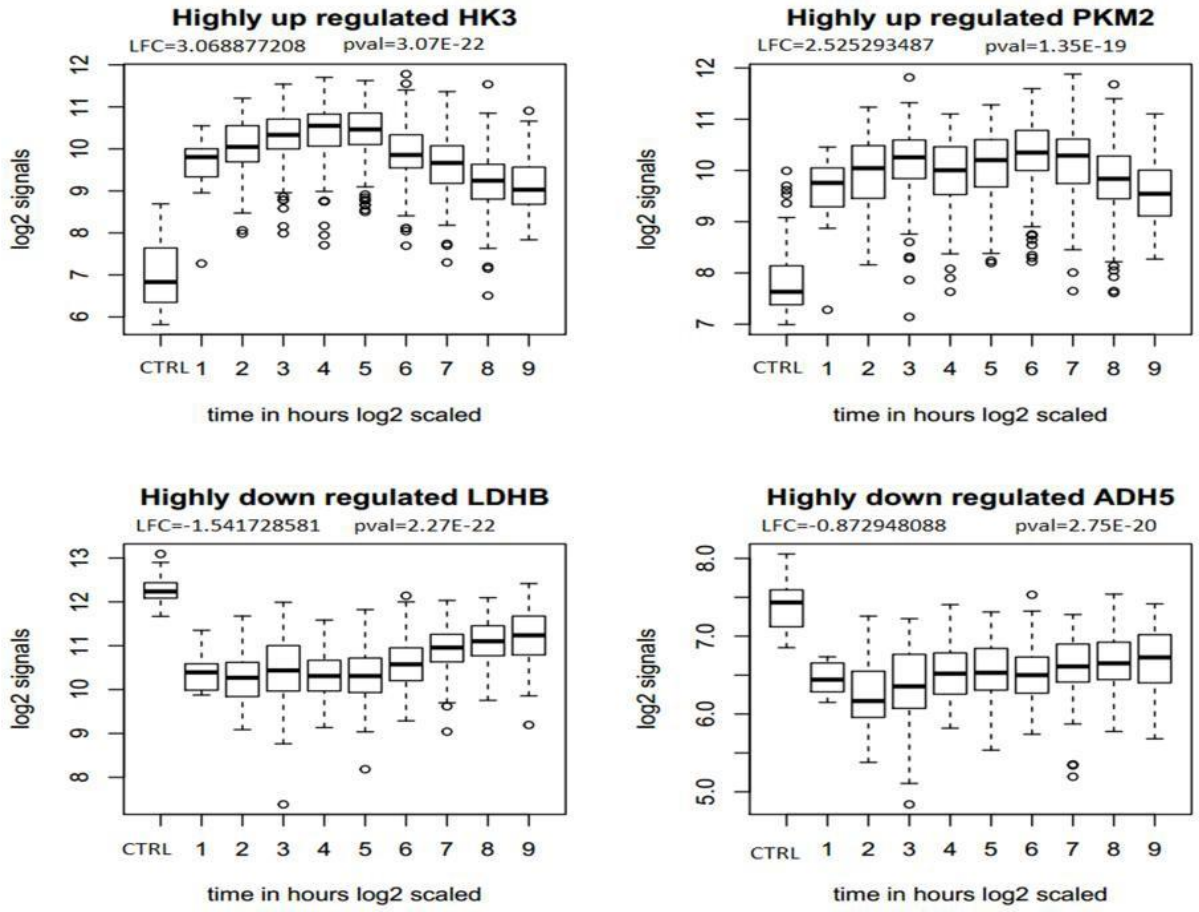


Figure 15: Box plot of highly up and down regulated genes of Glycolysis pathway

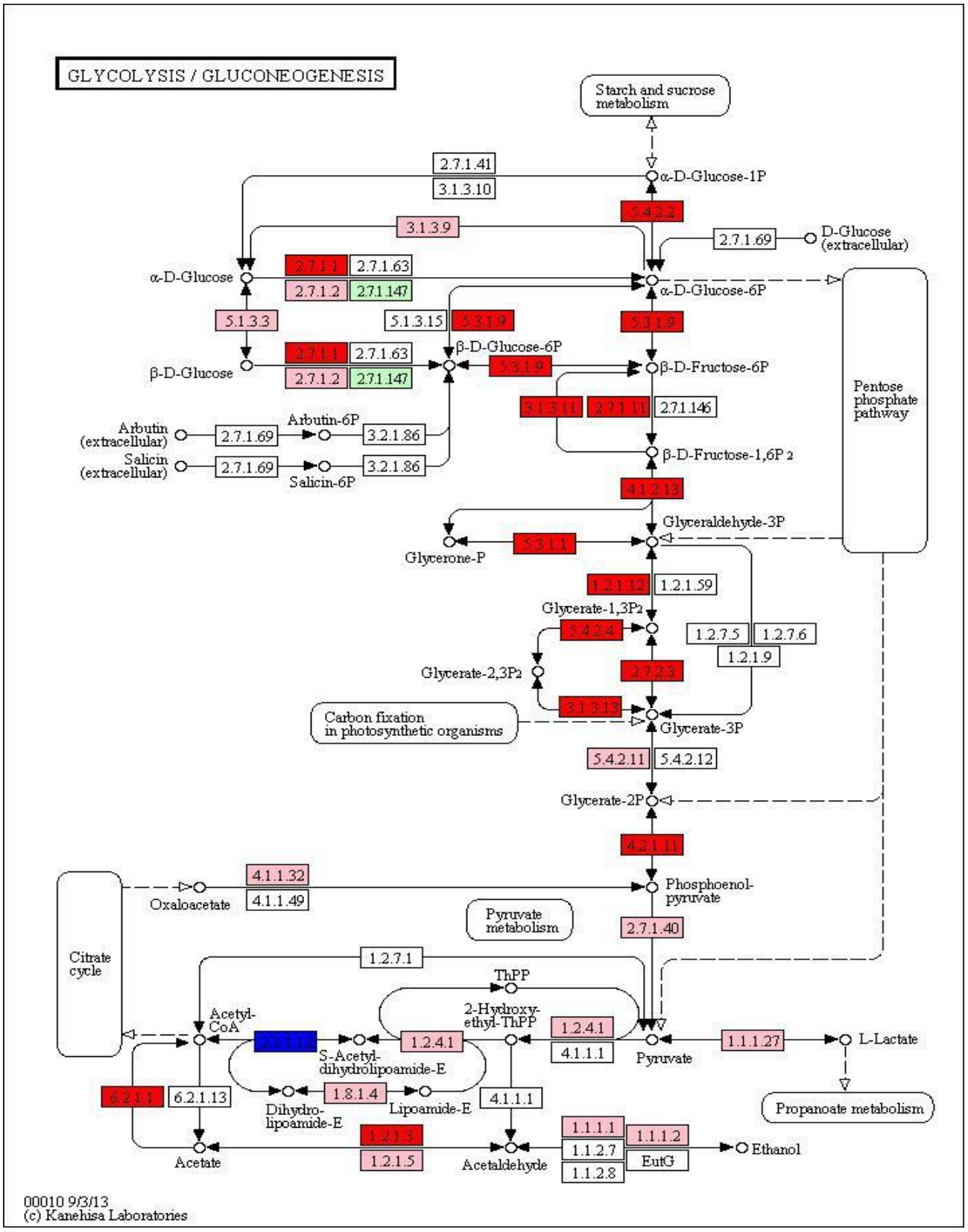


Figure 16: Glycolysis Gluconeogenesis pathway with genes regulation [54].

From literature survey, it is obvious that sepsis increases glucose utilization which could be measured by lactate and alanine production. However, the increased glucose uptake is not accompanied by corresponding increase in glucose oxidation. Instead, the glucose carbon is released from peripheral tissues into the venous blood as lactate and alanine production are normal or increased, it appears that glucose uptake and glycolysis is accelerated in Sepsis. Thus, glucose carbon is conserved by the body because oxidation would deplete the body stores of glucose carbon. Highly upregulated transcripts in glycolysis pathway include HK3 (Hexokinases) which has high affinity for glucose and phosphorylate glucose to produce glucose-6-phosphate, which is the very first step in most glucose metabolism pathways [55].

Ribosome

Ribosomes are the cellular factories responsible for making proteins. In eukaryotes, ribosome biogenesis involves the production and correct assembly of four rRNAs and about 80 ribosomal proteins. It requires hundreds of factors not present in the mature particle. In the absence of these proteins, ribosome biogenesis is stalled and cell growth is terminated even under optimal growth conditions [56]. Down-regulation of ribosomal pathway states that there are no more protein formation in the patients that is the cells are not dividing.

In addition to transcriptional regulation, posttranscriptional modifications such as mRNA stabilization may lead to cytokine superinduction via ribosomal inactivation in leukocytes and gut epithelial cells. RPS24, RPL31, RPL13, RPL22L1, RPL5, RPS27A, RPL4, RPL15, RPL14 and RPL22 transcripts were among the top down-regulated genes in the Ribosomal pathway, and only RPL8 and RPL10 were up-regulated.

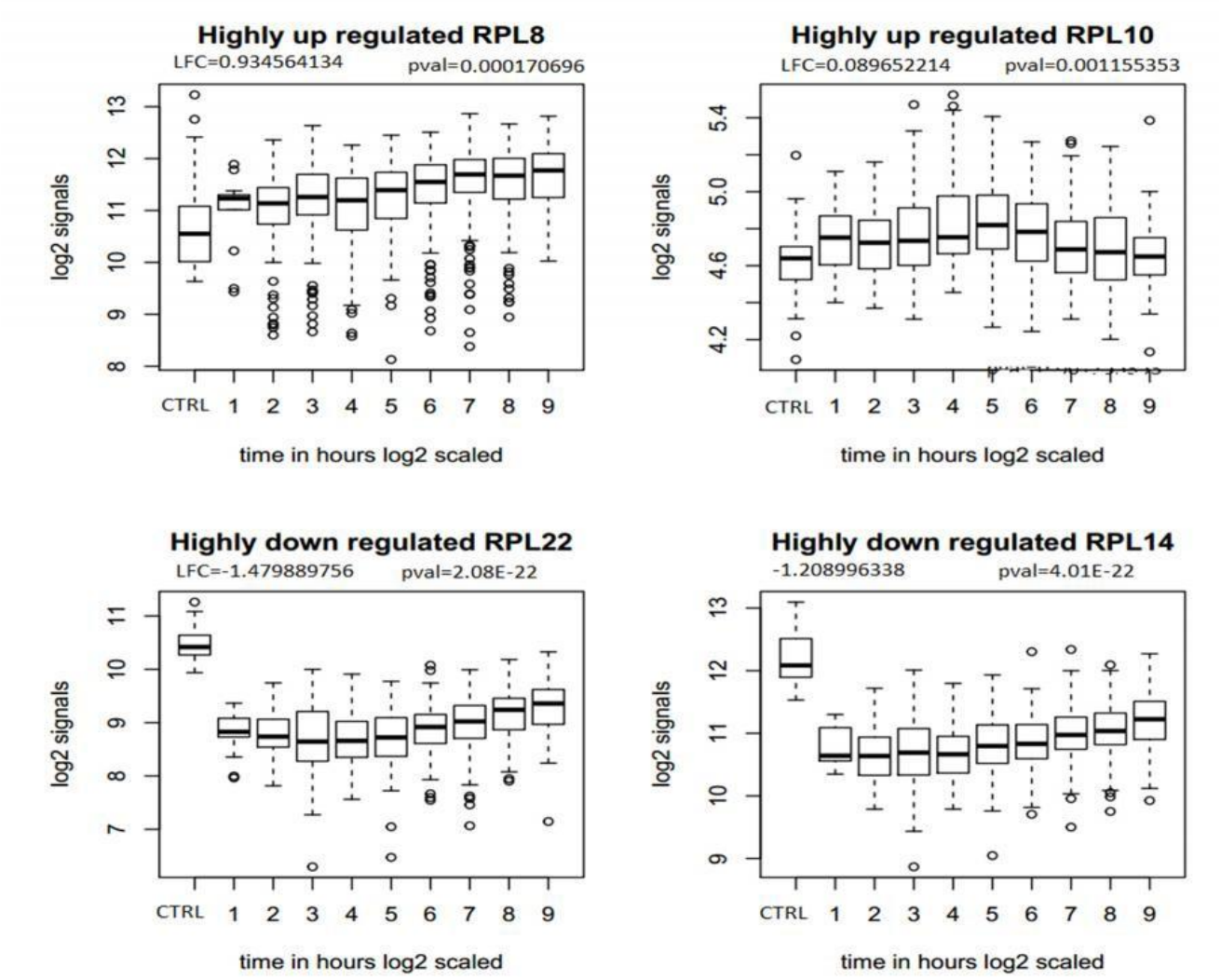


Figure 17: Box plot of highly up and down regulated genes of Ribosome pathway

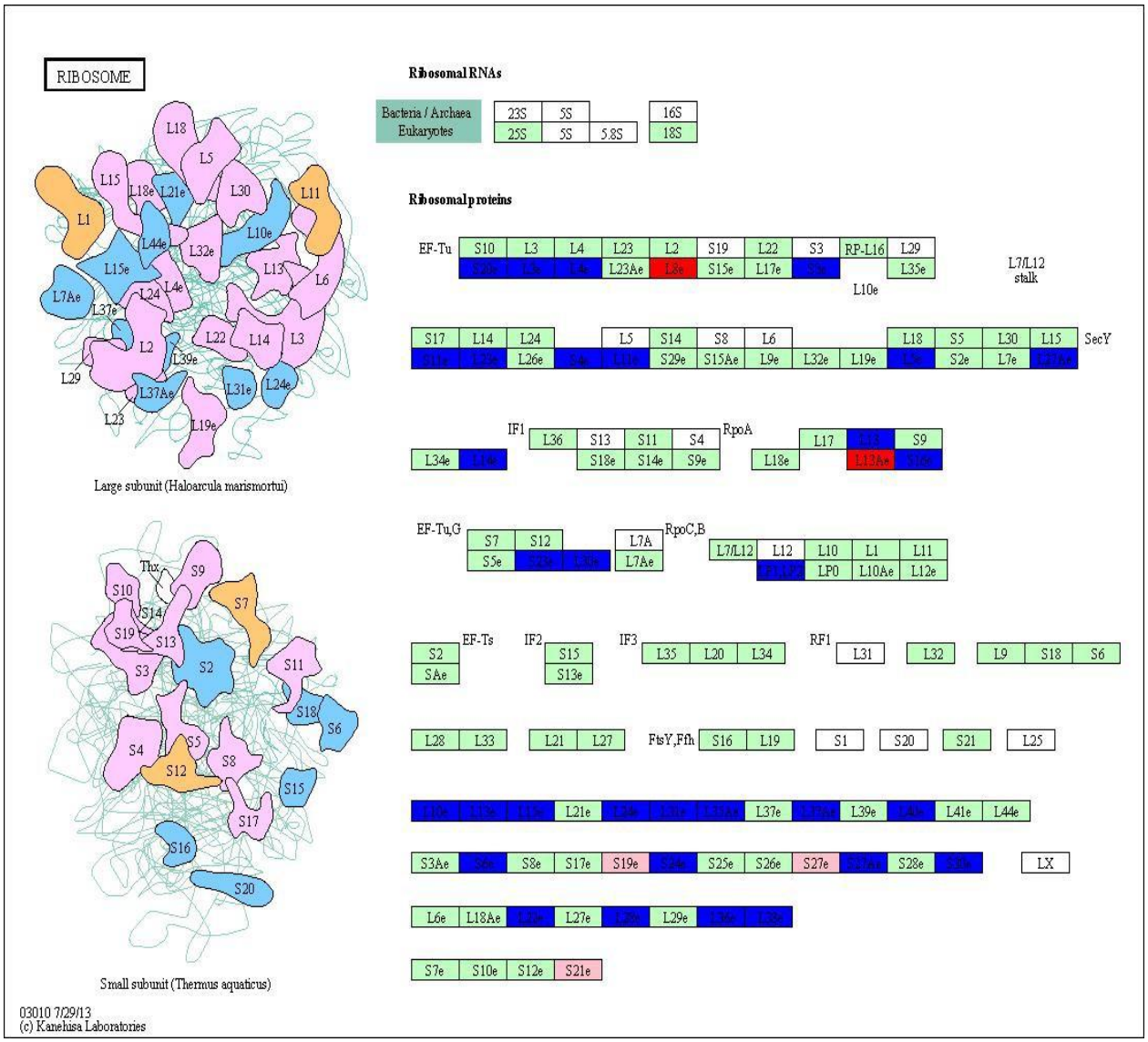
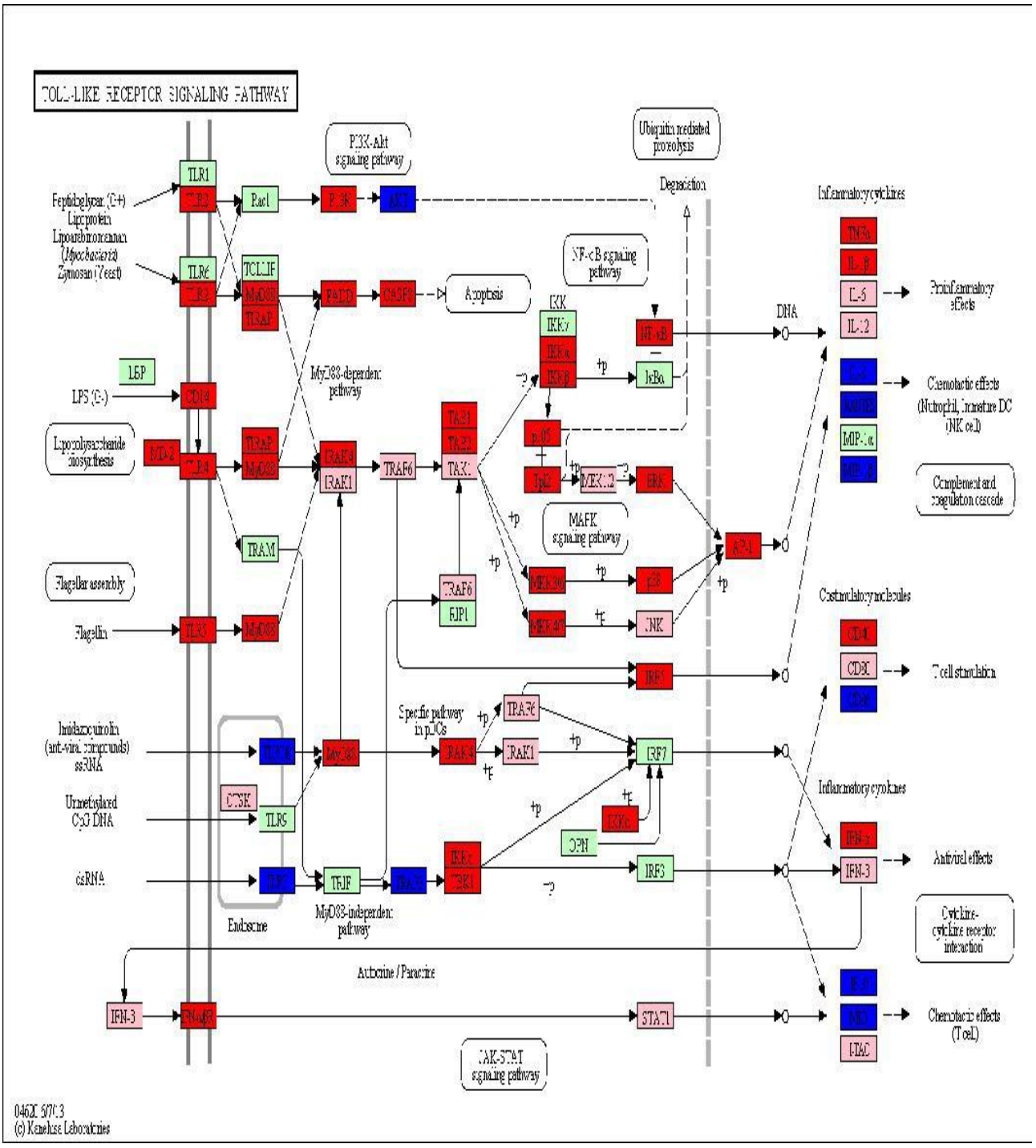


Figure 18: Ribosome pathway with genes regulation [57].

Toll like Receptors Signaling Pathway and Heat map

TLR are the specific families of pattern recognition receptors that are responsible for detecting microbial pathogens and generating innate immune responses. Toll-like receptors (TLRs) are membrane-bound receptors identified as homologs of Toll in Drosophila. TLR signaling pathways are separated into two groups: a MyD88-dependent pathway that leads to the production of pro-inflammatory cytokines with quick activation of NF-{kappa}B and MAPK, and a MyD88-independent pathway associated with the induction of IFN-beta and IFN-inducible genes, and maturation of dendritic cells with slow activation of NF-{kappa}B and MAPK [58].

Figure 19: TLR signalling pathway with genes regulation [59]



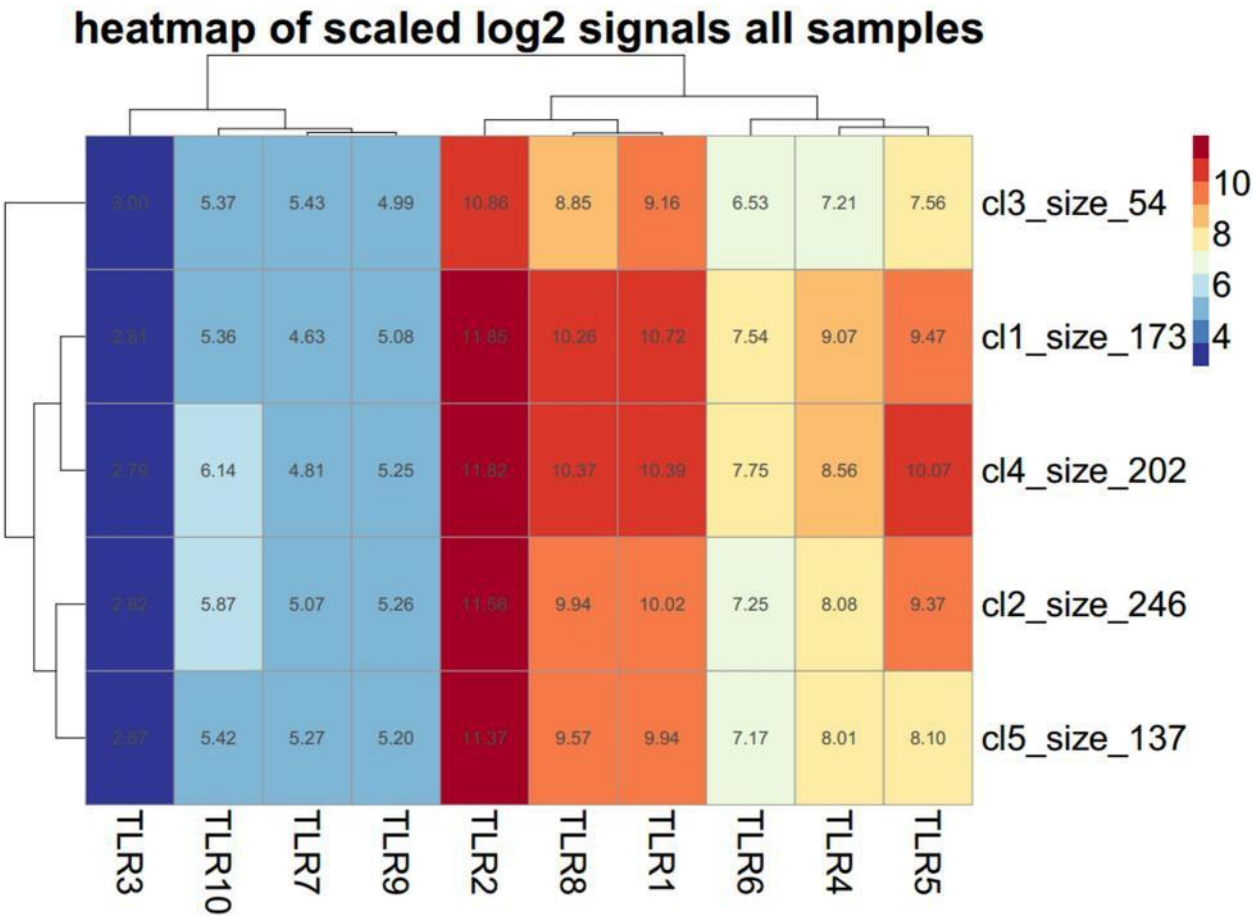


Figure 20: TLR genes heat map.

The expression of all of the Toll-like receptor (TLR) genes with the exception of TLR3, TLR 7 and TLR9, were increased after injury.

6. CONCLUSION:

High-throughput transcriptomic data enable researchers to monitor molecular dynamics on a broad scale and to determine promising diagnostic as well as interventional targets. A more comprehensive characterization of the genomic response to trauma is therefore required in order to increase our understanding of the molecular basis of clinical outcomes, leading to improvements in diagnosis and treatment.

The enrichment analysis offered several pathways, which seemed to be differentially expressed. The first one choose was Glycolysis, as of its specificity for sepsis. It is obvious that this pathway is enriched in Septic patients, as the rate of body metabolism increases during the disease and body works more rapidly to provide energy to the cells. This is evidence that the results of the enrichment analysis are constructive. Precisely, glycolysis is commonly known to be the most important energy source for cells. So this pathway could reveal disparities between the metabolism in patients and in healthy controls. The second selected pathway is Ribosome pathway. This could be an interesting pathway. The clinical data was available on the Glue Grant database (https://www.gluegrant.org/). The website had them and lots of more files about a year ago. Integrating the clinical variables give us more specific information regarding the conditions of a patient. A good correlation was found of IL5RA transcript with the Eosinophils ($r > 0.6$), so IL5RA could be a future biomarker. It is also possible to get the sample blood data, and perform some clinical tests and then incorporate the results with this data, possible, but not feasible (the clinicians don't like this associations for several reasons (e.g. after a long history of gene expression publications without this).

Some outliers were found in the sex linked genes. Outliers relate also to controls and hint towards some wrongly mapped samples, but in overall there is above 99% correct mapping, so close to perfect. If it would be more, than one could expect trauma kicking up a genosomal change. Just to reconfirm the activity of some genes during the Sepsis, the experiment was analyzed using one more dataset. Total RNA extracted from whole blood (lysed in Tempus tubes) drawn from pediatric patients with acute community acquired Staphylococcus aureus infection. This data contains the expression level of few genes (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30119). The main purpose to analyze this data is to evaluate the transcripts regulation during the Sepsis. It was very interesting that the LCN2 transcript was highly up regulated and the HLA-DMB was highly down regulated confirming the transcripts regulation in Genomic Storm data.

The DEG numbers which I get were very high and these numbers could have been the result of False Positives, as the number of control samples was very few. So, I tried to compare the

samples of first 24 samples and last 50 samples. Unfortunately, the no. of DEGs was still nearly the same in number.

The Models were developed with the linear modelling approach for multiple effects gene-wise and the conclusion out of it was that, patient ID, so individual response seems to be more influential than time. While evaluating the single patient effects by paired test, there were also some great results, for example, TXNIP (a gene responsible for type-1 and type-2 diabetes, pathway unknown) was found to be up regulated in older female. I believe that grouping the patients with of same age group and gender, could reveal many things.

7. REFERENCE:

[1] Mark Reimers, NCI,"An (opnionated) Guide to Microarray Data Analysis".

[2] http://www.people.vcu.edu/~mreimers/OGMDA/selecting.genes.html

[3] Jarno Tuimala, M. Minna Laine, DNA Microarray Data Analysis, CSC, Finnish Centre for Science; ISBN 952-9821-89-1; http://www.csc.fi/oppaat/siru/.

[4] Enuka Shay, 2003; Microarray cluster analysis and applications; University of Haifa.

[5] Angus DC, Linde-Zwirble WT, Lidicker J, Clermont G, Carcillo J, Pinsky MR: Epidemiology of severe sepsis in the United States: analysis of incidence, outcome and associated costs of care. Crit Care Med 2001, 29:1303-1310.

[6] Martin GS, Mennino DM, Eaton S, Moss M: The epidemiology of sepsis in the United States from 1979 through 2000. N Engl J Med 2003, 348:1546-1554.

[7] Alberti C, Brun-Buisson C, Burchardi H, Martin C, Goodman S, Artigas A, Sicignano A, Palazzo M, Moreno R, Boulmé R, Lepage E, et al.: Epidemiology of sepsis and infection in ICU patients from an international multicentre cohort study.

[8] http://www.cscc.uniklinikum-jena.de/en/Sepsis.html

[9] http://www.ccmtutorials.com/infection/sepsis/page3.htm

[10] American College of Chest Physicians/Society of Critical Care Medicine Consensus Conference: definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. Crit Care Med 1992; 20(6):864-874.

[11] Bone, R.C., Balk, R.A., Cerra, F.B., Dellinger,R.P.., Fein, A.M., Knaus, W.A., et al. (1992). Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. The ACCP/SCCM Concensus Conference Committee. American College of Chest Physicians/society of Critical Care Medicine. Chest, 101, 1644-1655.

[12] DeCamp MM, Demling RH (1988) Posttraumatic multisystem organ failure. J Am Med Assoc 260: 5304.

[13] Marshall JC, Vincent JL, Sibbald WJ (1995) Clinical Trials for the Treatment of Sepsis. Vincent JL, Sibbald WJ, eds. Berlin: Springer-Verlag. pp 122–138.

[14] Dewar D, Moore FA, Moore EE, Balogh Z (2009) Postinjury multiple organ failure. Injury

40: 912–918.

[15] Sasser, S.M., M. Varghese, M. Joshipura, and A. Kellermann. 2006. Preventing death and disability through the timely provision of pre-hospital trauma care. Bull. World Health Organ. 84:507. http://dx.doi.org/10.2471/BLT.06.033605.

[16] Probst, C., H.C. Pape, F. Hildebrand, G. Regel, L. Mahlke, P. Giannoudis, C. Krettek, and M.R. Grotz. 2009. 30 years of polytrauma care: An analysis of the change in strategies and results of 4849 cases treated at a single institution. Injury. 40:77–83. http://dx.doi.org/10.1016/j.injury.2008.10.004.

[17] Lindig S, et al. Age-independent co-expression of antimicrobial gene clusters in the blood of septic patients. Int J Antimicrob Agents (2013), http://dx.doi.org/10.1016/j.ijantimicag.2013.04.012.

[18] Opal SM (2003) Clinical trial design and outcomes in patients with severe sepsis. Shock 20: 295– 302.

[19] Baue AE (1997) Multiple organ failure, multiple organ dysfunction syndrome, and systemic inflammatory response syndrome. Why no magic bullets? Arch Surg 132: 703–707.

[20] Giannoudis, P.V. 2003. Current concepts of the inflammatory response after major trauma: An update. Injury. 34:397–404. http://dx.doi .org/10.1016/S0020-1383(02)00416-3

[21] DeLong, W.G. Jr., and C.T. Born. 2004. Cytokines in patients with poly-trauma. Clin.Orthop.Relat.Res.(422):57–65. http://dx.doi.org/10.1097/01.blo.0000130840.64528.1e

[22] Giannoudis, P.V., F. Hildebrand, and H.C. Pape. 2004. Inflammatory serum markers in patients with multiple trauma. Can they predict outcome? J. Bone Joint Surg.Br. 86:313–323. http://dx.doi.org/10.1302/0301-620X.86B3.15035

[23] Hotchkiss, R.S., A. Strasser, J.E. McDunn, and P.E. Swanson. 2009. Cell death. N. Engl. J. Med.361:1570–1583. http://dx.doi.org/10.1056/NEJMra0901217

[24] Hotchkiss, R.S., and I.E. Karl. 2003. The pathophysiology and treatment of sepsis. N. Engl. J. Med.348:138–150. http://dx.doi.org/10.1056/NEJMra021333

[25] Keel, M., and O. Trentz. 2005. Pathophysiology of poly-trauma. Injury. 36:691–709. http://dx.doi.org/10.1016/j.injury.2004.12.037

[26] http://www.scmm.utmb.edu/genomics/microarrays/results.asp

[27] http://www.ncbi.nlm.nih.gov/geo/

[28] http://en.wikipedia.org/wiki/KEGG

[29] Kanehisa M (1997). "A database for post-genome analysis". Trends Genet 13(9): 375–6.doi:10.1016/S0168-9525(97)01223-7. PMID 9287494

[30] Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, et al. (2006). "From genomics to chemical genomics: new developments in KEGG". Nucleic Acids Res 34 (Database issue): D354–7. doi:10.1093/nar/gkj102. PMC 1347464. PMID 16381885.

[31] Xiao W, Mindrinos MN, Seok J, Cuschieri J et al. A genomic storm in critically injured humans. J Exp Med 2011 Dec 19;208(13):2581-90. PMID: 22110166

[32] Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004;5:R80.

[33] Xiao W, Mindrinos MN, Seok J, Cuschieri J et al. A genomic storm in critically injured humans. J Exp Med 2011 Dec 19;208(13):2581-90. PMID: 22110166

[34] Sandberg R, Larsson O. Improved precision and accuracy for microarrays using updated probe set definitions. BMC Bioinformatics 2007;8:48.

[35] Shi L, Jones WD, Jensen RV, Harris SC, Perkins RG, Goodsaid FM, et al. The balance of reproducibility, sensitivity, and specificity of lists of differentially expressed genes in microarray studies. BMC Bioinformatics 2008;9(Suppl. 9):S10.

[36] Paradis E., Claude J. & Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20: 289-290.

[37] Yaari G, Bolen CR, Thakar J, Kleinstein SH. Quantitative set analysis for gene expression: a method to quantify gene set differential expression including gene-gene correlations. Nucleic Acids Res. 2013 Aug 5.

[38] A Basis for Brain and Adaptive Systems, by Zhe Chen, Simon Haykin, Jos J. Eggermont, and Suzanna Becker Copyright□2007 John Wiley & Sons, Inc.

[39] http://en.wikipedia.org/wiki/HLA-DMB

[40]http://www.lifetechnologies.com/in/en/home/life-science/cell-analysis/signaling-pathways/t-cell- receptor-tcr/t-cell-receptor-tcr-overview.html

[41] http://www.genome.jp/kegg-bin/show_pathway?hsa04612

[42] Kjeldsen L, Johnsen AH, Sengeløv H, Borregaard N (May 1993). "Isolation and primary structure of NGAL, a novel protein associated with human neutrophil gelatinase". J. Biol. Chem. 268 (14): 10425– 32. PMID 7683678

[43] Nelson AM, Zhao W, Gilliland KL, Zaenglein AL, Liu W, Thiboutot DM (April 2008). "Neutrophil gelatinase–associated lipocalin mediates 13-cis retinoic acid–induced apoptosis of human sebaceous gland cells". J. Clin. Invest. 118 (4): 1468–78. doi:10.1172/JCI33869. PMC 2262030. PMID 18317594

[44] Nelson AM, Zhao W, Gilliland KL, Zaenglein AL, Liu W, Thiboutot DM (March/April 2009). "Early gene changes induced by isotretinoin in the skin provide clues to its mechanism of action". Dermato- Endocrinology 1 (2): 100–1. doi:10.4161/derm.1.2.8107. PMC 2835899. PMID 20224692.

[45] Sánchez L, Calvo M, Brock JH (1992). "Biological role of lactoferrin". Arch. Dis. Child. 67 (5): 657–61. doi:10.1136/adc.67.5.657. PMC 1793702. PMID 1599309.

[46] Rogan MP, Geraghty P, Greene CM, O'Neill SJ, Taggart CC, McElvaney NG (2006). "Antimicrobial proteins and polypeptides in pulmonary innate defence". Respir. Res.7 (1): 29. doi:10.1186/1465- 9921-7-29. PMC 1386663. PMID 16503962

[47] http://www.ncbi.nlm.nih.gov/gene/6521

[48] http://www.uniprot.org/uniprot/P02730

[49] http://www.ncbi.nlm.nih.gov/gene?cmd=Retrieve&dopt=full_report&list_uids=3568

[50] Nurko, Samuel; Furuta, G T (2006). "Eosinophilic esophagitis". GI Motility online. doi:10.1038/gimo49

[51] Lahn BT, Page DC (Nov 1997). "Functional coherence of the human Y chromosome". Science 278 (5338): 675–80. doi:10.1126/science.278.5338.675. PMID 9381176.

[52] "Entrez Gene: DDX3Y DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, Y-linked". [53] http://www.genome.jp/dbget-bin/www_bget?pathway+hsa00010

[54] http://www.genome.jp/kegg-bin/show_pathway?13906663561694/hsa00010.args

[55]http://web.squ.edu.om/medLib/MED_CD/E_CDs/anesthesia/site/content/v05/050132r00.html

[56] http://www.genome.jp/dbget-bin/www_bget?hsa03008

[57]http://www.genome.jp/keggbin/show_pathway?map=ko03010&show_description=show

[58] http://www.wikipathways.org/index.php/Pathway:WP75

[59] http://www.genome.jp/kegg-bin/show_pathway?139072973331064/hsa04620.args.