Table of Contents

Mircoarray Analysis: Effects on Clioquional on Yeast

Template by [Fernando Ramirez Thinh Nguyen] Code by [Fernando Ramirez] adapted from fetch.m adapted from https://www.mathworks.com/help/bioinfo/ug/working-with-geo-series-data.html background -- Clioquional - family of durge hydroxyquinolines, inhibit particular enzymes related to DNA replication. Drugs dound to have activity against both virla and protozoal infections

```
% 00.) Analayze the microarray dataset made available by the following
 study
% https://www.ncbi.nlm.nih.gov/pubmed/21504115
% clioquinol.yeast.Li2010.pdf
```

01.) download the follwing GSE1757_series_matrix.txt file from https://ftp.ncbi.nlm.nih.gov/geo/series/GSE17nnn/GSE17257/matrix/ index. This part is done manually. Extract the the compress txt.gz file using winRAR, move file to the working file directory.

```
%EDA - early data analysis (exploration stage)

gseData=bmes_downloadandparsegse_thinh_fernando('GSE17257');
get(gseData.Data); %understanding the size of the Rownames, ColNames
d = gseData.Data; % Exploring GSE data, row names and column names

Downloading https://ftp.ncbi.nlm.nih.gov/geo/series/GSE17nnn/GSE17257/
matrix/GSE17257_series_matrix.txt.gz ...
Reading C:\Users\Fernando A. Ramirez\AppData\Local\Temp
\GSE17257.txt ...
            Name: ''
        RowNames: {10928×1 cell}
        ColNames: {1×6 cell}
           NRows: 10928
           NCols: 6
           NDims: 2
    ElementClass: 'double'


gpl_platform = gseData.Header.Series.platform_id; %saving pointer to
 gpl_platform
gpl =
 bmes_downloadandparsegpl_thinh_fernando(gpl_platform); %obtaining
 metadata for the gsl_platform
gpl.ColumnNames; %outputing to function, ensuring the cell array and
 metadata is read
```

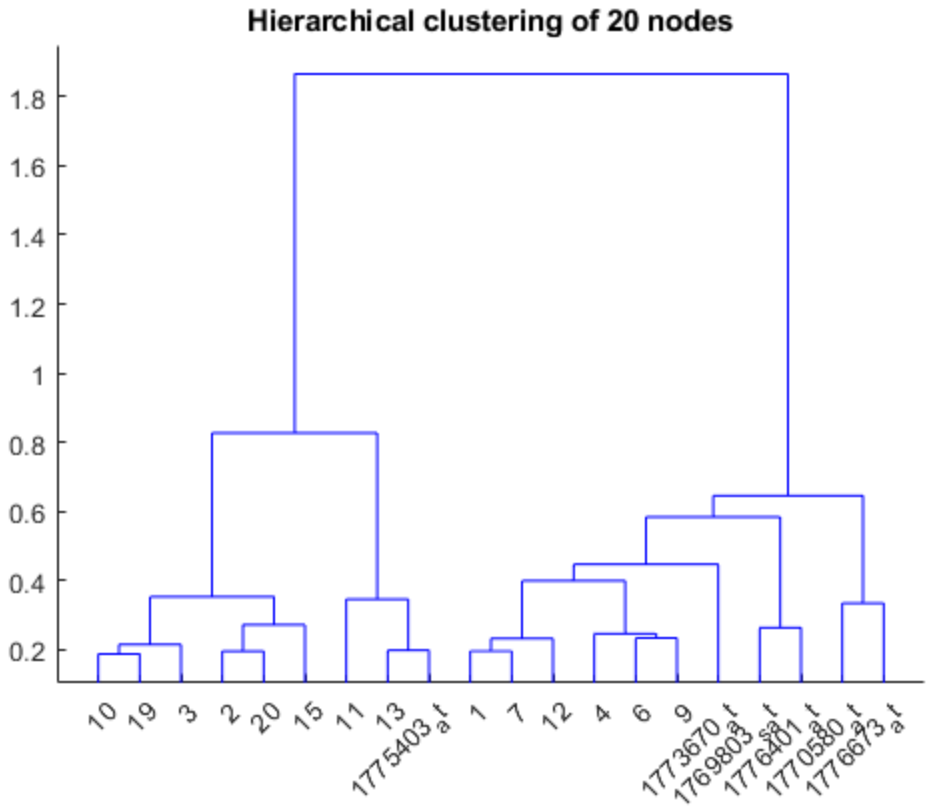
```
%Exploring the probsets to gene symbols, from the gplData, string
 comparison to the ID and Gene Symbol
gplProbesetIDs = gpl.Data(:, strcmp(gpl.ColumnNames, 'ID'));
geneSymbols = gpl.Data(:, strcmp(gpl.ColumnNames, 'Gene Symbol'));
gseprobes = d.rownames;
%row_change_geneSymbol = rownames(gse.Data.Data, ':', geneSymbols);
%the above code can be optimizing computationally by using a regex and
%saved to a variable, then variable is called.

%mapping the GSE to GPL values, intializing a zero non-vale matrix
MAP_GSE_GPL = zeros(numel(gseprobes),1);

%mapping of the geneSymbols same {} double as the gplProbesetIDs
%For each gseprobe, we need to search gplprobes and use the
 corresponding
%gene. Doing string comparison for each of them will be too slow.
 Let's
%use a Map container to speed this up.

map = containers.Map(gplProbesetIDs,1:numel(gplProbesetIDs));
for i = 1:numel(gseprobes)
    if map.isKey(gseprobes{i}); MAP_GSE_GPL(i)= map(gseprobes{i});
    end
end

gsegenes = gseprobes; %make a copy, so entries not found will keep the
 probe name.
%genenames = gseprobes;
gsegenes(find(MAP_GSE_GPL)) =
 gplProbesetIDs(MAP_GSE_GPL(find(MAP_GSE_GPL)));
%genenames(find(MAP_GSE_GPL)) =
 geneSymbols(MAP_GSE_GPL(find(MAP_GSE_GPL)));
d = d.rownames(':',gsegenes);
%datamatrix = d.rownames(':',genenames);
```

Data Analysis: Determine sample groups we'll work with

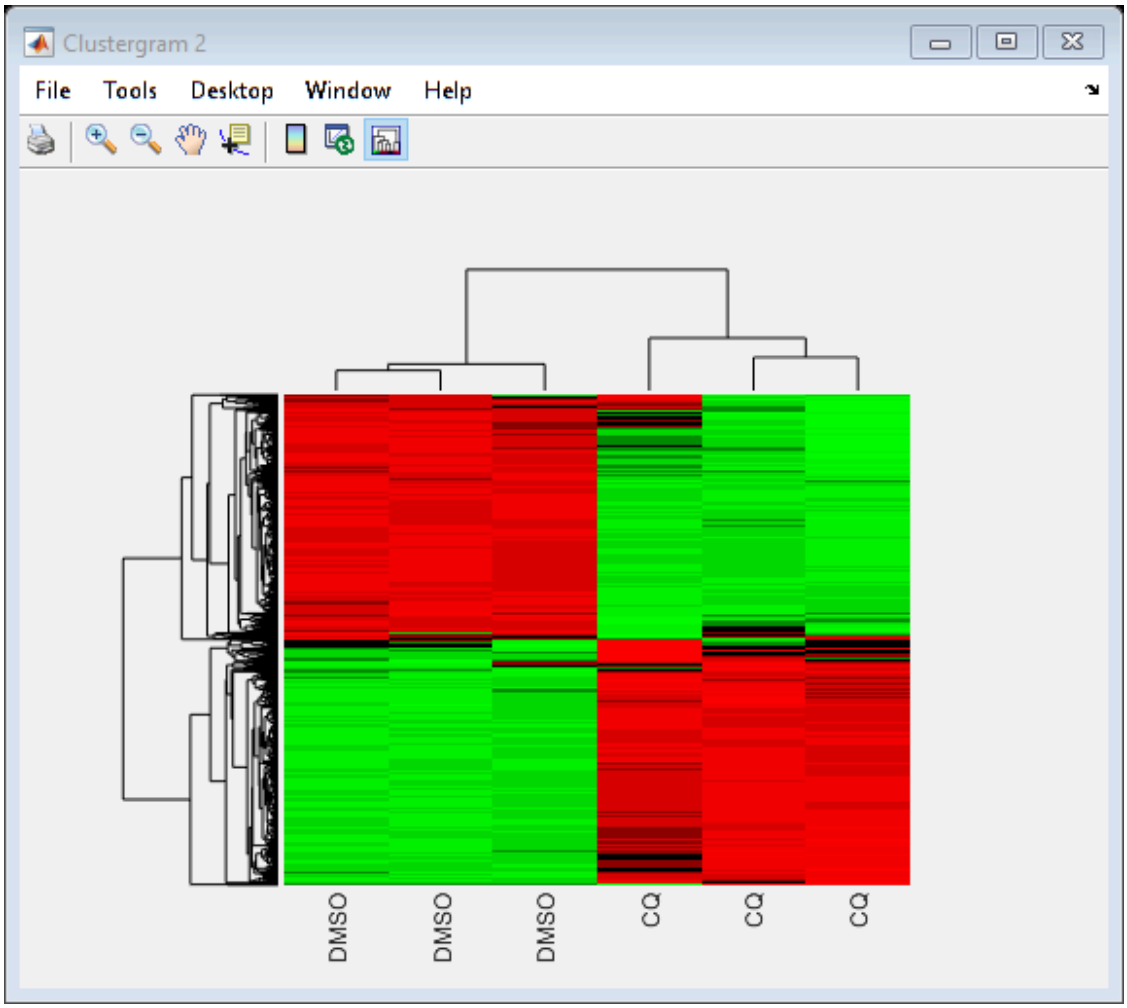
0.3) We are often interested in comparing groups of samples. We need to look at the header information and decide which information for samples we can use to group, Usually the Header.Samples structure usually contains what we need.

```
%taking a dive into the header information
samplegroups  = gseData.Header.Samples.characteristics_ch1(2,:);
unique_samplegroups = unique(samplegroups)';
%in total there are 6 samplegroups, however two uniques ones as
 follows
% 1% DMSO (dimethly sulfoxide)
% 80uM CQ (tumor development in chemotherapeutic agens, anticancer
 drug Chloroquine)
```

```
% create logical vectirs for sample groups of interest
Idmso = strcmpi(samplegroups,'media supplement: 1% DMSO');
Icq = strcmpi(samplegroups,'media supplement: 80 µM CQ');
%from a column indexing with logical vector

% create a numerical vector to assign each sample to a group 1-2. As
 `I`
% indexed groups.
Igroups=zeros(1,numel(samplegroups)); %initialize samplegroup size of
 6 as groups
Igroups(Idmso) = 1;
Igroups(Icq) = 2;
groupnames={'DMSO' 'CQ'};

colnames=d.colnames;
for i=1:2; colnames(Igroups==i) = groupnames(i); end
d=d.colnames(':',colnames); %this really means: "d.colnames=colnames;"

%04.) Show a hiearchical clustering of samples. (Just a hiearchical
%clustering (ie a dendrogram) of samples, not a heatmap of expression
%values.

% One idea is to only keep the genes that vary most across samples
% (ingoring sample groups.) This can be done using:
% lets try to to push this to computing to 90 variance level.
I =genevarfilter(d,'Percentile',90);
d2 = d( I, :);

% Let's create a distance matrix between pairs of genes.
% pdist() gives a vector (to save space). If you want the symmetric
 matrix,
% just pass the result through squareform().
% argument changed to the spearman
% spearman is produced as a vector,

genedist = pdist(d2,'corr');
%'spearman'

%linkage group under each group. these results contain information
 about
%which two groups are combine at each branch.
%Agglomerative hierarcvhical cluster tree. Using the average method.
%average = UPGMA -- Unweighted average distance.
tree = linkage(genedist,'average');

% visualize the tree, show only 20 nodes. Want to clear figure
% Groups of genes will have a numerical id for labels.
 %bmes_fig geneclust; clf
dendrogram(tree,20,'Labels',d2.rownames);
h=gca;
h.XTickLabelRotation=45;
title('Hierarchical clustering of 20 nodes')
%distance = pdist(genedist);
%leafOrder = optimalleaforder(tree,distance)
```




```
%05.) Show a clustergram (heatmap,combined with clustering of samples
 and
%clustering of genes of expression values

cg = clustergram(d2,'Standardize','Row');
```




```
%3.) Report the top 10 most different genes between the Clioquinol and
%control groups.
[dpvals] = mattest(d(:,Idmso), d(:, Icq),'permute',1000);
[dpvals2] = mattest(datamatrix(:,Idmso), d(:, Icq),'permute',1000);
%performing two-sample t-test to evaluate differential expression of
 genes
%from two experimental conditions or phenotypes, in this case it is
 DSMO
%and CQ medium conditions
signif_dpvals = dpvals(dpvals(:,1) <= 0.01,:);
signif_d = d(dpvals(:,1) <= 0.01,:);
d_sig = d(dpvals(:,1) <= 0.01,:);
%taking fold change
log2fc = log2(mean(d_sig(:,Idmso),2) ./ mean(d_sig(:,Icq),2));
%taking log2 scale to `compress scaling of values
%scatter(log2fc, -log10(signif_dpvals(:,1)), '.');
%xlabel('log_2(dsmo:CQ) media suppplements in yeast'), ylabel('-
log_{10}(pvalue)');

negfc = 2.^log2fc;
negfc(negfc<1) = - 1./negfc(negfc<1);
```

```
%in order to compute the 10 most different genes between the
 Clioquinol and
%control groups signif_d is needed
% Add the foldchange information to the dpvals object:
signif_dpvals=[signif_dpvals bioma.data.DataMatrix(negfc,'ColNames',
{'negfc'})];
% Select the genes with pvalue<=0.01 and FC>=1.5.

I = signif_dpvals(:,'p-values')<=0.01 &
 abs(signif_dpvals(:,'negfc'))>=1.5;
%I = abs(signif_dpvals(:,'negfc'))>=1.5;
%logical vector return
dsigfc = signif_dpvals(I,:);
dsigfc = dsigfc.sortrows('p-values');
fprintf('Found %d genes with pvalue<=0.01 and FC>=1.5. Showing top 10:
\n',size(dsigfc,1));
disp(dsigfc(1:10,:))
```

Found 906 genes with pvalue<=0.01 and FC>=1.5. Showing top 10:

	p-values	negfc
1777371_at	1.2783e-09	-4.5182
1777972_at	9.4241e-06	-6.8586
1777623_at	1.8843e-05	-3.5109
1771341_at	2.8282e-05	-22.45
1776525_at	3.7698e-05	2.9046
1771069_at	4.7146e-05	4.1521
1777661_at	5.654e-05	-2.6123
1771389_at	6.598e-05	-2.7784
1773288_at	7.5387e-05	3.451
1780031_at	8.7097e-05	-3.5482

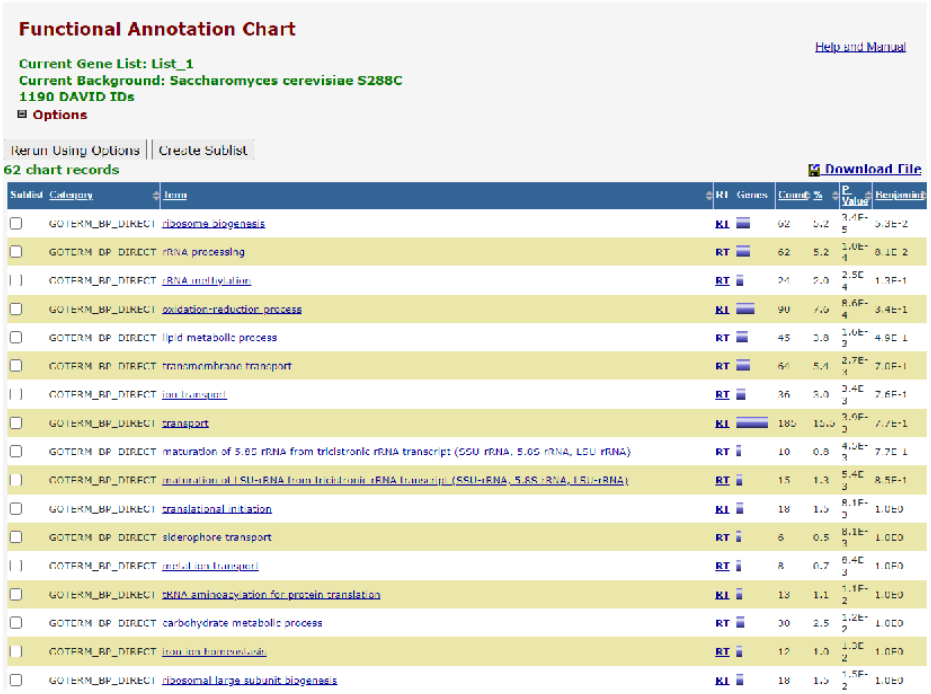
WRITING TO EXCEL to be used with DAVID

```
I=find(signif_dpvals(:,1)<=0.01);
nsig=numel(I);
xlsdata = cell(nsig, 3); %each row will contain
 genesymbol,pvalue,negfc
for i=1:nsig
 gene=signif_dpvals.rownames{I(i)};
 p=signif_dpvals.double(I(i), 1);
 nfc=signif_dpvals.double(I(i), 2);
 xlsdata(i,:) = {gene p nfc};
end

xlsdata=[ {'genesymbol' 'pvalue' 'negfc'}; xlsdata]; %add the header
 row.
xlswrite('cq.xlsx',xlsdata,'siggenesDMSO_cq');
```

Warning: Added specified worksheet.

```
%4.)Report the functional annotations (Go Biological Processses and
 KEGG
```

```
%Pathways) that are significantly different between the two groups.

% We have found the significantly different genes between two groups.
 But
% what do these genes do? Are there significant differences in
 biological
% functions between two groups? To answer these questions, we'll make
 use
% of the Gene Ontology terms, which annotate each gene to one or more
% Biological Processes, Cellular Components, and Molecular Functions.

%imgo processes
figure(1)
imshow(imread('gobp1.png'))
figure(2)
imshow(imread('gobp2.png'))
figure(3)
imshow(imread('gobp3.png'))
%KEGG pathways
figure(4)
imshow(imread('KEGG.png'))
```



Functional Annotation Chart

Help and Manual

Current Gene List: List_1
Current Background: Saccharomyces cerevisiae S288C
1190 DAVID IDs
Options

Rerun Using Options | Create Sublist

62 chart records

Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	GOTERM_BP_DIRECT	ribosome biogenesis	KI		62	5.2	3.4E-5	5.3E-2
	GOTERM_BP_DIRECT	rRNA processing	RT		62	5.2	1.0E-4	8.1E-2
	GOTERM_BP_DIRECT	rRNA methylation	RT		24	2.0	2.5E-4	1.3E-1
	GOTERM_BP_DIRECT	oxidation-reduction process	KI		90	7.6	8.6E-4	3.4E-1
	GOTERM_BP_DIRECT	lipid metabolic process	RT		45	3.8	1.6E-3	4.9E-1
	GOTERM_BP_DIRECT	transmembrane transport	RT		64	5.4	2.7E-3	7.0E-1
	GOTERM_BP_DIRECT	ion transport	RT		36	3.0	3.4E-3	7.6E-1
	GOTERM_BP_DIRECT	transport	KI		180	15.0	3.6E-3	7.7E-1
	GOTERM_BP_DIRECT	maturation of 5.8S rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)	RT		10	0.8	4.5E-3	7.7E-1
	GOTERM_BP_DIRECT	maturation of LSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)	RT		15	1.3	5.4E-3	8.5E-1
	GOTERM_BP_DIRECT	translational initiation	KI		18	1.5	8.1E-3	1.0E0
	GOTERM_BP_DIRECT	siderophore transport	RT		6	0.5	8.1E-3	1.0E0
	GOTERM_BP_DIRECT	metal ion transport	RT		8	0.7	8.4E-3	1.0E0
	GOTERM_BP_DIRECT	tRNA aminoacylation for protein translation	KI		13	1.1	1.1E-2	1.0E0
	GOTERM_BP_DIRECT	carbohydrate metabolic process	RT		30	2.5	1.2E-2	1.0E0
	GOTERM_BP_DIRECT	iron ion homeostasis	RT		12	1.0	1.3E-2	1.0E0
	GOTERM_BP_DIRECT	ribosomal large subunit biogenesis	KI		18	1.5	1.5E-2	1.0E0

☐	GOTERM_BP_DIRECT	phospholipid translocation	RT	6	0.5	5.9E-2	1.0E0
☐	GOTERM_BP_DIRECT	activation of GTPase activity	RT	6	0.5	5.9E-2	1.0E0
☐	GOTERM_BP_DIRECT	cell division	RT	47	3.9	6.6E-2	1.0E0
☐	GOTERM_BP_DIRECT	'de novo' pyrimidine nucleobase biosynthetic process	RT	5	0.4	7.2E-2	1.0E0
☐	GOTERM_BP_DIRECT	DNA-templated transcription, termination	RT	6	0.5	8.1E-2	1.0E0
☐	GOTERM_BP_DIRECT	fungal-type cell wall chitin biosynthetic process	RT	6	0.5	8.1E-2	1.0E0
☐	GOTERM_BP_DIRECT	mitochondrial electron transport, ubiquinol to cytochrome c	RT	6	0.5	8.1E-2	1.0E0
☐	GOTERM_BP_DIRECT	intracellular protein transport	RT	25	2.1	8.3E-2	1.0E0
☐	GOTERM_BP_DIRECT	sphingolipid biosynthetic process	RT	7	0.6	8.5E-2	1.0E0
☐	GOTERM_BP_DIRECT	transcription of nuclear large rRNA transcript from RNA polymerase I promoter	RT	7	0.6	8.5E-2	1.0E0
☐	GOTERM_BP_DIRECT	phospholipid transport	RT	8	0.7	8.5E-2	1.0E0
☐	GOTERM_BP_DIRECT	mitotic DNA replication checkpoint	RT	4	0.3	8.6E-2	1.0E0
☐	GOTERM_BP_DIRECT	heterochromatin organization involved in chromatin silencing	RT	4	0.3	8.6E-2	1.0E0
☐	GOTERM_BP_DIRECT	transcription factor import into nucleus	RT	4	0.3	8.6E-2	1.0E0
☐	GOTERM_BP_DIRECT	'de novo' UMP biosynthetic process	RT	4	0.3	8.6E-2	1.0E0
☐	GOTERM_BP_DIRECT	cellular amino acid biosynthetic process	RT	25	2.1	9.2E-2	1.0E0
☐	GOTERM_BP_DIRECT	phenylalanine transport	RT	3	0.3	9.4E-2	1.0E0
☐	GOTERM_BP_DIRECT	cadmium ion transport	RT	3	0.3	9.4E-2	1.0E0
☐	GOTERM_BP_DIRECT	endonucleolytic cleavage in ITS1 to separate SSU-rRNA from 5.8S rRNA and LSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)	RT	13	1.1	9.8E-2	1.0E0

☐	GOTERM_BP_DIRECT	regulation of translational initiation	RT	9	0.8	1.7E-2	1.0E0
☐	GOTERM_BP_DIRECT	cellular iron ion homeostasis	RT	15	1.3	1.9E-2	1.0E0
☐	GOTERM_BP_DIRECT	fatty acid biosynthetic process	RT	10	0.8	1.9E-2	1.0E0
☐	GOTERM_BP_DIRECT	tricarboxylic acid cycle	RT	12	1.0	2.2E-2	1.0E0
☐	GOTERM_BP_DIRECT	drug export	RT	4	0.3	2.3E-2	1.0E0
☐	GOTERM_BP_DIRECT	formation of cytoplasmic translation initiation complex	RT	4	0.3	2.3E-2	1.0E0
☐	GOTERM_BP_DIRECT	fructose 2,6 bisphosphate metabolic process	RT	4	0.3	2.3E-2	1.0E0
☐	GOTERM_BP_DIRECT	rRNA 2'-O-methylation	RT	4	0.3	2.3E-2	1.0E0
☐	GOTERM_BP_DIRECT	xenobiotic transport	RT	6	0.5	2.6E-2	1.0E0
☐	GOTERM_BP_DIRECT	fructose metabolic process	RT	5	0.4	2.7E-2	1.0E0
☐	GOTERM_BP_DIRECT	hydrogen ion transmembrane transport	RT	14	1.2	2.8E-2	1.0E0
☐	GOTERM_BP_DIRECT	transcription from RNA polymerase I promoter	RT	9	0.8	3.2E-2	1.0E0
☐	GOTERM_BP_DIRECT	cellular response to drug	RT	7	0.6	3.4E-2	1.0E0
☐	GOTERM_BP_DIRECT	translational termination	RT	7	0.6	3.4E-2	1.0E0
☐	GOTERM_BP_DIRECT	carbohydrate phosphorylation	RT	7	0.6	3.4E-2	1.0E0
☐	GOTERM_BP_DIRECT	maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)	RT	21	1.8	3.7E-2	1.0E0
☐	GOTERM_BP_DIRECT	sulfur compound metabolism	RT	6	0.5	4.0E-2	1.0E0
☐	GOTERM_BP_DIRECT	error-prone translesion synthesis	RT	6	0.5	4.0E-2	1.0E0
☐	GOTERM_BP_DIRECT	cell wall organization	RT	26	2.2	4.1E-2	1.0E0
☐	GOTERM_BP_DIRECT	rRNA modification	RT	5	0.4	4.7E-2	1.0E0
☐	GOTERM_BP_DIRECT	cellular manganese ion homeostasis	RT	5	0.4	4.7E-2	1.0E0
☐	GOTERM_BP_DIRECT	protein import into nucleus	RT	14	1.2	5.0E-2	1.0E0
☐	GOTERM_BP_DIRECT	regulation of Rho protein signal transduction	RT	4	0.3	5.0E-2	1.0E0
☐	GOTERM_BP_DIRECT	very long chain fatty acid biosynthetic process	RT	4	0.3	5.0E-2	1.0E0
☐	GOTERM_BP_DIRECT	positive regulation of GTPase activity	RT	26	2.2	5.2E-2	1.0E0
☐	GOTERM_BP_DIRECT	formation of translation preinitiation complex	RT	8	0.7	5.2E-2	1.0E0

Functional Annotation Chart

Help and Manual

Current Gene List: List_1
Current Background: Saccharomyces cerevisiae S288C
1190 DAVID IDs
Options

Rerun Using Options | Create Sublist

24 chart records

Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	KEGG_PATHWAY	Metabolic pathways	RT		171	14.4	1.4E-7	1.4E-5
	KEGG_PATHWAY	Biosynthesis of secondary metabolites	RT		77	6.5	4.6E-4	2.3E-2
	KEGG_PATHWAY	Biosynthesis of antibiotics	RT		56	4.7	3.7E-3	1.2E-1
	KEGG_PATHWAY	Biosynthesis of amino acids	RT		35	2.9	6.3E-3	1.2E-1
	KEGG_PATHWAY	Fatty acid elongation	RT		6	0.5	7.1E-3	1.2E-1
	KEGG_PATHWAY	Biosynthesis of unsaturated fatty acids	RT		7	0.6	7.5E-3	1.2E-1
	KEGG_PATHWAY	Amino sugar and nucleotide sugar metabolism	RT		13	1.1	8.3E-3	1.2E-1
	KEGG_PATHWAY	Pyrimidine metabolism	RT		22	1.8	1.1E-2	1.5E-1
	KEGG_PATHWAY	Pentose phosphate pathway	RT		11	0.9	2.2E-2	2.5E-1
	KEGG_PATHWAY	Cell cycle - yeast	RT		32	2.7	3.0E-2	3.0E-1
	KEGG_PATHWAY	Phenylalanine metabolism	RT		6	0.5	3.6E-2	3.3E-1
	KEGG_PATHWAY	2-Oxocarboxylic acid metabolism	RT		12	1.0	4.3E-2	3.4E-1
	KEGG_PATHWAY	RNA transport	RT		25	2.1	4.4E-2	3.4E-1
	KEGG_PATHWAY	Starch and sucrose metabolism	RT		13	1.1	5.0E-2	3.4E-1
	KEGG_PATHWAY	Valine, leucine and isoleucine biosynthesis	RT		6	0.5	5.3E-2	3.4E-1
	KEGG_PATHWAY	Citrate cycle (TCA cycle)	RT		11	0.9	5.5E-2	3.4E-1
	KEGG_PATHWAY	Carbon metabolism	RT		29	2.4	5.6E-2	3.4E-1
	KEGG_PATHWAY	Glycerophospholipid metabolism	RT		12	1.0	6.3E-2	3.6E-1
	KEGG_PATHWAY	MAPK signaling pathway - yeast	RT		15	1.3	7.1E-2	3.8E-1
	KEGG_PATHWAY	Ribosome biogenesis in eukaryotes	RT		23	1.9	8.0E-2	4.1E-1
	KEGG_PATHWAY	Alanine, aspartate and glutamate metabolism	RT		10	0.8	8.4E-2	4.1E-1
	KEGG_PATHWAY	Fatty acid metabolism	RT		8	0.7	9.3E-2	4.2E-1
	KEGG_PATHWAY	Oxidative phosphorylation	RT		19	1.6	9.8E-2	4.2E-1
	KEGG_PATHWAY	Purine metabolism	RT		24	2.0	9.9E-2	4.2E-1

```
%5.) Discuss whether your results align with the finding reported in
 the
% paper
%the results in the paper discuss gene of which fold changes were
%calculated using hte ratio of signals in C1-treated samples and DMSO
%treated controls, as shown in the analysis above. The genes that were
%found to have identical fold changes as stated in the paper and in
 the
%analysis above were FRE3, FET3, ENB1, ZPS1,ZRT1,ZRT3,PCA1, and SMF1.
 As
%discuss in the paper the genes were upregulated fold growth as
 analyed in
%our excel as abs(negfc). It would be nice to mapp these genes with
 the
%Affymetrix probeIds, as this would eliminate the need to reference
 the
%geneSymbols in the code above to ensure that each ProbeID is
 referecing
%the correct Gene in the paper. This improvement could be added in a
%revised iteration. Likewise a p-value threshold in the paper of
 p<0.01 was
%considered, likewise in this analysis. For the purposes of utilizing
 David
%bioinformatics database, probeset IDs were used initialize to map and
%report the significant functional annotations of Go Biological
 Processes
%and KEGG pathways,respectively.
```

Published with MATLAB® R2020b