Data Assignment 1

Name: Jyotirmaya Singh
Roll Number: 2021055

Q1.

Geo Accession Number: GSE14520

Title: Gene expression data of human hepatocellular carcinoma (HCC)

Platform: GPL3921 [HT_HG-U133A] Affymetrix HT Human Genome U133A Array

Q2.

Printing the summary statistics for the first five probe ids (can be changed as per requirement in the code)

[1] "1007_s_at"

Statistic	N	Mean	St. Dev.	Min	Max
probe_data_list	445	6.887	0.765	5.416	10.172

Data Assignment 1

[1] "1053_at"

Statistic	N	Mean	St. Dev.	Min	Max
probe_data_list	445	4.410	0.516	3.211	7.164

[1] "117_at"

Statistic	N	Mean	St. Dev.	Min	Max
probe_data_list	445	3.988	0.453	3.312	8.444

[1] "121_at"

Statistic	N	Mean	St. Dev.	Min	Max
probe_data_list	445	5.711	0.328	4.861	6.607

[1] "1255_g_at"

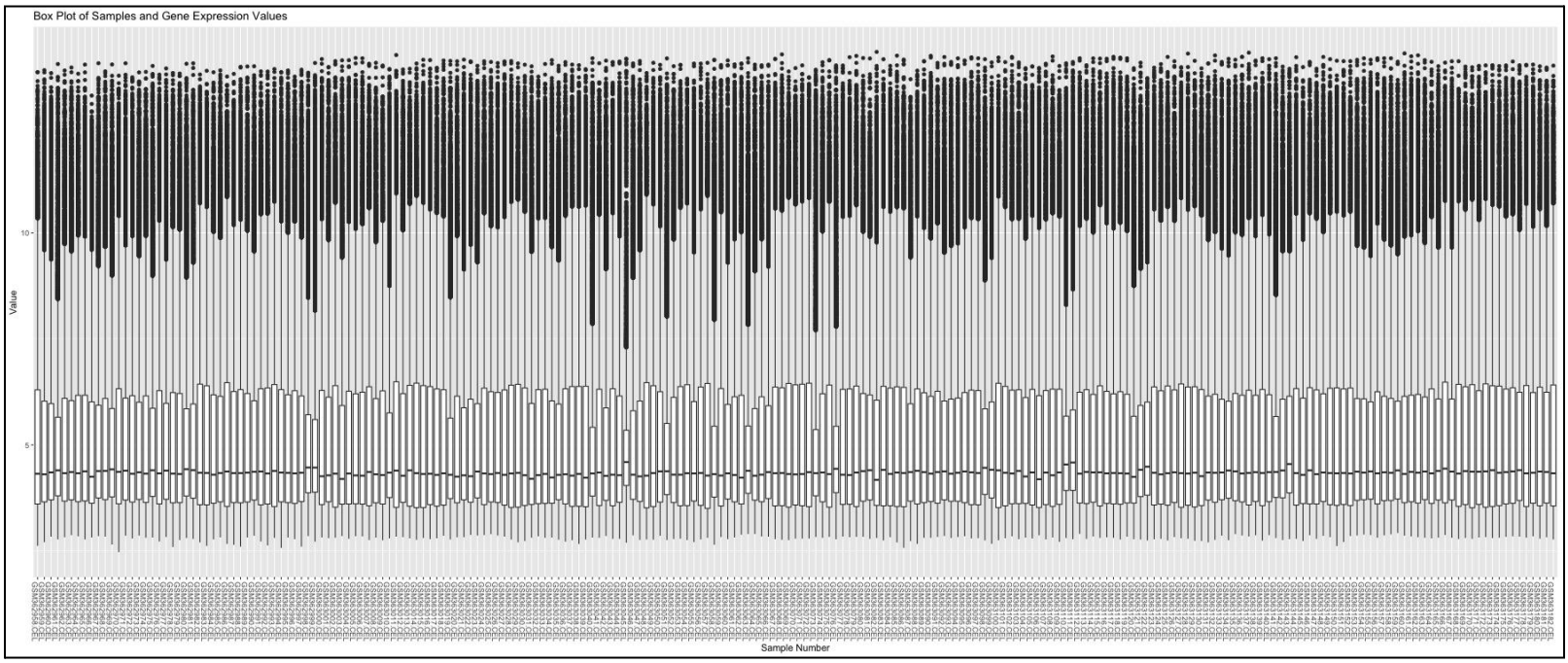
Statistic	N	Mean	St. Dev.	Min	Max
probe_data_list	445	3.197	0.120	2.927	3.682

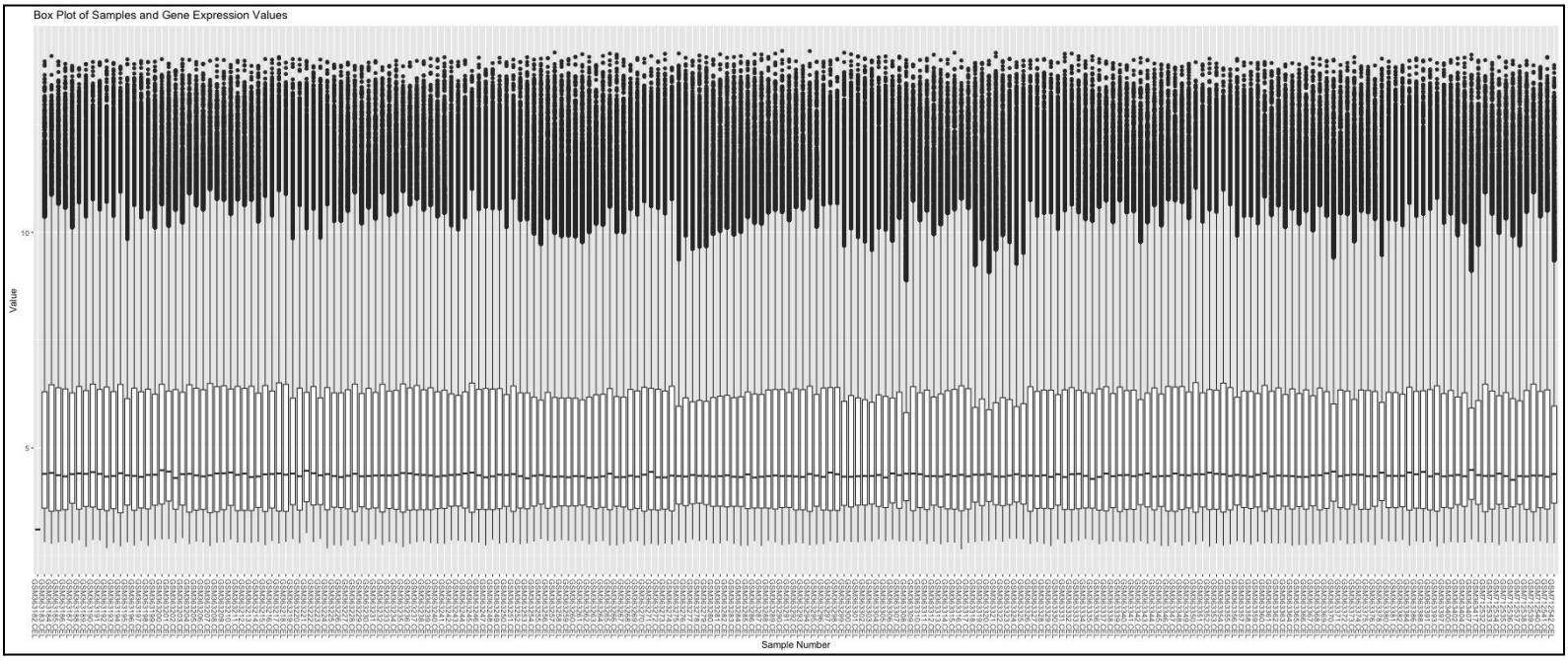
Data Assignment 1

Box Plot:

(Split in two for ease of reading)

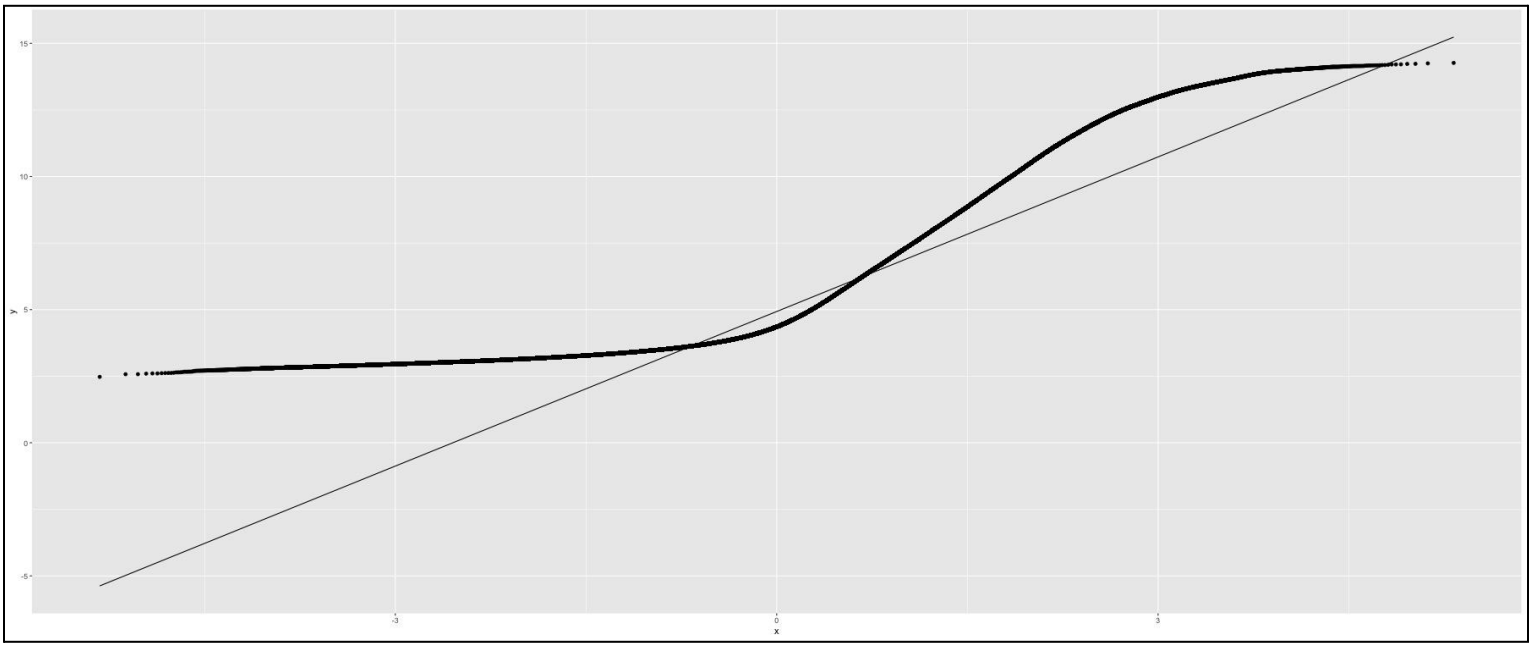
Plotted the gene values of each sample as a box, so we have 450 boxes (225 boxes in each plot).





Data Assignment 1

QQ Plot:



Skewness of all gene data = 1.33152
Kurtosis of all gene data = 4.287771

Data Assignment 1

Q3.

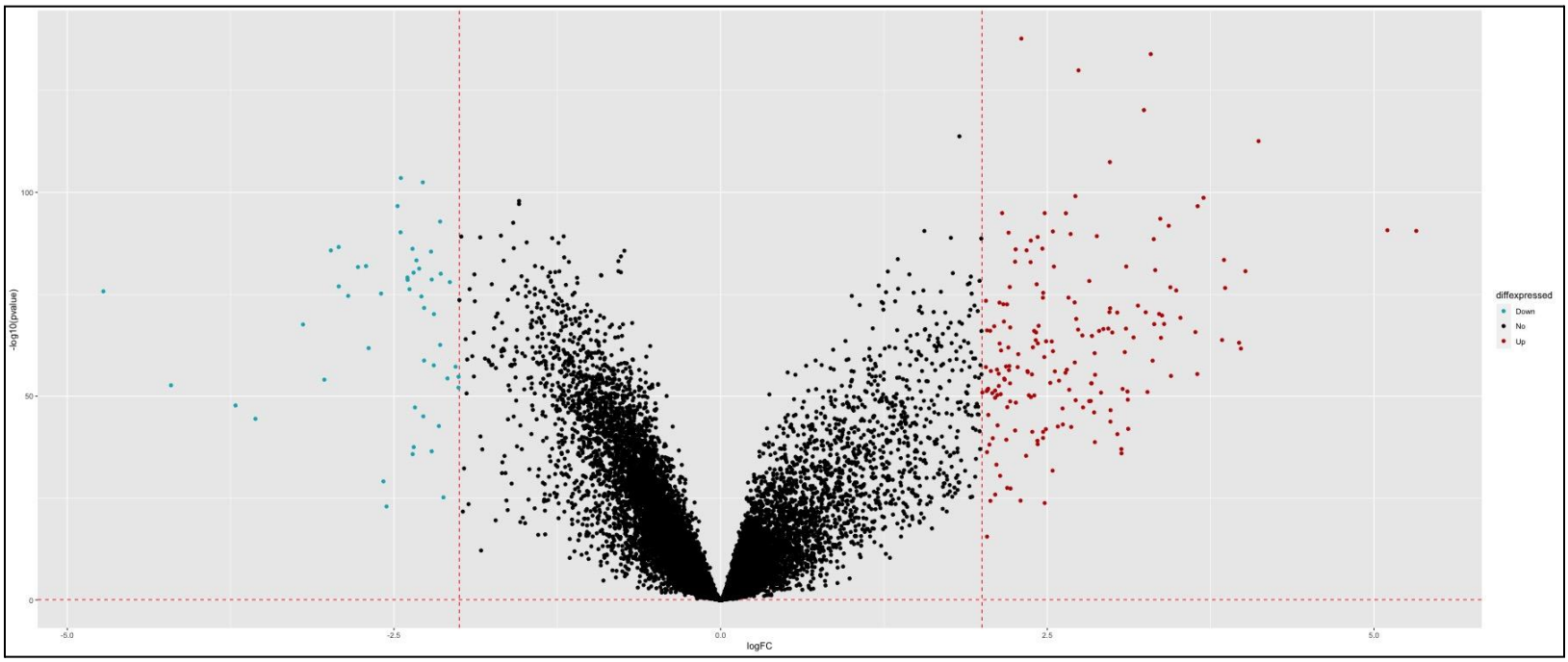
Since the microarray data is already normalised, we do not need to perform normalisation again.
However, if we were to get completely unprocessed data, taking the log transform would transform skew data to approximately conform to normality. The range of the values would change as well, the gene expression after log transform lies between 0 to 16.
If we were to take the log transform of already log transformed data (which is the case with us) the shape of the data would remain approximately the normally distributed. Only the range of values would change from 0 to 16, to, 0 to 4. It also makes data more comparable.

Note: I have not performed log transformation in the code (since data is already log transformed), however, I have written the code that would perform the log transformation (commented it out).

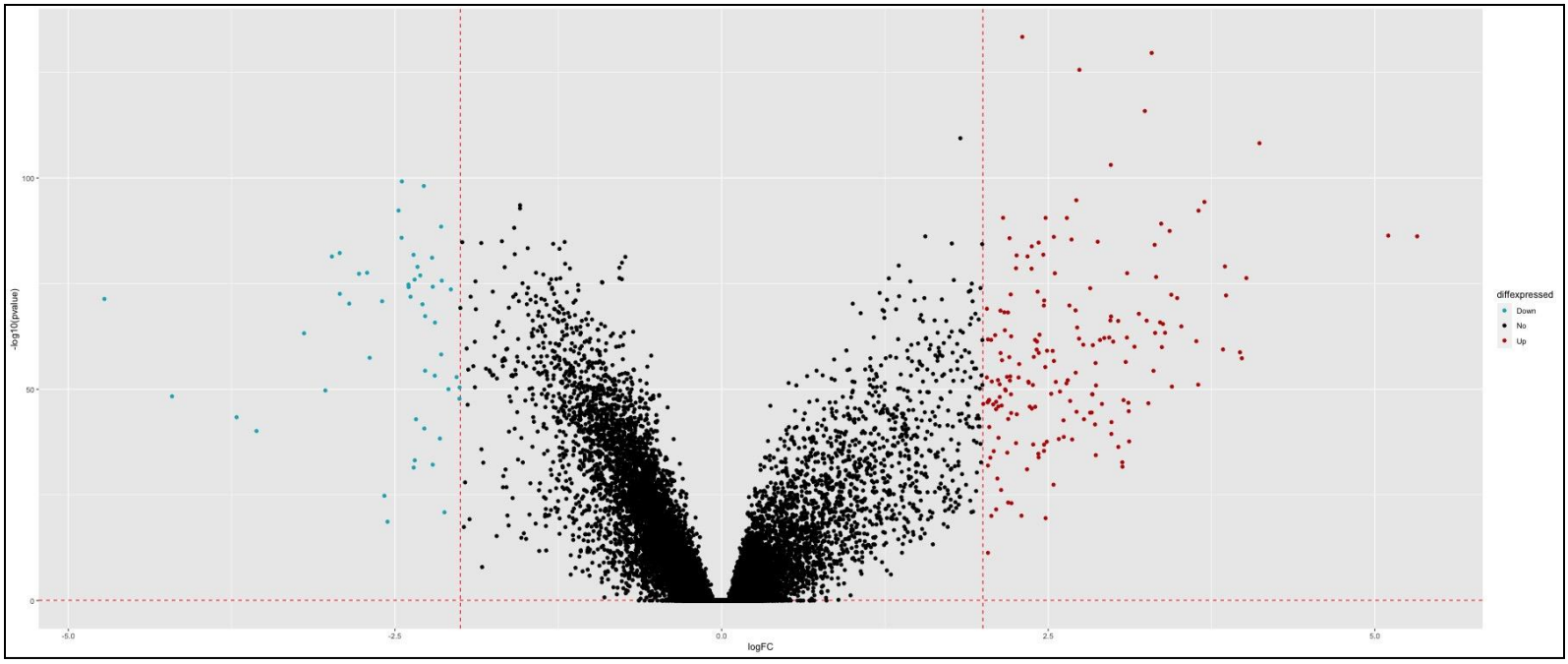
Data Assignment 1

Q4.

Volcano Plot:



(Without Holm Correction, Up)
(With Holm Correction, Down)



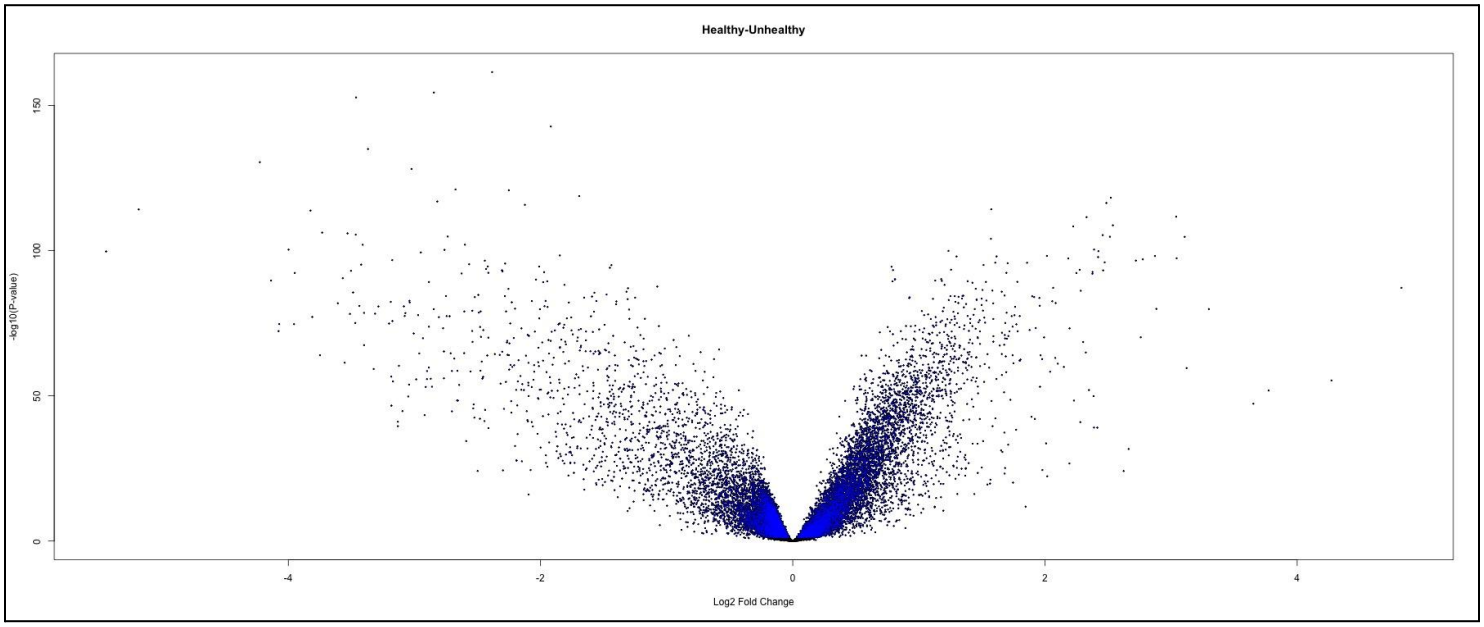
Data Assignment 1

Number of Differentially Expressed Genes Obtained: 230

Data Assignment 1

Q5.

Volcano Plot:



Number of Differentially Expressed Genes Obtained: 252

Data Assignment 1

Q6.

log(FoldChange) cutoff = |2|
p-value cutoff = 0.05

Normally we choose the fold change cutoff to be around 1 or 1.5, however I chose 2 as the cutoff since 1 and 1.5 were giving a large number of DEGs which did not seem correct to me. The difference between only a few genes is what leads to the difference between a person with cancer and a person without. The number of DEGs were fairly large even with a cutoff of 1.5, hence I chose 2 to be the cutoff value.

We generally we choose the p-value cutoff to be 0.05. Moreover, after Holm correction, a lot of p values were rounded up to 1. This meant that a good number of p-values had been adjusted for errors and therefore there was no need to change the cutoff to a lower value.
A p-value of 0.05 means that we expect an extreme statistic to show up in a chosen sample less than 5% of the time which is quite a low percentage anyway.

Data Assignment 1

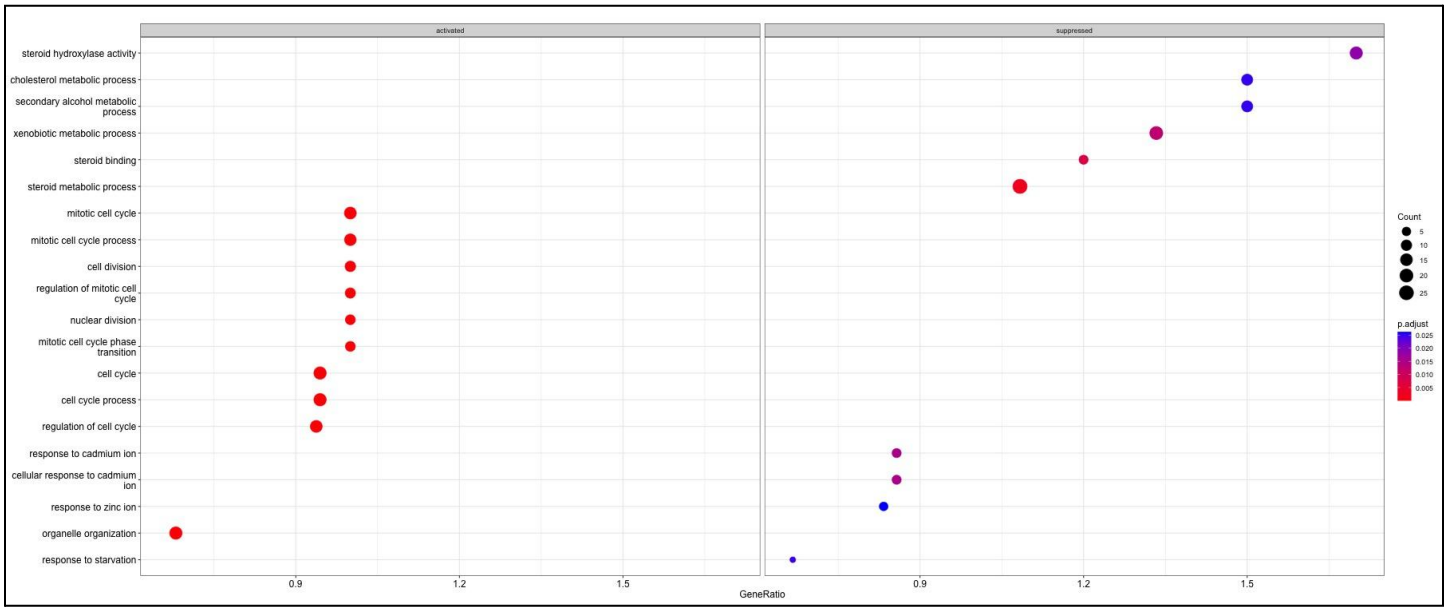
Q7.

Done in R Script

Q8.

GO Analysis

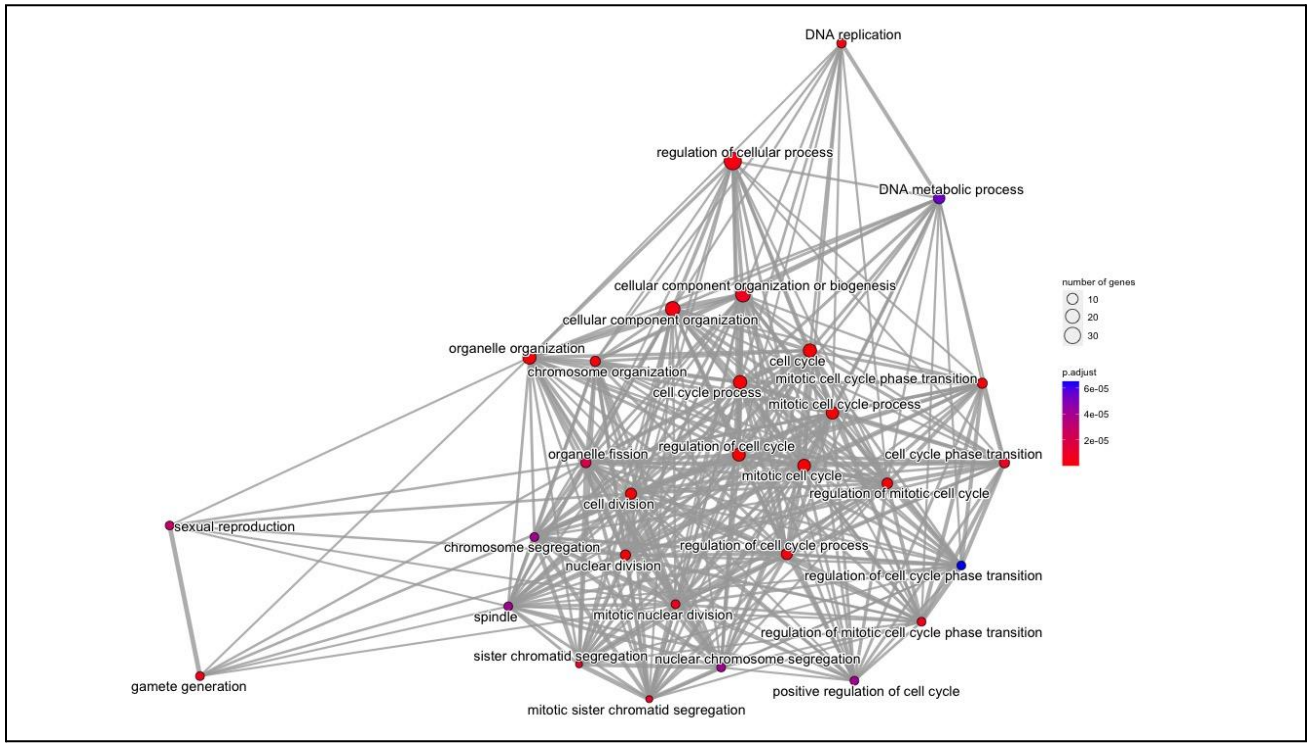
1. Dot Plot



Interpretation: The colour of the dot indicates the p-value with which the particular gene is determined. The count indicates the numbers of genes involved in that particular process.

Data Assignment 1

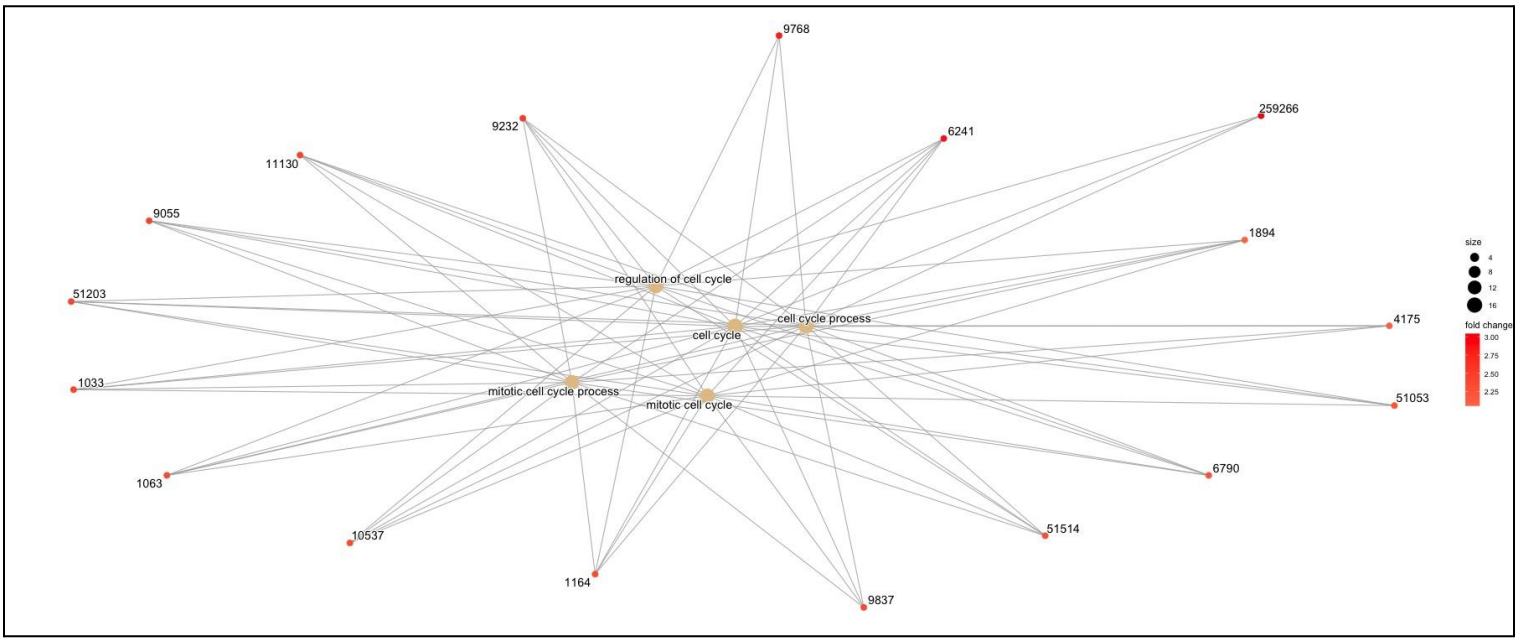
2. Enrichment Map



Interpretation: The map shows how functions are related to one another. The colour of the dots indicates the p-value with which that particular process is determined. The number of genes responsible for a particular process is indicated by the size of the dots.

Data Assignment 1

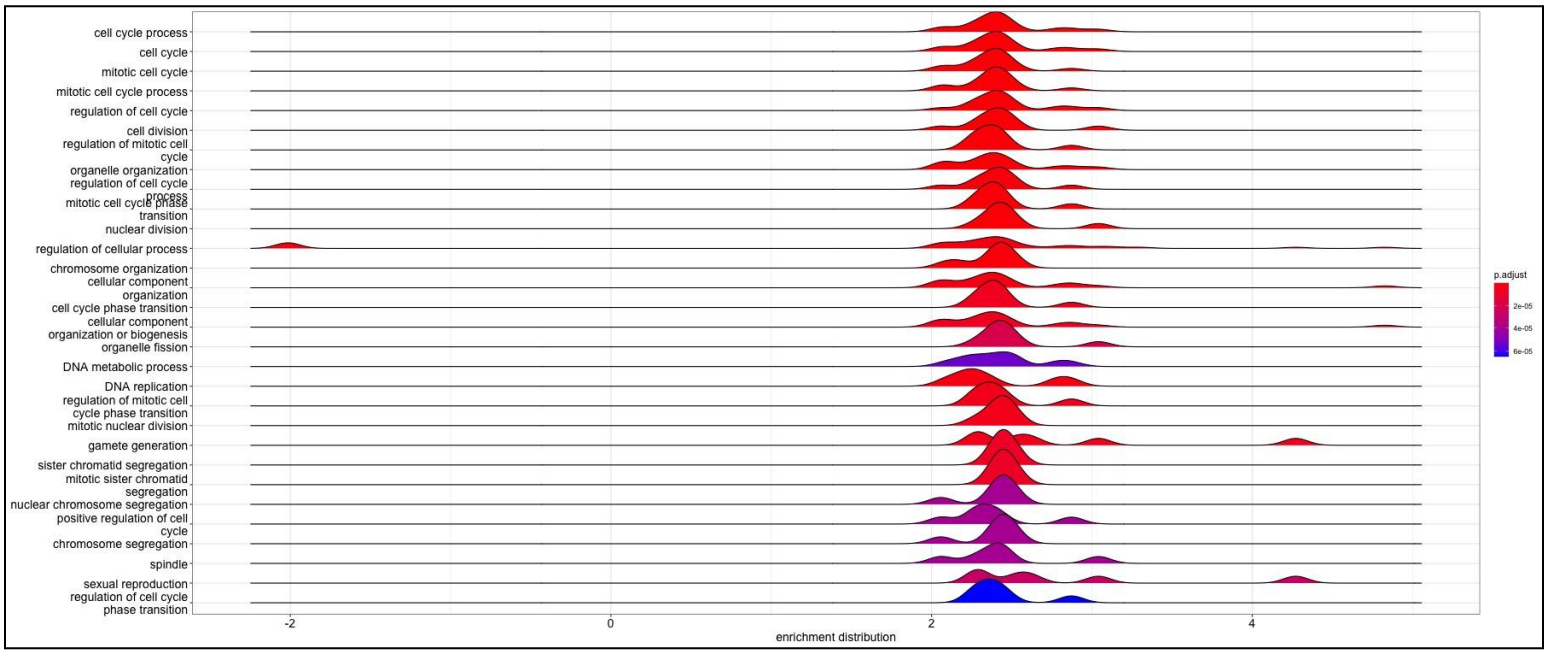
3. Category Netplot



Interpretation: The map shows how a particular gene is related to the above biological processes. The gene is identified by its ENTREZ ID. The colour of the gene indicates the calculated fold change. The size of the dot indicates the numbers of genes responsible for that particular process.

Data Assignment 1

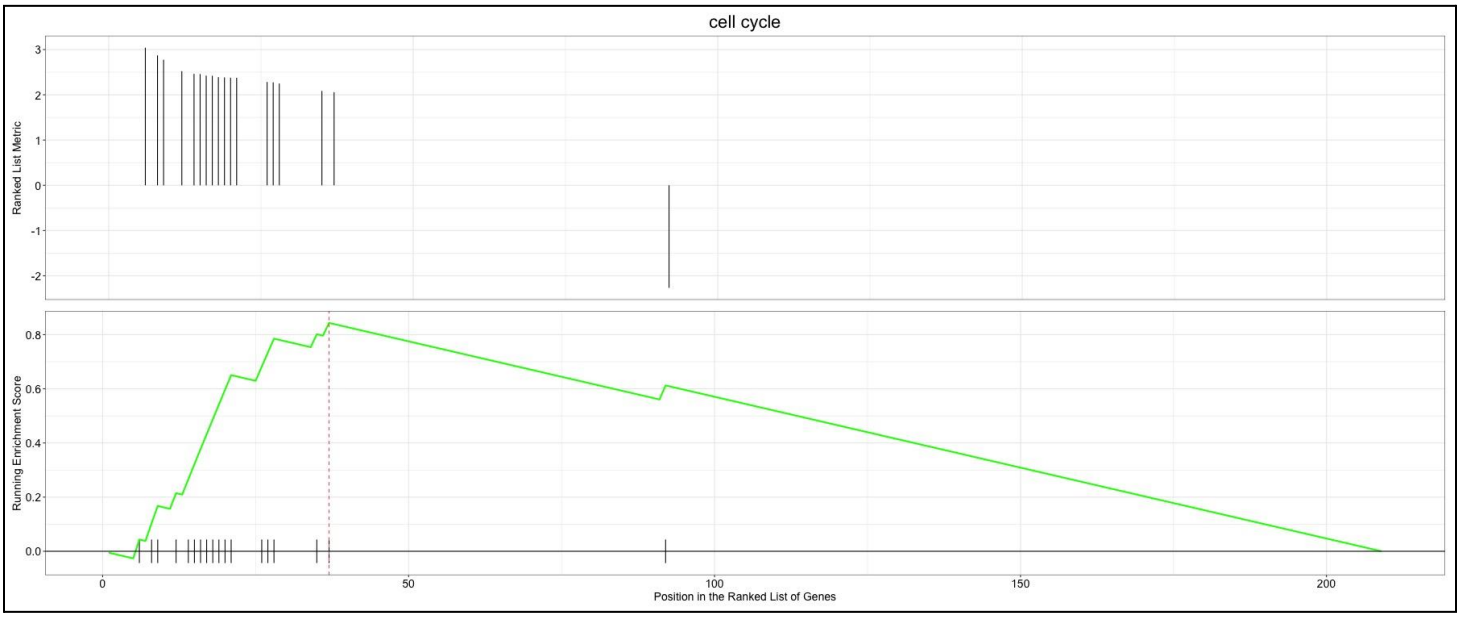
4. Ridge Plot



Interpretation: The above plot shows how functions are correlated with the genes. If the curve is to the right of 0, then the gene list is positively correlated to that particular function. Likewise, if the curve is to the left of 0 then the gene list is negatively correlated to that particular function.

Data Assignment 1

5. GSEA Plot

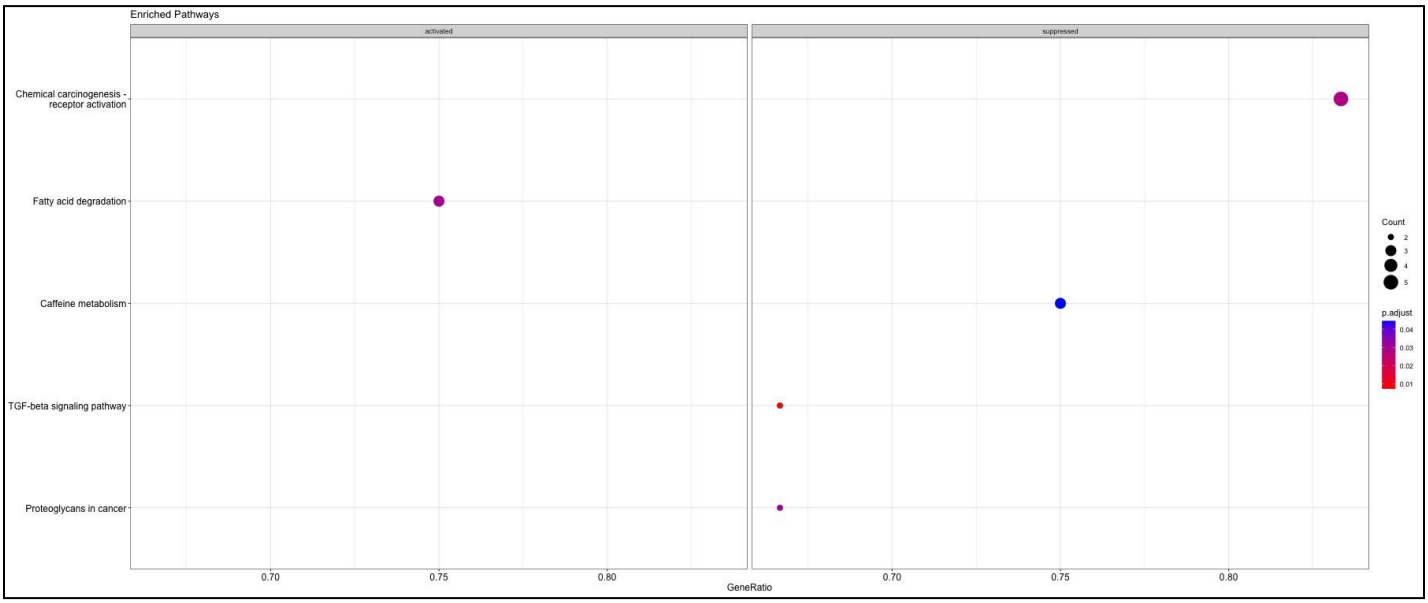


Interpretation: The above plot shows us the enrichment value (the peak of the curve is the enrichment value) for a particular process (cell cycle in our case). The higher the enrichment value, the more enriched the pathway/function is and vice versa.

Data Assignment 1

KEGG Analysis

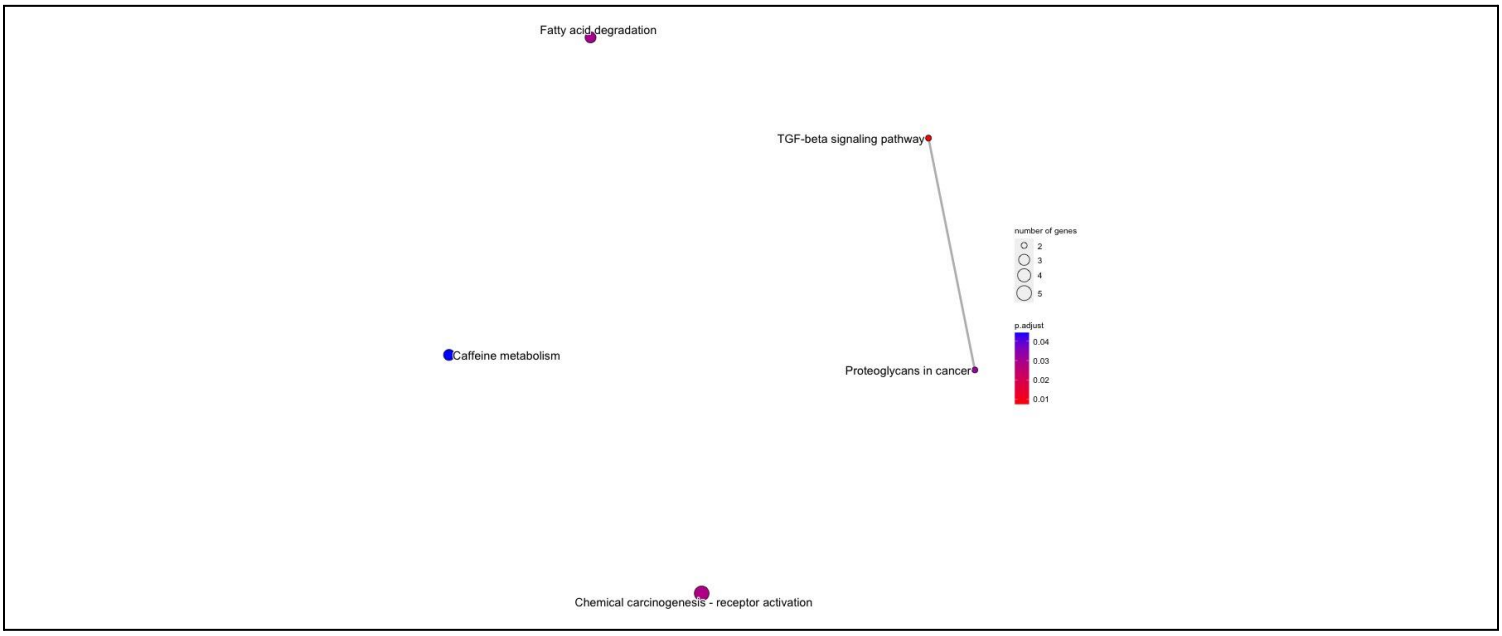
1. Dot Plot



Interpretation: The colour of the dot indicates the p-value with which the particular gene is determined. The count indicates the numbers of genes involved in that particular process.

Data Assignment 1

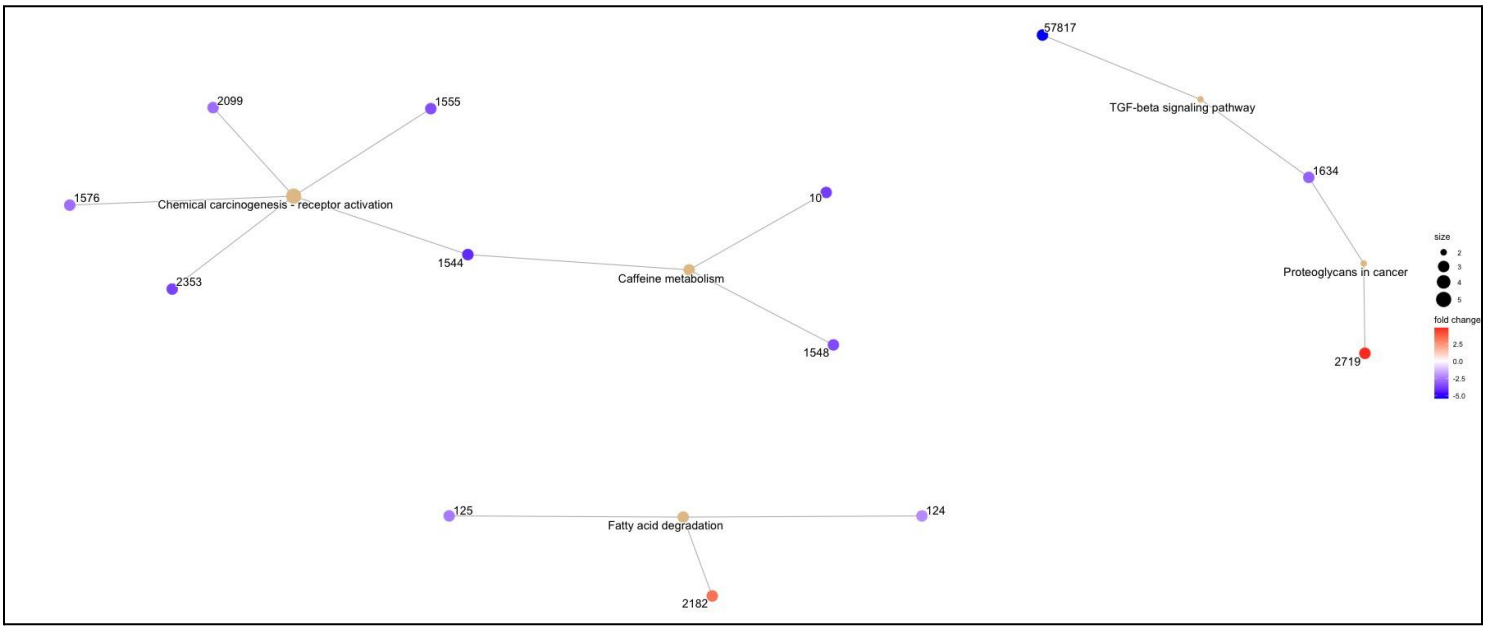
2. Enrichment Map



Interpretation: The map shows how functions are related to one another. The colour of the dots indicates the p-value with which that particular process is determined. The number of genes responsible for a particular process is indicated by the size of the dots.

Data Assignment 1

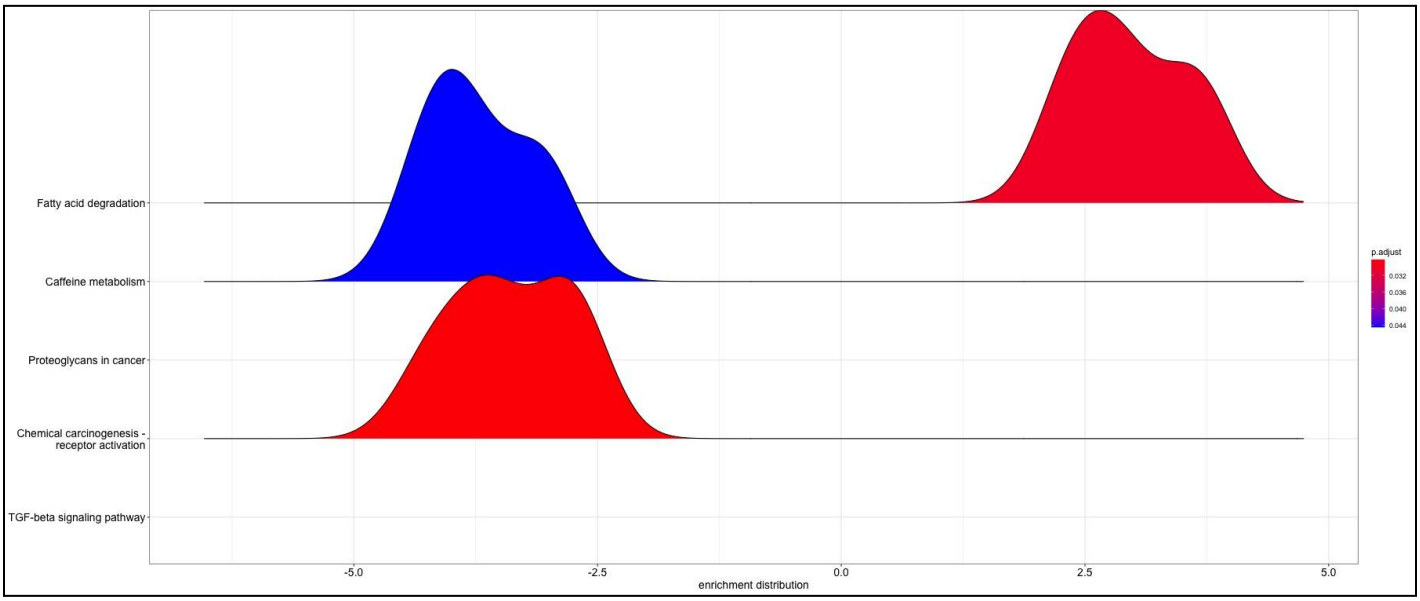
3. Category Netplot



Interpretation: The map shows how a particular gene is related to the above biological processes. The gene is identified by its ENTREZ ID. The colour of the gene indicates the calculated fold change. The size of the dot indicates the numbers of genes responsible for that particular process.

Data Assignment 1

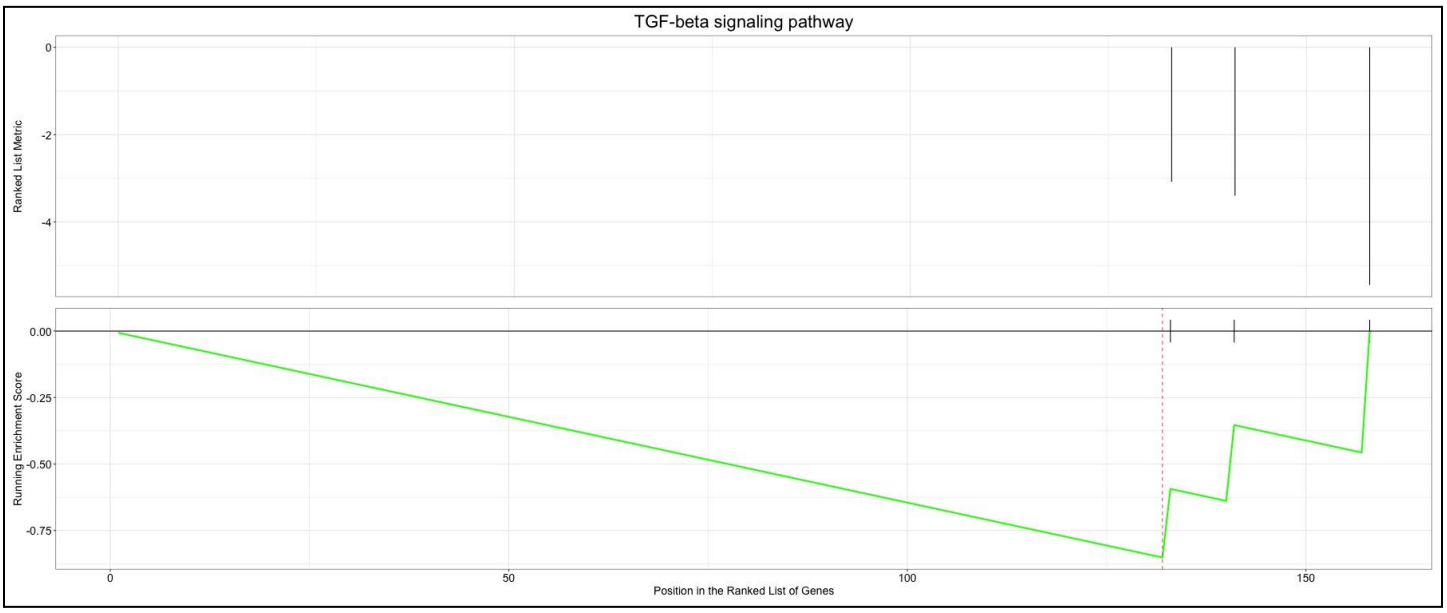
4. Ridge Plot



Interpretation: The above plot shows how functions are correlated with the genes. If the curve is to the right of 0, then the gene list is positively correlated to that particular function. Likewise, if the curve is to the left of 0 then the gene list is negatively correlated to that particular function.

Data Assignment 1

5. GSEA Plot



Interpretation: The above plot shows us the enrichment value (the valley of the curve is the enrichment value) for a particular pathway (TGF-beta signalling pathway in our case). The higher the enrichment value, the more enriched the pathway/function is and vice versa.

Data Assignment 1

Q9.

Pathway: Steroid metabolic process
NES Score: -2.039797
Since the NES Score < 0, it means that the above pathway is downregulated by our set of genes. So the above signalling pathway is negatively affected, which is a possible side effect of cancer.

Pathway: Cell division
NES Score: 3.268671
Since the NES Score > 0, it means that the above pathway is upregulated by our set of genes. So, degradation of fatty acids is higher, which is what happens in cancer (uncontrolled growth of cells).

The same can be done for the rest of the pathways (available in gse@result).

Go to file/function | Addins

pba1.R | pba.R | gse@result | original_gene_list | de_genes_new | kk2@result | groupA_exprs__dataframe | pbbac

Filter

	ONTOLOGY	ID	Description	setSize	enrichmentScore	NES	pvalue	p.ad
GO:0007049	BP	GO:0007049	cell cycle	18	0.8434531	4.202143	1.000000e-10	1.00
GO:0022402	BP	GO:0022402	cell cycle process	18	0.8434531	4.202143	1.000000e-10	1.00
GO:0000278	BP	GO:0000278	mitotic cell cycle	15	0.8865979	4.051040	1.000000e-10	1.00
GO:1903047	BP	GO:1903047	mitotic cell cycle process	14	0.8820513	3.914357	1.000000e-10	1.00
GO:0051726	BP	GO:0051726	regulation of cell cycle	16	0.8282846	3.868513	1.000000e-10	1.00
GO:0051301	BP	GO:0051301	cell division	10	0.8643216	3.268671	1.695905e-08	1.69
GO:0007346	BP	GO:0007346	regulation of mitotic cell cycle	9	0.9050000	3.236259	2.028419e-08	2.02
GO:0000280	BP	GO:0000280	nuclear division	8	0.9054726	3.053631	6.528593e-08	6.52
GO:0044772	BP	GO:0044772	mitotic cell cycle phase transition	8	0.9054726	3.053631	6.528593e-08	6.52
GO:0006996	BP	GO:0006996	organelle organization	25	0.5619746	3.205868	1.890252e-07	1.89
GO:0051276	BP	GO:0051276	chromosome organization	8	0.8656716	2.919406	6.665110e-07	6.66
GO:0010564	BP	GO:0010564	regulation of cell cycle process	11	0.7777757	3.079548	9.935103e-07	9.93
GO:0006260	BP	GO:0006260	DNA replication	7	0.8613861	2.682289	2.029742e-06	2.02
GO:0050794	BP	GO:0050794	regulation of cellular process	83	0.3242864	3.007898	2.296209e-06	2.29
GO:0016043	BP	GO:0016043	cellular component organization	43	0.3969368	2.875725	2.656693e-06	2.65
GO:0007276	BP	GO:0007276	gamete generation	6	0.8965517	2.638266	4.680059e-06	4.68
GO:0140014	BP	GO:0140014	mitotic nuclear division	6	0.8965517	2.638266	4.680059e-06	4.68
GO:1901990	BP	GO:1901990	regulation of mitotic cell cycle phase transition	6	0.8965517	2.638266	4.680059e-06	4.68
GO:0071840	BP	GO:0071840	cellular component organization or biogenesis	44	0.3874228	2.854129	6.239077e-06	6.23
GO:0044770	BP	GO:0044770	cell cycle phase transition	9	0.8007182	2.863350	6.279744e-06	6.27
GO:0000070	BP	GO:0000070	mitotic sister chromatid segregation	5	0.9362745	2.513424	8.280828e-06	8.28
GO:0000819	BP	GO:0000819	sister chromatid segregation	5	0.9362745	2.513424	8.280828e-06	8.28
GO:0048285	BP	GO:0048285	organelle fission	9	0.7715663	2.759103	2.004588e-05	2.00

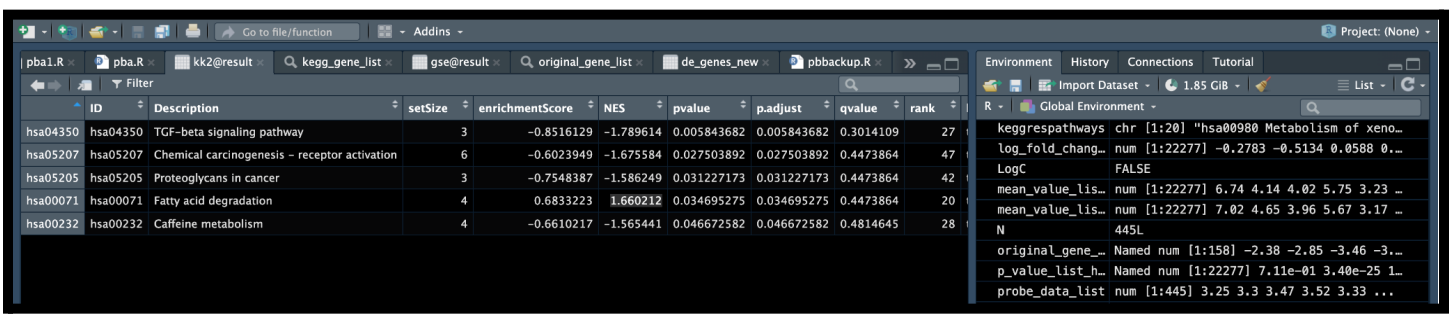
Showing 1 to 23 of 261 entries, 12 total columns

Data Assignment 1

Pathway: TGF-beta signaling pathway
NES Score: -1.789614
Since the NES Score < 0, it means that the above pathway is downregulated by our set of genes. So the above signalling pathway is negatively affected.

Pathway: Fatty acid degradation
NES Score: 1.660212
Since the NES Score > 0, it means that the above pathway is upregulated by our set of genes. So, degradation of fatty acids is higher.

The same can be done for the rest of the pathways (available in kk2@result).



	ID	Description	setSize	enrichmentScore	NES	pvalue	p.adjust	qvalue	rank
hsa04350	hsa04350	TGF-beta signaling pathway	3	-0.8516129	-1.789614	0.005843682	0.005843682	0.3014109	27
hsa05207	hsa05207	Chemical carcinogenesis - receptor activation	6	-0.6023949	-1.675584	0.027503892	0.027503892	0.4473864	47
hsa05205	hsa05205	Proteoglycans in cancer	3	-0.7548387	-1.586249	0.031227173	0.031227173	0.4473864	42
hsa00071	hsa00071	Fatty acid degradation	4	0.6833223	1.660212	0.034695275	0.034695275	0.4473864	20
hsa00232	hsa00232	Caffeine metabolism	4	-0.6610217	-1.565441	0.046672582	0.046672582	0.4814645	28

keggrespathways	chr [1:20] "hsa00980 Metabolism of xeno…
log_fold_chang…	num [1:22277] -0.2783 -0.5134 0.0588 0.…
LogC	FALSE
mean_value_lis…	num [1:22277] 6.74 4.14 4.02 5.75 3.23 …
mean_value_lis…	num [1:22277] 7.02 4.65 3.96 5.67 3.17 …
N	445L
original_gene_…	Named num [1:158] -2.38 -2.85 -3.46 -3.…
p_value_list_h…	Named num [1:22277] 7.11e-01 3.40e-25 1…
probe_data_list	num [1:445] 3.25 3.3 3.47 3.52 3.33 ...

References:

1. https://www.ncbi.nlm.nih.gov/geo/geo2r/?acc=GSE14520&platform=GPL3921
2. https://learn.gencore.bio.nyu.edu/rna-seq-analysis/gene-set-enrichment-analysis/