Midterm 2 - Gene Expression Data Analysis

Retrieve the NCBI GEO dataset: GSE96651

Answer the following questions:

1.(10 pts) **Use Geo2R boxplot to visualize the value distribution in each sample. Paste a screenshot below. What can you conclude from these value distributions of the samples – do the samples require global normalization? Explain your answer.**

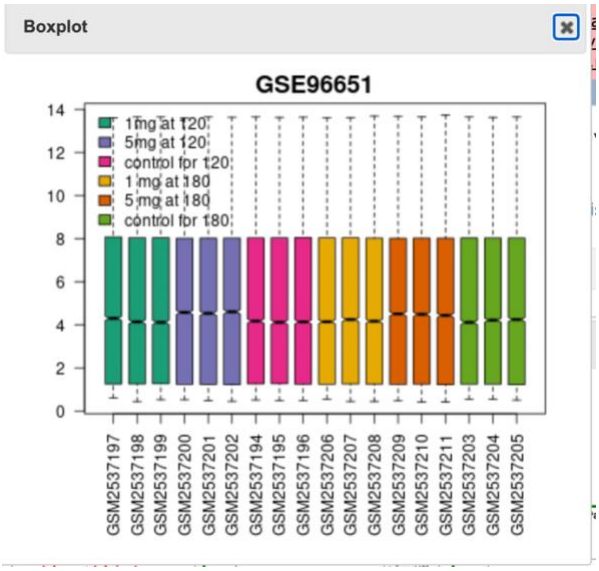


Figure 1

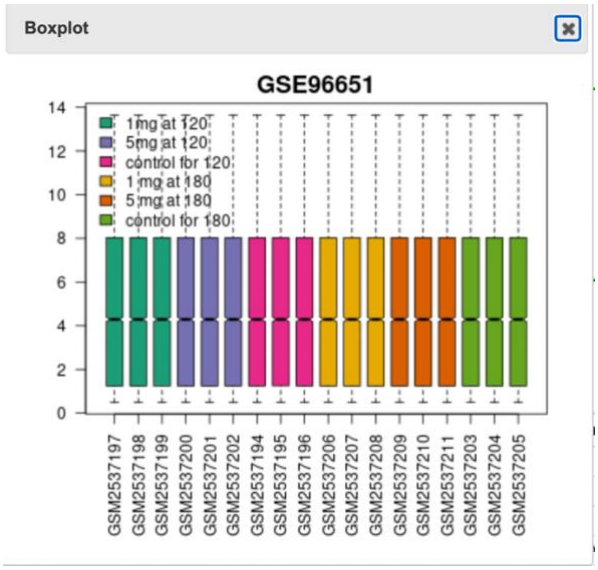


Figure 2

- While analyzing box plots, we can conclude that the median centered values are those where the data are normalized and can be cross – compared
- I have obtained the boxplot for the sample GSE96651 by dividing each of the two conditions (120 min and 180 min) with their unique conditions (1mg/L and 5mg/L of nonylphenol (NP) where each condition have three replicates into different categories of samples in Geo2R as indicated by the color-coded key. I further selected the analyze button to visualize the box plots.
- We can conclude that a sample is normalized if there are median centered values of gene expression indicating that the samples need not require global normalization
- Normalization is the process of correcting two or more datasets before comparing their gene expression values. Normalization is required in samples due to differences in the amount of input RNA, DNA quality, washing efficiency or signal detection.
- The x-axis on the plot indicates the samples and the y-axis indicates normalization values
- In this case, I would say the figure 1 suggests that the samples could offer scope for normalization just to ensure that all the data obtained will have identical normalization
- Considering this reasoning I performed a second iteration of analyze using the 'force normalize' option and the results are indicated on the figure 2.
- After force normalization, we can see an even more pronounced median centered value as indicated by figure 2.

2. In Geo2R, perform a comparison between the control samples and NP treated samples at 120 min and a second comparison at 180 min, for both 1 ug and 5 ug NP.

a) (10 pts) **Identify differentially expressed genes between control and treated samples at 120 minutes and 180 minutes, for 1 and 5 ug NP. Show volcano plots of these comparisons, and a table of the numbers of genes that are differentially expressed (up-regulated and down-regulated) in these two conditions at the two times points.**

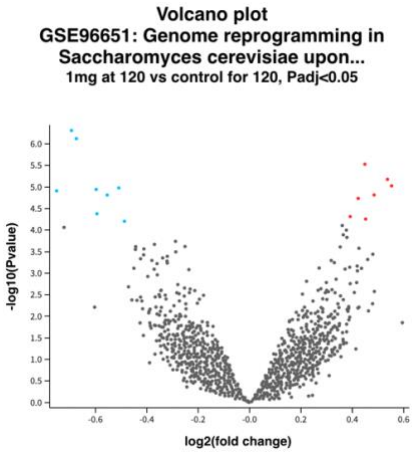


Fig1 : For 120 mins 1mg/L NP

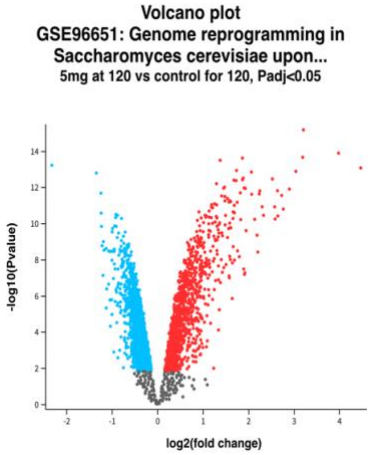


Fig2: For 120 mins 5mg/L NP

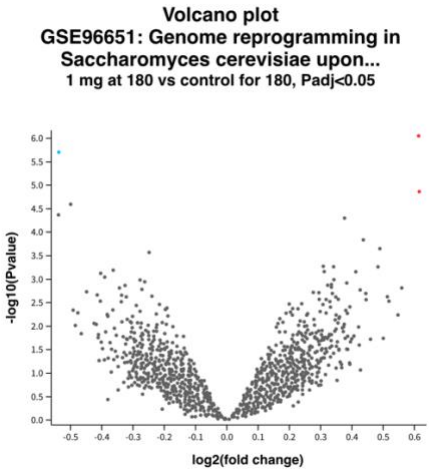


Fig1: For 180 mins 1mg /L NP

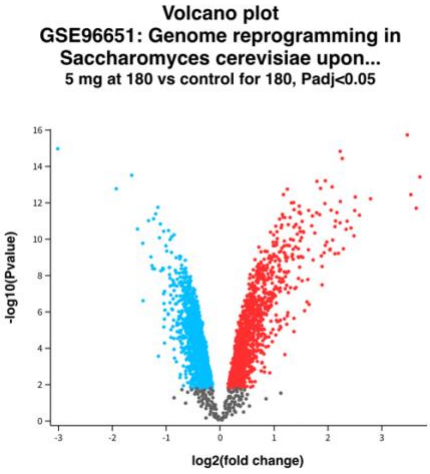


Fig2: For 180 mins 5mg/L NP

- The volcano plots shown above illustrate the exposure period and the amount of nonylphenol (NP) for each of the specified conditions.

- The table with the number of upregulated and downregulated genes in each of the condition can be summarized as:

Condition	**Number of upregulated genes**	**Number of downregulated genes**
For 120 mins 1mg/L	7	8
For 120 mins 5mg/L	1328	1808
For 180 mins 1mg/L	2	1
For 180 mins 5mg/L	1432	1856

b) (10 pts**) What is an appropriate statistical test to determine whether a given gene is differentially expressed genes between the control and treated samples? Be as specific as possible – name the test and appropriate variant or options to be used.**

Appropriate statistical methods which could be used to determine if a gene is differentially expressed between control and test samples depends on the number of samples in consideration for a particular experiment, this includes:

- **t – test** : used to compare two conditions where there is replication of samples. The t-test help determine signal to noise ratio. T distribution is present when both the conditions have equal mean. We can assign p-values to the T-test to determine if infact the possibility of means being equal is true (when p value is large)
- For 2 sets having unequal variance we can employ an alternate called as the **Welch's T-test**
- The T-test employed might be paired or unpaired depending on if the datasets are from independent replicates or not
- For testing more than one hypothesis at once we can use **ANOVA** – which has multiple statistical methods under it
- The **Family-wise error rate (FWER)** is one such test where it helps determine the probability of having false positives
- The **false discovery rates (FDR)** help measure the expected proportion of false positives among the significant hypotheses.
- If FDR includes cases where no hypothesis is significant – the proportion is set to 0
- **pFDR** considers cases where at least one significant hypothesis is found
- The **Benjamini and Hochberg FDR** is another variation for the FDR, where the test worries about whether an individual exceeds the adjusted thresholds or not rather than how much it exceeds a threshold.

Alternatively, we can use other methods to determine the differentially expressed genes between control and sample groups, these include :

- The **Z-test** picks all genes whose log-transformation deviates more than 2 standard deviations from mean
- The **P value** is a statistical value that indicates how likely the data has occurred by random chance
- The **Padjusted** value is where the level for statistical significance(alpha) is divided by number of hypotheses (genes or samples) – this is **Bonferroni** correction and is generally too stringent.

- The **log2 fold** change indicates how much the gene or transcript expression has changed between the comparison and control groups, this value is represented in logarithmic scale to base 2

c) (10 pts) **There are approximately 6000 yeast genes. If there are 20 genes found to be significantly differentially expressed at an FDR of 0.1, what does this mean regarding how many of the significant genes may be false positives?**

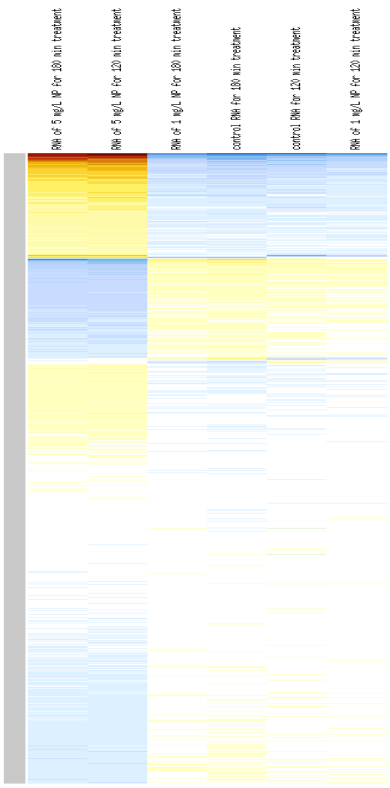
- The False discovery rate (FDR) = Number of false positives /(Total number of positive results)
- In the given question FDR is 0.1 and total number of positives is 20. Using the above formula, the number of false positives will be equal 2

d) (10 pts) **Given 6000 yeast genes tested on an expression microarray, how many false positives (rejected but true null hypotheses) would you expect at an alpha (p-value cutoff) of 0.05?**

- I would expect to have 300 false positives based on the formula that.The expected FDR equals the pvalue times the number of genes. In the given question the pvalue is 0.05 and the number of genes is 6000

3. (10 pts) **Using any suitable tool of your choice, generate a heat map of all the significantly differentially expressed genes in all the samples. Paste image below. Describe what tool and parameters you used.**

- I used the tool **ExAtlas, which is a software for meta-analysis of gene expression data** to generate a heatmap of all the significantly differentially expressed genes. I added the GEO accession numbers and loaded all the samples of my experiment to generate the heatmaps.
- The parameters that I used are as follows:
  a) The false discovery rate (adjusted Pvalue) that I used had a threshold value of 0.05, this is used instead of Pvalue to account for multiple hypothesis testing. The FDR value needs to be low to ensure that the gene which is differentially expressed does not occur due to chance event.
  b) The fold change threshold that I utilized was 1. The fold change indicates the change in transcription or gene expression between the control and the comparison groups.



4. (20 pts) **Generate a 2-D or 3-D PCA plot of all of the samples and paste image below. Also show a scree plot or show the percentage of variance captured on the 1st two principal components. What can you conclude from this PCA plot, about how the samples cluster, and which pairwise comparisons would yield the greatest number of differentially expressed genes?**

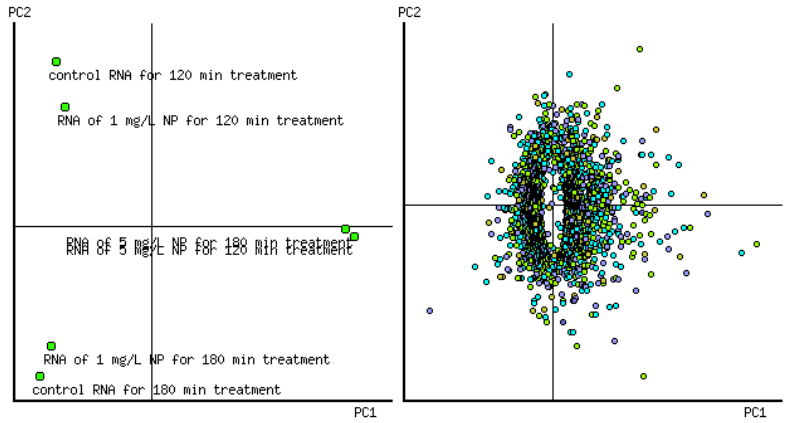


Figure: Depicting the PCA of all the experimental samples (left) and the genes (right) for each of the three distinct conditions

Table of eigenvalues for PCA without replicates for each condition

PC#	Value	Percent	Cumulative percent
1	0.02895	84.673	84.673
2	0.00349	10.207	94.88
3	0.00102	2.983	97.863
4	0.00045	1.316	99.179
5	0.00028	0.818	99.997

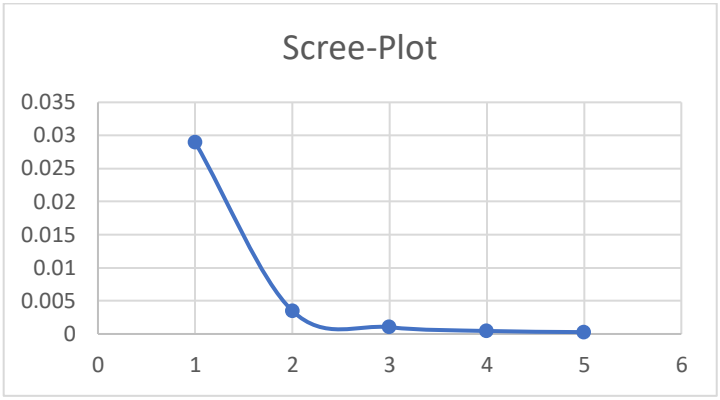


Figure: Depicting the Scree-plot for the eigenvalues as obtained in the PCA

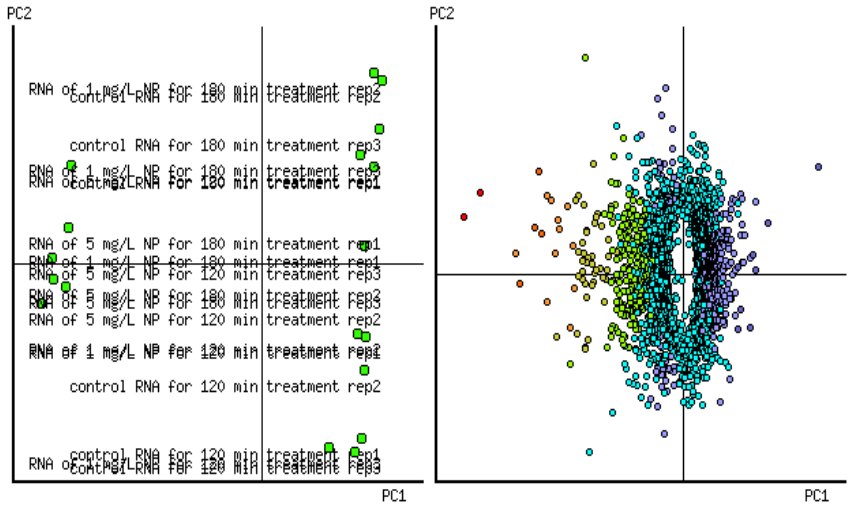


Figure: Depicting the PCA of all the experimental samples (left) and the genes (right) for each replicate of the three distinct conditions

Table of eigenvalues for PCA with replicates for each condition

PC#	Value	Percent	Cumulative percent
1	0.08718	75.382	75.382
2	0.01218	10.531	85.913
3	0.00454	3.925	89.838
4	0.00234	2.023	91.861
5	0.00158	1.366	93.227
6	0.00135	1.167	94.394
7	0.00128	1.106	95.5
8	0.00073	0.631	96.131
9	0.00066	0.57	96.701
10	0.00064	0.553	97.254

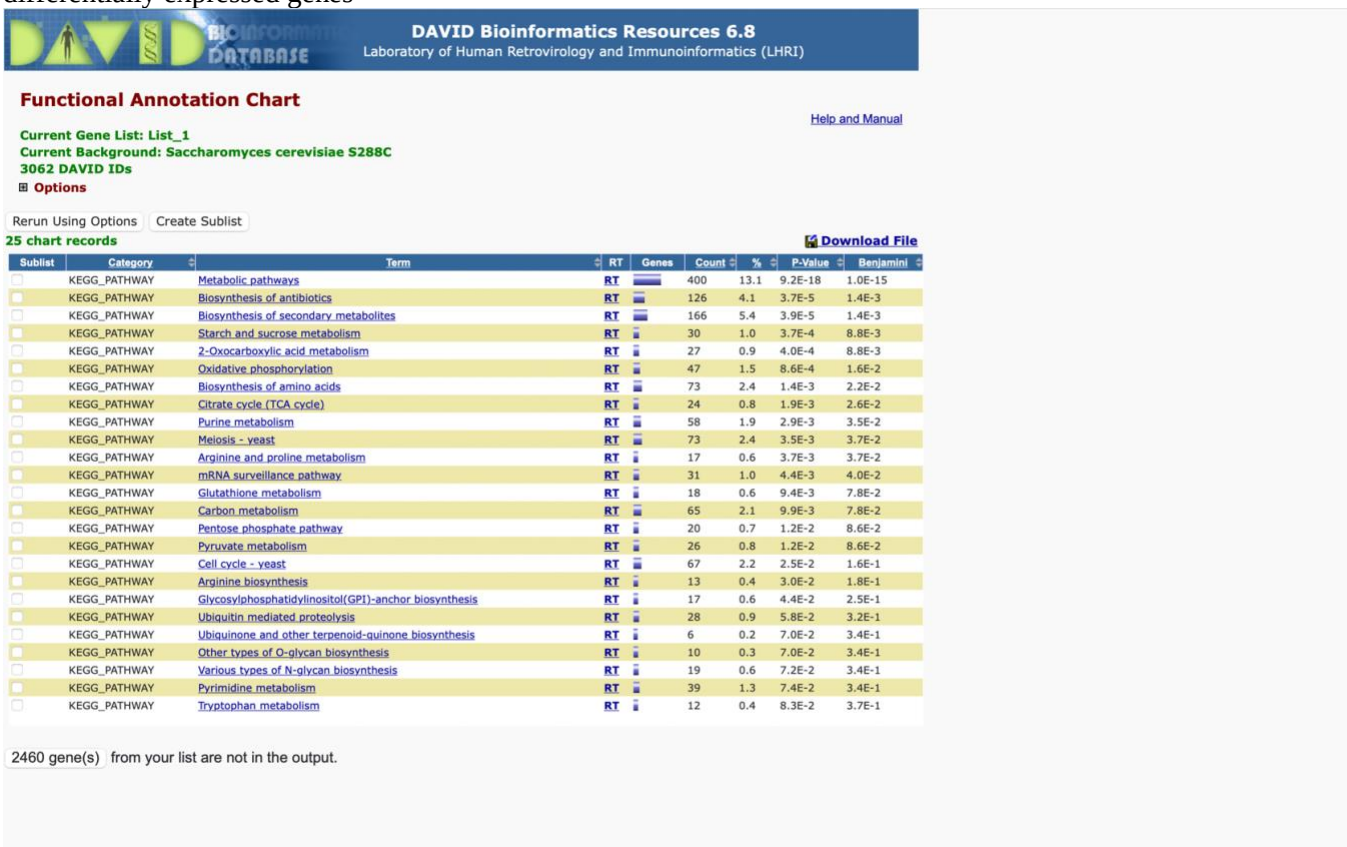
- The principal component analysis (PCA) helps visualize the components that contribute to the differences.
- The log2fold change was set at 1 and the co-relation was set at 0.07, which was the default setting.
- In our analysis, the variance with 1 principle component is around 84.673% and with 2 principle components we obtain a variance of around 10.207% (for the PCA without replicates) while we obtain a variance with 1 principle component to be around 75.382% and with two components we obtain a variance of around 10.531%. The other variance according to the # of PC's can be found in the tables.
- The PCA plot shows the clustering of samples based on similarity, this indicates that the samples which are treated with 5mg/L of NP cluster together and have a similar set of differentially expressed genes (in each of the two conditions 120 mins and 180 mins).
- For the 1mg/L at 120 mins and 1mg/L at 180 mins the clusters are much further away from each other which indicates that the intersection of differentially expressed genes between these two conditions are significantly lesser than our first condition. The same conclusion can be made for the controls for each of the experiment when viewed in a 3-dimensional aspect.
- The pairwise comparison which would result in the most number of differentially expressed genes in my opinion would be the 5mg/L vs control condition (for the condition 180 mins) followed by 5mg/L vs control condition (for the condition timed 120mins)

5. F**ind the pairwise comparison that gives the greatest number of differentially expressed genes. Use this list to perform pathway analysis. DAVID will convert your list to gene IDs that it can work with and identify the organism.**

a) (10 pts) **What KEGG pathways or functions are significantly over-represented in your list of differentially expressed genes? Paste below a table listing the pathways, with p-values and adjusted p-values. Also show a representative pathway diagram identifying the genes that are differentially expressed. This can be done in DAVID via Functional Annotation.**

- The condition I believe gave me the largest set of differentially expressed genes was 5mg/L at 180 mins
- I have taken the Gene ID obtained from geo2R used the gene list option in David to obtain the table pasted below which lists the pathways, p-values and adjusted p-values

- Based on the p-values it is clear that the metabolic pathway is the most significant in the list of differentially expressed genes



DAVID Bioinformatics Resources 6.8
Laboratory of Human Retrovirology and Immunoinformatics (LHRI)

Functional Annotation Chart

Help and Manual

Current Gene List: List_1
Current Background: Saccharomyces cerevisiae S288C
3062 DAVID IDs
Options

Rerun Using Options | Create Sublist

25 chart records

Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	KEGG_PATHWAY	Metabolic pathways	RT		400	13.1	9.2E-18	1.0E-15
	KEGG_PATHWAY	Biosynthesis of antibiotics	RT		126	4.1	3.7E-5	1.4E-3
	KEGG_PATHWAY	Biosynthesis of secondary metabolites	RT		166	5.4	3.9E-5	1.4E-3
	KEGG_PATHWAY	Starch and sucrose metabolism	RT		30	1.0	3.7E-4	8.8E-3
	KEGG_PATHWAY	2-Oxocarboxylic acid metabolism	RT		27	0.9	4.0E-4	8.8E-3
	KEGG_PATHWAY	Oxidative phosphorylation	RT		47	1.5	8.6E-4	1.6E-2
	KEGG_PATHWAY	Biosynthesis of amino acids	RT		73	2.4	1.4E-3	2.2E-2
	KEGG_PATHWAY	Citrate cycle (TCA cycle)	RT		24	0.8	1.9E-3	2.6E-2
	KEGG_PATHWAY	Purine metabolism	RT		58	1.9	2.9E-3	3.5E-2
	KEGG_PATHWAY	Meiosis - yeast	RT		73	2.4	3.5E-3	3.7E-2
	KEGG_PATHWAY	Arginine and proline metabolism	RT		17	0.6	3.7E-3	3.7E-2
	KEGG_PATHWAY	mRNA surveillance pathway	RT		31	1.0	4.4E-3	4.0E-2
	KEGG_PATHWAY	Glutathione metabolism	RT		18	0.6	9.4E-3	7.8E-2
	KEGG_PATHWAY	Carbon metabolism	RT		65	2.1	9.9E-3	7.8E-2
	KEGG_PATHWAY	Pentose phosphate pathway	RT		20	0.7	1.2E-2	8.6E-2
	KEGG_PATHWAY	Pyruvate metabolism	RT		26	0.8	1.2E-2	8.6E-2
	KEGG_PATHWAY	Cell cycle - yeast	RT		67	2.2	2.5E-2	1.6E-1
	KEGG_PATHWAY	Arginine biosynthesis	RT		13	0.4	3.0E-2	1.8E-1
	KEGG_PATHWAY	Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	RT		17	0.6	4.4E-2	2.5E-1
	KEGG_PATHWAY	Ubiquitin mediated proteolysis	RT		28	0.9	5.8E-2	3.2E-1
	KEGG_PATHWAY	Ubiquinone and other terpenoid-quinone biosynthesis	RT		6	0.2	7.0E-2	3.4E-1
	KEGG_PATHWAY	Other types of O-glycan biosynthesis	RT		10	0.3	7.0E-2	3.4E-1
	KEGG_PATHWAY	Various types of N-glycan biosynthesis	RT		19	0.6	7.2E-2	3.4E-1
	KEGG_PATHWAY	Pyrimidine metabolism	RT		39	1.3	7.4E-2	3.4E-1
	KEGG_PATHWAY	Tryptophan metabolism	RT		12	0.4	8.3E-2	3.7E-1

2460 gene(s) from your list are not in the output.

DAVID Bioinformatics Resources 6.8
Laboratory of Human Retrovirology and Immunoinformatics (LHRI)

Functional Annotation Table

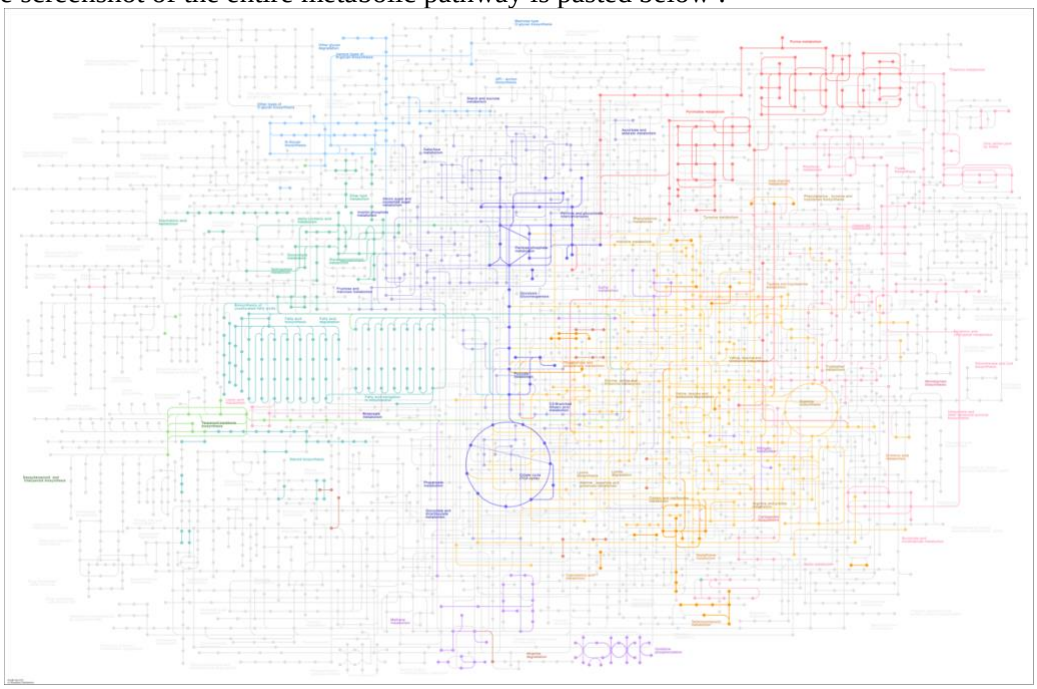
Help and Manual

Current Gene List: List_1
Current Background: Saccharomyces cerevisiae S288C
2987 DAVID IDs

977 record(s) Download File

1777350_at	(R,R)-butanediol dehydrogenase(BDH1)	Related Genes	Saccharomyces cerevisiae S288C
KEGG_PATHWAY	Butanoate metabolism,		
1772521_at	1,3-beta-D-glucan synthase(FKS1)	Related Genes	Saccharomyces cerevisiae S288C
KEGG_PATHWAY	Starch and sucrose metabolism, MAPK signaling pathway - yeast,		
1773780_at	1,3-beta-glucan synthase GSC2(GSC2)	Related Genes	Saccharomyces cerevisiae S288C
KEGG_PATHWAY	Starch and sucrose metabolism, MAPK signaling pathway - yeast,		
1772051_at	1,4-alpha-glucan branching enzyme(GLC3)	Related Genes	Saccharomyces cerevisiae S288C
KEGG_PATHWAY	Starch and sucrose metabolism, Metabolic pathways, Biosynthesis of secondary metabolites,		
1779230_at	1-acylglycerol-3-phosphate O-acyltransferase Slc1 (predicted)(slc1)	Related Genes	Schizosaccharomyces pombe 972h-
KEGG_PATHWAY	Glycerolipid metabolism, Glycerophospholipid metabolism, Metabolic pathways, Biosynthesis of secondary metabolites,		
1777661_at	1-phosphatidylinositol 4-kinase STT4(STT4)	Related Genes	Saccharomyces cerevisiae S288C
KEGG_PATHWAY	Inositol phosphate metabolism, Metabolic pathways, Phosphatidylinositol signaling system,		
1773708_at	1-phosphatidylinositol 4-kinase(PIK1)	Related Genes	Saccharomyces cerevisiae S288C
KEGG_PATHWAY	Inositol phosphate metabolism, Metabolic pathways, Phosphatidylinositol signaling system,		
1773117_at	1-pyrroline-5-carboxylate dehydrogenase(PUT2)	Related Genes	Saccharomyces cerevisiae S288C
KEGG_PATHWAY	Alanine, aspartate and glutamate metabolism, Arginine and proline metabolism, Metabolic pathways,		
1777706_at	18S rRNA pseudouridine methyltransferase(EMG1)	Related Genes	Saccharomyces cerevisiae S288C
KEGG_PATHWAY	Ribosome biogenesis in eukaryotes,		
1770894_at	19S proteasome regulatory subunit Rpt2(rpt2)	Related Genes	Schizosaccharomyces pombe 972h-
KEGG_PATHWAY	Proteasome,		
1778392_at	2,5-diamino-6-(ribosylamino)-4(3H)-pyrimidinone 5'-phosphate reductase(RIB7)	Related Genes	Saccharomyces cerevisiae S288C
KEGG_PATHWAY	Riboflavin metabolism, Metabolic pathways,		
1778282_at	2-isopropylmalate synthase LEU4(LEU4)	Related Genes	Saccharomyces cerevisiae S288C
KEGG_PATHWAY	Valine, leucine and isoleucine biosynthesis, Pyruvate metabolism, Metabolic pathways, Biosynthesis of secondary metabolites, 2-Oxocarboxylic acid metabolism, Biosynthesis of amino acids,		
1780138_at	2-isopropylmalate synthase LEU9(LEU9)	Related Genes	Saccharomyces cerevisiae S288C
KEGG_PATHWAY	Valine, leucine and isoleucine biosynthesis, Pyruvate metabolism, Metabolic pathways, Biosynthesis of secondary metabolites, 2-Oxocarboxylic acid metabolism, Biosynthesis of amino acids,		

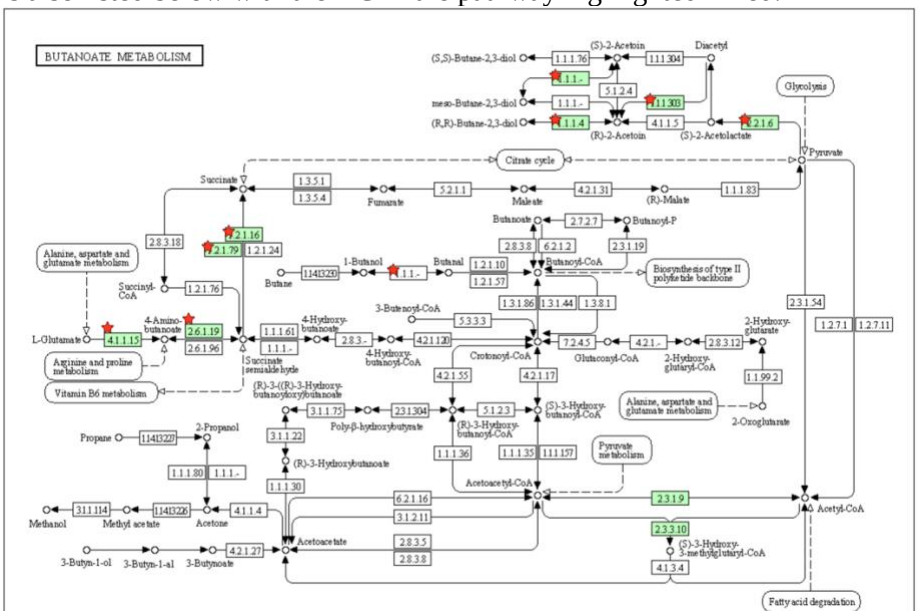
- The screenshot of the entire metabolic pathway is pasted below :



- The list of genes for this metabolic pathway is represented below and the total number of genes were ~ 400.

AFFYMETRIX_3PRIME_IVT_ID	GENE NAME	Related Genes	Species
1772051_at	1,4-alpha-glucan branching enzyme(GLC3)	RG	Saccharomyces cerevisiae S288C
1777661_at	1-phosphatidylinositol 4-kinase STT4(STT4)	RG	Saccharomyces cerevisiae S288C
1773708_at	1-phosphatidylinositol 4-kinase(PIK1)	RG	Saccharomyces cerevisiae S288C
1773117_at	1-pyrroline-5-carboxylate dehydrogenase(PUT2)	RG	Saccharomyces cerevisiae S288C
1778392_at	2,5-diamino-6-(ribosylamino)-4(3H)-pyrimidinone 5'-phosphate reductase(RIB7)	RG	Saccharomyces cerevisiae S288C
1778282_at	2-isopropylmalate synthase LEU4(LEU4)	RG	Saccharomyces cerevisiae S288C
1780138_at	2-isopropylmalate synthase LEU9(LEU9)	RG	Saccharomyces cerevisiae S288C
1771215_at	3'(2'),5'-bisphosphate nucleotidase(MET22)	RG	Saccharomyces cerevisiae S288C
1773765_at	3-deoxy-7-phosphoheptulonate synthase ARO4(ARO4)	RG	Saccharomyces cerevisiae S288C
1772019_at	3-hydroxyanthranilate 3,4-dioxygenase(BNA1)	RG	Saccharomyces cerevisiae S288C
1769683_at	3-isopropylmalate dehydrogenase(LEU2)	RG	Saccharomyces cerevisiae S288C
1776335_at	3-keto-steroid reductase(ERG27)	RG	Saccharomyces cerevisiae S288C
1769338_at	3-oxoacyl-[acyl-carrier-protein] reductase (NADPH)(OAR1)	RG	Saccharomyces cerevisiae S288C
1769913_s_at	4-amino-5-hydroxymethyl-2-methylpyrimidine phosphate synthase(THI5)	RG	Saccharomyces cerevisiae S288C
1774010_at	4-aminobutyrate transaminase(UGA1)	RG	Saccharomyces cerevisiae S288C
1774361_at	4-hydroxybenzoate octaprenyltransferase(COQ2)	RG	Saccharomyces cerevisiae S288C
1770451_at	5-formyltetrahydrofolate cyclo-ligase(FAU1)	RG	Saccharomyces cerevisiae S288C
1774859_at	5-methyltetrahydropteroyltriglutamate-homocysteine S-methyltransferase(MET6)	RG	Saccharomyces cerevisiae S288C
1774540_at	6-phosphogluconolactonase SOL4(SOL4)	RG	Saccharomyces cerevisiae S288C
1773695_at	ATP phosphoribosyltransferase(HIS1)	RG	Saccharomyces cerevisiae S288C
1770281_at	Anp1p(ANP1)	RG	Saccharomyces cerevisiae S288C
1775355_at	Bst1p(BST1)	RG	Saccharomyces cerevisiae S288C
1779961_at	C-22 sterol desaturase(ERG5)	RG	Saccharomyces cerevisiae S288C
1772336_at	C-5 sterol desaturase(ERG3)	RG	Saccharomyces cerevisiae S288C

- The pathway diagram for butanoate metabolism which is one of the sub-types under metabolism is also listed below with the DG in the pathway highlighted in red.



List genes are shown in red

DAVID Gene Name
(R,R)-butanediol dehydrogenase(BDH1)
4-aminobutyrate transaminase(UGA1)
acetolactate synthase catalytic subunit(ILV2)
acetolactate synthase regulatory subunit(ILV6)
acetyl-CoA C-acetyltransferase(ERG10)
glutamate decarboxylase GAD1(GAD1)
hydroxymethylglutaryl-CoA synthase(ERG13)
putative aryl-alcohol dehydrogenase(AAD10)
putative aryl-alcohol dehydrogenase(AAD14)
putative aryl-alcohol dehydrogenase(AAD3)
putative aryl-alcohol dehydrogenase(AAD4)
putative dehydrogenase BDH2(BDH2)
succinate-semialdehyde dehydrogenase (NAD(P)(+))(UGA2)

b) (10 pts) **What statistical test is used to calculate the p-value for pathway over-representation in DAVID or the pathway tool you used?**

- Since we are performing multiple testing, we are taking our gene list and doing an enrichment test for every pathway which is annotated by the tool – David.
- Higher the number of pathways in consideration then higher the chance that we will find significant p-values by random
- Therefore, we would require multiple correcting testing for multiple pathways such as using Benjamini-Hochberg procedure that decreases false discovery rate. DAVID requests adjusted *p*-values by using the linear step-up method of Benjamini and Hochberg [2]
- The Bonferroni in DAVID is the Bonferroni Šidák *p*-value [1] which is a technique slightly less conservative than Bonferroni.
- FDR in DAVID requests adaptive linear step-up adjusted *p*-values for approximate control of the false discovery rate, as discussed in Benjamini and Hochberg[3]. Use the lowest slope method to estimate the number of true NULL hypotheses.
- The Fisher's exact test is used to determine whether the proportions of those falling in each category differs by groups. This is also incorporated in David, to measure gene-enrichment in annotation terms. The values are computed by summing probabilities over defined sets

References:

- Šidák, Z. (1967). "Rectangular Confidence Regions for the Means of Multivariate Normal Distributions." Journal of the American Statistical Association 62:626-633.
- Benjamini, Y. and Hochberg, Y. (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." Journal of the Royal Statistical Society. Series B, Statistical Methodology 57:289-300.
- Benjamini, Y., and Hochberg, Y. (2000). "On the Adaptive Control of the False Discovery Rate in Multiple Testing with Independent Statistics." Journal of Educational and Behavioral Statistics 25:60-83.