R Team Project

HEATMAP

Presented by

Polatat Suwanit 6736156
Sorawan Tiratrakoonwichaya 6737929
Chakrit Jiarasatit 6737934

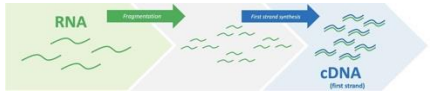
Heatmap: code

- Go to

You can scan this QR code to run code with us!!

https://github.com/Polatat/Heatmap_project

Workflow



RNA sequencing

mapped to the
human hg38 genome and
SARS-CoV-2 reference genome

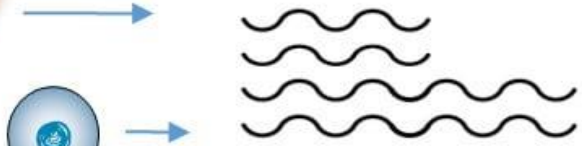
identify **differentially expressed genes** (DEGs) between infected and mock-infected samples

log2 fold change
(log2FC) values and
adjusted p-values (FDR)

default cutoff of P-adj < 0.05

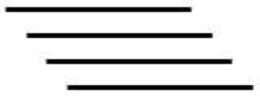
RNA sequencing

1. RNA isolation



2. RNA selection

3. cDNA library preparation



4. Fragmentation, size selection

5. Adapter ligation

6. Validation, quantification

7. Sequencing, analysis

```
ATGCAGCGACGACGAGCACG
CGACTACGACTCGACTCACG
GCACGACGAGTGTTGAGCAG
GGCAGCGTGCAGCGACGGAC
```

Single-cell RNA-Seq
scRNA-Seq analyzes gene expression of individual cells. Single cells are isolated prior to RNA extraction.

Total RNA-Seq
Whole transcriptome analysis omits selection in order to sequence coding and non-coding RNA

mRNA-Seq
mRNA can be enriched via poly-A selection or ribosomal RNA depletion

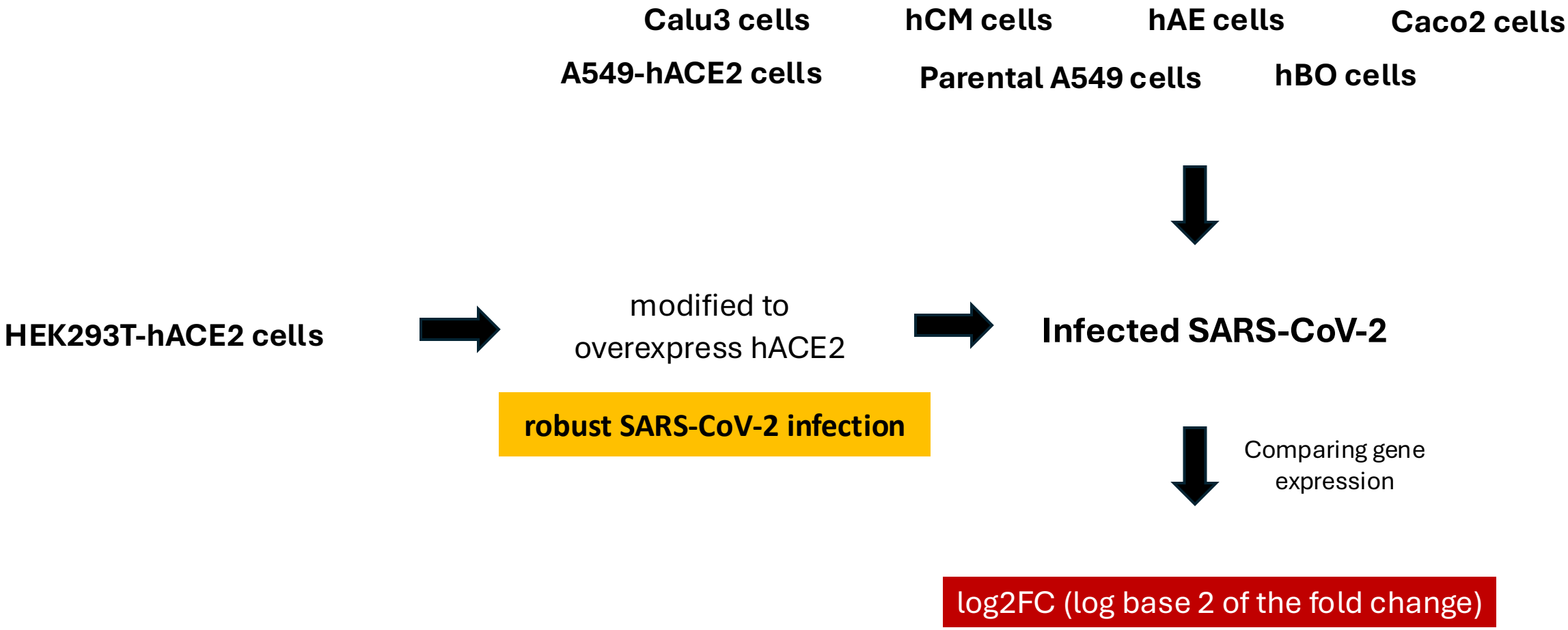
Small RNA-Seq
Size selection methods can be used to enrich for noncoding RNA such as miRNA

Targeted RNA-Seq
Specific RNA sequences can be selected using hybridization probes

Note: Details of the workflow can vary depending on the chosen sequencing platform.

Stimulated cells vs. Control cells

- HEK293T-hACE2 are HEK293T cells that have been genetically modified to overexpress hACE2 (human angiotensin-converting enzyme 2) to facilitate better infection by the SARS-CoV-2 virus.



log base 2 of the fold change

- log2FC is a measure used in the differential gene expression analysis performed by the DEseq2 package between the two conditions.

$$\text{Fold Change} = \frac{\text{Expression in Condition A}}{\text{Expression in Condition B}}$$

$$\log_2\text{Fold Change} = \log_2\left(\frac{\text{Expression in Condition A}}{\text{Expression in Condition B}}\right)$$

Fold Change (FC)

- Represents how much the expression of a gene has changed between the two conditions.

For example, the expression level of this gene is:

- In Condition A: **100 reads**
- In Condition B: **25 reads**

Fold Change (A/B) = 100 / 25 = 4

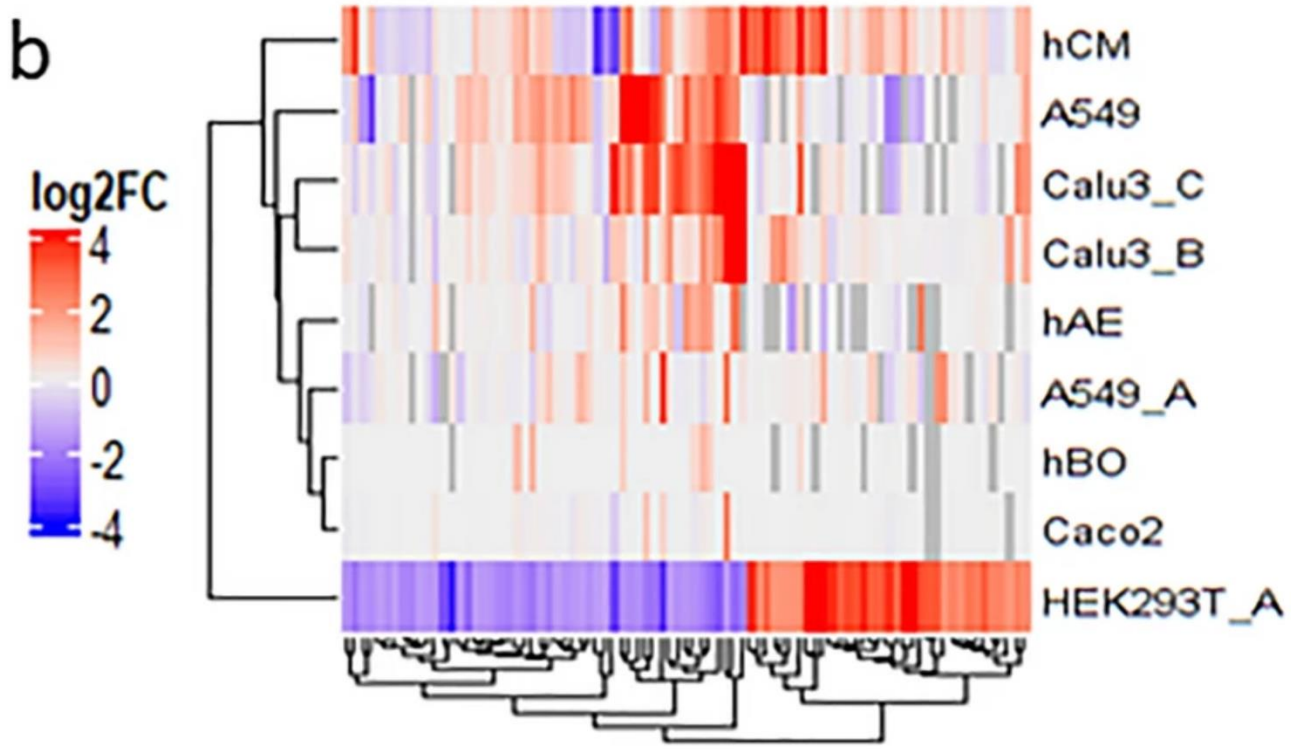
Log base 2 (log2)

- It makes up-regulation and down-regulation symmetrical around zero.
- This symmetry allows for easier comparison of the magnitude of changes in both directions.

- log2FC > 0 → Gene is upregulated
- log2FC < 0 → Gene is downregulated
- log2FC = 1 → Gene expression is increased 2-fold
- log2FC = -1 → Gene expression is decreased by half
- log2FC = 2 → Gene expression is increased 4-fold

Figure 4B

- Figure 4b shows the differential gene expression analysis of SARS-CoV-2 infected cell lines.
- The criteria for selecting genes in Figure 4b were based on
  - log2 fold change (log2FC) of ≥ 2 or ≤ -1.5
  - adjusted p-value of < 0.05



Up-regulated genes (log2FC > 2) are highlighted in red

down-regulated genes (log2FC < -1.5) are highlighted in blue.

Input data

log2FC

P-value adjustment

Genes

	baseMean	log2FoldCh	lfcSE	pvalue	padj
DDX11L1	0.1860746	-0.005964	0.2654724	0.9033765	NA
WASH7P	3.1456614	0.0045432	0.2616298	0.9217498	NA
MIR6859-3	0	NA	NA	NA	NA
MIR6859-2	0	NA	NA	NA	NA
MIR6859-1	0	NA	NA	NA	NA
MIR6859-4	0	NA	NA	NA	NA
MIR1302-2	0	NA	NA	NA	NA
MIR1302-11	0	NA	NA	NA	NA
MIR1302-9	0	NA	NA	NA	NA
MIR1302-10	0	NA	NA	NA	NA
FAM138A	0	NA	NA	NA	NA
FAM138C	0	NA	NA	NA	NA
FAM138F	0	NA	NA	NA	NA
OR4F5	0	NA	NA	NA	NA
LOC729737	0.9078218	-0.005848	0.2645607	0.8038859	NA
FAM138D	0	NA	NA	NA	NA
OR4F29	0	NA	NA	NA	NA
OR4F3	0	NA	NA	NA	NA
OR4F16	0	NA	NA	NA	NA

HEK293T-hACE2 | A549 | A549-hACE2 | Caco2 | Calu3_B | Calu3_C | hBO | hAE | hCM

Cell lines

Code: Expected input

log2FC for each cell line

Genes

	HEK293T_A	A549	A549_A	Caco2	Calu3_B	Calu3_C
ADM2	-1.776962	4.815869548	0.615639026	-0.031314354	0.803572999	1.9581481976
AKNA	-2.494399	1.201366551	-0.040966684	-0.023166505	0.411033283	0.9853411070
ALLC	5.264347	0.026761648	0.342792056	0.027446288	0.158372724	*NA*
ANKRD26P1	2.016181	0.193376857	0.031806242	0.004932063	1.867017774	0.0013403656
ANPEP	3.087821	0.559556521	-0.202357202	-0.019719257	0.122658391	-0.4705838978
ANXA1	-1.707163	-0.493931438	-1.145436981	0.281756679	-0.715508752	-0.5181789419
ASNS	-1.647729	3.125765397	-0.077341772	0.009665014	0.517034782	2.8756199032
ATF3	-1.640445	2.839344169	2.498324613	2.287292218	5.460378990	4.9677852136
ATP6V0D2	-2.984938	0.124259675	*NA*	-0.000082900	0.048100960	0.0007394331

Matrix

Code

```
# Filtering gene base on criteria
HEK293T_hACE2_filtered_with_padj <- HEK293T_hACE2_sheet %>%
                                   filter(padj < 0.05,
                                   log2FoldChange >= 2 | log2FoldChange <= -1.5)
```

```
# Change first column name to “Gene”
colnames(HEK293T_hACE2_filtered_with_padj)[1] <- "Gene"
```

Genes

log2FC

P-value adjustment

	...1	baseMean	log2FoldChange	lfcSE	pvalue	padj
1	TNFRSF9	44.88135	-2.416980	0.51756937	1.961096e-07	2.435489e-06
2	PTAFR	22.39896	2.355074	0.69179690	3.106075e-05	2.339727e-04
3	IL23R	11.17434	3.138083	1.12574478	2.988505e-04	1.728286e-03
4	CTH	1332.65734	-1.619297	0.09777830	8.726598e-63	6.113378e-60

```
# Select only “Gene” and “log2FoldChange” column
# Change second column name to “HEK293T_A”
HEK293T_hACE2_log2FC_filtered_with_padj <- HEK293T_hACE2_filtered_with_padj[,c("Gene","log2FoldChange")]
colnames(HEK293T_hACE2_log2FC_filtered_with_padj)[2] <- "HEK293T_A"
```

	Gene	HEK293T_A
1	TNFRSF9	-2.416980
2	PTAFR	2.355074
3	IL23R	3.138083

Code

```
# Unique gene name
gene_names <- HEK293T_hACE2_filtered_with_padj$Gene
filtered_gene_name <- unique(gene_names)
filtered_gene_name

filtered_gene <- c(
  "TNFRSF9", "PTAFR", "IL23R", "CTH", "PTGER3", "LOC100131564",
  "TBX15", "TXNIP", "MIR5087", "HSPA6", "HSPA7", "SLC9C2",
  "G0S2", "ATF3", "MRC1", "DDIT4", "P4HA1", "LGI1",
  "CYP2C18", "CYP2C19", "SNHG1", "RAPGEF3", "INHBE", "DDIT3",
  "SLC17A8", "PCK2", "DICER1-AS1", "CHAC1", "ANPEP", "NUPR1",
  "MVP", "ANKRD26P1", "TVP23C", "RDM1", "PECAM1", "SOX9-AS1",
  "CATSPERD", "KIAA1683", "LINC00663", "FUT1", "ALLC", "ERICH2",
  "TRIB3", "CEBPB-AS1", "PCK1", "C1QTNF6", "ADM2", "LAMB2P1",
  "SLC7A11-AS1", "SLC7A11", "NRN1", "HSPA1A", "HSPA1B", "C6orf48",
  "GCM1", "FLJ46906", "FRMD1", "INHBA", "ASNS", "TFR2",
  "MUC17", "DNAJB9", "MGAM", "MGAM2", "STC1", "SCARA5",
  "PPAPDC1B", "LINC01301", "C8orf46", "ATP6V0D2", "TG", "VLDLR-AS1",
  "LURAP1L", "TMEM215", "LINC00950", "ANXA1", "AKNA", "SAT1"
)
```

Code

```
# Import excel
Caco2_sheet <- read_excel("data/41598_2021_96462_MOESM1_ESM.xls",
                          sheet = "Caco2",
                          col_types = c("text", "numeric", "numeric",
                                        "numeric", "numeric", "numeric"))

# Filtering gene
colnames(Caco2_sheet)[1] <- "Gene"
Caco2_sheet_filtered <- Caco2_sheet%>%
                        filter(Gene %in% filtered_gene)

# Select only “Gene” and “log2FoldChange” column
# Change second column name to “HEK293T_A”
Caco2_log2FC_sheet <- Caco2_sheet_filtered[,c("Gene","log2FoldChange")]
colnames(Caco2_log2FC_sheet)[2] <- "Caco2"
```

	Gene	Caco2
1	TNFRSF9	0.075673751
2	PTAFR	0.028753926
3	IL23R	0.009636228

!! Do this for all the remaining cell lines

Code

```
# Create a list of data frames
log2FC_data <- list(HEK293T_hACE2_log2FC_filtered_with_padj,A549_sheet_log2Fc,A549_hACE2_log2FC_sheet,
                    Caco2_log2FC_sheet,Calu3_B_log2FC_sheet,
                    Calu3_C_log2FC_sheet,hBO_log2FC_sheet,
                    hAE_log2FC_sheet,hCM_log2FC_sheet)

# Merges multiple data frames in log2FC_data by the "Gene" column
merge_heatmap_table <- Reduce(function(x,y) merge(x,y, by ="Gene", all = TRUE),
                              log2FC_data)
```

	Gene	HEK293T_A
1	TNFRSF9	-2.416980
2	PTAFR	2.355074
3	IL23R	3.138083

+

	Gene	Caco2
1	TNFRSF9	0.075673751
2	PTAFR	0.028753926
3	IL23R	0.009636228

→

	Gene	HEK293T_A	Caco2
1	TNFRSF9	-2.416980	0.075673751
2	PTAFR	2.355074	0.028753926
3	IL23R	3.138083	0.009636228

Code

```
# Converts the "Gene" column into row names
# Converts the data frame merge_heatmap_table into a matrix
merge_heatmap_table <- column_to_rownames(merge_heatmap_table, var ="Gene")
matrix_heatmap_table <- as.matrix(merge_heatmap_table)
```

column

	Gene	HEK293T_A	A549	A549_A	Caco2
1	ADM2	-1.776962	4.815869548	0.615639026	-0.031314354
2	AKNA	-2.494399	1.201366551	-0.040966684	-0.023166505
3	ALLC	5.264347	0.026761648	0.342792056	0.027446288

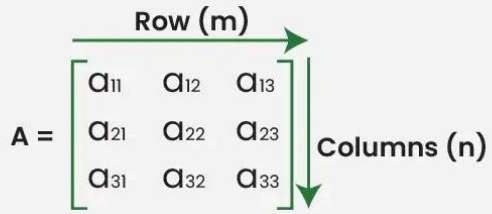
Data frame

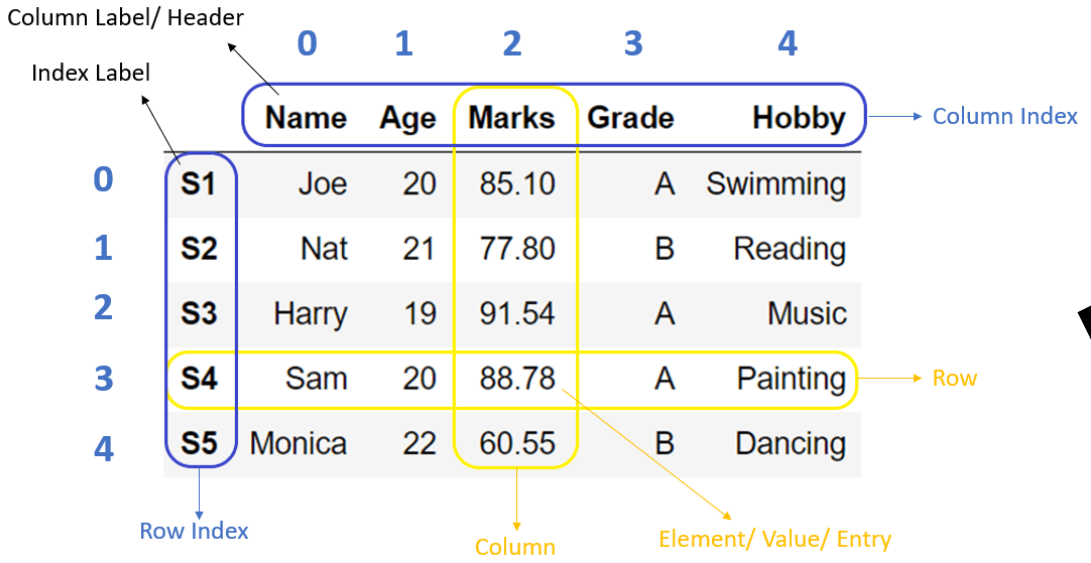
row name

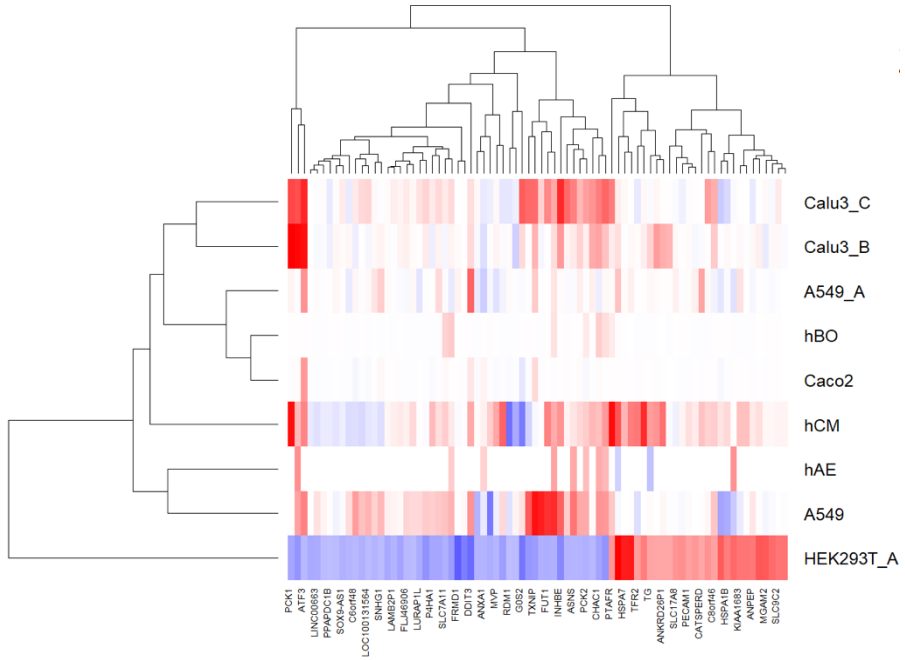
	HEK293T_A	A549	A549_A	Caco2
ADM2	-1.776962	4.815869548	0.615639026	-0.031314354
AKNA	-2.494399	1.201366551	-0.040966684	-0.023166505
ALLC	5.264347	0.026761648	0.342792056	0.027446288

Matrix

Matrix
Data Structure







Code

```
# Creates a color palette with 200 colors ranging from blue to white to red
bwr <- colorRampPalette(c("blue", "white", "red"))(200)

# Transposes the matrix, swapping rows and columns
# Generates a heatmap of the transposed matrix with additional customization
transpose_heatmap_table <- t(matrix_heatmap_table)

heatmap.2(transpose_heatmap_table,col=bwr, scale = "column", trace = "none", offsetCol=0,
          lmat=rbind( c(0, 3, 4), c(2,1,0 ) ), lwid=c(1.5, 4, 2 ))
```

Heatmap: Result with adjusted p-value

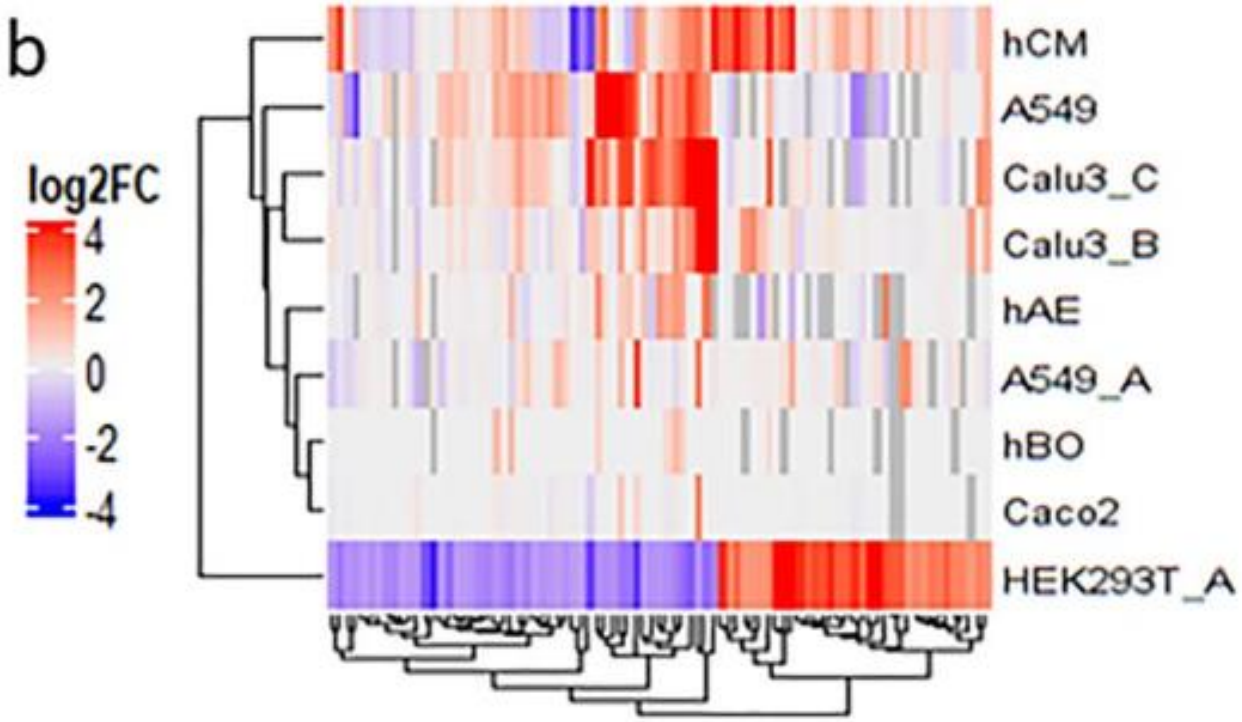
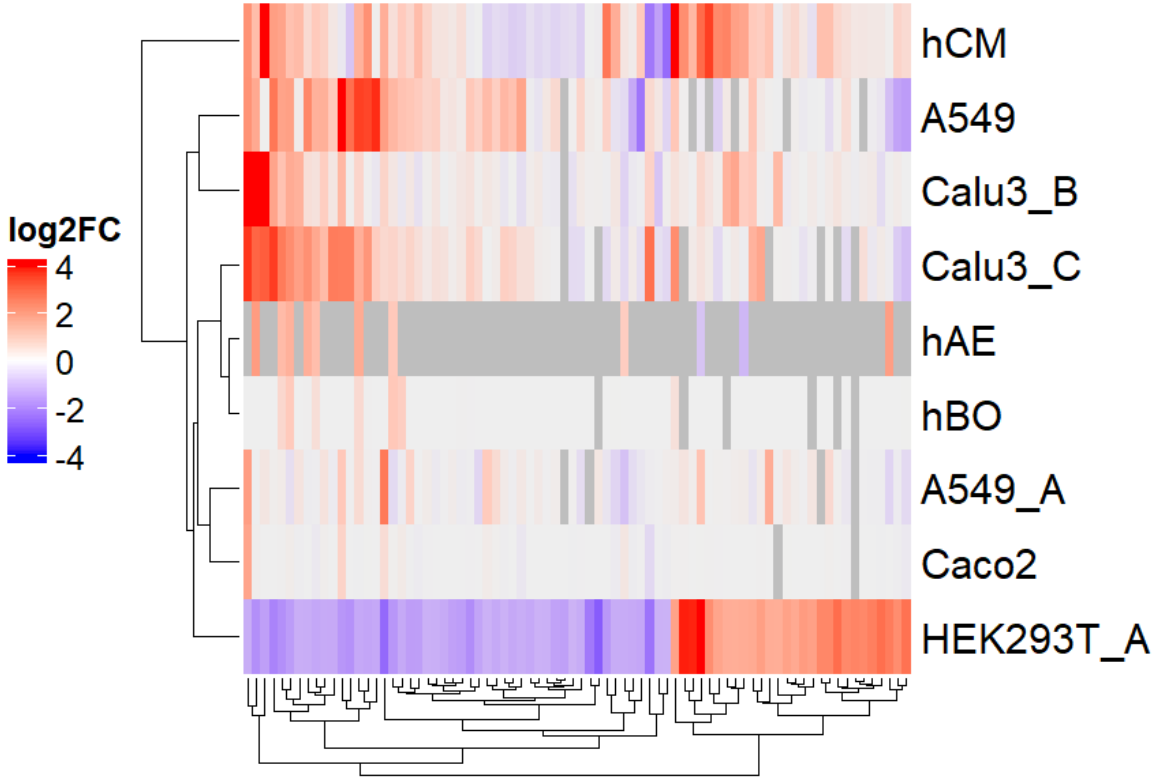
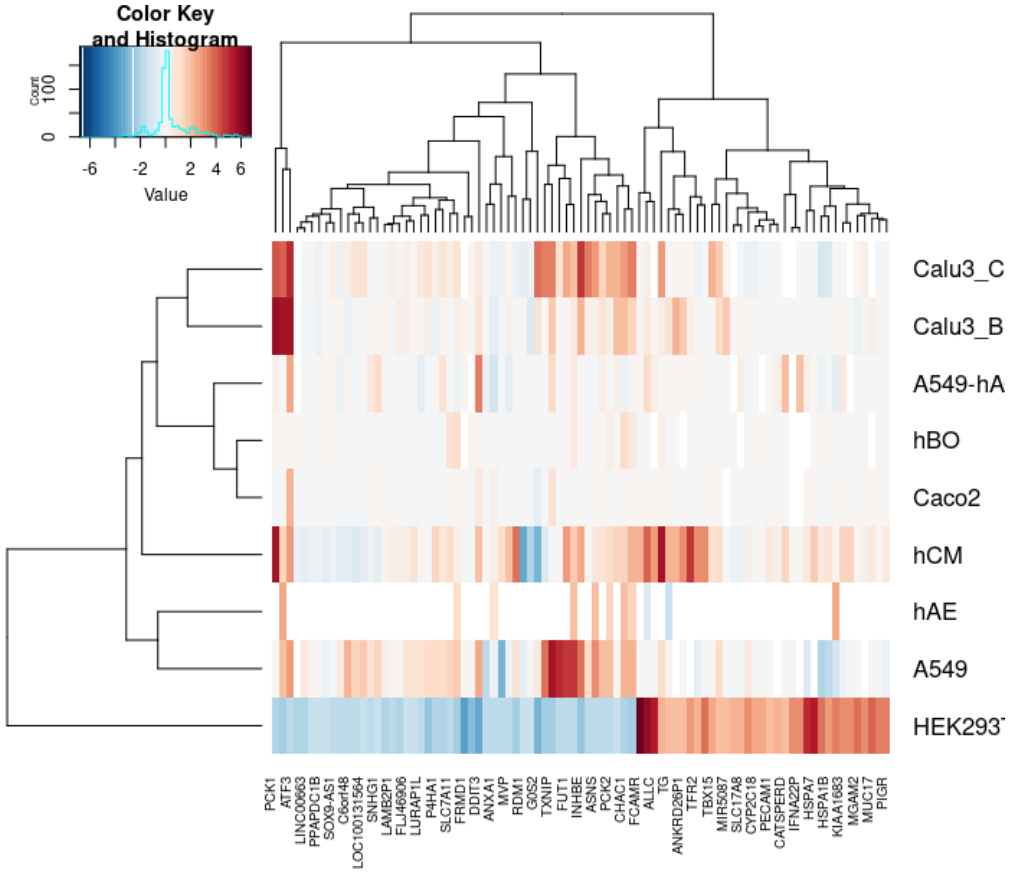


Figure 4B from paper

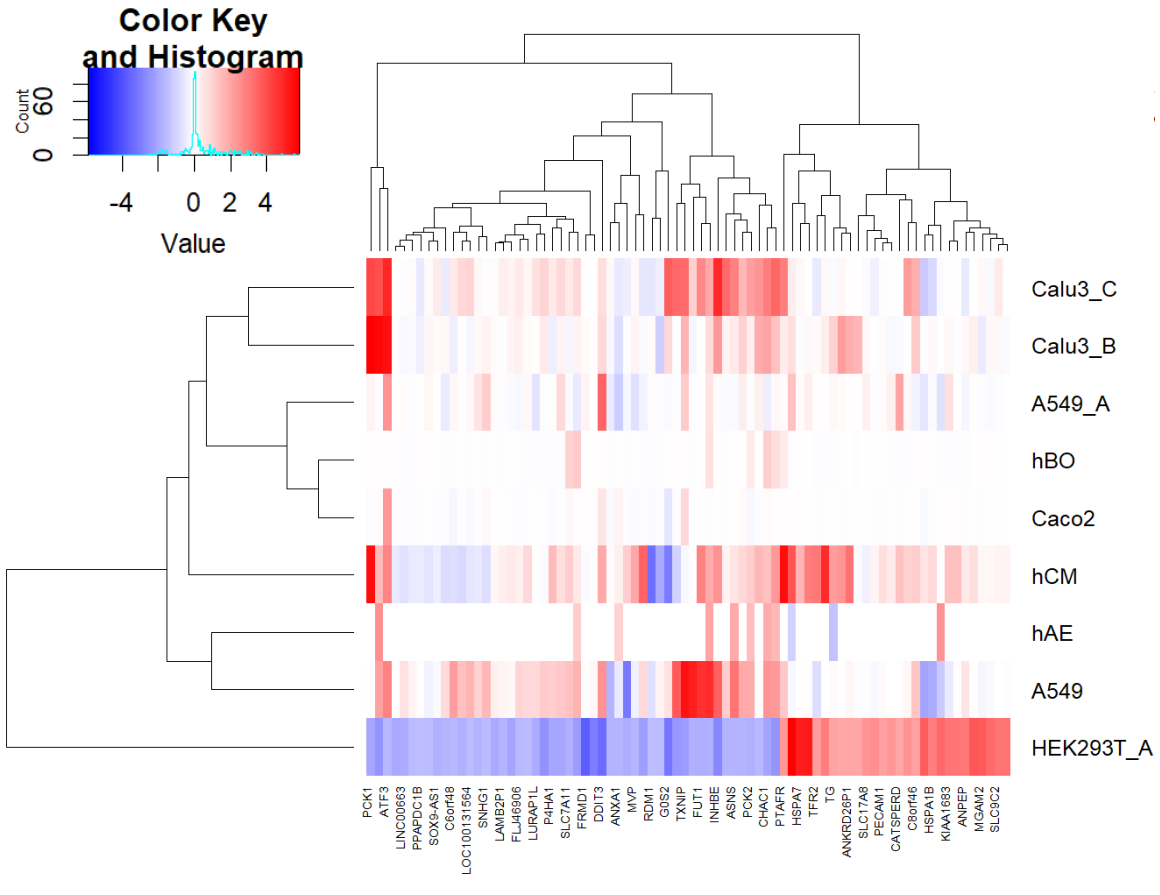


Our Heatmap

Heatmap: Result with adjusted p-value



Example Heatmap from teacher



Our Heatmap

Heatmap: Result without adjusted p-value

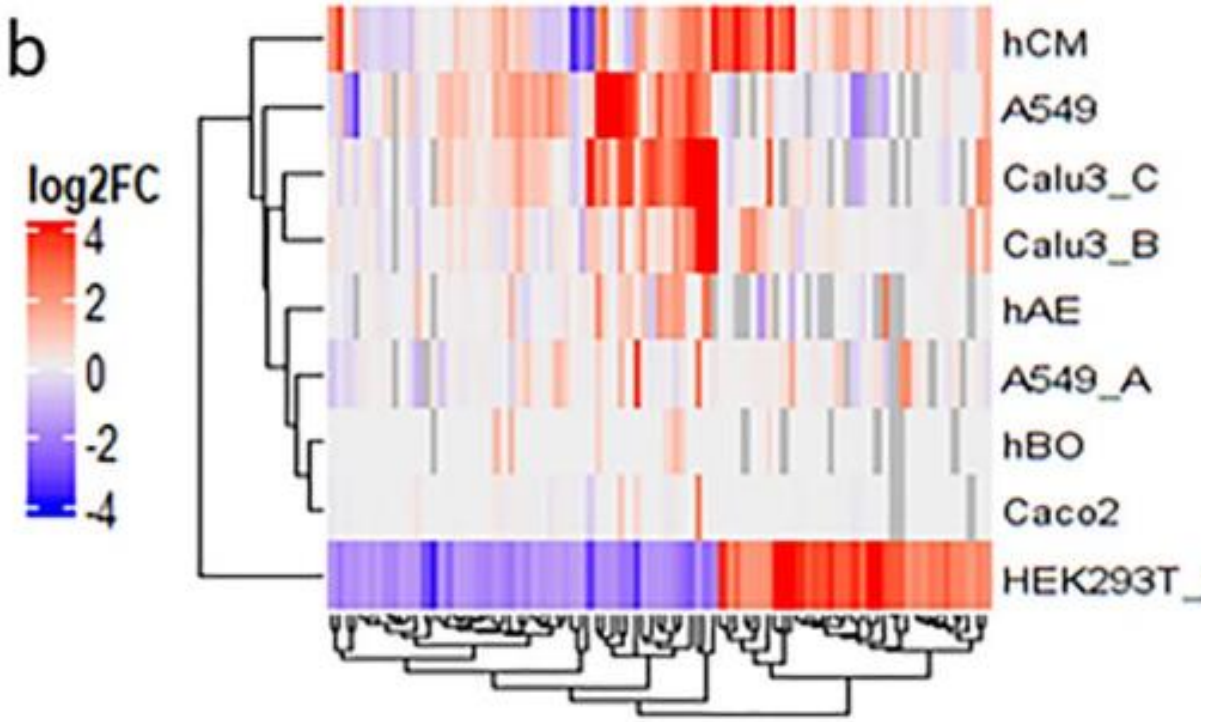
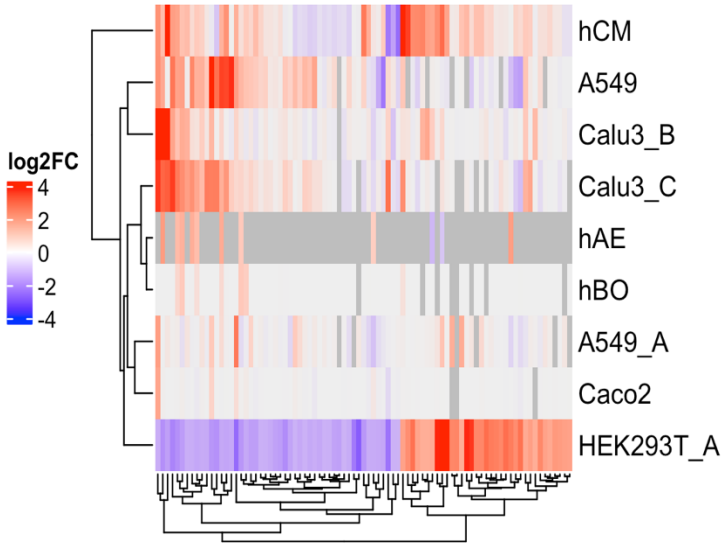
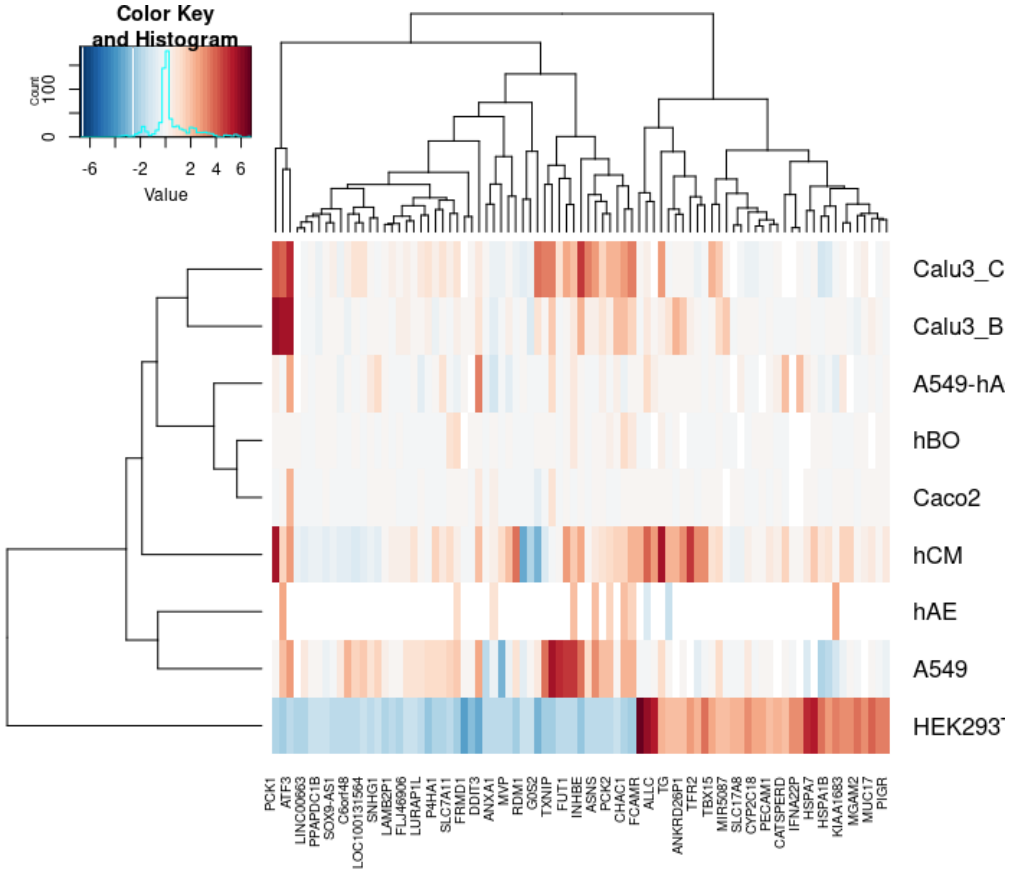


Figure 4B from paper

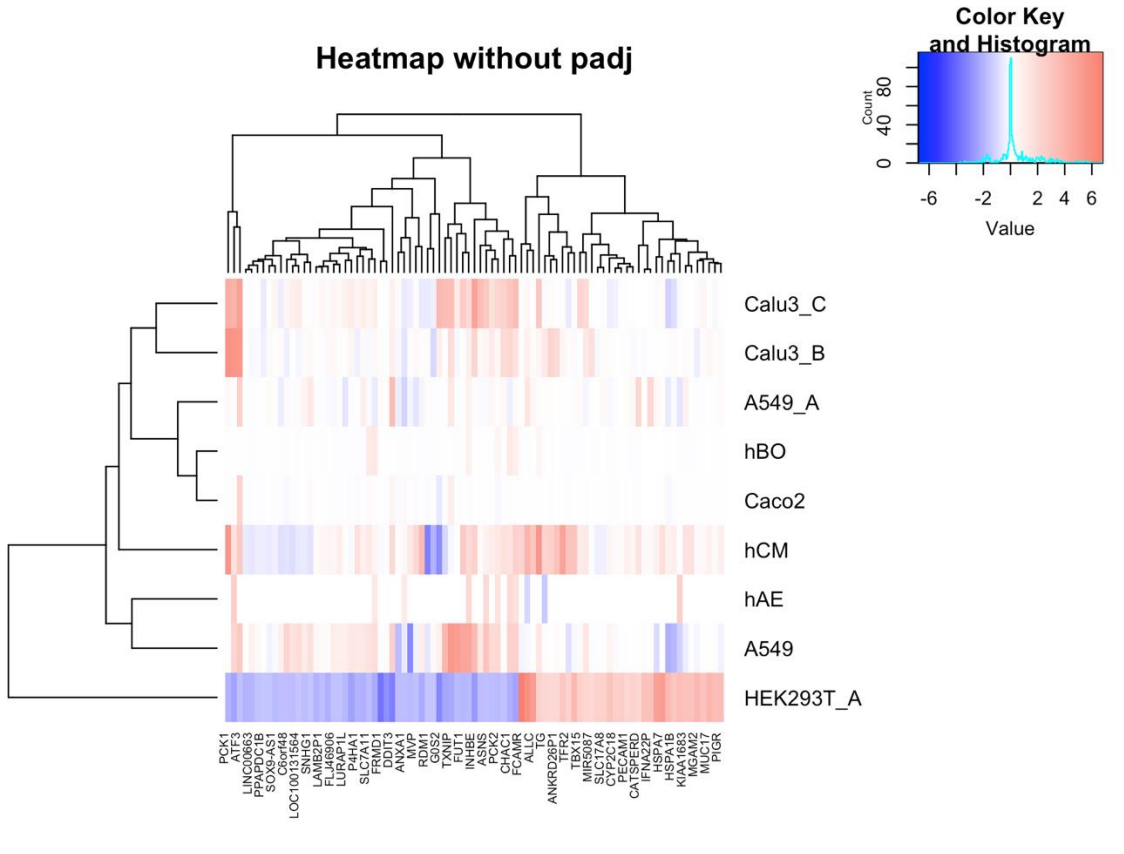


Our Heatmap

Heatmap: Result without adjusted p-value

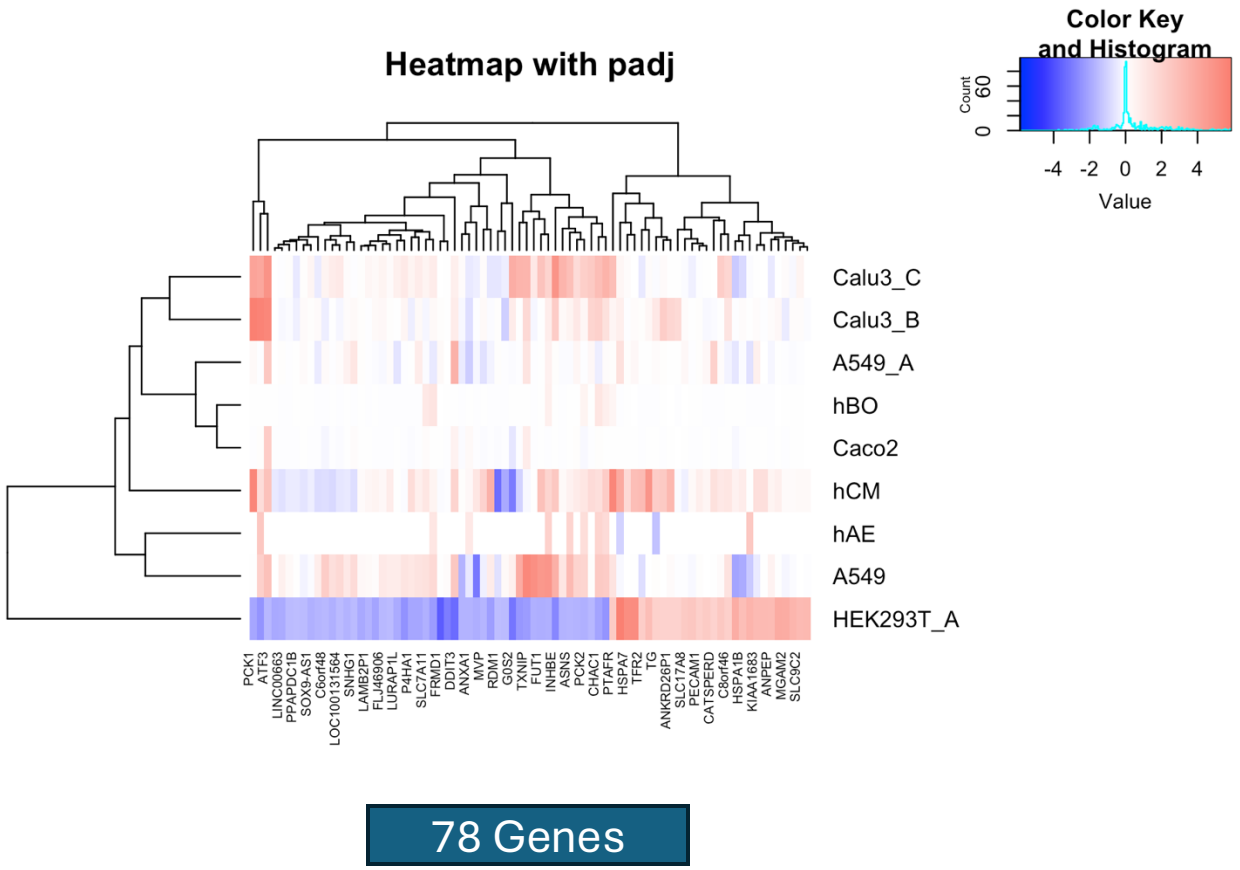


Example Heatmap from teacher

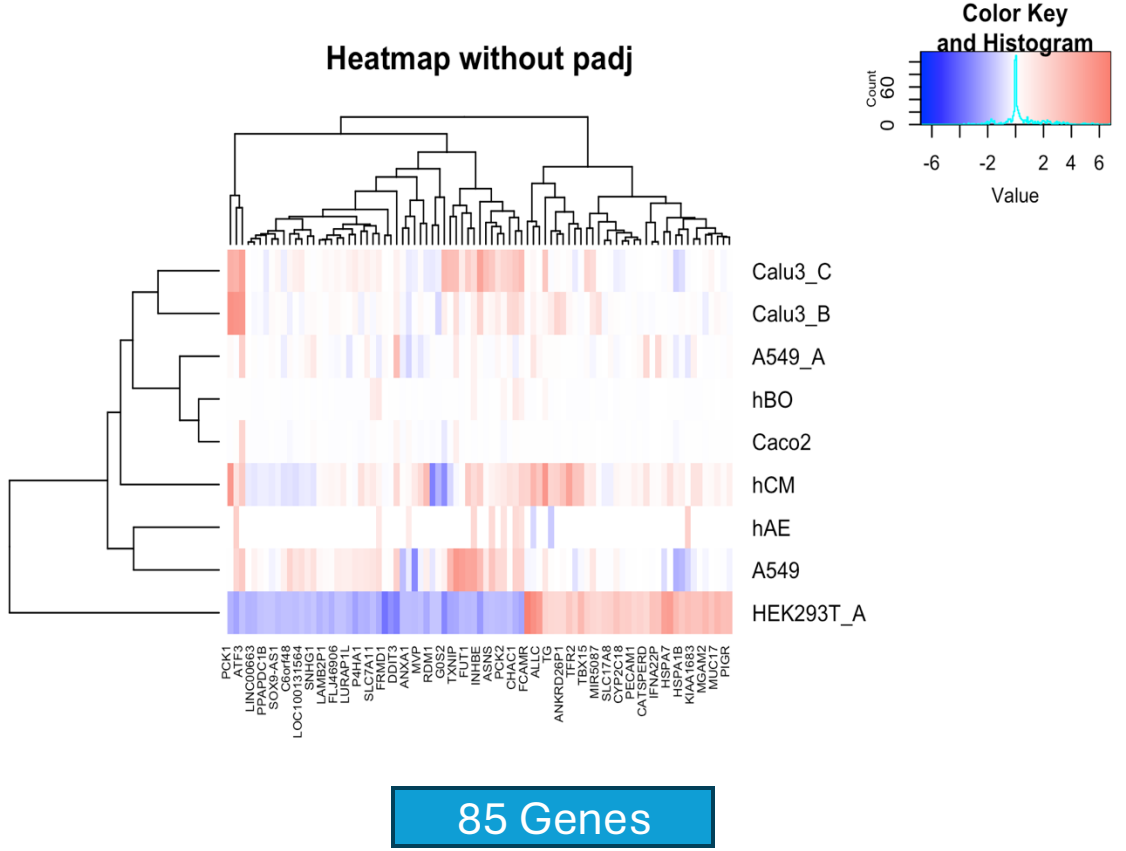


Our Heatmap

Heatmap:With VS Without Padj

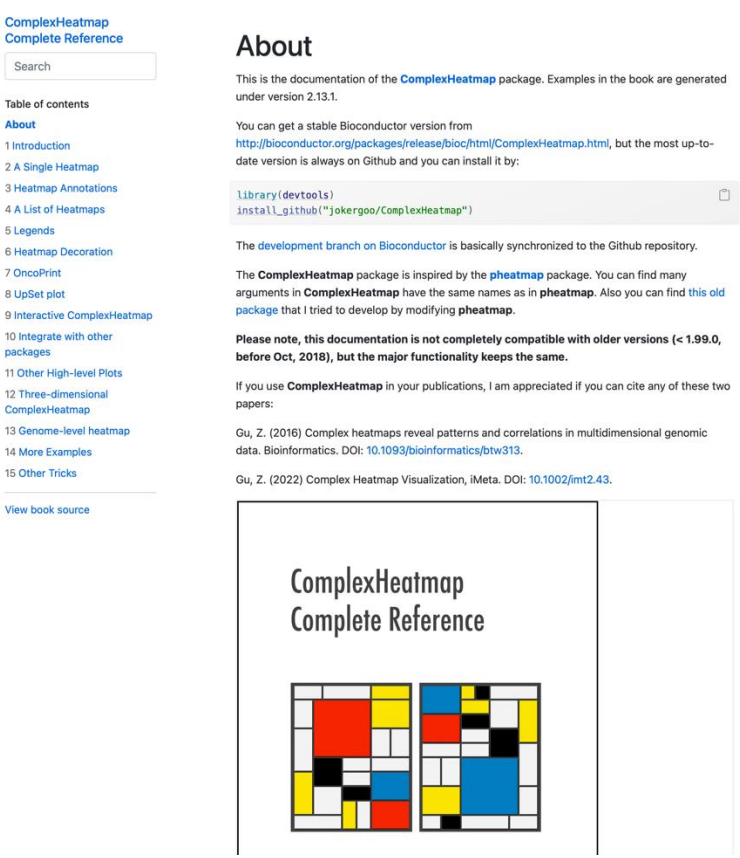


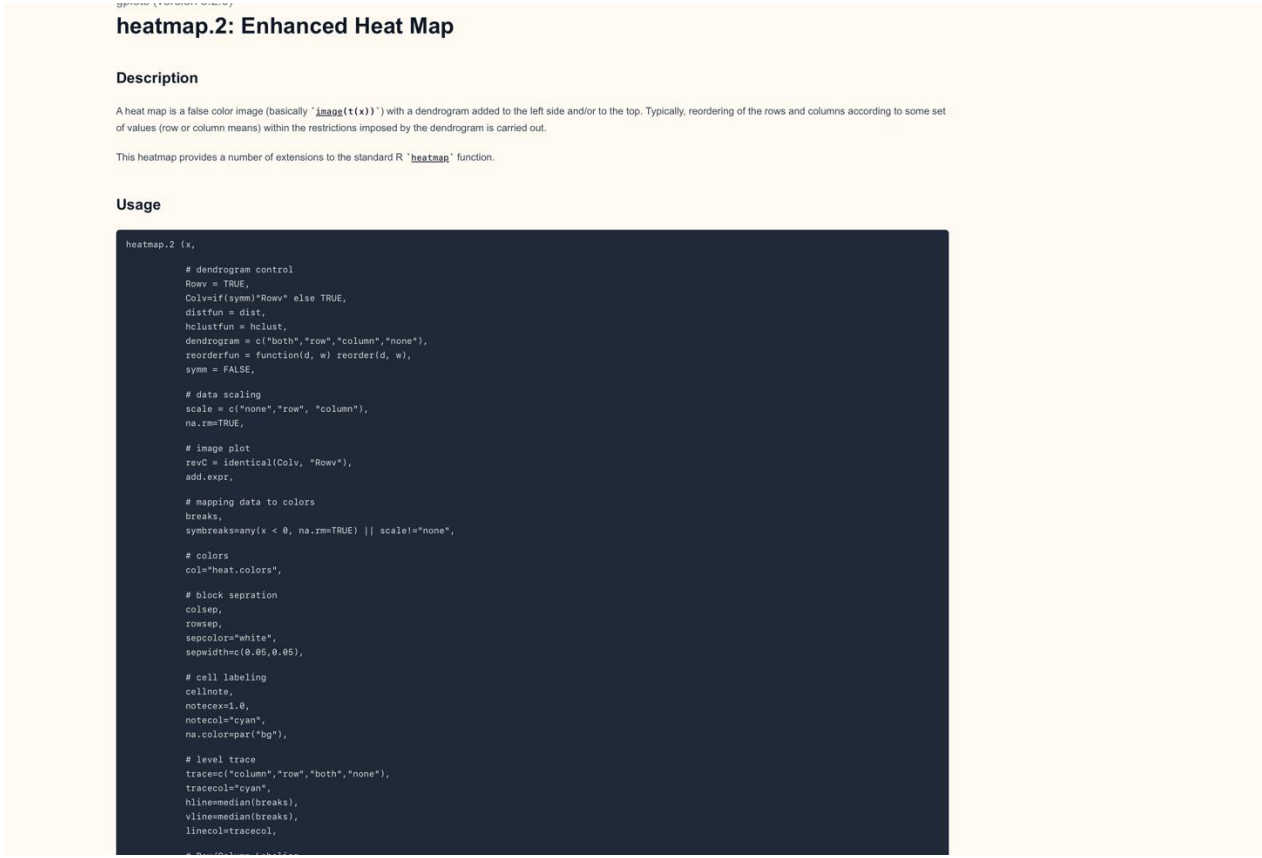
78 Genes



85 Genes

More Tutorial





https://jokergoo.github.io/ComplexHeatmap-reference/book/index.html

https://www.rdocumentation.org/packages/gplots/versions/3.2.0/topics/heatmap.2

THANK YOU