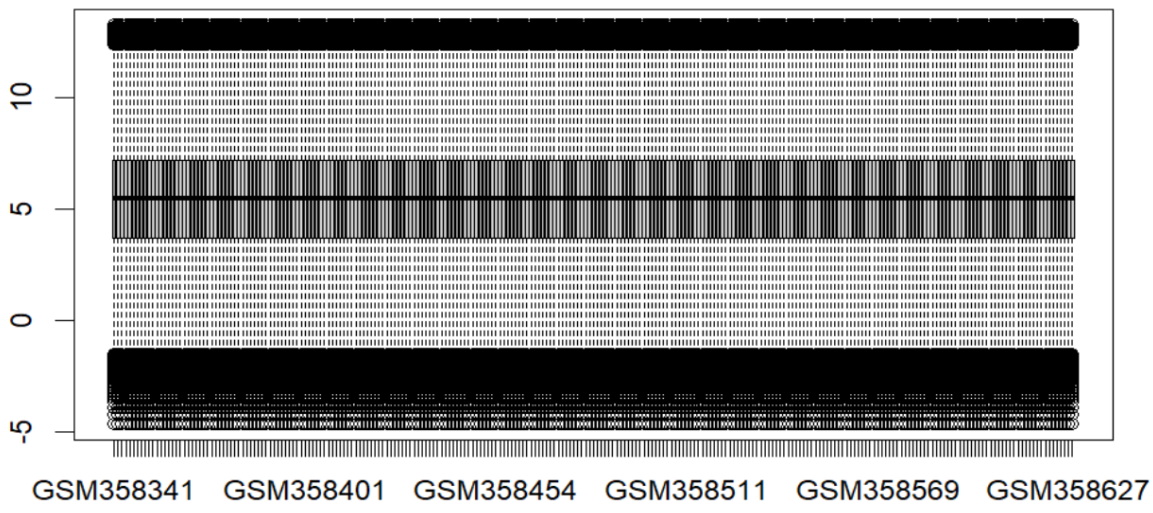
PB ASSIGNMENT 1 README

A1) To download the microarray data, we will use the **GEOquery package in R**. We will load series and platform data from **GEO using the getGEO() function**. We will select a dataset with at least 100 samples and two classes. In this example, we will use the **GSE14333 dataset**, which contains **290 colorectal cancer patient** samples and two classes (stage A and D). We will filter out samples marked as **"X" and perform log2 transformation**, if necessary, based on data distribution.

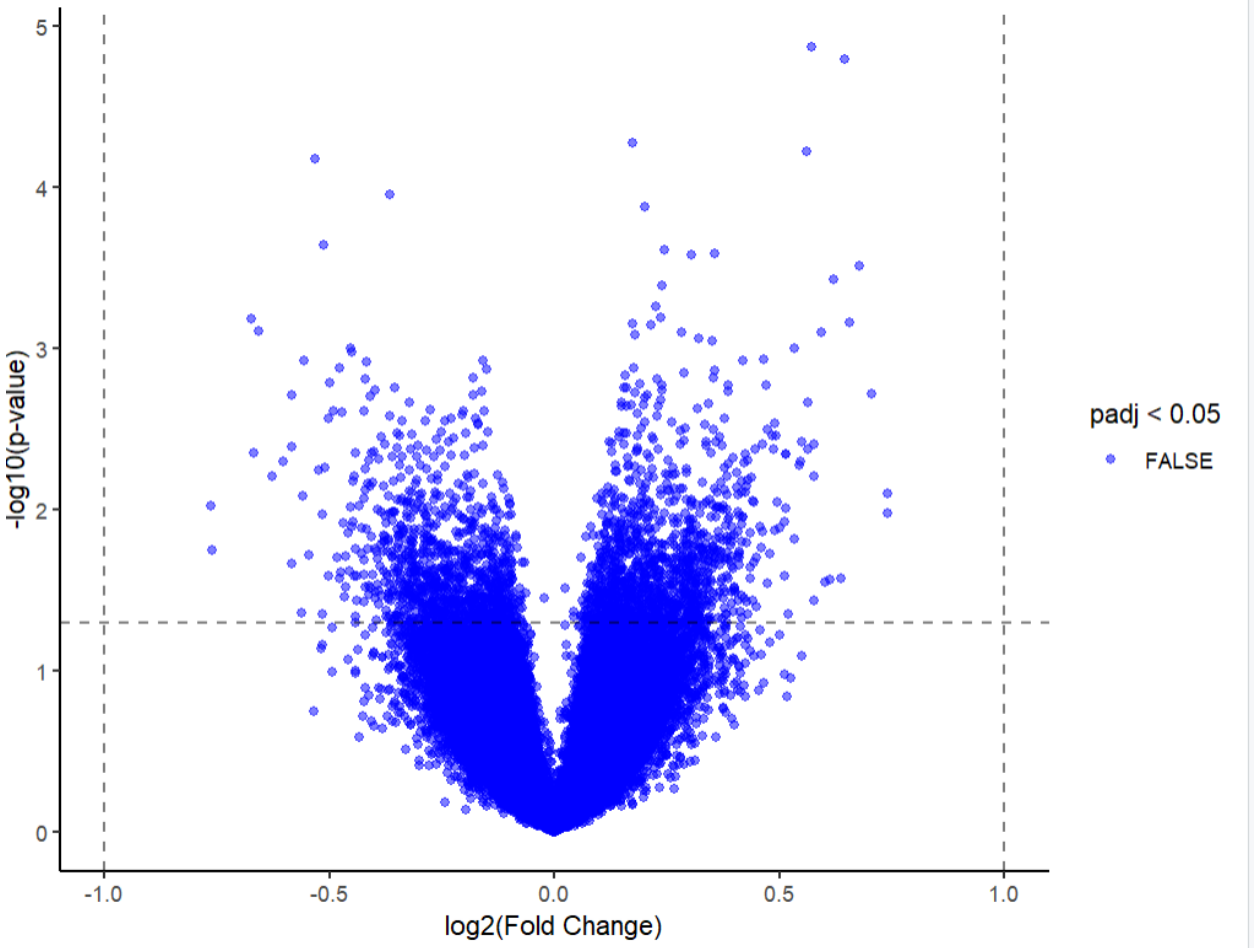
A2) After performing exploratory data analysis (EDA) and preprocessing on the gene expression data, the next step is to extract and list the data attributes. The code checks for missing values in the data using the is**.na() function**, and for outliers using the **boxplot() function**. The **normalizeBetweenArrays() function** from the **limma package** is used to normalize the data.
Next, an ExpressionSet object is created from the normalized data using the **ExpressionSet()** function. This object contains various attributes that can be extracted for further analysis. The **pData() and fData() functions are used to extract the phenotype and feature data**, respectively.
Finally, the attributes() function is used to list the data attributes. T**he eset object contains attributes such as the class, dimension, and data type, while the pdata and fdata objects contain the column names and data types of the phenotype and feature data, respectively.**



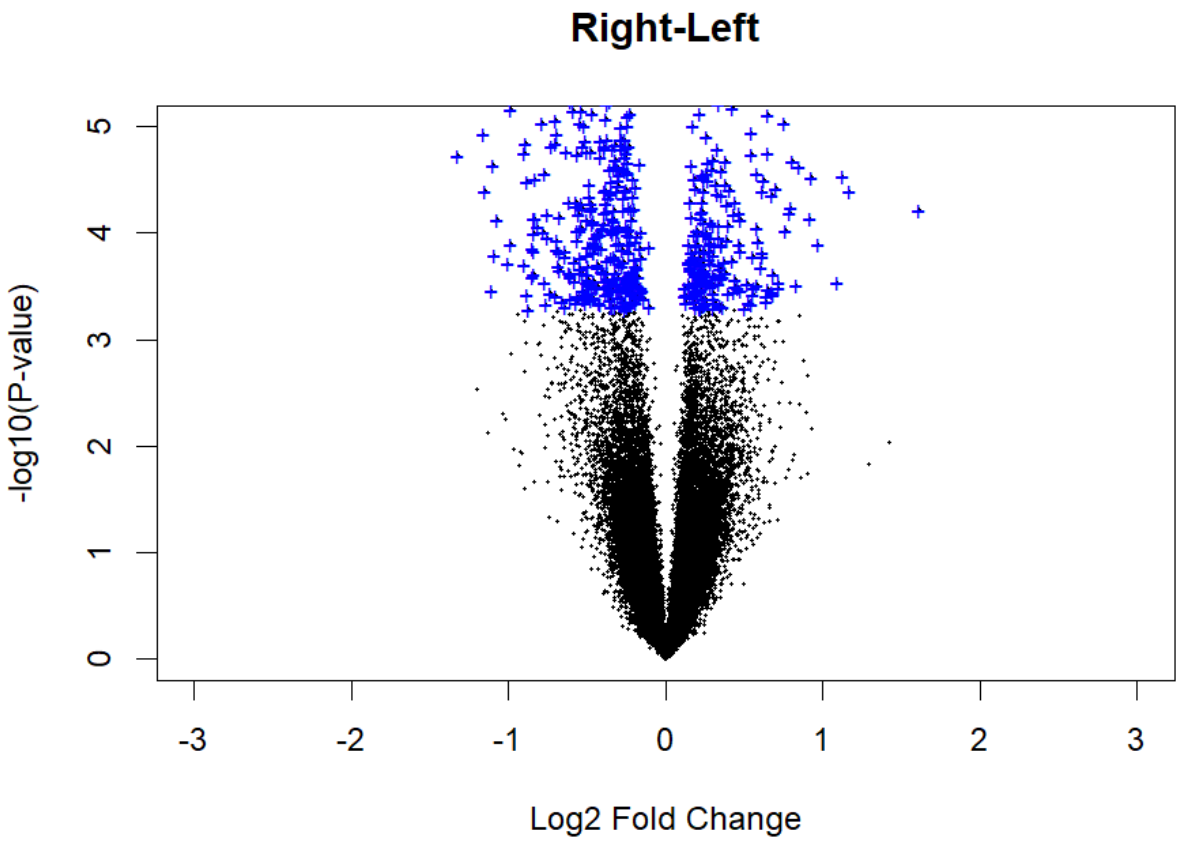
A3) The code checks **if a log2 transformation is necessary** for the microarray data based on its distribution. It calculates the quantiles of the data and checks if the difference between the **minimum and maximum values is greater than 50 and if the 99th percentile value is greater than 100**. If either of these conditions is true, it performs a log2 transformation on the data. If the data contains **any zero or negative values, they are set to NaN (Not a Number)** to

prevent issues during the log transformation.
Since the GSE14333 dataset had already undergone log2 transformation and had no considerable outliers, there was no requirement for further log transformation in this assignment.

A4) The code performs differential expression analysis using a **simple t-test and log fold change.** It assigns the samples to two groups and performs a t-test between the two groups for each gene. The results are stored in a **data frame and the p-values are corrected using the Holm correction method**. Finally, a volcano plot is created using the log fold change and corrected p-values, with significant genes highlighted in red. The plot also includes dashed lines representing a fold change of -1 and 1 and a significance level of p=0.05.



A5) The code uses the **limma package** to perform differential expression analysis. First, a design matrix is set up to model the groups, and a **linear model is fitted using lmFit()**. Contrasts of interest are set up, and the model coefficients are recalculated **using makeContrasts() and contrasts.fit(). eBayes()** is used to compute statistics, **and topTable() is used to generate a table of the top significant genes**, which is filtered using quantiles to remove outliers. Finally, a volcano plot is created using the filtered data, highlighting significant genes and showing the log2 fold change on the x-axis and -log10 p-value on the y-axis.



A6) Choosing a significant cutoff based on log(FC) and p-values depends on the specific study and its goals. Generally, a **p-value cutoff of 0.05 is commonly used to identify significant differential expressions**. For log(FC) cutoff, it depends on the context and the desired level of stringency. **A log(FC) cutoff of 1 (2-fold change) or 2 (4-fold change) is commonly used**. The choice of cutoff values should be justified based on the study's biological context, sample size, and other factors that affect the statistical power of the analysis.

A7) The code performs enrichment analysis using the set of genes obtained from the differential expression analysis. It uses the **"clusterProfiler" package in R**. First, the gene symbols from the **topTable** output are obtained. Then, the **"enrichGO" function** is used to perform gene set enrichment analysis using the **"org.Hs.eg.db" database** for the **Gene Ontology Biological Process ontology (ont = "BP")**. The enrichment results are corrected for multiple testing using the **Holm method (pAdjustMethod= "holm")**. The dotplot function is used to visualize the enriched pathways.

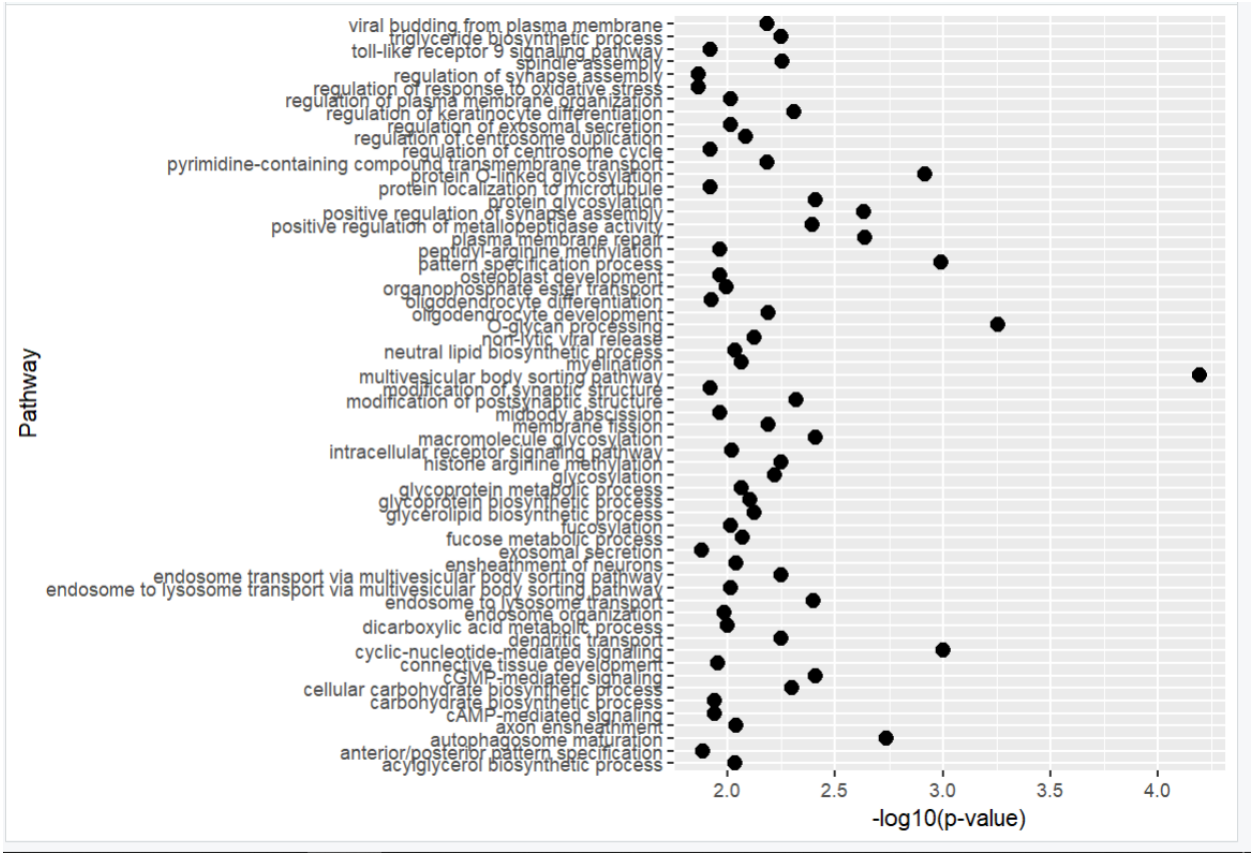
A8) Here is the explanation of the different parameters used in the Gene set enrichment analysis code:

1. **gene**: A vector of gene symbols from the topTable output that will be used as input for the gene set enrichment analysis.

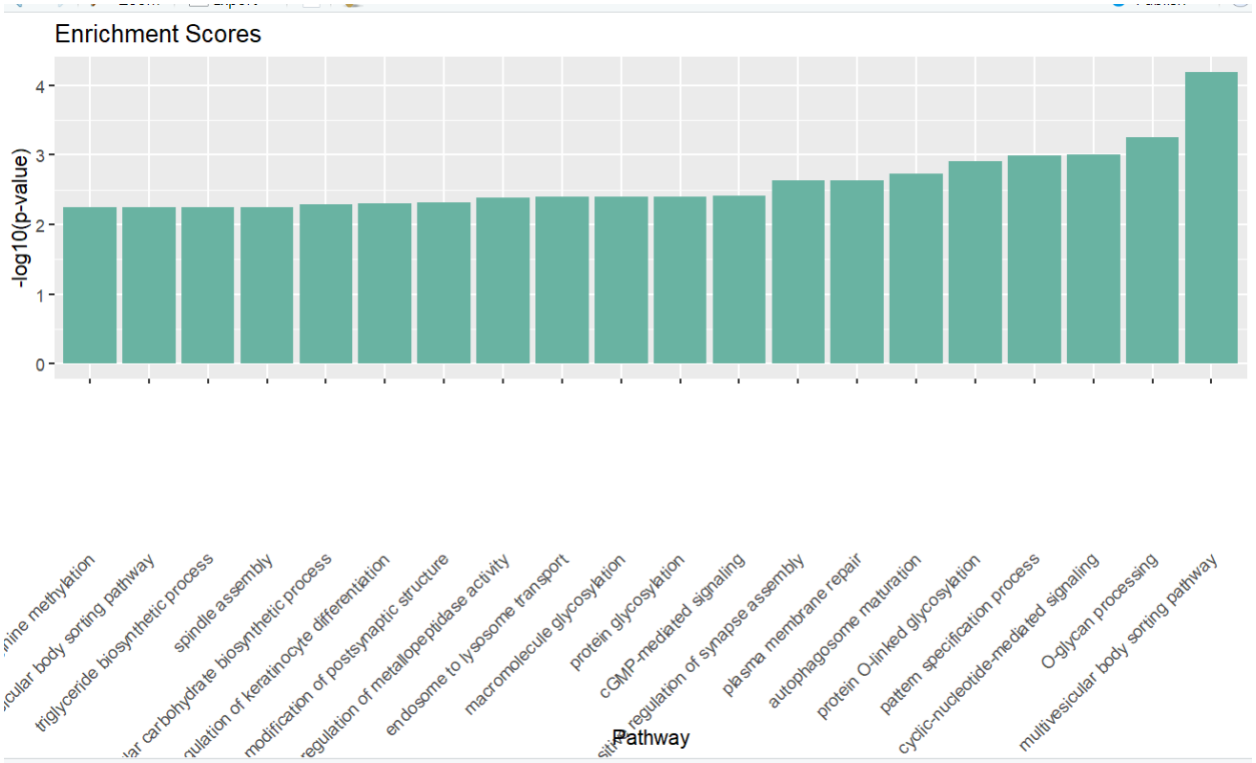
2. **OrgDb**: The organism-specific database that will be used to map the gene symbols to the corresponding gene identifiers. Here, we are using the "org.Hs.eg.db" database for human genes.
3. **keyType**: The type of gene identifiers used in the input gene vector. Here, we are using the "SYMBOL" type.
4. **ont**: The type of gene ontology that will be used for enrichment analysis. Here, we are using the "BP" (biological process) ontology.
5. **pAdjustMethod**: The method used for adjusting p-values for multiple testing. Here, we are using the "holm" method.
6. **pvalueCutoff**: The cutoff value for the p-value of enriched gene sets. Here, we have set it to 0.05.
7. **qvalueCutoff**: The cutoff value for the q-value of enriched gene sets. Here, we have set it to 0.1.
8. **readable**: A logical value indicating whether to include readable annotation in the output. Here, we have set it to TRUE.

Results of enrichment in various plots

Dot plot:-



Bar Plot:-



Snippets:-

ID	Description
GO:0071985 GO:0071985	multivesicular body sorting pathway
GO:0016266 GO:0016266	O-glycan processing
GO:0019935 GO:0019935	cyclic-nucleotide-mediated signaling
GO:0007389 GO:0007389	pattern specification process
GO:0006493 GO:0006493	protein O-linked glycosylation
GO:0097352 GO:0097352	autophagosome maturation
GO:0001778 GO:0001778	plasma membrane repair
GO:0051965 GO:0051965	positive regulation of synapse assembly
GO:0019934 GO:0019934	cGMP-mediated signaling
GO:0006486 GO:0006486	protein glycosylation
GO:0043413 GO:0043413	macromolecule glycosylation
GO:0008333 GO:0008333	endosome to lysosome transport
GO:1905050 GO:1905050	positive regulation of metallopeptidase activity
GO:0099010 GO:0099010	modification of postsynaptic structure
GO:0045616 GO:0045616	regulation of keratinocyte differentiation
GO:0034637 GO:0034637	cellular carbohydrate biosynthetic process
GO:0051225 GO:0051225	spindle assembly
GO:0019432 GO:0019432	triglyceride biosynthetic process
GO:0032509 GO:0032509	endosome transport via multivesicular body sorting pathway
GO:0034969 GO:0034969	histone arginine methylation
GO:0098935 GO:0098935	dendritic transport
GO:0070085 GO:0070085	glycosylation
GO:0014003 GO:0014003	oligodendrocyte development
GO:0090148 GO:0090148	membrane fission
GO:0046761 GO:0046761	viral budding from plasma membrane
GO:0072531 GO:0072531	pyrimidine-containing compound transmembrane transport
GO:0045017 GO:0045017	glycerolipid biosynthetic process
GO:0046753 GO:0046753	non-lytic viral release
GO:0009101 GO:0009101	glycoprotein biosynthetic process

	GeneRatio	BgRatio	pvalue	p.adjust	qvalue
GO:0071985	5/167	48/18903	6.383798e-05	0.1530196	0.1485745
GO:0016266	4/167	43/18903	5.543822e-04	1.0000000	0.4994615
GO:0019935	5/167	86/18903	9.916090e-04	1.0000000	0.4994615
GO:0007389	12/167	472/18903	1.020576e-03	1.0000000	0.4994615
GO:0006493	5/167	90/18903	1.216496e-03	1.0000000	0.4994615
GO:0097352	4/167	59/18903	1.832552e-03	1.0000000	0.4994615
GO:0001778	3/167	30/18903	2.307544e-03	1.0000000	0.4994615
GO:0051965	4/167	63/18903	2.333865e-03	1.0000000	0.4994615
GO:0019934	3/167	36/18903	3.904024e-03	1.0000000	0.4994615
GO:0006486	7/167	225/18903	3.916037e-03	1.0000000	0.4994615
GO:0043413	7/167	225/18903	3.916037e-03	1.0000000	0.4994615
GO:0008333	4/167	73/18903	3.983871e-03	1.0000000	0.4994615
GO:1905050	2/167	11/18903	4.049471e-03	1.0000000	0.4994615
GO:0099010	2/167	12/18903	4.831262e-03	1.0000000	0.4994615
GO:0045616	3/167	39/18903	4.901444e-03	1.0000000	0.4994615
GO:0034637	4/167	78/18903	5.046622e-03	1.0000000	0.4994615
GO:0051225	5/167	128/18903	5.604995e-03	1.0000000	0.4994615
GO:0019432	3/167	41/18903	5.644117e-03	1.0000000	0.4994615
GO:0032509	3/167	41/18903	5.644117e-03	1.0000000	0.4994615
GO:0034969	2/167	13/18903	5.676676e-03	1.0000000	0.4994615
GO:0098935	2/167	13/18903	5.676676e-03	1.0000000	0.4994615
GO:0070085	7/167	244/18903	6.059900e-03	1.0000000	0.4994615
GO:0014003	3/167	43/18903	6.450707e-03	1.0000000	0.4994615
GO:0090148	3/167	43/18903	6.450707e-03	1.0000000	0.4994615
GO:0046761	2/167	14/18903	6.584540e-03	1.0000000	0.4994615
GO:0072531	2/167	14/18903	6.584540e-03	1.0000000	0.4994615
GO:0045017	7/167	254/18903	7.490706e-03	1.0000000	0.4994615
GO:0046753	2/167	15/18903	7.553701e-03	1.0000000	0.4994615
GO:0009101	8/167	321/18903	7.832128e-03	1.0000000	0.4994615
GO:0010824	3/167	47/18903	8.260850e-03	1.0000000	0.4994615
GO:0006004	2/167	16/18903	8.583015e-03	1.0000000	0.4994615

	geneID	Count
GO:0071985	SYTL4/RAB27B/CHMP4B/VPS4B/RILP	5
GO:0016266	B3GNT7/POFUT1/GALNT11/B4GALT5	4
GO:0019935	PDE3A/ADGRG6/PRKG1/AHR/ADNP	5
GO:0007389	HOXC6/FOXD1/HOXB6/MIB1/POFUT1/VANGL2/MSX2/GALNT11/HOXC4/HHIP/SATB2/KIF3B	12
GO:0006493	DPM1/B3GNT7/POFUT1/GALNT11/B4GALT5	5
GO:0097352	MAP1LC3A/CHMP4B/VPS4B/EPG5	4
GO:0001778	SYTL4/CHMP4B/VPS4B	3
GO:0051965	FLRT3/PTPRD/SRPX2/ADNP	4
GO:0019934	PDE3A/PRKG1/ADNP	3
GO:0006486	ST6GAL2/DPM1/B3GNT7/FUT8/POFUT1/GALNT11/B4GALT5	7
GO:0043413	ST6GAL2/DPM1/B3GNT7/FUT8/POFUT1/GALNT11/B4GALT5	7
GO:0008333	CHMP4B/VPS4B/EPG5/RILP	4
GO:1905050	MBP/DDRGK1	2
GO:0099010	STAU1/WASF3	2
GO:0045616	ROCK2/MSX2/NCOA3	3
GO:0034637	B3GNT7/SORBS1/B4GALT5/PCK1	4
GO:0051225	MAPRE2/CHMP4B/VPS4B/MAPRE1/KIF3B	5
GO:0019432	FITM2/PNPLA3/PCK1	3
GO:0032509	CHMP4B/VPS4B/RILP	3
GO:0034969	PRMT5/NR1H4	2
GO:0098935	STAU1/KIF3B	2
GO:0070085	ST6GAL2/DPM1/B3GNT7/FUT8/POFUT1/GALNT11/B4GALT5	7
GO:0014003	WASF3/B4GALT5/EIF2B2	3
GO:0090148	SH3GLB1/CHMP4B/VPS4B	3
GO:0046761	CHMP4B/VPS4B	2

A9) The pathway analysis identified several enriched pathways in the input gene set. The top 15 pathways based on the lowest q-values are listed below:

1. Multivesicular body sorting pathway
2. O-glycan processing
3. Cyclic-nucleotide-mediated signaling
4. Pattern specification process
5. Protein O-linked glycosylation
6. Autophagosome maturation
7. Plasma membrane repair
8. Positive regulation of synapse assembly
9. cGMP-mediated signaling
10. Protein glycosylation
11. Macromolecule glycosylation
12. Endosome to lysosome transport
13. Positive regulation of metallopeptidase activity
14. Modification of postsynaptic structure
15. Regulation of keratinocyte differentiation

These pathways suggest potential biological processes and molecular mechanisms that could be involved in the studied condition or experimental system. For instance, the multivesicular body sorting pathway and the endosome-to-lysosome transport pathway are related to protein degradation and trafficking, which could be relevant to the clearance of misfolded or abnormal proteins. Overall, these pathways provide useful insights into the potential biological processes and molecular mechanisms that may be relevant to the input gene set and could guide further experimental investigations.

Snippets:-

```
$KEGG_2019_Human
                                                Term Overlap    P.value Adjusted.P.value Old.P.value Old.Adjusted.P.value
1                                N-Glycan biosynthesis    3/50 0.01243547        0.8503892           0                     0
2                 Other types of O-glycan biosynthesis    2/22 0.01885182        0.8503892           0                     0
3                                         Necroptosis   5/162 0.02059493        0.8503892           0                     0
4                               PPAR signaling pathway    3/74 0.03480413        0.8503892           0                     0
5                   Mucin type O-glycan biosynthesis    2/31 0.03586700        0.8503892           0                     0
6                               Pyruvate metabolism    2/39 0.05437927        0.8503892           0                     0
7                                       Ferroptosis    2/40 0.05688744        0.8503892           0                     0
8                 Porphyrin and chlorophyll metabolism    2/42 0.06202188        0.8503892           0                     0
9                       Glycosaminoglycan biosynthesis    2/53 0.09276028        0.8503892           0                     0
10                Taurine and hypotaurine metabolism    1/11 0.10119749        0.8503892           0                     0
11                Thyroid hormone signaling pathway   3/116 0.10211376        0.8503892           0                     0
12                          Fatty acid biosynthesis    1/13 0.11847053        0.8503892           0                     0
13                          FoxO signaling pathway   3/132 0.13553815        0.8503892           0                     0
14                Glycolysis / Gluconeogenesis    2/68 0.13995252        0.8503892           0                     0
15                           Pathways in cancer   8/530 0.14180752        0.8503892           0                     0
16              Adipocytokine signaling pathway    2/69 0.14326091        0.8503892           0                     0
17                    Insulin signaling pathway   3/137 0.14667150        0.8503892           0                     0
18                            Adherens junction    2/72 0.15328252        0.8503892           0                     0
```

	Odds.Ratio	Combined.Score	Genes
1	6.6382979	29.1235546	DPM1;FUT8;ST6GAL2
2	10.3596859	41.1398217	ST6GAL2;POFUT1
3	3.3287031	12.9243897	H2AFJ;VPS4B;CHMP4B;SPATA2;BIRC3
4	4.3890289	14.7384431	ACSL6;SORBS1;PCK1
5	7.1413613	23.7660042	GALNT11;B4GALT5
6	5.5950191	16.2914219	ME2;PCK1
7	5.4475062	15.6162611	MAP1LC3A;ACSL6
8	5.1746073	14.3867954	FECH;BLVRA
9	4.0562571	9.6447114	FUT8;B3GNT7
10	10.3109375	23.6190724	GGT7
11	2.7518398	6.2787842	THRB;NCOA3;PFKP
12	8.5915799	18.3266219	ACSL6
13	2.4085679	4.8135282	USP7;STK4;PCK1
14	3.1320006	6.1589291	PCK1;PFKP
15	1.5975976	3.1205628	DAPK1;PTGER1;ROCK2;GNG4;HHIP;NCOA3;STK4;BIRC3
16	3.0850981	5.9946162	ACSL6;PCK1
17	2.3181068	4.4497449	PRKAR2A;SORBS1;PCK1
18	2.9524308	5.5372028	SORBS1;WASF3
19	2.0019015	3.2593161	VANGL2;ROCK2;WIF1

```
 [ reached 'max' / getOption("max.print") -- omitted 35 rows ]

>
> summary(kegg_enrich)
                 Length Class      Mode
KEGG_2019_Human 9      data.frame list
>
```