A2_Sabbir_Hossain

BCB420 - Computational Systems Biology: Assignment 2 - Differential Gene Expression and Preliminary ORA

Sabbir Hossain

2022-04-16

Overview

The purpose of this assignment is to take the normalized expression data that was created in Assignment #1 and then rank genes according to differential expression. With that ranked list perform thresholded over-representation analysis(ORA) to highlight dominant genes/themes in the top set of genes.

The next two sections will deal with providing background information from the paper and previous assignments.

The link to the journal can be found at the end of the document, or by clicking the related subheading in the table of contents above.

Background information here, paraphased from the paper.

This was a really interesting paper, but I will save that discussion for the later part of this assignment. A prior transcriptome meta-analysis found significantly decreased levels of corticotropin-releasing hormone (CRH) mRNA in corticolimbic brain areas in MDD patients, indicating that cortical CRH-expressing (CRH+) cells are impaired in MDD. Although rodent studies reveal that cortical CRH is predominantly expressed in GABAergic interneurons, little is known about the characteristics of CRH+ cells in the human cerebral cortex and their relationship to MDD. Human volunteers without brain illnesses had their subgenual anterior cingulate cortex (sgACC) identified for CRH and markers of excitatory (SLC17A7), inhibitory (GAD1), and other interneuron subpopulations using fluorescent in situ hybridization (FISH) (PVALB, SST, VIP). Changes in CRH+ cell density and cellular CRH expression (n = 6/group) were investigated in MDD patients. RNA-sequencing was done on sgACC CRH+ interneurons from comparison and MDD participants (n = 6/group) to see if there were any variations between the two groups. In mice with TrkB function suppressed, the effect of decreased BDNF on CRH expression was investigated. GABAergic cells made up 80 percent of CRH+ cells, whereas glutamatergic cells made up 17.5 percent. VIP (52%) and SST (7%), as well as PVALB, were co-expressed by CRH+ GABAergic interneurons (7 percent). MDD patients had lower CRH mRNA levels in GABAergic interneurons than control participants, despite no differences in cell density. The transcriptome profile of CRH+ interneurons suggests decreased excitability and less GABA release and reuptake. Further research revealed that these molecular alterations are not caused by altered glucocorticoid feedback, but rather occur downstream of a common neurotrophic function modulator.

Essentially, there was a strong relationship between the gene expression or lack thereof for individuals who suffered from MDD.

Here is a direct link to the query for this dataset. (GSE193417 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE193417)).

Here is a direct link to the paper that is associated with the dataset above. (PMID: 35280164 (https://www.ncbi.nlm.nih.gov/pubmed/35280164)) (PMCID: PMC8913899 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8913899/))

Ideas, Interpretation, basic statistics and analysis from A1

In Assignment 1, we created log density graphs, MDS plots, MA plots after we chose a expression dataset from GEO database so that we may use them for analysis and processing as one would do in the field of bioinformatics. After selecting this expression set, we are to retrieve it, map it, normalize it, and finally interpret it, by use of graphs and plots as mentioned earlier. The conclusion that I came to in Assignment 1 was that there was in fact a strong relationship between the genes, their expressions and the results the authors of the paper had.

The dataset (GSE193417 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE193417)), started with 19961 genes for each of the 12

Samples; CRH-Hu1001 sgACC_MDD, CRH-Hu103 sgACC_control, CRH-Hu1047 sgACC_control, CRH-Hu1086 sgACC_control, CRH-Hu513 sgACC_MDD, CRH-Hu600 sgACC_MDD, CRH-Hu615 sgACC_control, CRH-Hu789 sgACC_control, CRH-Hu809 sgACC_MDD, CRH-Hu852 sgACC_control. Where sgACC_control groups are individuals have unimpaired CRH+ cells whereas MDD individuals have impaired CRH+ cells. After cleaning and normalizing the data, by removing genecounts that were fewer than 6 due to each sample size having 6 particpants each, the dataset had the ensemble_gene_ids mapped to HGNC symbols for easy gene identification afterwhich the genecounts were normalized and plotted. Assignment 1 removed 23.11% of the original genes, which left me with a final gene count of 15349. Which is slightly lower than the paper's genecount but that was most likely due to the authors using different cleaning and limitation methods. Which was my final data frame object , 'FinalGeneFilter'. What I found in the normalized dataset after plotting was that there was a causal relationship, the same conclusions as the author. For more information about this please see the Figure 1, an MDS below and Figure 2, the density distribution curve, both using the normalized dataset as mentioned before.

Setup

Set up all the data we used from Assignment 1. Reasoning for doing it this way…is because there could have been edits to the original dataset that is being used. Also, using Assignment as a child node for this Assignment is assumiming that every that was done, and the manner it was done in Assignment 1 was correct. In some cases this is not correct as my MDS plot, did not express the full numerical range that it could have. I have addressed that below in this assignment, by properly selecting the datatype and addressing some minor formatting issues in the data from Assignment 1 as well.

This section is responsible for downloading the dataset, cleaning the imported dataset, applying normalization and saving it along any milestone steps. Not all of these are required, but I did this as a form of redundancy in the event that some data accession problems occur, for instance the ensemble site and its varios mirrors not being accessible. Causing a significant delay and frustration in compilation as well as during analysis. This way having all the require files being downloaded and saved once makes for simple and quick accession and analysis.

Error Check

```
#check to see if the files are actually there. If not delete everything and try again.
if (!file.exists("GSE193417_normalized_datastruct.rds")) {
  print("The required files do not exist, one moment while all the data is being configured.
        If the notebook stops running, you might need to delete all the
        associated files with this notebook and run it from the begining.")
} else {
print("The required files exist, you may continue to run this R-Notebook")
}
```

```
## [1] "The required files exist, you may continue to run this R-Notebook"
```

Just a quick aside and sidenote here, for me the output will always print as "The required files exist, you may continue to run this R-Notebook", because I have all the required RDS configured and downloaded. However, for the first run it will print the error that the files don't exist, in which case you will have to delete everything and start over as the print statement suggests.

Differential Gene Expression

Construction of the Design Matrix

```r
samples <- data.frame(
  lapply(colnames(testFinalGeneFilter), FUN=function(x){
    unlist(strsplit(x, split = FALSE))[c(2,3)]
  }
))
colnames(samples) <- (colnames(testFinalGeneFilter)[3:14])
#removes the empty column
samples <- subset(samples, select = -c(13, 14))
rownames(samples) <- c("sample_expression", "sample_group")
samples[is.na(samples)] <- 0
#fill in the sample expression type, if they are sgACC or MDD
samples[1,] <- c("sgACC_MDD", "sgACC_control", "sgACC_control", "sgACC_control", "sgACC_MDD",   "sgACC_MDD",    "sgACC_control", "sgACC_control", "sgACC_MDD", "sgACC_control", "sgACC_MDD",    "sgACC_MDD")
#fill in the sample_cell type, if they are a control or not.
samples[2,] <- c("MDD","control","control","control","MDD","MDD","control","control","MDD","control","MDD","MDD")
samplesMat <- as.matrix(samples)
geneSymID <- testFinalGeneFilter[1:2]
expDesign <- testFinalGeneFilter
# create groups matrix:
expGroups <- as.data.frame(cbind(response = expDesign$header))
# Some experimental design information was incorrectly entered in the downloaded metadata file, so this information was manually entered based on paper figures:
expGroups <- as.data.frame(rbind(expGroups, samples))
designedd <- data.frame(lapply(colnames(normalized_counts_pmil), function(x) {
  gsub("\\d", "", unlist(strsplit(x, "_")))
}))
rownames(designedd) <- rownames(samples[1,])
rownames(designedd) <- rownames(samples[2,])
colnames(designedd) <- colnames(normalized_counts_pmil)
expDesign <- testFinalGeneFilter
expGroups <- as.data.frame(cbind(response = expDesign$header))
expGroups <- as.data.frame(rbind(expGroups, samples))
designedd <-expGroups
design2 <- data.frame(t(designedd))
knitr::kable(design2, type="html")
```

	sample_expression	sample_group
CRH-Hu1001	sgACC_MDD	MDD
CRH-Hu1031	sgACC_control	control
CRH-Hu1047	sgACC_control	control
CRH-Hu1086	sgACC_control	control
CRH-Hu513	sgACC_MDD	MDD
CRH-Hu600	sgACC_MDD	MDD
CRH-Hu615	sgACC_control	control
CRH-Hu789	sgACC_control	control
CRH-Hu809	sgACC_MDD	MDD
CRH-Hu852	sgACC_control	control
CRH-Hu863	sgACC_MDD	MDD
CRH-Hu943	sgACC_MDD	MDD

Table 1. Shows the design matrix for the samples, stating the sample’s expression type and its grouping. The controls are control individuals who do not have Major Depressive Disorder (MDD), while those labelled with MDD have MDD.

Corrected MDS plot

```
snames <- colnames(testFinalGeneFilter)[3:14]
d <- testFinalGeneFilter[3:14]
groupss <- interaction(snames, samplesMat["sample_group", ])

plotMDS(d, col=as.numeric(groupss), main="MDS plot of Norm. RNASeq Samples for GSE193417")
```

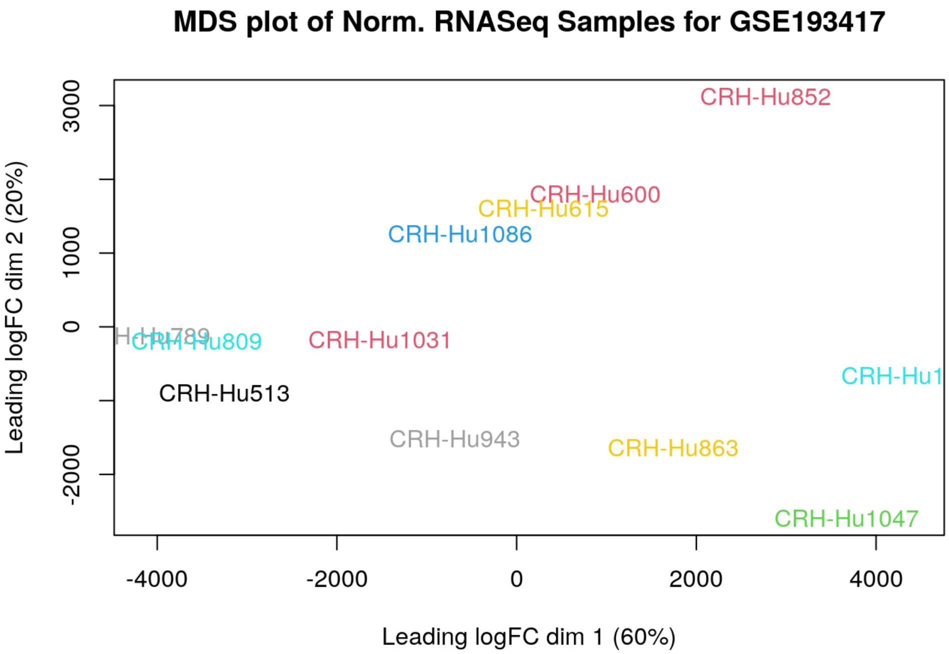


Figure 1. An MDS plot, i.e. multidimensional scaling points, which have already been accounted for with the correct log base due to the “edge” fixing, are a way of showing how the NR and R groups of any experiment interact with one another, as well as with other samples from the same organism, different organisms, different ages, and so on. There is over a half-fold difference between the two values, the first being the control group for this study, those who are not affected by Major Depressive Disorder (MDD), and the second being those who are affected by MDD. CRH+ cells in human sgACC are a diverse population of GABAergic interneurons, despite the fact that they predominantly co-express VIP. Our findings suggest that MDD is associated with decreased inhibitory function indicators in sgACC CRH+ interneurons, and they add to the growing body of evidence that MDD causes changes in GABAergic function in the cortex. We can see from the results that there is a lot of variation between controls and non-controls, but those who have MDD or don’t have MDD tend to cluster around the same area of the plot. This does imply that there are genes responsible for the responses observed by individuals and researchers. One thing that would be great to see in a plot like this is obviously a lot more samples, as 6 samples per group seems insufficient, as the researchers pointed out in their paper.

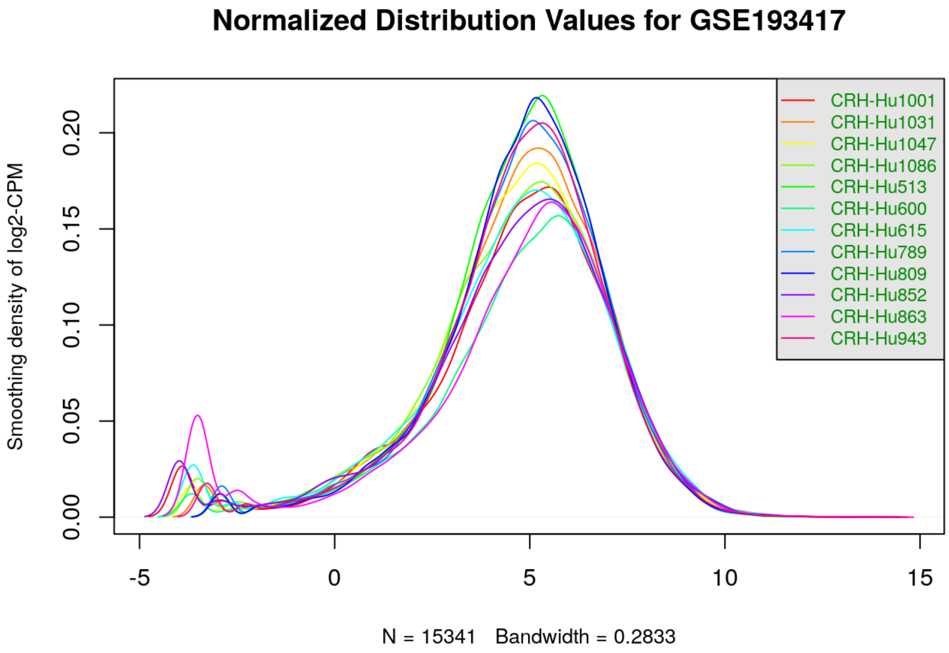
Density Curve for Normalized Distribution Values

```
ctsDenct <- apply(log2(edgeR::cpm(normalized_datastruct[,3:14])), 2, density)
dataplot2 <- c(colnames(samples))

xlim <- 0; ylim <- 0
for (i in 1:length(ctsDenct)) {
  xlim <- range(c(xlim, ctsDenct[[i]]$x));
  ylim <- range(c(ylim, ctsDenct[[i]]$y))
}
cols <- rainbow(length(ctsDenct))
ltys <- rep(1, length(ctsDenct))

plot(ctsDenct[[1]], xlim = xlim, ylim = ylim, type = "n", ylab = "Smoothing density of log2-CPM", main = "Normalized Dist
ribution Values for GSE193417", cex.lab = 0.85)
for (i in 1:length(ctsDenct))
  lines(ctsDenct[[i]], col = cols[i], lty = ltys[i])

legend("topright", legend = dataplot2, col=cols, lty=ltys, cex=0.75, border="blue", text.col = "green4", merge = TRUE, bg
= "gray90")
```



```
#Remember to add in the figure name, number and description.
```

Figure 2. A normalized distribution curve The density curves are another way to analyse our data before and after the normalisation process; because they are proportional to the size of the data set, they work particularly well with larger data sets. The more data we can add to this curve, the better we'll be able to predict and model it. While outlines and data that do not match the average are more than likely to affect the extreme ends of the curve, we can still deduce that those who suffer from MDD will find someone with whom they can comfortably align themselves. The more data we can add to this curve, the better we'll be able to predict and model it. The higher curves for the youngest and oldest of the group in this data set and these samples specifically. While there are a few points where the difference in the variable values mapped against the curve falls short of the non-normalized graph, the normalised graph lacks extra values, repetitions are the most likely culprit because they count for twice as much as they should.

Normalized Data Strucutre

```
knitr::kable(t(normalized_datastruct[1:5,]), type="html",)
```

	1	2	3	4	5
ensembl_gene_id	ENSG00000000003	ENSG00000000419	ENSG00000000457	ENSG00000000460	ENSG00000000971
hgnc_symbol	TSPAN6	DPM1	SCYL3	C1orf112	CFH
CRH-Hu1001	0.000000	9.575111	0.000000	0.000000	0.000000
CRH-Hu1031	10.657725	14.498347	46.951601	3.168513	0.000000
CRH-Hu1047	0.6299757	0.0000000	15.2769115	4.0160953	11.1820693
CRH-Hu1086	0.00000000	0.08828009	31.16287262	2.64840277	0.17656018
CRH-Hu513	15.561174	5.227582	45.954092	18.965181	1.945147
CRH-Hu600	11.3993735	0.0000000	5.4588549	0.1605546	0.0000000
CRH-Hu615	0.000000	18.436284	58.255289	15.994949	1.346943
CRH-Hu789	8.393148	0.000000	25.688120	37.260491	24.670769
CRH-Hu809	9.2910323	8.1761084	0.1238804	21.5551949	6.1940215
CRH-Hu852	0.0000000	4.6425067	0.8500364	13.1428712	0.0000000
CRH-Hu863	5.491062	41.787916	31.829548	0.000000	1.489102
CRH-Hu943	0.00000	15.83270	22.83562	21.82071	27.80871

Table 2. A table that shows the completed normalized data structure for this dataset, with matching ensemble gene IDs, and HGNC symbols for each of the samples. This data structure was one of the intended final products of Assignment 1. The object is called normalized_datastruct.

Differential Gene Expression on Normalized Dataset

```r
design_model <- model.matrix(~ design2$sample_expression)
minimal_set <- ExpressionSet(normalized_counts_pmil)
fit <- limma::lmFit(minimal_set, design_model)
fit2 <- limma::eBayes(fit, trend=TRUE)
topfit <- limma::topTable(fit2, coef=ncol(design_model), adjust.method="BH", number=nrow(normalized_counts_pmil))
# Add the gene symbols
output_hits <- merge(
  data.frame(gene = geneSymID, row.names = rownames(normalized_counts_pmil)),
  topfit,
  by=0,
  all=TRUE
)
# Sort by unadjusted p-value
output_hits <- output_hits[order(output_hits$P.Value, decreasing = FALSE),]
output_hits
```

	Row.names <I<chr>>	gene.ensembl_gene_id <chr>	gene.hgnc_symbol <chr>	logFC <dbl>	AveExpr <dbl>	▸
3017	12823	ENSG00000183087	GAS6	-85.690230	9.514635e+01	
10023	5068	ENSG00000125633	CCDC93	-69.007096	1.167344e+02	
7304	2568	ENSG00000103260	METRN	-28.325229	1.749028e+01	
13061	7882	ENSG00000147133	TAF1	-81.657048	1.485376e+02	
14016	8761	ENSG00000156076	WIF1	-50.018930	4.572789e+01	
2509	12340	ENSG00000178950	GAK	-65.472325	1.309363e+02	
11764	6681	ENSG00000138029	HADHB	42.269884	2.762715e+01	
12123	7018	ENSG00000140386	SCAPER	85.231951	1.950236e+02	
8938	407	ENSG00000023902	PLEKHO1	-63.376206	1.008727e+02	
11154	6114	ENSG00000134644	PUM1	78.423039	2.191940e+02	

1-10 of 10,000 rows | 1-6 of 10 columns Previous **1** 2 3 4 5 6 ... 1000 Next

Table 3. A table that shows the first 10 genes with the lowest P-values of the dataset. This table is ordered, by the P.value attribute and starts with the lowest P.value and goes to the largest P.value.

Volcano Plot

```
vol_plot <- output_hits

vol_plot$diffexpressed <- "Not Significant"
# if logFC > 1 and P.Value < 0.05, set as "Up Regulated"
vol_plot$diffexpressed[vol_plot$logFC > 1 & vol_plot$P.Value < 0.05] <- "Up Regulated"
# if logFC < -1 and P.Value < 0.05, set as "Down Regulated"
vol_plot$diffexpressed[vol_plot$logFC < 1 & vol_plot$P.Value < 0.05] <- "Down Regulated"

# Create a new column "delabel" to vol_plot, that will contain the name of genes differentially expressed (NA in case the
y are not)
vol_plot$delabel <- NA
vol_plot$delabel[vol_plot$diffexpressed != "Not Significant"] <- vol_plot$gene.hgnc_symbol[vol_plot$diffexpressed != "Not
Significant"]

ggplot(data=vol_plot, aes(x=logFC, y=-log10(P.Value), col=diffexpressed, label=delabel)) +
        geom_point() +
        theme_minimal() +
        geom_text_repel() +
        scale_color_manual(values=c("blue", "black", "purple")) +
        geom_vline(xintercept=c(-1, 1), col="red") +
        geom_hline(yintercept=-log10(0.05), col="red") +
        labs(title="All differentially expressed genes with a p-value < 0.05 for GSE193417")
```

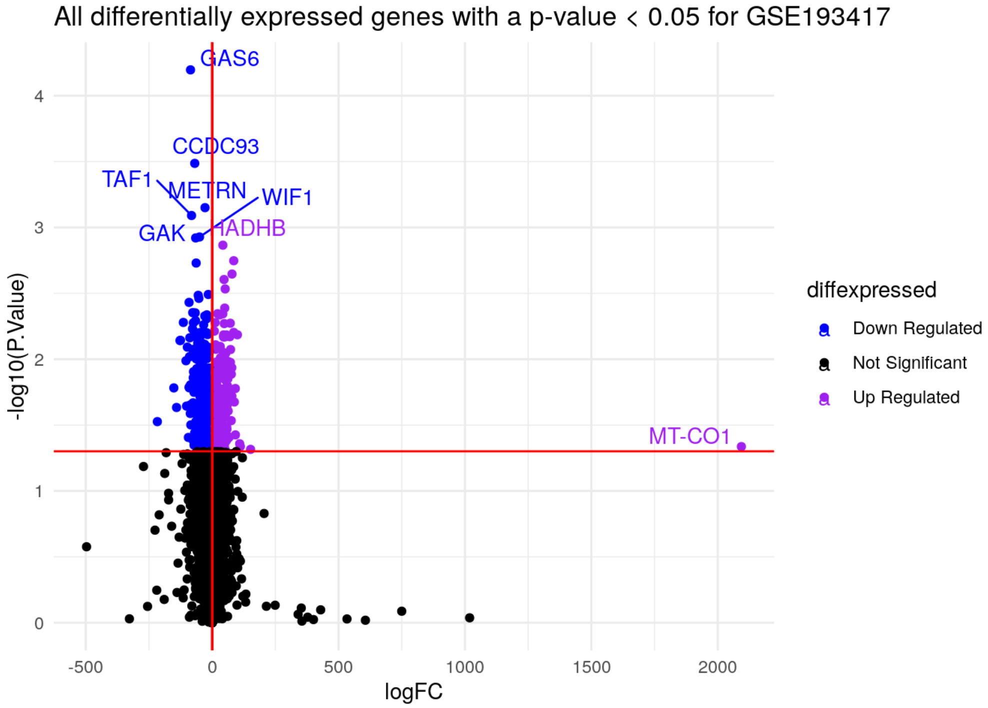


Figure 3. A volcano plot that shows the single upregulated gene, MT-CO1 (https://www.uniprot.org/uniprot/P00395) and the 6 downregulated genes GAS6 (https://www.uniprot.org/uniprot/Q14393), CCDC93 (https://www.uniprot.org/uniprot/Q567U6), METRN (https://www.uniprot.org/uniprot/Q9UJH8), TAF1 (https://www.uniprot.org/uniprot/P21675), GAK (https://www.uniprot.org/uniprot/O14976), WIF1 (https://www.uniprot.org/uniprot/Q9Y5W5).

MA plot

```
limma::plotMA(minimal_set, status=output_hits$P.Value < 0.05, main="Differentially expressed genes with a p-value < 0.05
for GSE193417")
```

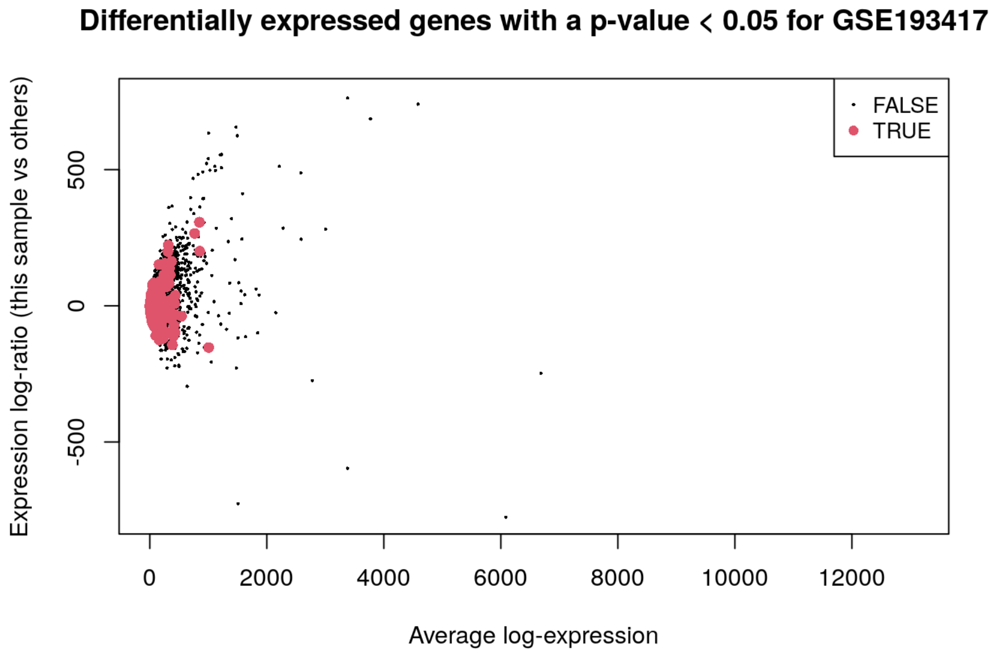


Figure 4. Matches up with the volc plot, they are just graphical translations of each other. The points, and spread all match up.

Heatmap

```
#Get the rownames of each of the genes that satisfy the threshold.
topHitHM <- output_hits$Row.names[output_hits$P.Value < 0.05]
#Get the sections of the data that are also in the specified threshold.
heatmap_matrix_tophits <- t(scale(t(normalized_counts_pmil[which(rownames(normalized_counts_pmil) %in% topHitHM),])))
#
heatmap_col <- circlize::colorRamp2(c(min(heatmap_matrix_tophits), 0, max(heatmap_matrix_tophits)), c("blue", "white", "red"))

heatmap <- ComplexHeatmap::Heatmap(as.matrix(heatmap_matrix_tophits), name = "gene expr. value",
  cluster_rows = TRUE, show_row_dend = TRUE,
  cluster_columns = TRUE,show_column_dend = TRUE,
  col=heatmap_col,show_column_names = TRUE,
  show_row_names = FALSE,show_heatmap_legend = TRUE,
  column_title = "Heatmap of diff. expressed genes with a p-value < 0.05 for GSE193417")
heatmap
```

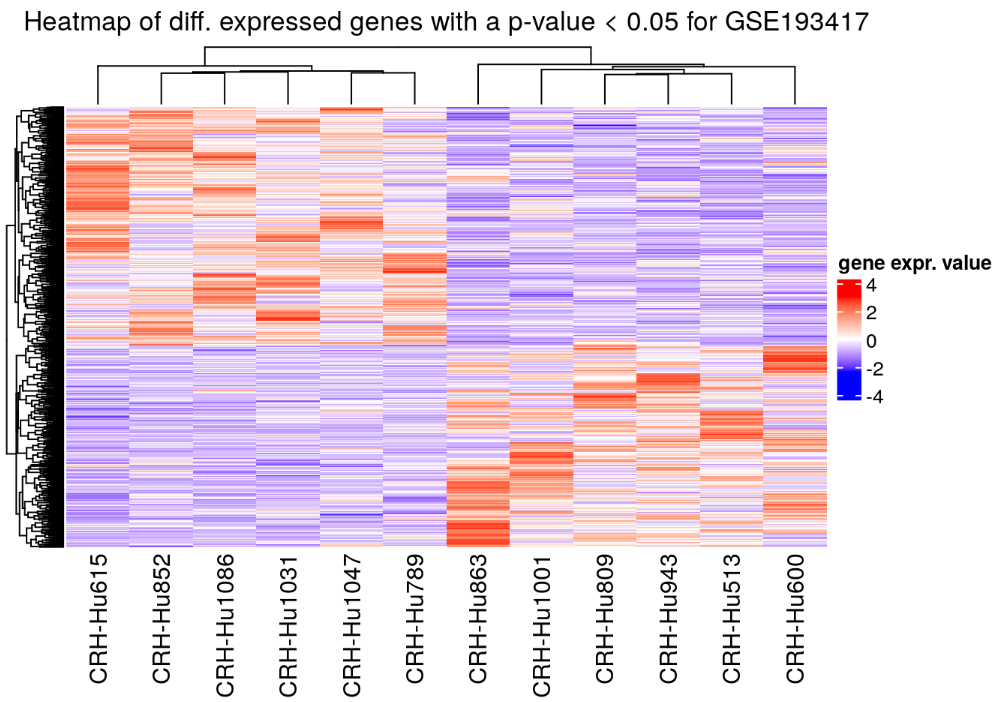


Figure 5 is a heatmap of differentially expressed genes within the dataset, with the samples being represented along the bottom of the heatmap. Good clustering and consistent clustering between the samples. There is consistent banding and clustering of colours as well. If you divide the heatmap horizontally, right along the 0 of the heatmap, you can separate it into two halves for every sample, a top half that has a significant colour palette and a bottom half with another different signficant colour palette. For the most part, each half of the sample is dominated by either purple-ish or orange-ish colour. For example, CRH-Hu1001 is a sample that suffers with MDD, the top half is primarily purple, whilst the bottom half is primarily orange. This pattern where the top is purple and the bottom is orange is consistent across all MDD samples, and for controls, like CRH-Hu1031 the top half is primarily orange, whilst the bottom half is primarily purple.

Answers to Questions

1. a. Calculate p-values for each of the genes in your expression set. b) How many genes were significantly differentially expressed? c) What thresholds did you use and why?

a. For the p-values for each of the genes in the expression set, please see Table 3.
b. 608 or 3.9632358% of the 15341 normalized and cleaned genes.
c. I did not use any thresholds to limit the data in this portion. My reasoning for doing this is because based on the number of genes that passed the significant differential expression already being under the P-value threshold of 5%, I found it unnecessary. It doesn't make sense to further limit the values and data when it already is under the range to reject the null hypothesis. Meaning that there are genes with strong differential expression, and in this case it is 608 of them. Now, we can adjust for any irregularities and perform some hypothesis corrections. Though to clarify, I am making my P.Value threshold of 0.05, as this is 5%, enough to reject the null hpothesis, signifying strong differential expression as I stated before, but beyond this threshold, I did not have another as it was not required as stated earlier. So, threshold was P.Value < 0.05, but nothing further.

2. Multiple hypothesis testing - correct your p-values using a multiple hypothesis correction method.

a. Which method did you use? And Why? b) How many genes passed correction?
b. The method I used was the Benjamini-Hochberg (BH) correction method. This is mainly due to the fact that it is the most widely used of the methods, it works really well for smaller sample sizes, which is the case for us, with data set, and lastly it is the most lenient of the P-value adjustment methods. Since the sample size is so small, we might not have much variation. Though, due to this method, 0 genes passes correction, which is interesting to say the least.
c. The number of genes that passed correction is: 0.

3. Show the amount of differentially expressed genes using an MA Plot or a Volcano plot. Highlight genes of interest. For this question please see Figure 4 and Figure 5 above, an MA Plot and Volcano plot respectively. Please pay special note to the Volcano plot, where it highlights the genes; downregulated, GAS6, CCDC93, METRN, TAF1, GAK, WIF1 and up regulated, MT-CO1 data points. This is really interesting because these "stars" perse are much further out that the rest of the cluster of the dataset. In the MA plot, you can see the outliers simply signifed as dots, as the MA plot and Volcano plot are graph translations, graphed on the anti-axis of one another. You can see the very similar grouping structure around the origin of each graph. In the MA plot, the scatter is more spread out due to the larger scale resulting from the change base, of log2 to log10, and the average of said log expressions. Both essentially represent the same thing, that there are a major number of genes that are expressed with a P-value of less than 0.05. For both plot, False or insignificant values are represented in black colour. The Volcano plot is just differentiating between up or down regulatiion for the True values. With

this in mind, and the redundancy further supporting my claims, we can conclude that there is most likely gene strong differential gene expression.

4. Visualize your top hits using a heatmap. Do you conditions cluster together? Explain why or why not. Yes there is a clear clustering due to the conditions of the experiment, and sample type. Those that have MDD and those that are controls have very similar gene expressions. Please see Figure 5's description for more information. In summation of the description, those with MDD have a unique colour pattern on the heatmap, and individuals with said pattern can be identified that way (purple-ish top, orange-ish bottom). Similarly, individuals that do not suffer from MDD have a unique pattern purple-ish bottom, orange-ish top. What this basically states is that there is a clear relationship between the genes and their effects. Pretty interesting that there is also a visual similarity. However, I should note that this might just be a coincidence and like stated in the paper as well as in A1, a larger sample size could have made the relationship more or less apparent. Still a pretty neat thing! So in short, the purple means there is downregulated expression and the orange means there is upregulated expression for the respective genes.

Thresholded over-representation analysis

Create lists of allreupregulated, downregulated genes:

```
all_up_ens <- output_hits$gene.ensembl_gene_id[output_hits$P.Value <= 0.05 & output_hits$logFC >= 0]
all_up_hgnc <- output_hits$gene.hgnc_symbol[output_hits$P.Value <= 0.05 & output_hits$logFC >= 0]
all_down_ensg <- output_hits$gene.ensembl_gene_id[output_hits$P.Value <= 0.05 & output_hits$logFC <= 0]
all_down_hgnc <- output_hits$gene.hgnc_symbol[output_hits$P.Value <= 0.05 & output_hits$logFC <= 0]
write.table(all_up_ens, file = "all_up_ens.txt", sep="\n", row.names = FALSE,col.names = FALSE,quote = FALSE)
write.table(all_up_hgnc, file = "all_up_hgnc.txt", sep="\n", row.names = FALSE,col.names = FALSE,quote = FALSE)
write.table(all_down_ensg, file = "all_down_ensg.txt", sep="\n", row.names = FALSE,col.names = FALSE,quote = FALSE)
write.table(all_down_hgnc, file = "all_down_hgnc.txt", sep="\n", row.names = FALSE,col.names = FALSE,quote = FALSE)
all_up <- append(all_up_ens, all_up_hgnc)
all_down <- append(all_down_ensg, all_down_hgnc)
allvalues <- append(all_up_hgnc, all_down_hgnc)
allvaluesnamed <- append(all_up, all_down)
```

g:profiler for ORA differentually expressed genes:

```
gostrestp1 <- gost(allvaluesnamed, sources=c("GO:BP", "GO:MF", "GO:CC", "KEGG", "HPA", "HP", "REAC"))
gostplot(gostrestp1, capped = FALSE, interactive = TRUE)
```

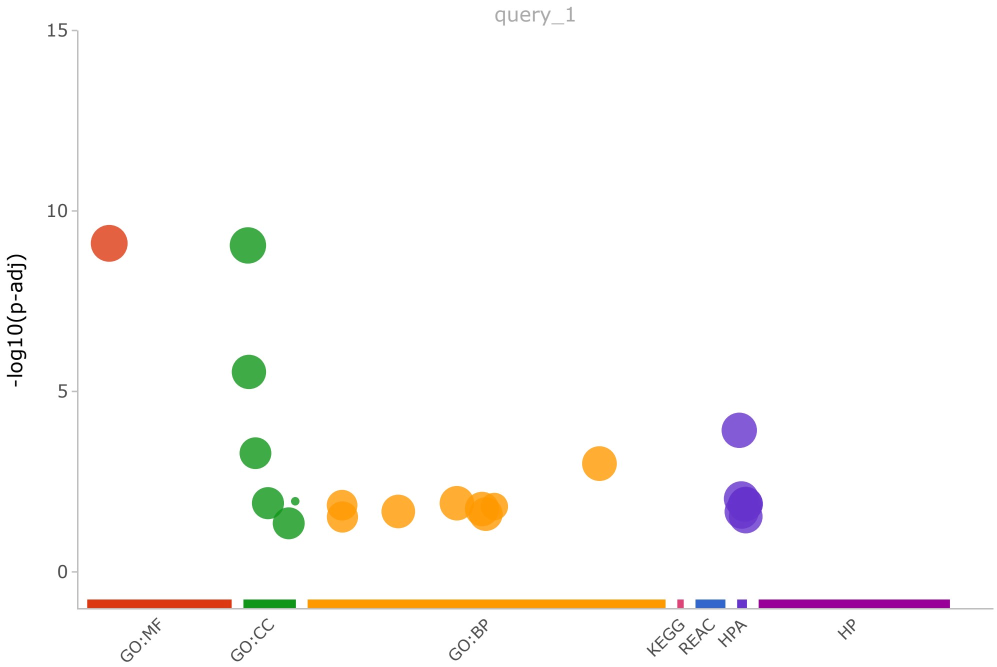


Figure 6A: Manhattan Plot of ORA results for all differentially expressed genes

Figure 6A is a Manhattan Plot what shows the hits and genes from the different annotation databases. This includes both downregulated and upregulated genes for this dataset. The large numbers of hits from HPA, GO:BP (GO:BP), GO:CC (GO:CC). GO:MF (GO:MF) make sense because there are a number of genes that account for different processes, i.e. molecular function genes, cellular components, human proteins etc. and the fact that we have returns from these annotations databases make sense. We can verify this by just looking at some of the hits from our volcano plot earlier and see that many of the genes that are regulated do indeed fall under the functions and databases they have pinged from. Here is a list of them for reference; MT-CO1 (https://www.uniprot.org/uniprot/P00395) GAS6 (https://www.uniprot.org/uniprot/Q14393), CCDC93 (https://www.uniprot.org/uniprot/Q567U6), METRN (https://www.uniprot.org/uniprot/Q9UJH8), TAF1 (https://www.uniprot.org/uniprot/P21675), GAK (https://www.uniprot.org/uniprot/O14976), WIF1 (https://www.uniprot.org/uniprot/Q9Y5W5).

Here is the link to gprofiler ((https://biit.cs.ut.ee/gprofiler/gost?
organism=hsapiens&query=HADHB%0ASCAPER%0APUM1%0ACEP41%0APCGF5%0APPP1R21%0AIGFBP6%0AHMGXB4%0APDLIM7%0ABRD
DPA1%0AGSTA4%0ACCR9%0ARASSF2%0APHKG1%0ASMOC2%0ASKA2%0ACTNNA3%0ARPSA%0AATP5F1B%0ARGMA%0AKLHDC1%0ASC
CO1%0AOSBPL9%0AHAX1%0ANOP10%0AAIP%0AGPATCH2%0AHCN1%0AMSR1%0ATMEM126B%0ASLC19A3%0ALSM14B%0AZBTB25%0AE
1%0ANOMO2%0APARG%0ARAB6C%0AARSD1%0ACDH13%0ASELENBP1%0AABHD14A%0AALDH3B1%0AGCM2%0AELP3%0AGAS6%0ACC
1%0AUBQLN4%0AABCF1%0AZNF592%0AC6orf141%0AMESP1%0ANUDT18%0AMAN2C1%0AERO1A%0APOU6F1%0AARL4C%0ARAP2C%0AC
which shows my query for the entire ordered gene list. Which looks almost identical, in terms of spread, and hits. The settings and query information can be found in their respective tabs by following this link. This link is for references to make sure that what I am doing is logical.

g:profiler for ORA differentually up regulated expressed genes:

```
gostrestp1Up <- gost(all_up, sources=c("GO:BP", "GO:MF", "GO:CC", "KEGG", "HPA", "HP", "REAC"))
gostplot(gostrestp1Up, capped = FALSE, interactive = TRUE)
```

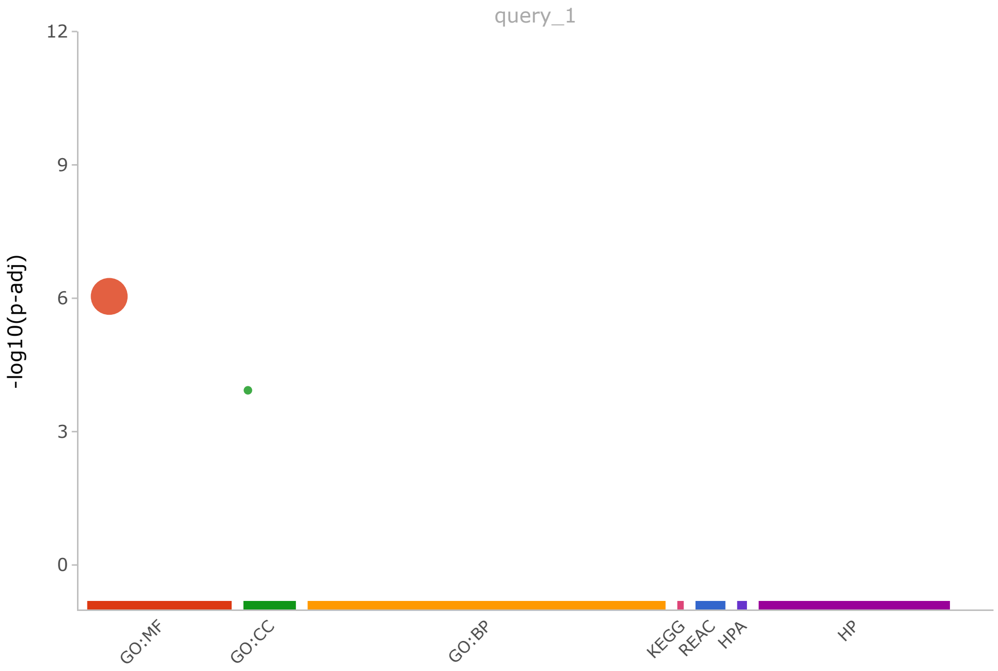


Figure 6B: Manhattan Plot of ORA results for all up-regulated differentially expressed genes

Figure 6B is a Manhattan Plot what shows the hits and genes from the different annotation databases. This includes only upregulated genes for this dataset. The only his we get back are from GO:MF (GO:MF) and GO:CC (GO:CC) annotation databases. When looking up the gene hMT-CO1 (https://www.uniprot.org/uniprot/P00395) on uniprot we can see under the Functions heading and the GO subheadings that it indeed is a gene that encodes for a structure that is responsible for molecular and is indeed a cellular component since it is found in the mitochondria. Which is very similar to the result that I got running this input on gprofiler ((https://biit.cs.ut.ee/gprofiler/gost?
organism=hsapiens&query=HADHB%0ASCAPER%0APUM1%0ACEP41%0APCGF5%0APPP1R21%0AIGFBP6%0AHMGXB4%0APDLIM7%0ABRD
DPA1%0AGSTA4%0ACCR9%0ARASSF2%0APHKG1%0ASMOC2%0ASKA2%0ACTNNA3%0ARPSA%0AATP5F1B%0ARGMA%0AKLHDC1%0ASC
CO1%0AOSBPL9%0AHAX1%0ANOP10%0AAIP%0AGPATCH2%0AHCN1%0AMSR1%0ATMEM126B%0ASLC19A3%0ALSM14B%0AZBTB25%0AE
1%0ANOMO2%0APARG%0ARAB6C%0AARSD1%0ACDH13%0ASELENBP1%0AABHD14A%0AALDH3B1%0AGCM2%0AELP3&ordered=true&all_
with the settings and ideas from lecture. These can be found by navigating to the “Query Info tab”. This link is for references to make sure that what I am doing is logical.

g:profiler for ORA differentually down regulated expressed genes:

```
gostrestp1Down <- gost(all_down, sources=c("GO:BP", "GO:MF", "GO:CC", "KEGG", "HPA", "HP", "REAC"))
gostplot(gostrestp1Down, capped = FALSE, interactive = TRUE)
```

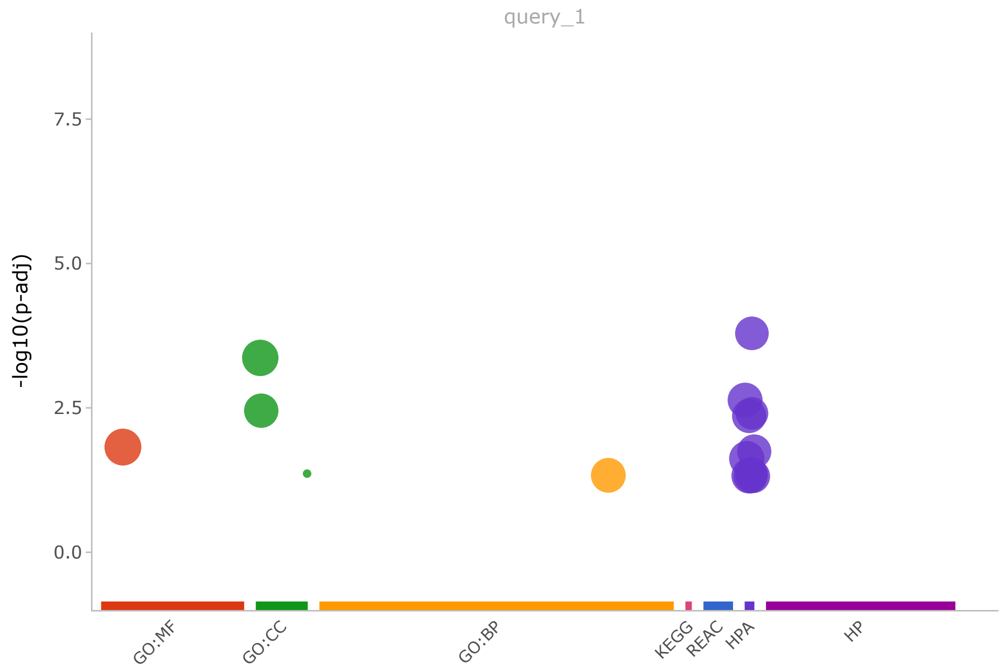


Figure 6A: Manhattan Plot of ORA results for all down-regulated differentially expressed genes

Figure 6C is a Manhattan Plot what shows the hits and genes from the different annotation databases. This includes only downregulated genes for this dataset. Here we can see a large number of the hits from HPA, GO:BP (GO:BP), GO:CC (GO:CC). GO:MF (GO:MF). This also has a similar database spread, although the number of hits is not the same, the goprofiler website had different options, we are only using the genes of interest. They are using the entire negative differentially expressed list. The query can be found here ((https://biit.cs.ut.ee/gprofiler/gost?
organism=hsapiens&query=GAS6%0ACCDC93%0AMETRN%0ATAF1%0AWIF1%0AGAK%0APLEKHO1%0ASNRPB%0AMAPK4%0APACSIN2%0A
1%0AUBQLN4%0AABCF1%0AZNF592%0AC6orf141%0AMESP1%0ANUDT18%0AMAN2C1%0AERO1A%0APOU6F1%0AARL4C%0ARAP2C%0A
The settings and information about the query can be found by navigating to the “Query info” tab.

Answers to Questions

With your significantly up-regulated and down-regulated set of genes run a thresholded gene set enrichment analysis:

1. Which method did you choose and why? I chose to use Over-Representation Analysis (ORA). I am using this method because it is the one I most familiar with for the gene set enrichment analysis. It is one of the most widely used and simple (Huang et al., 2009). For this assignment and my ORAs, I chose to use the gprofiler2 R package. I found g:Profiler very convenient to use and was somewhat familiar with it from BCH441. Also g:profiler is extremely robust and thorough, it allows you to be very creative in how you choose your values, or whatever else you need. Additionally, I used it during my last assignment and for plotting it seems to be the most widely used and recommended. One of the best features in my opinion is to quickly and easily compare a number of genes to a number of different databases in one expression/line of code. There are also other customization that you can do regarding the plots you make as well.

2. a. What annotation data did you use and why? b) What version of the annotation are you using?

a. The annotation sources used for this analysis of this dataset in this assignment are as follows; GO: Molecular Functions, GO: Cellular Components, GO: Biological Pathways, and KEGG, HPA + HP (Human Protean Atlas). All of the data that was acquired was from human cells, so it makes sense that the sources would also reference and be human, mainly the HPA database, since a vast majority of the cells are human proteins. The other hits on the other databases also make sense as in the experiment as the authors were looking at cells that are related to function and inhibition of different molecular structures, cellular components and biologicall pathways as the affected cells behave different from normal cells. Addiontionally, a lot of the annotation data was mentioned in the paper by the authors, for example GO:BP (GO:BP), and GO:MF (GO:MF).
b. The version of annotation data is associated and updated based on g:Profiler, with the most recent gprofiler2 package version being 0.2.1. This is the version that I am using. You can verify this by looking at the packages tab and scrolling or searching for gprofiler2 and looking for the version number.

3. a. How many genesets were returned with what thresholds? The order of these answers will coincide with the order of the plots for consistency sake.

i. The number for all differentially expressed genes: 608 .
ii. The number for all differentially upregulated genes: 280 .
iii. The number for all differentially downregulated genes: 328 . The thresholds for all these was the same value that I used in my entire assignment, a P-Value < 0.05, this allowed for genes that had a strong correlation with the experiment to be selected and further processed.

4. Run the analysis using the up-regulated set of genes, and the down-regulated set of genes separately. How do these results compare to using the whole list (i.e all differentially expressed genes together vs. the up-regulated and down regulated differentially expressed genes separately)? In the analysis of just the up-regulated set of genes, 2 genepathways were plotted, but I had 560 upregulated genes. One was from GO:MF (GO:MF) and the other was from GO:CC (GO:CC). For the down-regulated genes, there were 328 genesets returned. I had genesets from; GO:MF (GO:MF), GO:CC (GO:CC), GO:BP (GO:BP), and a few between HPA and HP. For all the differentially expressed genes, I had genesets from; GO:MF (GO:MF), GO:CC (GO:CC), GO:BP (GO:BP), as well as some plotted between HPA and HP. Which makes sense as if I am using the full list for differentially expressed genes, I should have results that entail both the upregulated and down regulated geneset lists. The results as a whole make sense as in the paper there was use and mention of all these databases. As stated in the background of this paper, CRH+ cells are impaired in persons with MDD, and not in the control group. Since there are a number of pathways, functions and cellular components that are related to cortisol, it would make sense that we would see hits on these databases for genes that are related to these functions. I would also like to note, that due to it being related to proteins, and how proteins are imperative in biological function we see a lot of the database hits there too. I included the links to the website for gprofiler for each of the respective queries because it is a means of verifying that the steps taken thus far are correct. Since, we have not manually filtered the values, as I only used their gene names, the thresholds and limitations don’t do much. But the trend and overal dispertion of the results are very similar, which is why I wanted to include them.

Interpretation and Discussion

The most important aspect of the analysis is relating your results back to the initial data and question.

1. Do the over-representation results support conclusions or mechanism discussed in the original paper?

Yes they completely support the conclusions in the original paper by the authors,(Oh, Hyunjung et al., 2022) because they themselves also performed a differential expression analysis of the genes with similar results. They had an total differential expression of 835 genes, whereas my total was 608. The paper mentions how they used ClueGo, a Cytoscape plugin. The authors of the paper identified 307 genes showing significant group differences 168 upregulated, 139 downregulated and gene sets were significantly altered in those with MDD, 528 gene sets, 267 upregulated, 261 downregulated. These are around the same ballpark as the number of genes that I had 608 for all differentially expressed genes, 328 for all down regulated differentially expressed genes, 280 for all up regulated differentially expressed genes. Which is close to the total upregulated total of 435 and the downregulated total of 400 for the authors. Now, there is a slight variation but I think that makes sense as the authors had a different method for limiting and choosing genes, different from how we were doing it in this assignment. The authors of the paper used Gene Set Enrichment Analysis (GSEA) with whole transcriptomic data to identify altered biological pathways in these interneurons. Additionally as mentioned in Assignment 1, the authors of the paper also had more genes than I did at the end of my analysis. I would like to point out again that they did not remove any outliers, or used a completely different method for doing so, but I cannot find that in the paper as it is not mentioned. Not in the same context that I was looking for anyway.”

2. Can you find evidence, i.e. publications, to support some of the results that you see. How does this evidence support your results.

Link between genes and depression have been found out in the recent years. There are many papers that support the notion of gene expression being tied to the regulation and well being of mental health, it makes sense thinking about it from a purely scientific perspective as well. For instance, everything our body does or can do, is coded by for DNA, all the chemistry, behaviours, interactions between cellular processes are due to genes and their interations. So it makes sense that brain chemistry and dependence would also be determined by genes and their differential expression or lack thereof. Here are a few examples of publications that can support the idea that was presented in this paper by (Hyunjung et. al., 2022). A paper titled “Reduced brain somatostatin in mood disorders: a common pathophysiological substrate and drug target?” (Lin et. al. 2013) deteremined that Somaticaostatin is a neuropeptide that acts as an inhibitory modulator and it is crucial in cortical local inhibitory circuit abnormalities that result in aberrant corticolimbic network activity. This shows that somatostatin-expressing neurons have a unique cellular vulnerability. This paper and the literature I am using dealt with cortisol regulation effects. So it is pretty easy to say that this evidence directly supports my results, since both papers were trying to determine corticol effects on differential expression. Here is a summary and paraphrase of their abstract; an inhibitory modulatory neuropeptide called somaticaostatin may play a key role in cortical local inhibitory circuit abnormalities that contribute to aberrant corticolimbic network activity and clinical mood symptoms in a variety of neurological illnesses. Although our understanding of the biology of affect dysregulation has improved over time, pharmaceutical therapies remain insufficient. We focus on direct data from the postmortem brain of humans and examine rodent genetic and pharmacological research that investigate the function of the somatostatin system in mood. Another paper titled “Reduced glial cell density and neuronal size in the anterior cingulate cortex in major depressive disorder” by (Cotter et. al., 2001) concluded that there is reduced frontal cortical glial cell density and neuronal size in major depressive disorder, the decreased cell density was due to gene expression of certain cells like glyceraldehyde phosphate dehydrogenase’s messenger RNA levels or subgenual anterior cingulate cortex (ACC) having reduced gilial cell density, due to glutamatergic pathways to subcortical structures having regulation issues. Here is a summary of this paper’s ideas; Glial cells make up more than half of all brain cells and are assumed to play a key role in many nervous system activities. Glial cell dysfunction is thought to have a role in the pathogenesis of major mental diseases including MDD and schizophrenia. Lastly the paper titled “Altered expression of genes involved in inflammation and apoptosis in frontal cortex in major depression” by (Shelton et. al., 2011) which revealed that depressed people have higher levels of inflammatory and apoptotic stress in BA10, including elevated levels of particular cytokines and anti-apoptotic proteins. Although the exact causes of these anomalies are unknown, oxidative stress has been linked to them. Clearly, further study is needed to confirm these findings and examine causative pathways in more depth. To summarize their abstract and paraphrase their abstract; the cause of major depression (MDD) is unknown. Local inflammatory, apoptotic, and oxidative stress are seen in post-mortem brain tissue samples in MDD patients. A gene set analysis revealed that a number of pro- and anti-inflammatory cytokines were upregulated. Metallothionein 1M (MT1M), a zinc binding protein involved in the control of oxidative stress, was one of the genes with lower expression. So, to reiterate, there is a strong relationship between some genes and their differential expression and pathways being incorrectly regulated. This can lead to individuals suffering from MDD or other illnesses that affect the quality of life. Both my assignment, and paper have concluded that there is a strong differential expression in these 12 samples with some of the more major genes being highlighted in the plots and figures above.

Link to Journal

Click Me! (https://github.com/bcb420-2022/Sabbir_Hossain/wiki/Journal-Entry-Assignment-%232:--Differential-Gene-expression-and-Preliminary-ORA)

Citations

1. Bonnin, S. (2022). 19.11 Volcano plots | Introduction to R. Biocorecrg.github.io. Retrieved 13 April 2022, from https://biocorecrg.github.io/CRG_RIntroduction/volcano-plots.html (https://biocorecrg.github.io/CRG_RIntroduction/volcano-plots.html).
2. Cotter, D., Mackay, D., Landau, S., Kerwin, R., & Everall, I. (2001). Reduced Glial Cell Density and Neuronal Size in the Anterior Cingulate Cortex in Major Depressive Disorder. Archives Of General Psychiatry, 58(6), 545. https://doi.org/10.1001/archpsyc.58.6.545 (https://doi.org/10.1001/archpsyc.58.6.545)
3. Differential Expression with Limma-Voom. Ucdavis-bioinformatics-training.github.io. (2022). Retrieved 13 April 2022, from https://ucdavis-bioinformatics-training.github.io/2018-June-RNA-Seq-Workshop/thursday/DE.html (https://ucdavis-bioinformatics-training.github.io/2018-June-RNA-Seq-Workshop/thursday/DE.html).
4. Duan, E. (2022). R|Py notes: Volcano plots with ggplot2. R|Py notes. Retrieved 13 April 2022, from https://erikaduan.github.io/posts/2021-01-02-volcano-plots-with-ggplot2/ (https://erikaduan.github.io/posts/2021-01-02-volcano-plots-with-ggplot2/).
5. Falcon, S., & Gentleman, R. (2006). Using GOstats to test gene lists for GO term association. Bioinformatics, 23(2), 257-258. https://doi.org/10.1093/bioinformatics/btl567 (https://doi.org/10.1093/bioinformatics/btl567)
6. Geistlinger, L., Csaba, G., Santarelli, M., Ramos, M., Schiffer, L., Turaga, N., Law, C., Davis, S., Carey, V., Morgan, M., Zimmer, R., & Waldron, L. (2021). Toward a gold standard for benchmarking gene set enrichment analysis. Briefings in bioinformatics, 22(1), 545–556. https://doi.org/10.1093/bib/bbz158 (https://doi.org/10.1093/bib/bbz158)
7. Huang, D., Sherman, B., & Lempicki, R. (2008). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Research, 37(1), 1-13. https://doi.org/10.1093/nar/gkn923 (https://doi.org/10.1093/nar/gkn923)
8. Lin, L., & Sibille, E. (2013). Reduced brain somatostatin in mood disorders: a common pathophysiological substrate and drug target?. Frontiers In Pharmacology, 4. https://doi.org/10.3389/fphar.2013.00110 (https://doi.org/10.3389/fphar.2013.00110)
9. Oh, H., Newton, D., Lewis, D., & Sibille, E. (2022). Lower Levels of GABAergic Function Markers in Corticotropin-Releasing Hormone-Expressing Neurons in the sgACC of Human Subjects With Depression. Frontiers In Psychiatry, 13. https://doi.org/10.3389/fpsyt.2022.827972 (https://doi.org/10.3389/fpsyt.2022.827972)
10. Peng, R. (2022). R Programming for Data Science. Bookdown.org. Retrieved 13 April 2022, from https://bookdown.org/rdpeng/rprogdatascience/ (https://bookdown.org/rdpeng/rprogdatascience/).
11. Shelton, R., Claiborne, J., Sidoryk-Wegrzynowicz, M., Reddy, R., Aschner, M., Lewis, D., & Mirnics, K. (2010). Altered expression of genes involved in inflammation and apoptosis in frontal cortex in major depression. Molecular Psychiatry, 16(7), 751-762. https://doi.org/10.1038/mp.2010.52 (https://doi.org/10.1038/mp.2010.52)
12. Steipe, B., & Isserlin, R. (2022). BCB420 - Computational System Biology. Bcb420-2022.github.io. Retrieved 13 April 2022, from https://bcb420-2022.github.io/General_course_prep/index.html#attributions (https://bcb420-2022.github.io/General_course_prep/index.html#attributions).
13. Steipe, B., & Isserlin, R. (2022). BCB420 - Computational System Biology. Bcb420-2022.github.io. Retrieved 13 April 2022, from https://bcb420-2022.github.io/R_basics/ (https://bcb420-2022.github.io/R_basics/).
14. Steipe, B., & Isserlin, R. (2022). BCB420 - Computational System Biology. Bcb420-2022.github.io. Retrieved 13 April 2022, from https://bcb420-2022.github.io/Bioinfo_Basics/ (https://bcb420-2022.github.io/Bioinfo_Basics/).
15. Lecture modules: https://q.utoronto.ca/courses/248455/modules (https://q.utoronto.ca/courses/248455/modules)
16. Uku Raudvere, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, Jaak Vilo: g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update) Nucleic Acids Research 2019; doi:10.1093/nar/gkz369 (doi:10.1093/nar/gkz369) [PDF].

```
citation("tidyverse")
```

```
##
##   Wickham et al., (2019). Welcome to the tidyverse. Journal of Open
##   Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
##
## A BibTeX entry for LaTeX users is
##
##   @Article{,
##     title = {Welcome to the {tidyverse}},
##     author = {Hadley Wickham and Mara Averick and Jennifer Bryan and Winston Chang and Lucy D'Agostino McGowan and Rom
ain François and Garrett Grolemund and Alex Hayes and Lionel Henry and Jim Hester and Max Kuhn and Thomas Lin Pedersen an
d Evan Miller and Stephan Milton Bache and Kirill Müller and Jeroen Ooms and David Robinson and Dana Paige Seidel and Vit
alie Spinu and Kohske Takahashi and Davis Vaughan and Claus Wilke and Kara Woo and Hiroaki Yutani},
##     year = {2019},
##     journal = {Journal of Open Source Software},
##     volume = {4},
##     number = {43},
##     pages = {1686},
##     doi = {10.21105/joss.01686},
##   }
```

```
citation("edgeR")
```

```
##
## See Section 1.2 in the User's Guide for more detail about how to cite
## the different edgeR pipelines.
##
##   Robinson MD, McCarthy DJ and Smyth GK (2010). edgeR: a Bioconductor
##   package for differential expression analysis of digital gene
##   expression data. Bioinformatics 26, 139-140
##
##   McCarthy DJ, Chen Y and Smyth GK (2012). Differential expression
##   analysis of multifactor RNA-Seq experiments with respect to
##   biological variation. Nucleic Acids Research 40, 4288-4297
##
##   Chen Y, Lun ATL, Smyth GK (2016). From reads to genes to pathways:
##   differential expression analysis of RNA-Seq experiments using
##   Rsubread and the edgeR quasi-likelihood pipeline. F1000Research 5,
##   1438
##
## To see these entries in BibTeX format, use 'print(<citation>,
## bibtex=TRUE)', 'toBibtex(.)', or set
## 'options(citation.bibtex.max=999)'.
```

```
citation("GEOmetadb")
```

```
##
## Please cite the following if utilizing the GEOmetadb software:
##
##   Zhu Y, Davis S, Stephens R, Meltzer PS, Chen Y. GEOmetadb: powerful
##   alternative search engine for the Gene Expression Omnibus.
##   Bioinformatics. 2008 Dec 1;24(23):2798-800. doi:
##   10.1093/bioinformatics/btn520. Epub 2008 Oct 7. PubMed PMID:
##   18842599; PubMed Central PMCID: PMC2639278.
##
## A BibTeX entry for LaTeX users is
##
##   @Article{,
##     author = {Yuelin Zhu and Sean Davis and Robert Stephens and Paul S. Meltzer and Yidong Chen},
##     title = {GEOmetadb: powerful alternative search engine for the Gene Expression Omnibus.},
##     journal = {Bioinformatics (Oxford, England)},
##     year = {2008},
##     month = {Dec},
##     day = {01},
##     volume = {24},
##     number = {23},
##     pages = {2798--2800},
##     abstract = {The NCBI Gene Expression Omnibus (GEO) represents the largest public repository of microarray data. Ho
wever, finding data in GEO can be challenging. We have developed GEOmetadb in an attempt to make querying the GEO metadat
a both easier and more powerful. All GEO metadata records as well as the relationships between them are parsed and stored
in a local MySQL database. A powerful, flexible web search interface with several convenient utilities provides query cap
abilities not available via NCBI tools. In addition, a Bioconductor package, GEOmetadb that utilizes a SQLite export of t
he entire GEOmetadb database is also available, rendering the entire GEO database accessible with full power of SQL-based
queries from within R.},
##     issn = {1367-4811},
##     doi = {10.1093/bioinformatics/btn520},
##     url = {http://www.ncbi.nlm.nih.gov/pubmed/18842599},
##     language = {eng},
##   }
```

```
citation("RColorBrewer")
```

```
##
## To cite package 'RColorBrewer' in publications use:
##
##   Erich Neuwirth (2022). RColorBrewer: ColorBrewer Palettes. R package
##   version 1.1-3.
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {RColorBrewer: ColorBrewer Palettes},
##     author = {Erich Neuwirth},
##     year = {2022},
##     note = {R package version 1.1-3},
##   }
```

```
citation("ggplot2")
```

```
##
## To cite ggplot2 in publications, please use:
##
##   H. Wickham. ggplot2: Elegant Graphics for Data Analysis.
##   Springer-Verlag New York, 2016.
##
## A BibTeX entry for LaTeX users is
##
##   @Book{,
##     author = {Hadley Wickham},
##     title = {ggplot2: Elegant Graphics for Data Analysis},
##     publisher = {Springer-Verlag New York},
##     year = {2016},
##     isbn = {978-3-319-24277-4},
##     url = {https://ggplot2.tidyverse.org},
##   }
```

```
citation("readxl")
```

```
##
## To cite package 'readxl' in publications use:
##
##   Hadley Wickham and Jennifer Bryan (2022). readxl: Read Excel Files.
##   https://readxl.tidyverse.org, https://github.com/tidyverse/readxl.
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {readxl: Read Excel Files},
##     author = {Hadley Wickham and Jennifer Bryan},
##     year = {2022},
##     note = {https://readxl.tidyverse.org, https://github.com/tidyverse/readxl},
##   }
```

```
citation("dplyr")
```

```
##
## To cite package 'dplyr' in publications use:
##
##   Hadley Wickham, Romain François, Lionel Henry and Kirill Müller
##   (2022). dplyr: A Grammar of Data Manipulation.
##   https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr.
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {dplyr: A Grammar of Data Manipulation},
##     author = {Hadley Wickham and Romain François and Lionel Henry and Kirill Müller},
##     year = {2022},
##     note = {https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr},
##   }
```

```
citation("AnnotationDbi")
```

```
##
## To cite package 'AnnotationDbi' in publications use:
##
##   Hervé Pagès, Marc Carlson, Seth Falcon and Nianhua Li (2021).
##   AnnotationDbi: Manipulation of SQLite-based annotations in
##   Bioconductor. R package version 1.56.2.
##   https://bioconductor.org/packages/AnnotationDbi
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor},
##     author = {Hervé Pagès and Marc Carlson and Seth Falcon and Nianhua Li},
##     year = {2021},
##     note = {R package version 1.56.2},
##     url = {https://bioconductor.org/packages/AnnotationDbi},
##   }
##
## ATTENTION: This citation information has been auto-generated from the
## package DESCRIPTION file and may need manual editing, see
## 'help("citation")'.
```

```
citation("limma")
```

```
##
## Please cite the paper below for the limma software itself.  Please also
## try to cite the appropriate methodology articles that describe the
## statistical methods implemented in limma, depending on which limma
## functions you are using.  The methodology articles are listed in
## Section 2.1 of the limma User's Guide.
##
##   Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and
##   Smyth, G.K. (2015). limma powers differential expression analyses for
##   RNA-sequencing and microarray studies. Nucleic Acids Research 43(7),
##   e47.
##
## A BibTeX entry for LaTeX users is
##
##   @Article{,
##     author = {Matthew E Ritchie and Belinda Phipson and Di Wu and Yifang Hu and Charity W Law and Wei Shi and Gordon K
Smyth},
##     title = {{limma} powers differential expression analyses for {RNA}-sequencing and microarray studies},
##     journal = {Nucleic Acids Research},
##     year = {2015},
##     volume = {43},
##     number = {7},
##     pages = {e47},
##     doi = {10.1093/nar/gkv007},
##   }
```

```
citation("Biobase")
```

```
##
##   Orchestrating high-throughput genomic analysis with Bioconductor. W.
##   Huber, V.J. Carey, R. Gentleman, ..., M. Morgan Nature Methods,
##   2015:12, 115.
##
## A BibTeX entry for LaTeX users is
##
##   @Article{,
##     author = {W. Huber and V. J. Carey and R. Gentleman and S. Anders and M. Carlson and B. S. Carvalho and H. C. Brav
o and S. Davis and L. Gatto and T. Girke and R. Gottardo and F. Hahne and K. D. Hansen and R. A. Irizarry and M. Lawrence
and M. I. Love and J. MacDonald and V. Obenchain and A. K. {Ole's} and H. {Pag`es} and A. Reyes and P. Shannon and G. K.
Smyth and D. Tenenbaum and L. Waldron and M. Morgan},
##     title = {{O}rchestrating high-throughput genomic analysis with {B}ioconductor},
##     journal = {Nature Methods},
##     year = {2015},
##     volume = {12},
##     number = {2},
##     pages = {115--121},
##     url = {http://www.nature.com/nmeth/journal/v12/n2/full/nmeth.3252.html},
##   }
```

```
citation("BiocManager")
```

```
##
## To cite package 'BiocManager' in publications use:
##
##   Martin Morgan (2021). BiocManager: Access the Bioconductor Project
##   Package Repository. R package version 1.30.16.
##   https://CRAN.R-project.org/package=BiocManager
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {BiocManager: Access the Bioconductor Project Package Repository},
##     author = {Martin Morgan},
##     year = {2021},
##     note = {R package version 1.30.16},
##     url = {https://CRAN.R-project.org/package=BiocManager},
##   }
```

```
citation("biomaRt")
```

```
##
## To cite the biomaRt package in publications use:
##
##   Mapping identifiers for the integration of genomic datasets with the
##   R/Bioconductor package biomaRt. Steffen Durinck, Paul T. Spellman,
##   Ewan Birney and Wolfgang Huber, Nature Protocols 4, 1184-1191 (2009).
##
##   BioMart and Bioconductor: a powerful link between biological
##   databases and microarray data analysis. Steffen Durinck, Yves Moreau,
##   Arek Kasprzyk, Sean Davis, Bart De Moor, Alvis Brazma and Wolfgang
##   Huber, Bioinformatics 21, 3439-3440 (2005).
##
## To see these entries in BibTeX format, use 'print(<citation>,
## bibtex=TRUE)', 'toBibtex(.)', or set
## 'options(citation.bibtex.max=999)'.
```

```
citation("magrittr")
```

```
##
## To cite package 'magrittr' in publications use:
##
##   Stefan Milton Bache and Hadley Wickham (2022). magrittr: A
##   Forward-Pipe Operator for R. https://magrittr.tidyverse.org,
##   https://github.com/tidyverse/magrittr.
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {magrittr: A Forward-Pipe Operator for R},
##     author = {Stefan Milton Bache and Hadley Wickham},
##     year = {2022},
##     note = {https://magrittr.tidyverse.org,
## https://github.com/tidyverse/magrittr},
##   }
```

```
citation("GEOquery")
```

```
##
## Please cite the following if utilizing the GEOquery software:
##
##   Davis, S. and Meltzer, P. S. GEOquery: a bridge between the Gene
##   Expression Omnibus (GEO) and BioConductor. Bioinformatics, 2007, 14,
##   1846-1847
##
## A BibTeX entry for LaTeX users is
##
##   @Article{,
##     author = {Sean Davis and Paul Meltzer},
##     title = {GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor},
##     journal = {Bioinformatics},
##     year = {2007},
##     volume = {14},
##     pages = {1846--1847},
##   }
```

```
citation("RSQLite")
```

```
##
## To cite package 'RSQLite' in publications use:
##
##   Kirill Müller, Hadley Wickham, David A. James and Seth Falcon (2022).
##   RSQLite: SQLite Interface for R. https://rsqlite.r-dbi.org,
##   https://github.com/r-dbi/RSQLite.
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {RSQLite: SQLite Interface for R},
##     author = {Kirill Müller and Hadley Wickham and David A. James and Seth Falcon},
##     year = {2022},
##     note = {https://rsqlite.r-dbi.org, https://github.com/r-dbi/RSQLite},
##   }
```

```
citation("limma")
```

```
##
## Please cite the paper below for the limma software itself.  Please also
## try to cite the appropriate methodology articles that describe the
## statistical methods implemented in limma, depending on which limma
## functions you are using.  The methodology articles are listed in
## Section 2.1 of the limma User's Guide.
##
##   Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and
##   Smyth, G.K. (2015). limma powers differential expression analyses for
##   RNA-sequencing and microarray studies. Nucleic Acids Research 43(7),
##   e47.
##
## A BibTeX entry for LaTeX users is
##
##   @Article{,
##     author = {Matthew E Ritchie and Belinda Phipson and Di Wu and Yifang Hu and Charity W Law and Wei Shi and Gordon K
Smyth},
##     title = {{limma} powers differential expression analyses for {RNA}-sequencing and microarray studies},
##     journal = {Nucleic Acids Research},
##     year = {2015},
##     volume = {43},
##     number = {7},
##     pages = {e47},
##     doi = {10.1093/nar/gkv007},
##   }
```

```
citation("org.Hs.eg.db")
```

```
##
## To cite package 'org.Hs.eg.db' in publications use:
##
##   Marc Carlson (2021). org.Hs.eg.db: Genome wide annotation for Human.
##   R package version 3.14.0.
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {org.Hs.eg.db: Genome wide annotation for Human},
##     author = {Marc Carlson},
##     year = {2021},
##     note = {R package version 3.14.0},
##   }
##
## ATTENTION: This citation information has been auto-generated from the
## package DESCRIPTION file and may need manual editing, see
## 'help("citation")'.
```

```
citation('umap')
```

```
##
## To cite package 'umap' in publications use:
##
##   Tomasz Konopka (2022). umap: Uniform Manifold Approximation and
##   Projection. R package version 0.2.8.0.
##   https://github.com/tkonopka/umap
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {umap: Uniform Manifold Approximation and Projection},
##     author = {Tomasz Konopka},
##     year = {2022},
##     note = {R package version 0.2.8.0},
##     url = {https://github.com/tkonopka/umap},
##   }
```

```
citation("vegan")
```

```
##
## To cite package 'vegan' in publications use:
##
##   Jari Oksanen, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt,
##   Pierre Legendre, Dan McGlinn, Peter R. Minchin, R. B. O'Hara, Gavin
##   L. Simpson, Peter Solymos, M. Henry H. Stevens, Eduard Szoecs and
##   Helene Wagner (2020). vegan: Community Ecology Package.
##   https://cran.r-project.org, https://github.com/vegandevs/vegan.
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {vegan: Community Ecology Package},
##     author = {Jari Oksanen and F. Guillaume Blanchet and Michael Friendly and Roeland Kindt and Pierre Legendre and Da
n McGlinn and Peter R. Minchin and R. B. O'Hara and Gavin L. Simpson and Peter Solymos and M. Henry H. Stevens and Eduard
Szoecs and Helene Wagner},
##     year = {2020},
##     note = {https://cran.r-project.org, https://github.com/vegandevs/vegan},
##   }
##
## ATTENTION: This citation information has been auto-generated from the
## package DESCRIPTION file and may need manual editing, see
## 'help("citation")'.
```

```
citation('gprofiler2')
```

```
##
## To cite gprofiler2 in publications, please use:
##
## Kolberg L, Raudvere U, Kuzmin I, Vilo J, Peterson H (2020).
## "gprofiler2- an R package for gene list functional enrichment analysis
## and namespace conversion toolset g:Profiler." _F1000Research_, *9
## (ELIXIR)*(709). R package version 0.2.1.
##
## A BibTeX entry for LaTeX users is
##
##   @Article{,
##     title = {gprofiler2-- an R package for gene list functional enrichment analysis and namespace conversion toolset
g:Profiler},
##     journal = {F1000Research},
##     author = {Liis Kolberg and Uku Raudvere and Ivan Kuzmin and Jaak Vilo and Hedi Peterson},
##     volume = {9 (ELIXIR)},
##     number = {709},
##     year = {2020},
##     note = {R package version 0.2.1},
##   }
```

```
citation('ggrepel')
```

```
##
## To cite package 'ggrepel' in publications use:
##
##   Kamil Slowikowski (2021). ggrepel: Automatically Position
##   Non-Overlapping Text Labels with 'ggplot2'. R package version 0.9.1.
##   https://github.com/slowkow/ggrepel
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {ggrepel: Automatically Position Non-Overlapping Text Labels with
## 'ggplot2'},
##     author = {Kamil Slowikowski},
##     year = {2021},
##     note = {R package version 0.9.1},
##     url = {https://github.com/slowkow/ggrepel},
##   }
```