Genomic Information Science & Technology

Project: Computational Analysis of TCGA Cancer Data

Wanying Li

December 12, 2016

Contents

1 Introduction

The project combines some tasks of analyzing gene expression data to derive associations among genes as well as disease phenotypes, co-expression signatures, and survival analysis. The literature search in the biological/medical literature provides an understanding of the underlying biology behind results reached by computational methods.

2 Signatures and Mutual Information (M.I.)

Given three seed genes (TPX2, COL1A2, and LCP2), all genes in the data are ranked in terms of their association with the seed genes. The association is measured by mutual information (M.I). A fictitious "metagene" whose expression is the average expression of the top-ranked genes is used in a iterative loop to find the new top-ranked genes. The loop will eventually converge and the "signature" is defined to be the set of the converged top ten genes.

If the seeded gene is TPX2, it takes 2 iterations until the top ten associated genes to converge. The result of iterative loop, ie. the signature represented by the ten genes, is shown in table 1. If the seed gene is COL1A2, it takes 2 iterations until the top ten associated genes to converge. The result is shown in table 2. If the seed gene is LCP2, it takes 3 iterations until the top ten associated genes to converge. The result is shown in table 3.

Seed gene = TPX2	
Top 10 Genes	M.I.
KIF4A	0.8699
PLK1	0.8501
TPX2	0.8453
NCAPG	0.8407
MYBL2	0.8307
HJURP	0.8201
DLGAP5	0.8087
SPC25	0.8029
NCAPH	0.7963
ASPM	0.7952

Table 1. The top 10 genes associated with TPX2.

Seed gene = COL1A2	
Top 10 Genes	M.I.
COL1A2	0.9376
COL1A1	0.9158
COL6A3	0.9080
COL3A1	0.8784
COL5A1	0.8660
THBS2	0.8148
BGN	0.8140
SPARC	0.8103
COL5A2	0.8099
NID2	0.8052

Table 2. The top 10 genes associated with COL1A2.

Seed gene = LCP2	
Top 10 Genes	M.I.
NCKAP1L	0.9228
PTPRC	0.9164
LAIR1	0.9145
SASH3	0.9071
LCP2	0.9047
HAVCR2	0.9028
GPR65	0.9021
CD53	0.9002
CYBB	0.8983
GIMAP4	0.8902

Table 3. The top 10 genes associated with LCP2.

3 Signature Strength and Co-expression

The lowest of mutual information is a measure of the strength of co-expression of the signature. The signature with the highest strength is LCP2 (strength = 0.8902).

Figure 1 presents a color-coded scatter plot for the top three genes among the ten that are associated with LCP2. The top three genes are NCKAP1L, PTPRC and LAIR1. This plot shows that the

expression level of these three genes go up and down together, indicated by the linear relationship, which confirms the co-expression in these three genes.

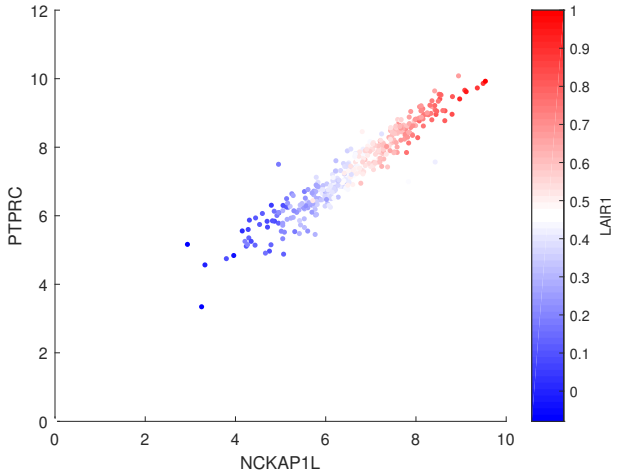

Figure 1. The scatter plot for the top three genes among the ten that are associated with LCP2. Gene NCKAP1L is represented by the x axis, gene PTPRC is represented by the y axis, and gene LAIR1 is color-coded.

Figure 2, 3 shows the color-coded scatter plots for the top three genes that are associated with COL1A2 (the 2nd highest strength signature) and TXP2 (the lowest strength signature) respectively. Both plots present a linear relationship between the top three genes that are associated with TXP2 and COL1A2. However, it can be observed that the plot for TPX2 has more points scattering around the line, which indicates a weaker co-expression comparing to that of the LCP2.

4 Lymphocyte Infiltration Signature: LCP2 Associated Genes

The re-occurring words in the descriptions (of the ten genes associated with LCP2) include "lymphocytes", "T-cell", "immune response", "signal transduction", "regulation" and "growth control". The gene descriptions are included in Appendix C. This leads us to conclude that the presence of these genes have an influence on the signal transduction and the regulation of T-cells (a type of lymphocytes), which can affect the immune response and play a role in the growth control. Given that tumor is an abnormal growth of tissue, these set of genes is likely to have a correlation with the growth of tumor. Therefore, this signature is called lymphocyte infiltration, representing the infiltration of possibly a specific T-lymphocytes in a tumor.

The top-ranked genes seem to point to a specific type of lymphocytes in the leukocyte infiltration. Research has hypothesized that the infiltration lymphocytes are T cells that have undergone a certain type of co-stimulation, and this may serve as a foundation for related adoptive transfer therapy [1, 2]. Given that the gene membership of the attractor signature gaves us some insight into the underlying immune mechanism, it could be valuable towards generating hypotheses for potential immunotherapies:

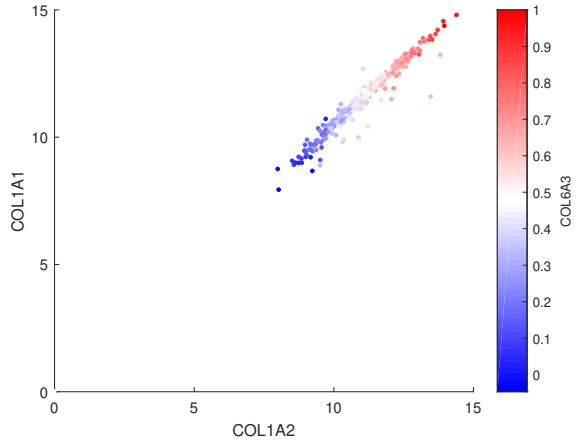

Figure 2. The scatter plot for the top three genes among the ten that are associated with COL1A2. Gene COL1A2 is represented by the x axis, gene COL1A1 is represented by the y axis, and gene COL6A3 is color-coded.

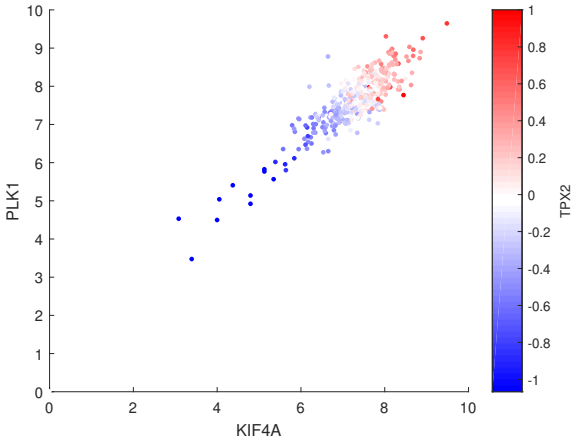

Figure 3. The scatter plot for the top three genes among the ten that are associated with TPX2. Gene KIF4A is represented by the x axis, gene PLK1 is represented by the y axis, and gene TPX2 is color-coded.

For example, the presence of the signal-transducing LCP2 (aka SLP-76) gene, together with the adaptor FYB (aka ADAP), suggests the formation of the SLP-76-ADAP adaptor module, which is known to regulate lymphocyte co-stimulation mediated by integrin ITGB2 (aka LFA-1), another prominent gene in the attractor [3].

5 Survival Analysis on Breast Cancer Data

By sorting according to the concordance index, the top five protective genes for breast cancer are founded to be:

Top 5 Protective Genes	Concordance Index
LOC100128977	0.3100
FGD3	0.3226
RLN1	0.3270
LOC100130148	0.3284
SUSD3	0.3284

Table 4. The top 5 protective genes from the breast cancer data set.

Gene LOC100128977 and LOC100130148 are on the same chromosome – chromosome 17 of the human genome. These genes are very close, in fact, they are located right next to each other. LOC100128977 is at 545,477-597,641 and LOC100130148 is located at 597,911-600,926. On the other hand, gene FGD3, RLN1 and SUSD3 are on the same chromosome – chromosome 9 of the human genome. FGD3 and SUSD3 are closer to each other than to RLN1. FGD3 locates at

92,947,451-93,036,236, SUSD3 locates at 93,058,688-93,085,132, whereas RLN1 locates at an earlier point 5,334,969-5,339,873.

The Kaplan-Meier plot (figure 4) for the top-ranked (ie. most protective) gene LOC100128977 suggests the subpopulation with low LOC100128977 expression (red line) has worse prognosis than that with high LOC100128977 expression (blue line). In other words, the low expression of gene LOC100128977 is associated with poor survival in breast cancer and vice versa. Comparing to the Kaplan-Meier plot for FGD3 (figure 5), both exhibit similar trends. We know that FGD3 is a protective gene from lecture notes, confirming the result obtained for LOC100128977.

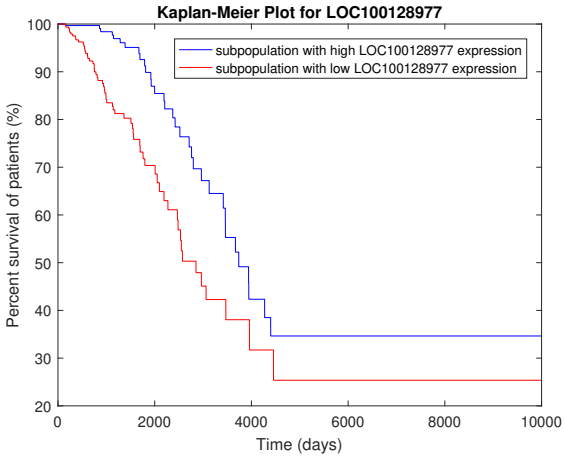

Figure 4. The Kaplan-Meier plot for the most protective gene – LOC100128977.

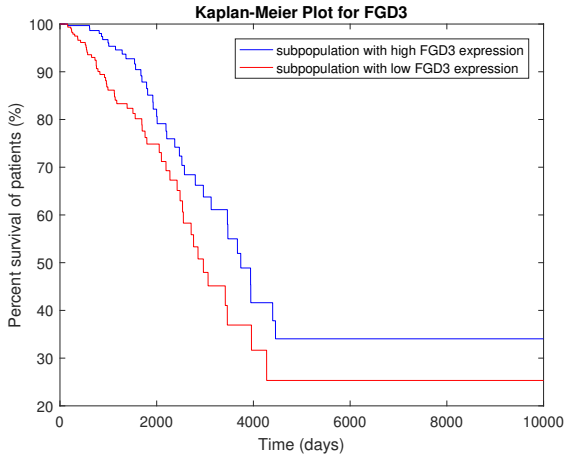

Figure 5. The Kaplan-Meier plot for the second most protective gene – FGD3.

6 Heat Map

In the heat map of LUSC (figure 6), we can observe three main red blocks, with some noise shown in dark red and in black colors. Note that the results are not as prominent as that of the breast cancer (figure 7). Red represents high expression level of genes and green represent low expression level. A red block can be interpreted as a specific set of genes together have high expression level in a certain group of patients. The three red blocks that we saw cannot happen due to chance, which suggests that these three sets of genes (ie. signatures) are correlated with tumor in certain ways.

Research has shown that the signatures in these red squares are associated with mitotic chromosomal instability (CIN), lymphocyte infiltration (LYM) and mesenchymal transition (MES) [2]. The CIN signature contains genes TPX2, KIF4A, KIFC1, NCAPG, BUB1, NCAPH, CDCA5,KIF2C, PLK1, CENPA, and etc. The LYM signature includes genes SASH3, CD53, NCKAP1L, LCP2, IL10RA, PTPRC, EVI2B, BIN2, WAS, HAVCR2, and etc. The MES signature has genes COL3A1, COL5A2, COL1A2, THBS2, COL5A1, VCAN,COL6A3, SPARC, AEBP1, FBN1 and etc.

The set of patients who has is CIN signature is associated with tumor grade and genomic instability [4]. Mitotic chromosomal instability is the inability to segregate equal chromosome complements to two daughter cells during mitosis, which is thought to serve as the fuel for tumorigenic progression

[5]. The patients with a LYM signature is associated with the infiltration of possibly a specific T-lymphocytes in a tumor (details are explained in section 5) [4]. The MES signature is suggested to be associated with tumor invasiveness. It contains mostly kinetochore-associated genes [3]. This is a grade-associated attractor, in which the signature is significantly present only when an intermediate level of tumor grade is reached [3].

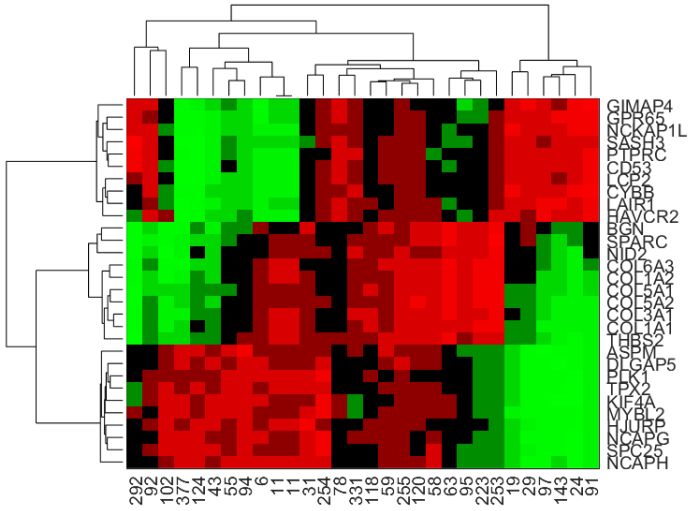

Figure 6. The heat map for LUSC.

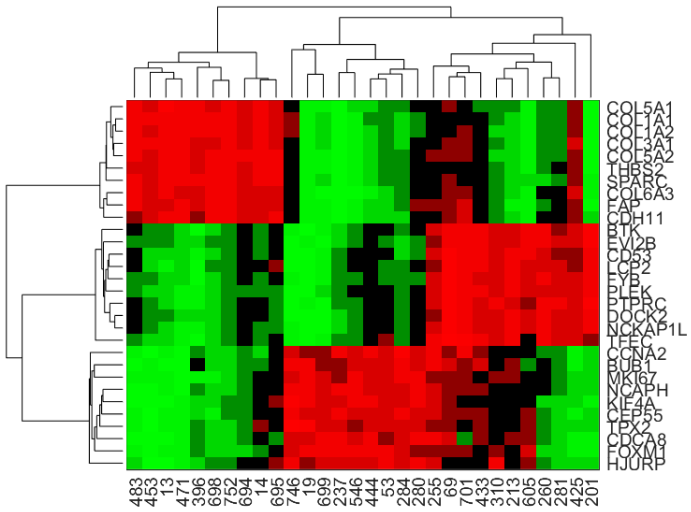

Figure 7. The heat map for BRCA.

References

[1] Steven A. Rosenberg, Nicholas P. Restifo, James C. Yang, Richard A. Morgan, and Mark E. Dudley. Adoptive cell transfer: a clinical path to effective cancer immunotherapy. 8(4):299–308.

[2] Wei-Yi Cheng, Tai-Hsien Ou Yang, Hui Shen, Peter W. Laird, Dimitris Anastassiou, and the Cancer Genome Atlas Research Network. Multi-cancer molecular signatures and their interrelationships.

[3] Wei-Yi Cheng, Tai-Hsien Ou Yang, and Dimitris Anastassiou. Biomolecular events in cancer revealed by attractor metagenes. 9(2):1–14.

[4] Wei-Yi Cheng. Attractor molecular signatures and their applications for prognostic biomarkers.

[5] Juan-Manuel Schvartzman, Rocio Sotillo, and Robert Benezra. Mitotic chromosomal instability and cancer: mouse modelling of the human disease. 10(2):102–115.

[6] Michael H. Malone, Zhengqi Wang, and Clark W. Distelhorst. The glucocorticoid-induced gene tdag8 encodes a pro-apoptotic g protein-coupled receptor whose activation promotes glucocorticoid-induced apoptosis. 279(51):52850–52859.

Appendices

A Data Used

LUSC (Lung Squamous Cell) data in `Pan10.mat`, except for section 6 where `TCGA_PanCan12.mat` is used.

B MATLAB Program

```
%% Load the cancer data
clear; close all;
load('Pan10.mat');
E = log(1+LUSC_ge);      % BRCA_ge
[num_gene,num_patient] = size(LUSC_ge);


%% TPX2
% find the top 10 related genes to TPX2
seed1 = 'TPX2';
choice = findgene(Gene,seed1);
TPX2_top10 = topgenesmi_fcn(E,Gene,choice);

% iteration loop
TPX2_count = 0; % number of iterations it takes to converge
while 1
    TPX2_count = TPX2_count + 1;

    % Create a fictitious "metagene" whose expression level is the average
    % of the expression levels of the ten genes that were found
    TPX2_metagene = zeros(1,num_patient);
    for i=1:10
        TPX2_metagene = TPX2_metagene+E(findgene(Gene,TPX2_top10{i}),:);
    end
    TPX2_metagene = TPX2_metagene/10;

    % Find the list of the top ten genes most associated with that metagene
    F = [E;TPX2_metagene];      % create F by appending the metagene to E
    index_top10 = 2:2+10-1;     % top 10 genes excluding the 1st metagene (ie. 2nd
    to 11th genes)
    choice = num_gene+1;        % choose the data for metagene (ie. last row of F)
    TPX2_top10_new = topgenesmi_fcn(F,Gene,choice,index_top10);

    % Continue iterating until it converges
    if isequal(TPX2_top10,TPX2_top10_new) == 0
        TPX2_top10 = TPX2_top10_new;
    else
        break;
    end
end

% Compute the mutual information of the final ten genes with the final
% metagene. The final ten genes are already ranked in order of m.i.
TPX2_top10_new_mi = zeros(1,10);
for i=1:10
```

```
    TPX2_top10_new_mi(i) = mi(E(findgene(Gene,TPX2_top10{i}),:),TPX2_metagene);
end

% result: final genes are similar but not exactly the same as before,
% significant discovery that points to the core of expression


%% COL1A2
% find the top 10 related genes to COL1A2
seed2 = 'COL1A2';
choice = findgene(Gene,seed2);
COL_top10 = topgenesmi_fcn(E,Gene,choice);

% iteration loop
COL_count = 0; % number of iterations it takes to converge
while 1
    COL_count = COL_count + 1;

    % Create a fictitious "metagene" whose expression level is the average
    % of the expression levels of the ten genes that were found
    COL_metagene = zeros(1,num_patient);
    for i=1:10
        COL_metagene = COL_metagene+E(findgene(Gene,COL_top10{i}),:);
    end
    COL_metagene = COL_metagene/10;

    % Find the list of the top ten genes most associated with that metagene
    F = [E;COL_metagene];        % create F by appending metagene to E
    index_top10 = 2:2+10-1;      % top 10 genes excluding the 1st metagene (ie. 2nd
    to 11th genes)
    choice = num_gene+1;         % choose the data for metagene (ie. last row of F)
    COL_top10_new = topgenesmi_fcn(F,Gene,choice,index_top10);

    % Continue iterating until it converges
    if isequal(COL_top10,COL_top10_new) == 0
        COL_top10 = COL_top10_new;
    else
        break;
    end
end

% Compute the mutual information of the final ten genes with the final
% metagene. The final ten genes are already ranked in order of m.i.
COL_top10_new_mi = zeros(1,10);
for i=1:10
    COL_top10_new_mi(i) = mi(E(findgene(Gene,COL_top10{i}),:),COL_metagene);
end



%% LCP2
% find the top 10 related genes to LCP2
seed3 = 'LCP2';
choice = findgene(Gene,seed3);
LCP2_top10 = topgenesmi_fcn(E,Gene,choice);

```

```
% iteration loop
LCP2_count = 0; % number of iterations it takes to converge
while 1
    LCP2_count = LCP2_count + 1;

    % Create a fictitious "metagene" whose expression level is the average
    % of the expression levels of the ten genes that were found
    LCP2_metagene = zeros(1,num_patient);
    for i=1:10
        LCP2_metagene = LCP2_metagene+E(findgene(Gene,LCP2_top10{i}),:);
    end
    LCP2_metagene = LCP2_metagene/10;

    % Find the list of the top ten genes most associated with that metagene
    F = [E;LCP2_metagene];      % create F by appending the metagene info to E
    index_top10 = 2:2+10-1;     % top 10 genes excluding the 1st metagene (ie. 2nd
    to 11th genes)
    choice = num_gene+1;        % choose the data for metagene (ie. last row of F)
    LCP2_top10_new = topgenesmi_fcn(F,Gene,choice,index_top10);

    % Continue iterating until it converges
    if isequal(LCP2_top10,LCP2_top10_new) == 0
        LCP2_top10 = LCP2_top10_new;
    else
        break;
    end
end

% Compute the mutual information of the final ten genes with the final
% metagene. The final ten genes are already ranked in order of m.i.
LCP2_top10_new_mi = zeros(1,10);
for i=1:10
    LCP2_top10_new_mi(i) = mi(E(findgene(Gene,LCP2_top10{i}),:),LCP2_metagene);
end


%% Scatter plot of 3 genes for the max-strength signature
% LCP2 has the highest strength = 0.8902 (lowest/tenth of m.i.)
LCP2_g1 = LCP2_top10_new_sort(1);LCP2_g1_loc = find(strcmp(Gene,LCP2_g1) == 1);
LCP2_g2 = LCP2_top10_new_sort(2);LCP2_g2_loc = find(strcmp(Gene,LCP2_g2) == 1);
LCP2_g3 = LCP2_top10_new_sort(3);LCP2_g3_loc = find(strcmp(Gene,LCP2_g3) == 1);

sc3(LCP2_g1_loc, LCP2_g2_loc, LCP2_g3_loc, E);
xlabel(LCP2_g1);ylabel(LCP2_g2);


%% Scatter plot for other signatures
% COL1A2
COL_g1 = COL_top10_new_sort(1);COL_g1_loc = find(strcmp(Gene,COL_g1) == 1);
COL_g2 = COL_top10_new_sort(2);COL_g2_loc = find(strcmp(Gene,COL_g2) == 1);
COL_g3 = COL_top10_new_sort(3);COL_g3_loc = find(strcmp(Gene,COL_g3) == 1);
sc3(COL_g1_loc, COL_g2_loc, COL_g3_loc, E);
xlabel(COL_g1);ylabel(COL_g2);

% TPX2
TPX2_g1 = TPX2_top10_new_sort(1);TPX2_g1_loc = find(strcmp(Gene,TPX2_g1) == 1);
```

```
TPX2_g2 = TPX2_top10_new_sort(2);TPX2_g2_loc = find(strcmp(Gene,TPX2_g2) == 1);
TPX2_g3 = TPX2_top10_new_sort(3);TPX2_g3_loc = find(strcmp(Gene,TPX2_g3) == 1);
sc3(TPX2_g1_loc, TPX2_g2_loc, TPX2_g3_loc, E);
xlabel(TPX2_g1);ylabel(TPX2_g2);


%% Survival analysis of breast cancer data
load('TCGA_PanCan12.mat');
E = log(1+BRCA_ge);
Time = BRCA_time;
Status = BRCA_status;
[num_gene,num_patient] = size(BRCA_ge);

% Compute concordance index (CI) for all genes
CI = zeros(1,num_gene);
for i=1:num_gene
    CI(i) = concordanceindex(E(i,:),Time,Status);
end

% Rank all genes in terms of their concordance index (CI)
[CI_sort,I] = sort(CI);
gene_sort = Gene(I);
top5_protective_genes = gene_sort(1:5);


%% KM curve for LOC100128977
most_protective_loc = findgene(Gene,'LOC100128977');
xx = find(E(most_protective_loc,:) > median(E(most_protective_loc,:)));
yy = find(E(most_protective_loc,:) <= median(E(most_protective_loc,:)));

figure;
kaplan_meier(Time(xx),Status(xx),'b');  % high LOC100128977 expression
hold on;
kaplan_meier(Time(yy),Status(yy),'r');  % low LOC100128977 expression
title('Kaplan-Meier Plot for LOC100128977');
xlabel('Time (days)');
ylabel('Percent survival of patients (%)');
legend('subpopulation with high LOC100128977 expression','subpopulation with low
    LOC100128977 expression');


%% Heat map
E = log(1 + LUSC_ge);

% Initiate signature genes' location
sig1 = zeros(1,10);
sig2 = sig1;
sig3 = sig1;

% Location of the genes of each signature
for i = 1:10
    sig1(i) = findgene(Gene,TPX2_top10_new_sort{i});
    sig2(i) = findgene(Gene,COL_top10_new_sort{i});
    sig3(i) = findgene(Gene,LCP2_top10_new_sort{i});
end

```

```
% function "mean" outputs the mean values of each column
[~,I1] = sort(mean(E(sig1,:)),'descend');
[~,I2] = sort(mean(E(sig2,:)),'descend');
[~,I3] = sort(mean(E(sig3,:)),'descend');

selected_genes=[sig1 sig2 sig3];
selected_samples=[I1(1:10) I2(1:10) I3(1:10)];
G = E(selected_genes,selected_samples);
clustergram(G,'Standardize','row','RowLabels',Gene(selected_genes),'ColumnLabels',
    selected_samples);
```

C Gene Descriptions

NCKAP1L
The encoded protein is a part of the Scar/WAVE complex which plays an important role in regulating cell shape in both metazoans and plants.

PTPRC
The protein encoded by this gene is a member of the protein tyrosine phosphatase (PTP) family. PTPs are known to be signaling molecules that regulate a variety of cellular processes including cell growth, differentiation, mitosis, and oncogenic transformation. This PTP contains an extracellular domain, a single transmembrane segment and two tandem intracytoplasmic catalytic domains, and thus is classified as a receptor type PTP. This PTP has been shown to be an essential regulator of T- and B-cell antigen receptor signaling. It functions through either direct interaction with components of the antigen receptor complexes, or by activating various Src family kinases required for the antigen receptor signaling. This PTP also suppresses JAK kinases, and thus functions as a regulator of cytokine receptor signaling.

LAIR1
The protein encoded by this gene is an inhibitory receptor found on peripheral mononuclear cells, including natural killer cells, T cells, and B cells. Inhibitory receptors regulate the immune response to prevent lysis of cells recognized as self. The gene is a member of both the immunoglobulin superfamily and the leukocyte-associated inhibitory receptor family. The gene maps to a region of 19q13.4 called the leukocyte receptor cluster, which contains at least 29 genes encoding leukocyte-expressed receptors of the immunoglobulin superfamily. The encoded protein has been identified as an anchor for tyrosine phosphatase SHP-1, and may induce cell death in myeloid leukemias.

SASH3
The protein encoded by this gene contains a Src homology-3 (SH3) domain and a sterile alpha motif (SAM), both of which are found in proteins involved in cell signaling. This protein may function as a signaling adapter protein in lymphocytes.

LCP2
This gene encodes an adapter protein that acts as a substrate of the T cell antigen receptor (TCR)-activated protein tyrosine kinase pathway. The encoded protein associates with growth factor receptor bound protein 2, and is thought to play a role TCR-mediated intracellular signal transduction. A similar protein in mouse plays a role in normal T-cell development and activation. Mice lacking this gene show subcutaneous and intraperitoneal fetal hemorrhaging, dysfunctional platelets and impaired viability.

HAVCR2
The protein encoded by this gene belongs to the immunoglobulin superfamily, and TIM family of proteins. CD4-positive T helper lymphocytes can be divided into types 1 (Th1) and 2 (Th2) on the basis of their cytokine secretion patterns. Th1 cells are involved in cell-mediated immunity to intracellular pathogens and delayed-type hypersensitivity reactions, whereas, Th2 cells are involved in the control of extracellular helminthic infections and the promotion of atopic and allergic diseases. This protein is a Th1-specific cell surface protein that regulates macrophage activation, and inhibits Th1-mediated auto- and alloimmune responses, and promotes immunological tolerance.

GPR65
Using gene expression profiles of lymphoma cell lines and primary thymocytes treated with the synthetic glucocorticoid dexamethasone, we discovered that induction of TDAG8 (T-cell death-associated gene 8) was a common event in each model system investigated [6].

CD53
The protein encoded by this gene is a member of the transmembrane 4 superfamily, also known as the tetraspanin family. Most of these members are cell-surface proteins that are characterized by the presence of four hydrophobic domains. The proteins mediate signal transduction events that play a role in the regulation of cell development, activation, growth and motility. This encoded protein is a cell surface glycoprotein that is known to complex with integrins. It contributes to the transduction of CD2-generated signals in T cells and natural killer cells and has been suggested to play a role in growth regulation. Familial deficiency of this gene has been linked to an immunodeficiency associated with recurrent infectious diseases caused by bacteria, fungi and viruses.

CYBB
Cytochrome b (-245) is composed of cytochrome b alpha (CYBA) and beta (CYBB) chain. It has been proposed as a primary component of the microbicidal oxidase system of phagocytes. CYBB deficiency is one of five described biochemical defects associated with chronic granulomatous disease (CGD). In this disorder, there is decreased activity of phagocyte NADPH oxidase; neutrophils are able to phagocytize bacteria but cannot kill them in the phagocytic vacuoles. The cause of the killing defect is an inability to increase the cell's respiration and consequent failure to deliver activated oxygen into the phagocytic vacuole.

GIMAP4
This gene encodes a protein belonging to the GTP-binding superfamily and to the immuno-associated nucleotide (IAN) subfamily of nucleotide-binding proteins. The encoded protein of this gene may be negatively regulated by T-cell acute lymphocytic leukemia 1 (TAL1). In humans, the IAN subfamily genes are located in a cluster at 7q36.1. [provided by RefSeq, Jul 2008]