310_Project_Codebase

2023-12-09

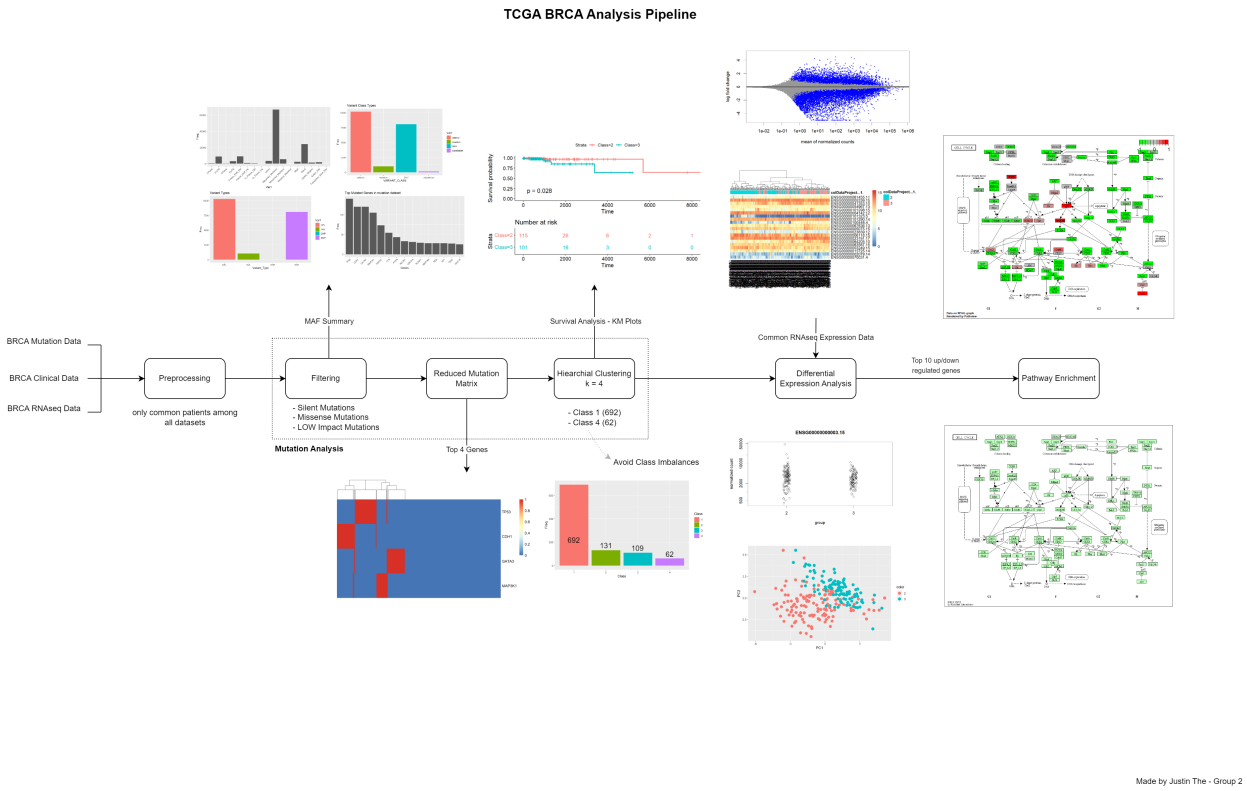
Group Members

Justin_The_48875413 \ Evan_Wong_69389104 \ Andrew Wang_97658025 \

Installing packages

```
library(ggplot2)
library(pheatmap)
library(maftools)
library(dplyr)
library(purrr)
library(TCGAbiolinks)
library(survival)
library(survminer)
library("DESeq2")
library("dplyr")
```

Pipeline Overview



Loading Dataset

```
data.clinical <- read.delim("data/data_clinical_patient.txt", header = TRUE, sep = "\t",
    skip = 4)
# head(data.clinical)
```

```
data.mut <- read.delim("data/data_mutations.txt", header = TRUE, sep = "\t")
# head(data.mutation)
```

```
RNAseq <- read.csv("data/RNAseq_BRCA.csv", header = TRUE)
# head(RNAseq)
```

```
data.mut2 <- data.mut[-which(data.mut$Variant_Classification %in% c("Silent", "Missense_Mutation")),
    ]
# data.mutation <- data.mut2[which(data.mut2$IMPACT=='HIGH'),]
data.mutation <- data.mut2[which(data.mut2$IMPACT %in% c("HIGH", "MODERATE")), ]
```

Common patients

```
# Unique patients in data.clinical and data.mutations
patient_ID <- unique(gsub(".{3}$", "", data.mutation$Tumor_Sample_Barcode))
```

```r
# as.data.frame(patient_ID)
data.mutations.num <- length(patient_ID)
data.clinical.num <- length(unique(data.clinical$PATIENT_ID))

# Unique patients in RNAseq
cnames <- colnames(RNAseq)
cnames <- gsub(".{16}$", "", cnames)
cnames <- gsub("\\.", "-", cnames)
RNAseq.num <- length(unique(cnames[-1]))

commonPatients <- Reduce(intersect, list(patient_ID, data.clinical$PATIENT_ID, cnames))
```

```r
data.mutation[, 112] <- gsub(".{3}$", "", data.mutation$Tumor_Sample_Barcode)
length(unique(data.mutation$V112))
```

```
## [1] 997
```

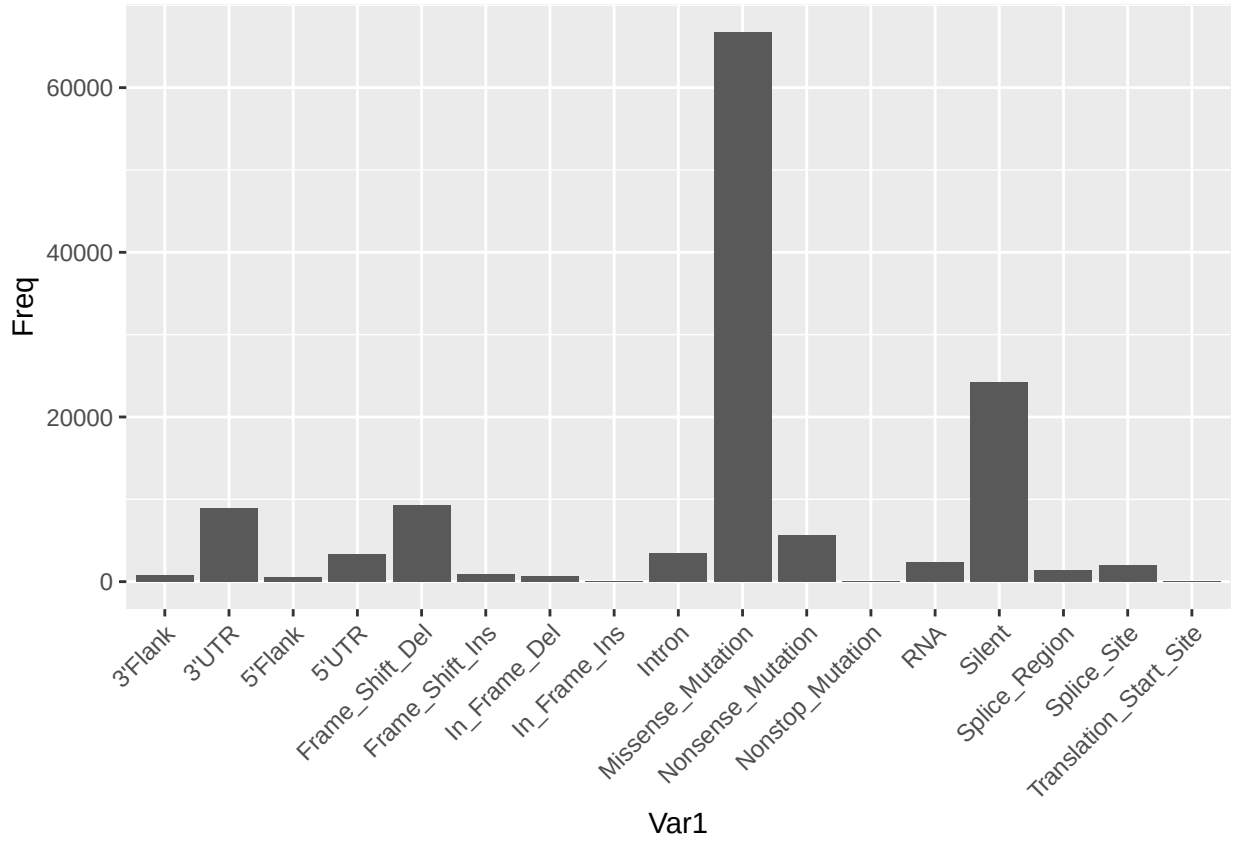
```r
commonPatients_mutation_idx <- which(data.mutation$V112 %in% commonPatients)
commonPatients_mutation <- data.mutation[commonPatients_mutation_idx, ]
# length(unique(commonPatients_mutation$V112)) dim(commonPatients_mutation)
commonPatients_clinical_idx <- which(data.clinical$PATIENT_ID %in% commonPatients)
commonPatients_clinical <- data.clinical[commonPatients_clinical_idx, ]
dim(commonPatients_clinical)
```

```
## [1] 994  38
```

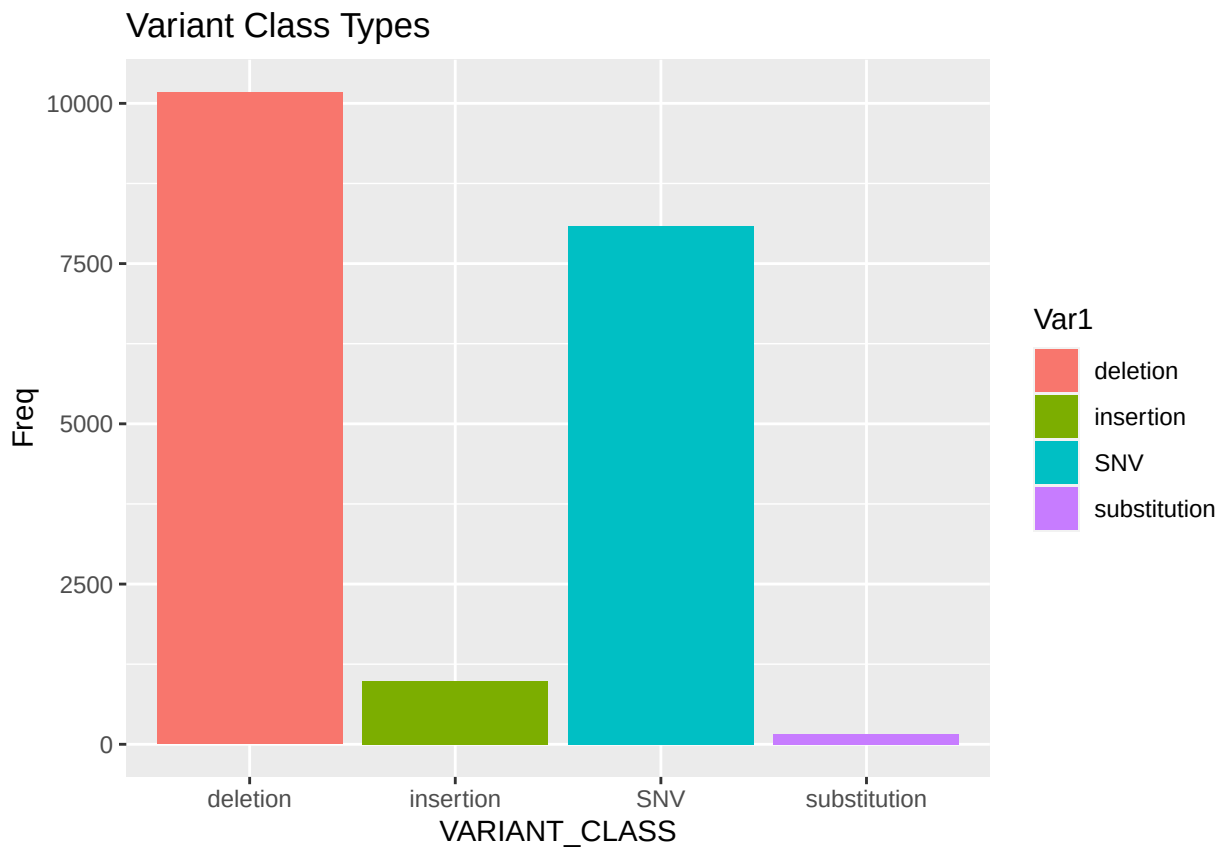
Now we have a mutation dataset with only common patients across all datasets.

MAF Summary

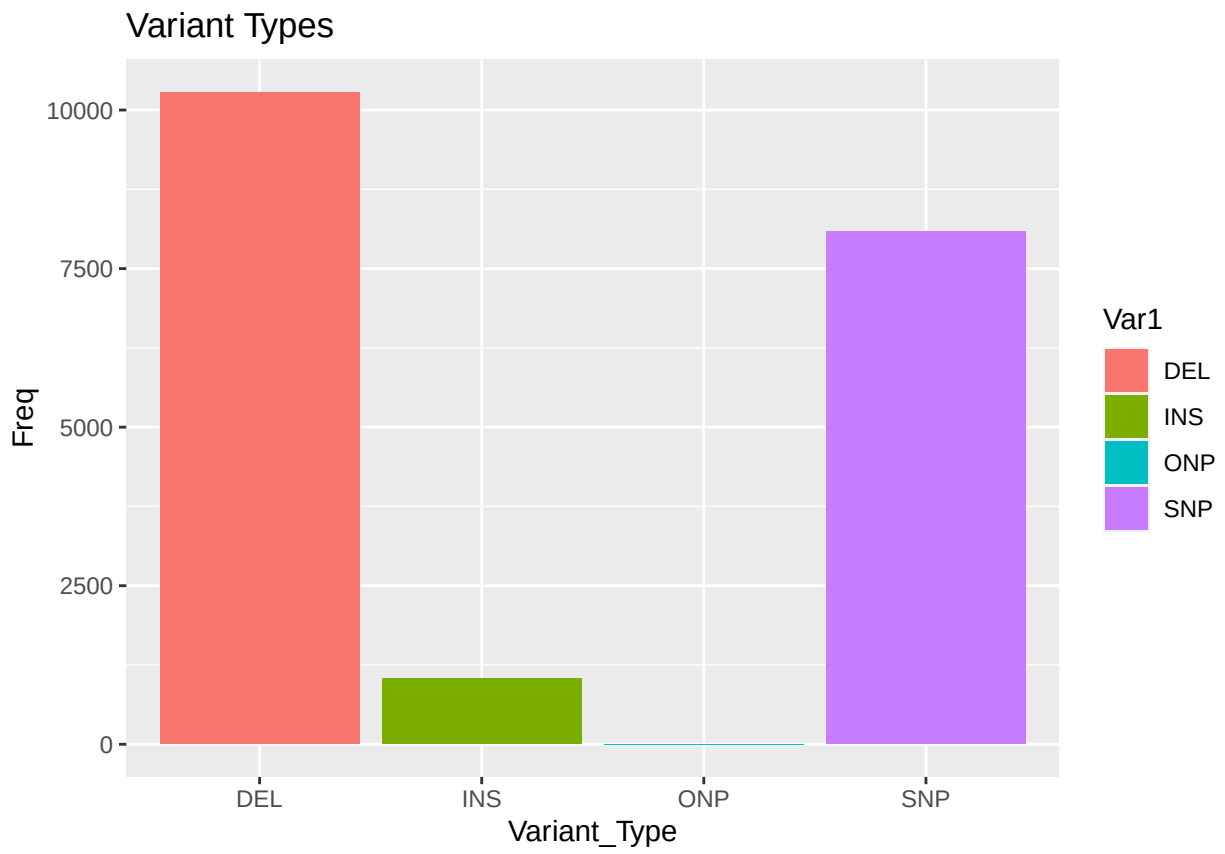
```r
var.class <- as.data.frame(table(data.mut$Variant_Classification))

ggplot(data = var.class, aes(x = Var1, y = Freq)) + geom_col() + theme(axis.text.x = element_text(angle
    hjust = 1))
```



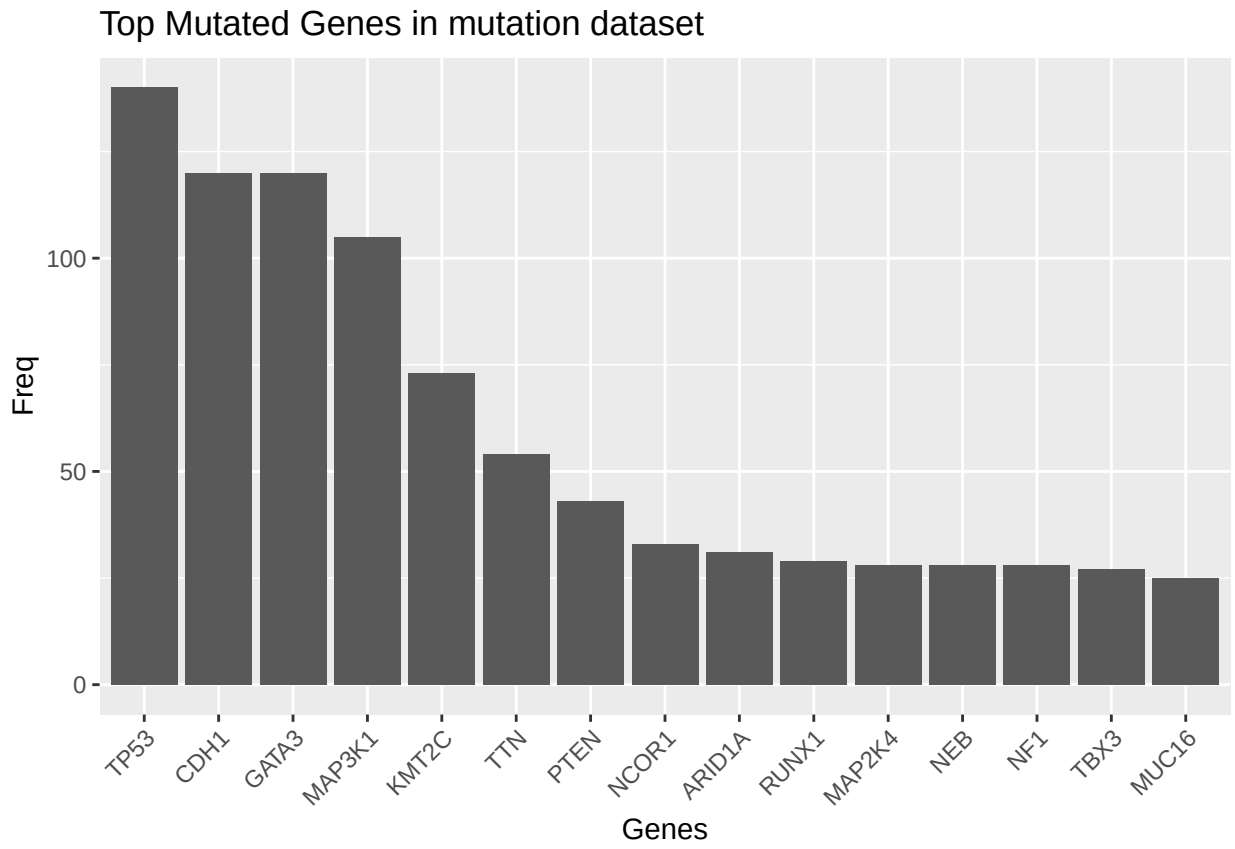

```
var.class2 <- as.data.frame(table(commonPatients_mutation$VARIANT_CLASS))

ggplot(data = var.class2, aes(x = Var1, y = Freq)) + geom_col(aes(fill = Var1)) +
    ggtitle("Variant Class Types") + xlab("VARIANT_CLASS")
```




```
var.type <- as.data.frame(table(commonPatients_mutation$Variant_Type))

ggplot(data = var.type, aes(x = Var1, y = Freq)) + geom_col(aes(fill = Var1)) + ggtitle("Variant Types")
    xlab("Variant_Type")
```




```
# sample.name <-
# as.data.frame(table(commonPatients_mutation$Tumor_Sample_Barcode))

hugo <- as.data.frame(table(commonPatients_mutation$Hugo_Symbol))

hugo.ordered <- hugo[order(-hugo$Freq), ]

ggplot(data = hugo.ordered[1:15, ], aes(x = Var1, y = Freq)) + geom_col() + theme(axis.text.x = element_
    hjust = 1)) + scale_x_discrete(limits = hugo.ordered[1:15, ]$Var1) + ggtitle("Top Mutated Genes in m
    xlab("Genes")
```

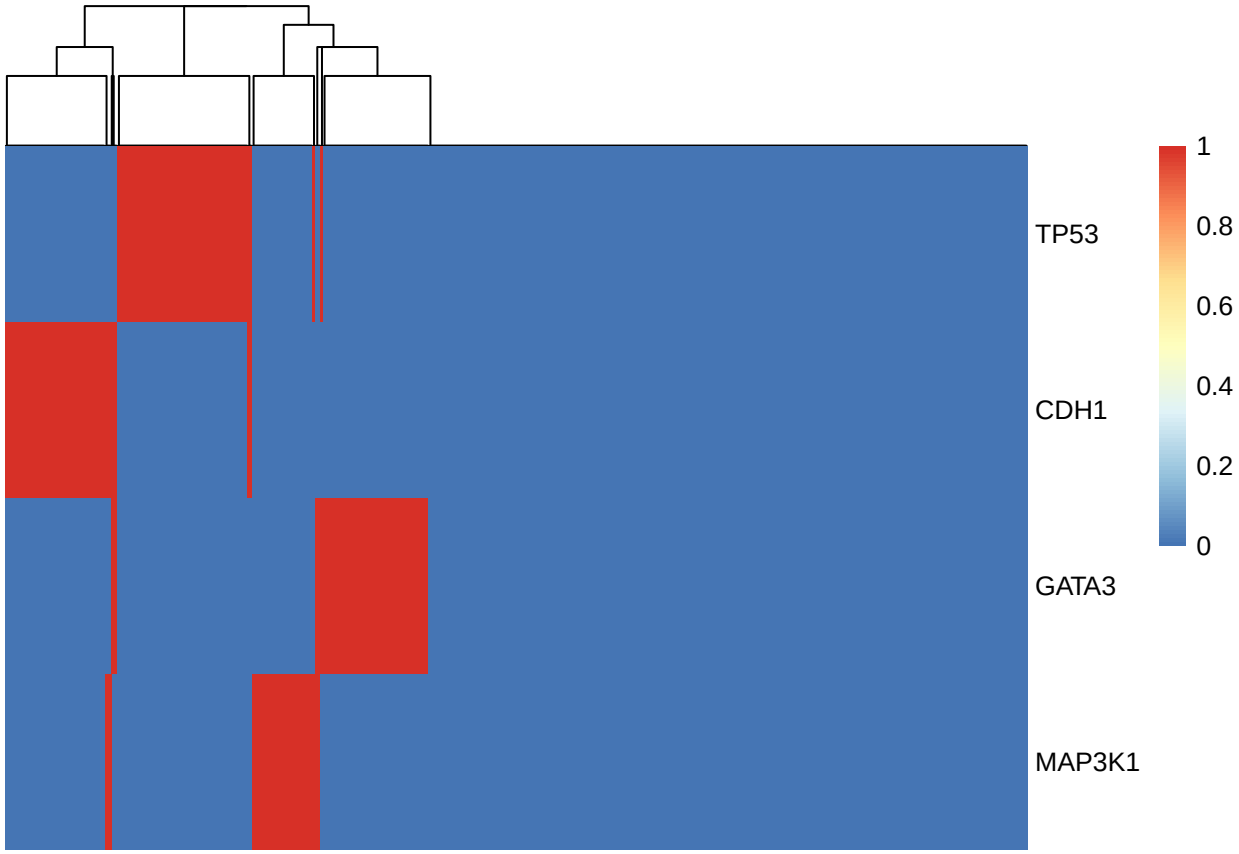


Generate Oncoplot

```
commonPatients_mutation <- commonPatients_mutation[-112]
cnv_events <- unique(commonPatients_mutation$Variant_Classification)
oncomat <- reshape2::dcast(data = commonPatients_mutation, formula = Hugo_Symbol ~
    Tumor_Sample_Barcode, fun.aggregate = function(x, cnv = cnv_events) {
    x = as.character(x)  # >= 2 same/distinct variant classification = Multi_Hit
    xad = x[x %in% cnv]
    xvc = x[!x %in% cnv]

    if (length(xvc) > 0) {
        xvc = ifelse(test = length(xvc) > 1, yes = "Multi_Hit", no = xvc)

    }
    x = ifelse(test = length(xad) > 0, yes = paste(xad, xvc, sep = ";"), no = xvc)
    x = gsub(pattern = ";$", replacement = "", x = x)
    x = gsub(pattern = "^;", replacement = "", x = x)
    return(x)
}, value.var = "Variant_Classification", fill = "", drop = FALSE)
```

```
rownames(oncomat) = oncomat$Hugo_Symbol
oncomat <- oncomat[, -1]

oncomat.ordered <- oncomat[order(-hugo$Freq), ]
```

```
mat <- oncomat.ordered
mat[mat != ""] = 1
mat[mat == ""] = 0
mat <- apply(mat, 2, as.numeric)
mat <- as.matrix(mat)
rownames(mat) <- row.names(oncomat.ordered)

reduce.mat <- mat[1:4, ]
res <- pheatmap(reduce.mat, cluster_rows = F, show_colnames = FALSE, )
```



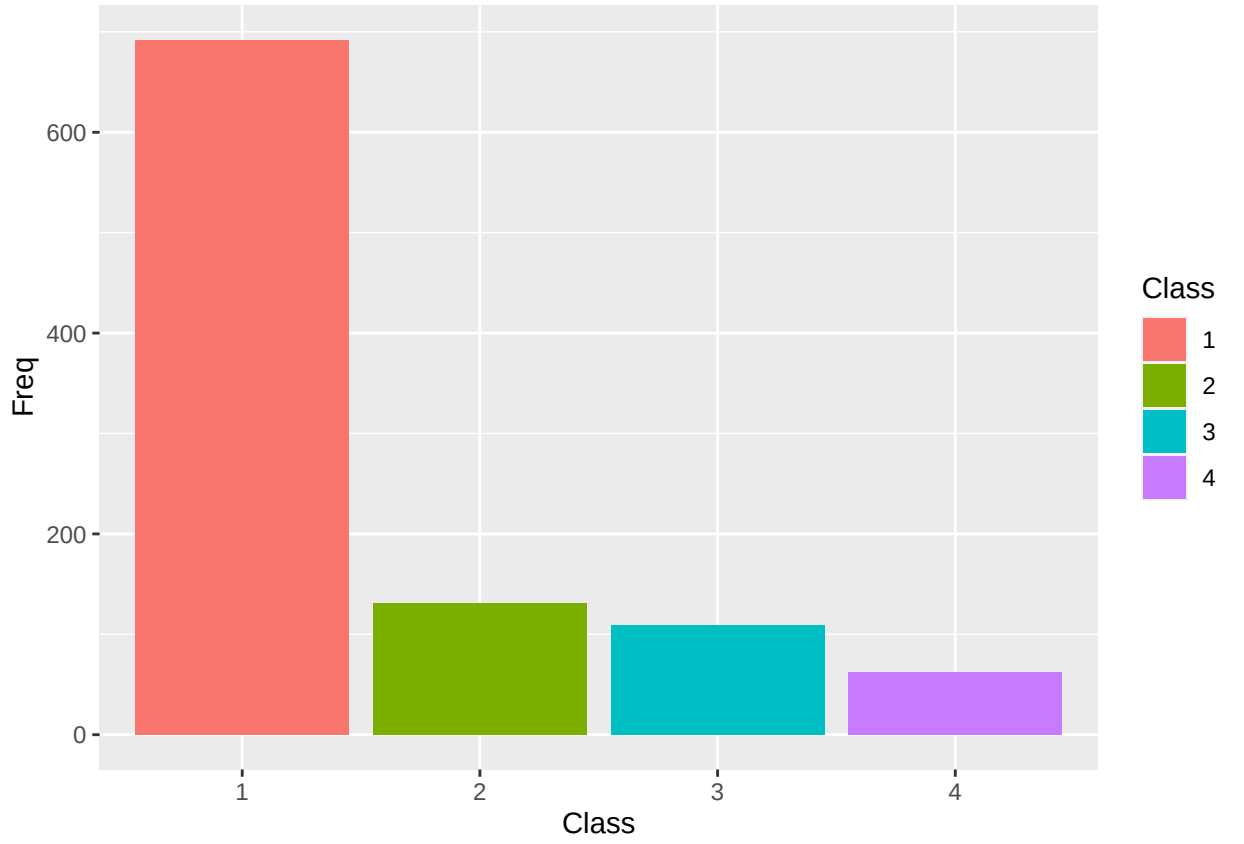

```
cluster <- as.data.frame(cutree(res$tree_col, k = 4))  #choose k

rownames(cluster) <- gsub(".{3}$", "", rownames(cluster))
colnames(cluster) <- "Class"
cluster_frequency <- as.data.frame(table(cluster))
cluster_frequency
```

```
##   Class Freq
## 1     1  692
## 2     2  131
## 3     3  109
## 4     4   62
```

```
head(cluster)
```

```
##              Class
## TCGA-3C-AAAU     1
## TCGA-3C-AALI     2
## TCGA-3C-AALJ     1
## TCGA-3C-AALK     1
## TCGA-4H-AAAK     3
## TCGA-5L-AAT0     3
```

```
ggplot(data = cluster_frequency, aes(x = Class, y = Freq)) + geom_col(aes(fill = Class))
```




```
# class_1 <- rownames(clustercluster$Class=='1') class_1 <-
# rownames(cluster[cluster$Class==1, , drop=FALSE]) class_2 <-
# rownames(cluster[cluster$Class==2, , drop=FALSE]) class_3 <-
# rownames(cluster[cluster$Class==3, , drop=FALSE]) class_4 <-
# rownames(cluster[cluster$Class==4, , drop=FALSE])

# data.clinical_class1 <-
# data.clinical[which(data.clinical$PATIENT_ID%in%class_1),]
# data.clinical_class2 <-
# data.clinical[which(data.clinical$PATIENT_ID%in%class_2),]
# data.clinical_class3 <-
# data.clinical[which(data.clinical$PATIENT_ID%in%class_3),]
```

```
# data.clinical_class4 <-
# data.clinical[which(data.clinical$PATIENT_ID%in%class_4),]
```

Survival Analysis

```
clinical_df <- cbind(commonPatients_clinical, cluster$Class)

clinical_df_colnames <- colnames(clinical_df)
clinical_df_colnames[39] <- "Class"
colnames(clinical_df) <- clinical_df_colnames

head(clinical_df)
```

```
##     PATIENT_ID   SUBTYPE CANCER_TYPE_ACRONYM
## 1 TCGA-3C-AAAU BRCA_LumA                BRCA
## 2 TCGA-3C-AALI BRCA_Her2                BRCA
## 3 TCGA-3C-AALJ BRCA_LumB                BRCA
## 4 TCGA-3C-AALK BRCA_LumA                BRCA
## 5 TCGA-4H-AAAK BRCA_LumA                BRCA
## 6 TCGA-5L-AAT0 BRCA_LumA                BRCA
##                       OTHER_PATIENT_ID AGE    SEX AJCC_PATHOLOGIC_TUMOR_STAGE
## 1 6E7D5EC6-A469-467C-B748-237353C23416  55 Female                     STAGE X
## 2 55262FCB-1B01-4480-B322-36570430C917  50 Female                   STAGE IIB
## 3 427D0648-3F77-4FFC-B52C-89855426D647  62 Female                   STAGE IIB
## 4 C31900A4-5DCD-4022-97AC-638E86E889E4  52 Female                    STAGE IA
## 5 6623FC5E-00BE-4476-967A-CBD55F676EA6  50 Female                  STAGE IIIA
## 6 86C6F993-327F-4525-9983-29C55625593A  42 Female                   STAGE IIA
##   AJCC_STAGING_EDITION DAYS_LAST_FOLLOWUP DAYS_TO_BIRTH
## 1                  6TH               4047        -20211
## 2                  6TH               4005        -18538
## 3                  7TH               1474        -22848
## 4                  7TH               1448        -19074
## 5                  7TH                348        -18371
## 6                  7TH               1477        -15393
##   DAYS_TO_INITIAL_PATHOLOGIC_DIAGNOSIS             ETHNICITY
## 1                                    0 Not Hispanic Or Latino
## 2                                    0 Not Hispanic Or Latino
## 3                                    0 Not Hispanic Or Latino
## 4                                    0 Not Hispanic Or Latino
## 5                                    0 Not Hispanic Or Latino
## 6                                    0     Hispanic Or Latino
##   FORM_COMPLETION_DATE HISTORY_NEOADJUVANT_TRTYN ICD_10 ICD_O_3_HISTOLOGY
## 1              1/13/14                       No  C50.9           8520/3
## 2              7/28/14                       No  C50.9           8500/3
## 3              7/28/14                       No  C50.9           8500/3
## 4              7/28/14                       No  C50.9           8500/3
## 5             11/13/14                       No  C50.9           8520/3
## 6              8/15/14                       No  C50.9           8520/3
##   ICD_O_3_SITE INFORMED_CONSENT_VERIFIED
## 1        C50.9                      Yes
## 2        C50.9                      Yes
```

```
## 3        C50.9                       Yes
## 4        C50.9                       Yes
## 5        C50.9                       Yes
## 6        C50.9                       Yes
##   NEW_TUMOR_EVENT_AFTER_INITIAL_TREATMENT PATH_M_STAGE PATH_N_STAGE
## 1                                     No           MX           NX
## 2                                     No           M0          N1A
## 3                                     No           M0          N1A
## 4                                     No           M0      N0 (I+)
## 5                                     No           M0          N2A
## 6                                     No           M0           N0
##   PATH_T_STAGE PERSON_NEOPLASM_CANCER_STATUS
## 1           TX                    With Tumor
## 2           T2                    Tumor Free
## 3           T2                    Tumor Free
## 4          T1C                    Tumor Free
## 5           T2                    Tumor Free
## 6           T2                    Tumor Free
##   PRIMARY_LYMPH_NODE_PRESENTATION_ASSESSMENT PRIOR_DX                    RACE
## 1                                        Yes       No                   White
## 2                                        Yes       No Black or African American
## 3                                        Yes       No Black or African American
## 4                                        Yes       No Black or African American
## 5                                        Yes       No                   White
## 6                                        Yes      Yes                   White
##   RADIATION_THERAPY WEIGHT IN_PANCANPATHWAYS_FREEZE OS_STATUS OS_MONTHS
## 1                No     NA                      Yes  0:LIVING 133.05060
## 2               Yes     NA                      Yes  0:LIVING 131.66979
## 3                No     NA                      Yes  0:LIVING  48.45974
## 4                No     NA                      Yes  0:LIVING  47.60496
## 5                No     NA                      Yes  0:LIVING  11.44097
## 6               Yes     NA                      Yes  0:LIVING  48.55837
##                   DSS_STATUS DSS_MONTHS            DFS_STATUS DFS_MONTHS
## 1 0:ALIVE OR DEAD TUMOR FREE  133.05060 1:Recurred/Progressed   59.44044
## 2 0:ALIVE OR DEAD TUMOR FREE  131.66979         0:DiseaseFree  131.66979
## 3 0:ALIVE OR DEAD TUMOR FREE   48.45974         0:DiseaseFree   48.45974
## 4 0:ALIVE OR DEAD TUMOR FREE   47.60496                              NA
## 5 0:ALIVE OR DEAD TUMOR FREE   11.44097         0:DiseaseFree   11.44097
## 6 0:ALIVE OR DEAD TUMOR FREE   48.55837                              NA
##      PFS_STATUS PFS_MONTHS GENETIC_ANCESTRY_LABEL Class
## 1 1:PROGRESSION   59.44044                   EUR     1
## 2    0:CENSORED  131.66979                   AFR     2
## 3    0:CENSORED   48.45974              AFR_ADMIX     1
## 4    0:CENSORED   47.60496                   AFR     1
## 5    0:CENSORED   11.44097                   EUR     3
## 6    0:CENSORED   48.55837                           3
```

```
clinical_df[, 30] <- gsub("^(.).*", "\\1", clinical_df$OS_STATUS)
clinical_df[, 30] <- ifelse(clinical_df$OS_STATUS == "1", TRUE, FALSE)
head(clinical_df)
```

```
##     PATIENT_ID   SUBTYPE CANCER_TYPE_ACRONYM
## 1 TCGA-3C-AAAU BRCA_LumA                BRCA
## 2 TCGA-3C-AALI BRCA_Her2                BRCA
```

```
## 3 TCGA-3C-AALJ BRCA_LumB                BRCA
## 4 TCGA-3C-AALK BRCA_LumA                BRCA
## 5 TCGA-4H-AAAK BRCA_LumA                BRCA
## 6 TCGA-5L-AAT0 BRCA_LumA                BRCA
##                       OTHER_PATIENT_ID AGE    SEX AJCC_PATHOLOGIC_TUMOR_STAGE
## 1 6E7D5EC6-A469-467C-B748-237353C23416  55 Female                    STAGE X
## 2 55262FCB-1B01-4480-B322-36570430C917  50 Female                  STAGE IIB
## 3 427D0648-3F77-4FFC-B52C-89855426D647  62 Female                  STAGE IIB
## 4 C31900A4-5DCD-4022-97AC-638E86E889E4  52 Female                   STAGE IA
## 5 6623FC5E-00BE-4476-967A-CBD55F676EA6  50 Female                 STAGE IIIA
## 6 86C6F993-327F-4525-9983-29C55625593A  42 Female                   STAGE IIA
##   AJCC_STAGING_EDITION DAYS_LAST_FOLLOWUP DAYS_TO_BIRTH
## 1                  6TH               4047        -20211
## 2                  6TH               4005        -18538
## 3                  7TH               1474        -22848
## 4                  7TH               1448        -19074
## 5                  7TH                348        -18371
## 6                  7TH               1477        -15393
##   DAYS_TO_INITIAL_PATHOLOGIC_DIAGNOSIS             ETHNICITY
## 1                                    0 Not Hispanic Or Latino
## 2                                    0 Not Hispanic Or Latino
## 3                                    0 Not Hispanic Or Latino
## 4                                    0 Not Hispanic Or Latino
## 5                                    0 Not Hispanic Or Latino
## 6                                    0     Hispanic Or Latino
##   FORM_COMPLETION_DATE HISTORY_NEOADJUVANT_TRTYN ICD_10 ICD_O_3_HISTOLOGY
## 1              1/13/14                       No  C50.9            8520/3
## 2              7/28/14                       No  C50.9            8500/3
## 3              7/28/14                       No  C50.9            8500/3
## 4              7/28/14                       No  C50.9            8500/3
## 5             11/13/14                       No  C50.9            8520/3
## 6              8/15/14                       No  C50.9            8520/3
##   ICD_O_3_SITE INFORMED_CONSENT_VERIFIED
## 1        C50.9                      Yes
## 2        C50.9                      Yes
## 3        C50.9                      Yes
## 4        C50.9                      Yes
## 5        C50.9                      Yes
## 6        C50.9                      Yes
##   NEW_TUMOR_EVENT_AFTER_INITIAL_TREATMENT PATH_M_STAGE PATH_N_STAGE
## 1                                      No           MX           NX
## 2                                      No           M0          N1A
## 3                                      No           M0          N1A
## 4                                      No           M0      N0 (I+)
## 5                                      No           M0          N2A
## 6                                      No           M0           N0
##   PATH_T_STAGE PERSON_NEOPLASM_CANCER_STATUS
## 1           TX                   With Tumor
## 2           T2                   Tumor Free
## 3           T2                   Tumor Free
## 4          T1C                   Tumor Free
## 5           T2                   Tumor Free
## 6           T2                   Tumor Free
##   PRIMARY_LYMPH_NODE_PRESENTATION_ASSESSMENT PRIOR_DX                    RACE
```

```
## 1                                     Yes       No                     White
## 2                                     Yes       No Black or African American
## 3                                     Yes       No Black or African American
## 4                                     Yes       No Black or African American
## 5                                     Yes       No                     White
## 6                                     Yes      Yes                     White
##   RADIATION_THERAPY WEIGHT IN_PANCANPATHWAYS_FREEZE OS_STATUS OS_MONTHS
## 1                No     NA                      Yes     FALSE 133.05060
## 2               Yes     NA                      Yes     FALSE 131.66979
## 3                No     NA                      Yes     FALSE  48.45974
## 4                No     NA                      Yes     FALSE  47.60496
## 5                No     NA                      Yes     FALSE  11.44097
## 6               Yes     NA                      Yes     FALSE  48.55837
##                  DSS_STATUS DSS_MONTHS            DFS_STATUS DFS_MONTHS
## 1 0:ALIVE OR DEAD TUMOR FREE  133.05060 1:Recurred/Progressed   59.44044
## 2 0:ALIVE OR DEAD TUMOR FREE  131.66979         0:DiseaseFree  131.66979
## 3 0:ALIVE OR DEAD TUMOR FREE   48.45974         0:DiseaseFree   48.45974
## 4 0:ALIVE OR DEAD TUMOR FREE   47.60496                              NA
## 5 0:ALIVE OR DEAD TUMOR FREE   11.44097         0:DiseaseFree   11.44097
## 6 0:ALIVE OR DEAD TUMOR FREE   48.55837                              NA
##      PFS_STATUS PFS_MONTHS GENETIC_ANCESTRY_LABEL Class
## 1 1:PROGRESSION   59.44044                    EUR     1
## 2    0:CENSORED  131.66979                    AFR     2
## 3    0:CENSORED   48.45974              AFR_ADMIX     1
## 4    0:CENSORED   47.60496                    AFR     1
## 5    0:CENSORED   11.44097                    EUR     3
## 6    0:CENSORED   48.55837                            3
```

```
# clinical_df[, 36] <- gsub('^(.).*', '\\1', clinical_df$PFS_STATUS)
# clinical_df[, 36] <- ifelse(clinical_df$PFS_STATUS=='1', TRUE, FALSE)
# head(clinical_df)
```

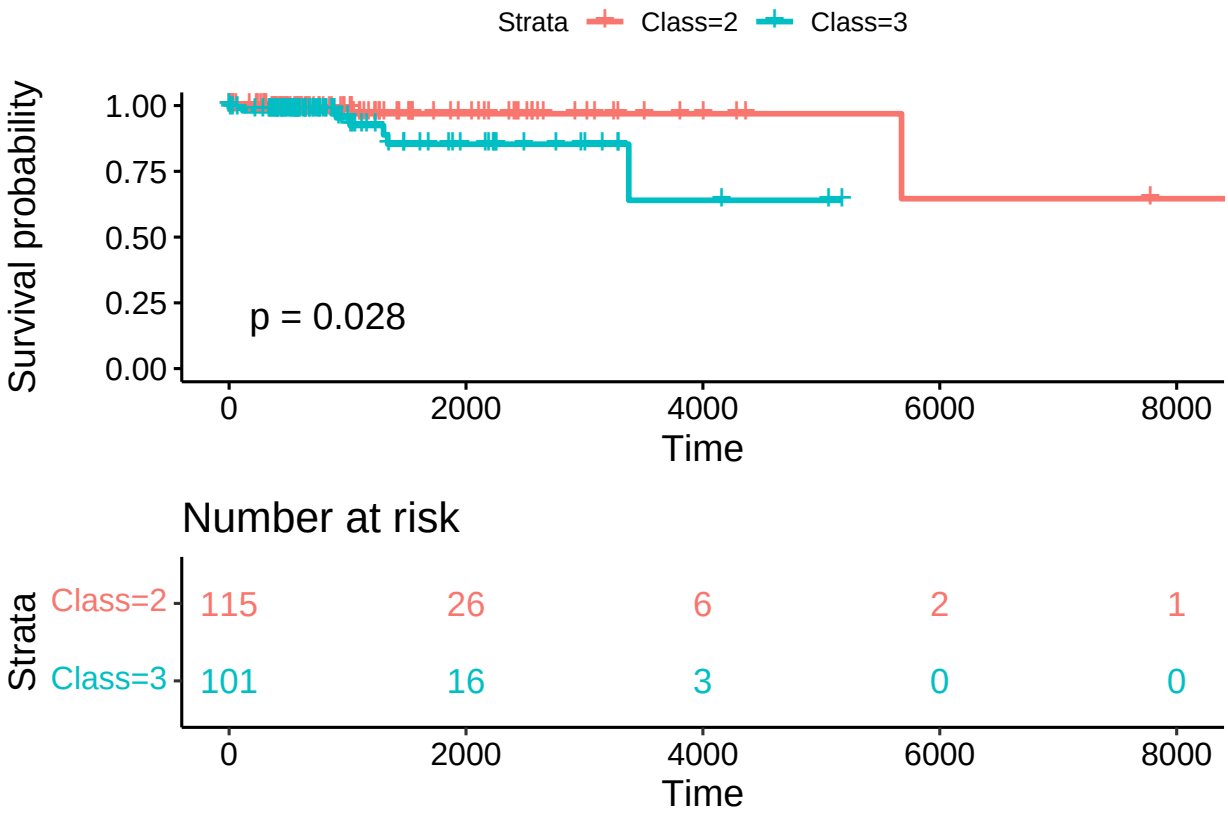
```
clinical_df_subset <- clinical_df[which(clinical_df$Class %in% c(2, 3)), ]
# head(clinical_df_subset)

as.data.frame(table(clinical_df_subset$Class))
```

```
##   Var1 Freq
## 1    2  131
## 2    3  109
```

```
fit <- survfit(Surv(DAYS_LAST_FOLLOWUP, OS_STATUS) ~ Class, data = clinical_df_subset)
print(fit)
```

```
## Call: survfit(formula = Surv(DAYS_LAST_FOLLOWUP, OS_STATUS) ~ Class,
##     data = clinical_df_subset)
##
##    24 observations deleted due to missingness
##           n events median 0.95LCL 0.95UCL
## Class=2 115      3     NA    5677      NA
## Class=3 101      7     NA    3374      NA
```

```
ggsurvplot(fit, data = clinical_df_subset, pval = T, risk.table = T, risk.table.col = "strata",
    risk.table.height = 0.4)
```



DE Analysis

```
countData.Project <- RNAseq
countData.Project_df <- as.data.frame(countData.Project)

colDataMatrix <- matrix(clinical_df_subset$Class)
patient_id <- matrix(clinical_df$PATIENT_ID)

colDataProject <- clinical_df_subset[, c("PATIENT_ID", "Class")]
head(colDataProject)
```

```
##      PATIENT_ID Class
## 2  TCGA-3C-AALI     2
## 5  TCGA-4H-AAAK     3
## 6  TCGA-5L-AAT0     3
## 20 TCGA-A1-A0SP     2
## 23 TCGA-A2-A04P     2
## 27 TCGA-A2-A04U     2
```

```
rownames_data_colData <- rownames(colDataProject)
rownames_data_colData <- colDataProject$PATIENT_ID
rownames(colDataProject) <- rownames_data_colData
colDataProject <- colDataProject[-1]

rownames_data <- rownames(countData.Project)
rownames_data <- countData.Project$X
rownames(countData.Project) <- rownames_data
countData.Project <- countData.Project[-1]
cnames <- colnames(countData.Project)
cnames <- gsub(".{16}$", "", cnames)
cnames <- gsub("\\.", "-", cnames)
colnames(countData.Project) <- cnames
```

```
countData.Project <- sapply(countData.Project, as.numeric)
rownames(countData.Project) <- rownames_data

commonPatients_countData_idx <- which(rownames_data_colData %in% colnames(countData.Project))
commonPatients_countData <- countData.Project[, commonPatients_countData_idx]
commonPatients_countData <- as.data.frame(commonPatients_countData)


col_order <- c(rownames(colDataProject))
countData.Project.Ordered <- countData.Project[, col_order]
commonPatients_countData_idx <- which(rownames_data_colData %in% colnames(countData.Project.Ordered))
commonPatients_countData <- countData.Project.Ordered[, commonPatients_countData_idx]


colDataProject <- data.matrix(colDataProject, rownames.force = rownames_data_colData)
rownames(colDataProject) <- rownames_data_colData
```

```
# colDataProject[['Class']] <- as.factor(colDataProject[['Class']])

colDataProject <- as.data.frame(colDataProject)
colDataProject$Class <- as.factor(colDataProject$Class)
# colDataProject[[Class]]

dds = DESeqDataSetFromMatrix(countData = commonPatients_countData, colData = colDataProject,
    design = ~Class)
```

```
## converting counts to integer mode
```

```
dds = DESeq(dds)
```

```
## estimating size factors

## estimating dispersions

## gene-wise dispersion estimates

## mean-dispersion relationship
```

```
## final dispersion estimates

## fitting model and testing

## -- replacing outliers and refitting for 9624 genes
## -- DESeq argument 'minReplicatesForReplace' = 7
## -- original counts are preserved in counts(dds)

## estimating dispersions

## fitting model and testing
```

```
dds
```

```
## class: DESeqDataSet
## dim: 60660 240
## metadata(1): version
## assays(6): counts mu ... replaceCounts replaceCooks
## rownames(60660): ENSG00000000003.15 ENSG00000000005.6 ...
##   ENSG00000288674.1 ENSG00000288675.1
## rowData names(23): baseMean baseVar ... maxCooks replace
## colnames(240): TCGA-3C-AALI TCGA-4H-AAAK ... TCGA-XX-A89A TCGA-Z7-A8R5
## colData names(3): Class sizeFactor replaceable
```

```
res <- results(dds)
res
```

```
## log2 fold change (MLE): Class 3 vs 2
## Wald test p-value: Class 3 vs 2
## DataFrame with 60660 rows and 6 columns
##                     baseMean log2FoldChange     lfcSE       stat      pvalue
##                    <numeric>      <numeric> <numeric>  <numeric>   <numeric>
## ENSG00000000003.15  3803.9086      -0.386649 0.1241558   -3.11422 1.84430e-03
## ENSG00000000005.6     66.4615       1.686901 0.2968787    5.68212 1.33035e-08
## ENSG00000000419.13  2359.1799      -0.758268 0.0665859  -11.38781 4.80938e-30
## ENSG00000000457.14  1563.6911       0.534840 0.0763157    7.00825 2.41319e-12
## ENSG00000000460.17   825.1319      -0.612438 0.1032444   -5.93193 2.99394e-09
## ...                       ...            ...       ...        ...         ...
## ENSG00000288669.1    0.163025     0.14288816 0.8614397  0.1658713  0.86825820
## ENSG00000288670.1  432.579633     0.00472643 0.0932553  0.0506827  0.95957840
## ENSG00000288671.1    0.000000              NA        NA         NA          NA
## ENSG00000288674.1    9.236421     0.40548661 0.1281080  3.1651944  0.00154979
## ENSG00000288675.1   32.920308    -0.33966613 0.1362861 -2.4923023  0.01269180
##                          padj
##                     <numeric>
## ENSG00000000003.15 4.85427e-03
## ENSG00000000005.6  8.77841e-08
## ENSG00000000419.13 3.40312e-28
## ENSG00000000457.14 2.68098e-11
## ENSG00000000460.17 2.18769e-08
## ...                       ...
## ENSG00000288669.1          NA
```

```
## ENSG00000288670.1   0.97197822
## ENSG00000288671.1           NA
## ENSG00000288674.1   0.00414861
## ENSG00000288675.1   0.02730046
```

```
summary(res)
```

```
##
## out of 57078 with nonzero total read count
## adjusted p-value < 0.1
## LFC > 0 (up)       : 9374, 16%
## LFC < 0 (down)     : 14573, 26%
## outliers [1]       : 0, 0%
## low counts [2]     : 14355, 25%
## (mean count < 0)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

p-values and adjusted p-values

```
res.05 <- results(dds, alpha = 0.05)
table(res.05$padj < 0.05)
```

```
##
## FALSE  TRUE
## 18810 21723
```

```
resLFC1 <- results(dds, lfcThreshold = 1)
table(resLFC1$padj < 0.1)
```

```
##
## FALSE  TRUE
## 37012  2419
```

```
res <- res[order(res$pvalue), ]
summary(res)
```

```
##
## out of 57078 with nonzero total read count
## adjusted p-value < 0.1
## LFC > 0 (up)       : 9374, 16%
## LFC < 0 (down)     : 14573, 26%
## outliers [1]       : 0, 0%
## low counts [2]     : 14355, 25%
## (mean count < 0)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

```r
sum(res$padj < 0.1, na.rm = TRUE)
```

```
## [1] 23947
```

Multiple Testing

```r
sum(res$pvalue < 0.05, na.rm = TRUE)
```

```
## [1] 23549
```

```r
sum(!is.na(res$pvalue))
```

```
## [1] 57094
```

```r
sum(res$padj < 0.06, na.rm = TRUE)
```

```
## [1] 22141
```

```r
resSig <- subset(res, padj < 0.06)
head(resSig[order(resSig$log2FoldChange), ])
```

```
## log2 fold change (MLE): Class 3 vs 2
## Wald test p-value: Class 3 vs 2
## DataFrame with 6 rows and 6 columns
##                     baseMean log2FoldChange     lfcSE      stat      pvalue
##                    <numeric>      <numeric> <numeric> <numeric>   <numeric>
## ENSG00000204019.5   171.2979       -9.47260  0.632924  -14.9664 1.21715e-50
## ENSG00000147381.11  365.1390       -8.43767  0.701479  -12.0284 2.52035e-33
## ENSG00000275216.2    82.9871       -8.42500  0.588804  -14.3087 1.93131e-46
## ENSG00000213401.10  113.4469       -8.37587  0.694486  -12.0605 1.70651e-33
## ENSG00000239311.1    15.9034       -8.31479  0.786492  -10.5720 4.01827e-26
## ENSG00000197172.10  342.1113       -8.19660  0.772792  -10.6065 2.78034e-26
##                          padj
##                     <numeric>
## ENSG00000204019.5  6.50248e-48
## ENSG00000147381.11 2.47058e-31
## ENSG00000275216.2  6.55099e-44
## ENSG00000213401.10 1.71610e-31
## ENSG00000239311.1  1.95823e-24
## ENSG00000197172.10 1.37534e-24
```

```r
# 20 genes with largest positive log2fold change
genes_up <- order(resSig$log2FoldChange, decreasing = TRUE)[1:10]

# 20 genes with largest negative log2fold change
genes_down <- order(resSig$log2FoldChange, decreasing = FALSE)[1:10]

# combine top 20 upregulated and top 20 downregulated genes
```

```r
combined_genes <- c(genes_up, genes_down)

# or select top 20 significant genes genes_20_sig <- order(reSig$padj,
# decreasing = TRUE)[1:20]
```

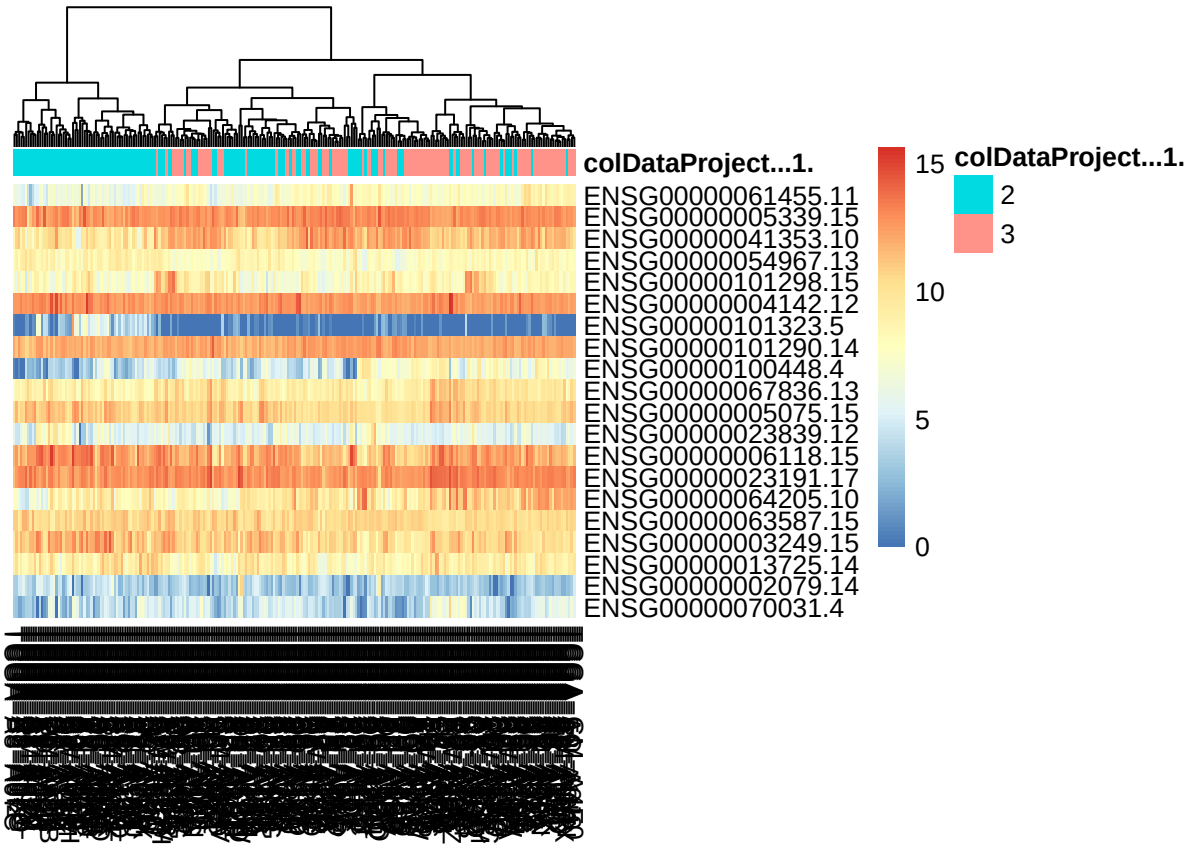
```r
# Top down regulated genes
head(resSig[order(resSig$log2FoldChange), ])
```

```
## log2 fold change (MLE): Class 3 vs 2
## Wald test p-value: Class 3 vs 2
## DataFrame with 6 rows and 6 columns
##                      baseMean log2FoldChange     lfcSE      stat      pvalue
##                     <numeric>      <numeric> <numeric> <numeric>   <numeric>
## ENSG00000204019.5   171.2979       -9.47260  0.632924  -14.9664 1.21715e-50
## ENSG00000147381.11  365.1390       -8.43767  0.701479  -12.0284 2.52035e-33
## ENSG00000275216.2    82.9871       -8.42500  0.588804  -14.3087 1.93131e-46
## ENSG00000213401.10  113.4469       -8.37587  0.694486  -12.0605 1.70651e-33
## ENSG00000239311.1    15.9034       -8.31479  0.786492  -10.5720 4.01827e-26
## ENSG00000197172.10  342.1113       -8.19660  0.772792  -10.6065 2.78034e-26
##                          padj
##                     <numeric>
## ENSG00000204019.5  6.50248e-48
## ENSG00000147381.11 2.47058e-31
## ENSG00000275216.2  6.55099e-44
## ENSG00000213401.10 1.71610e-31
## ENSG00000239311.1  1.95823e-24
## ENSG00000197172.10 1.37534e-24
```

```r
# Top up regulated genes
head(resSig[order(resSig$log2FoldChange, decreasing = TRUE), ])
```

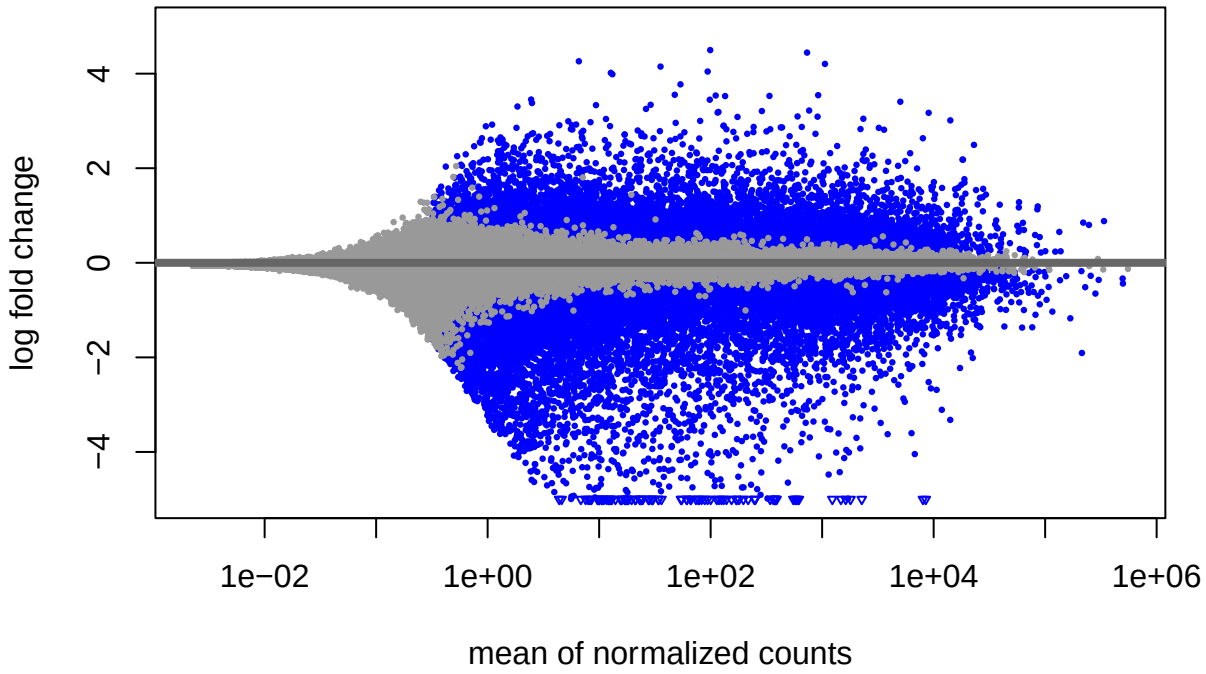
```
## log2 fold change (MLE): Class 3 vs 2
## Wald test p-value: Class 3 vs 2
## DataFrame with 6 rows and 6 columns
##                      baseMean log2FoldChange     lfcSE      stat      pvalue
##                     <numeric>      <numeric> <numeric> <numeric>   <numeric>
## ENSG00000249203.2    99.16309        4.49783  0.419975  10.70976 9.16014e-27
## ENSG00000108576.10  732.18197        4.44492  0.300045  14.81415 1.18682e-49
## ENSG00000286935.1     6.58015        4.26132  0.371299  11.47678 1.72598e-30
## ENSG00000141668.10 1060.08092        4.20550  0.384442  10.93922 7.48376e-28
## ENSG00000180269.8    35.52170        4.14879  0.511819   8.10597 5.23272e-16
## ENSG00000188869.13   94.06559        4.04477  0.257209  15.72562 1.00961e-55
##                          padj
##                     <numeric>
## ENSG00000249203.2  4.72250e-25
## ENSG00000108576.10 5.39610e-47
## ENSG00000286935.1  1.28289e-28
## ENSG00000141668.10 4.32228e-26
## ENSG00000180269.8  8.96901e-15
## ENSG00000188869.13 8.80609e-53
```

```r
# this gives log2(n + 1)
ntd <- normTransform(dds)
# Variance stabilizing transformation
vsd <- vst(dds)

#colDataProject_factor <- as.factor(colDataProject)

annot_col = data.frame(colDataProject[,1])
row.names(annot_col) <- rownames(colDataProject)

sampleDistMatrix = assay(ntd)[combined_genes, ]

rownames(sampleDistMatrix) = rownames(commonPatients_countData[combined_genes, ])
colnames(sampleDistMatrix) = colnames(commonPatients_countData)

pheatmap(sampleDistMatrix,
         cluster_rows = FALSE,
         show_rownames = TRUE,
         #show_colnames = FALSE,
         cluster_cols = TRUE,
         annotation_col = annot_col,
         clustering_method = "ward.D2")
```



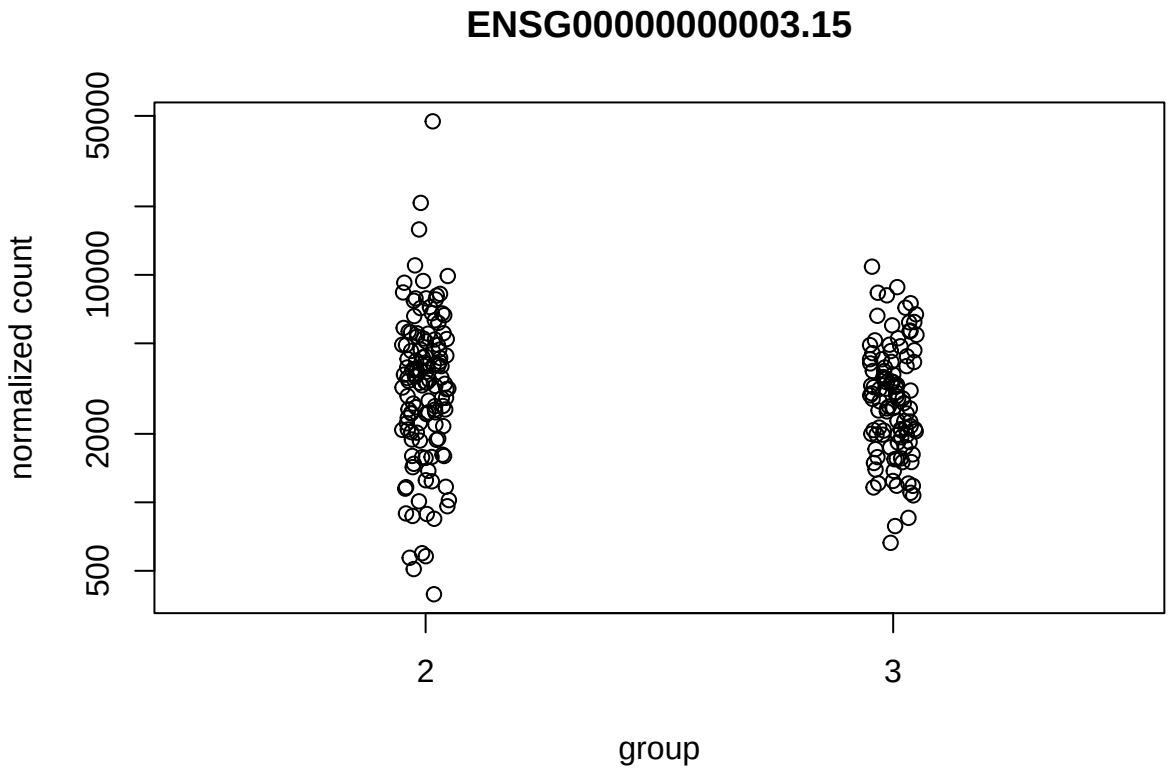
MA-plot

```
plotMA(res, ylim = c(-5, 5))
```



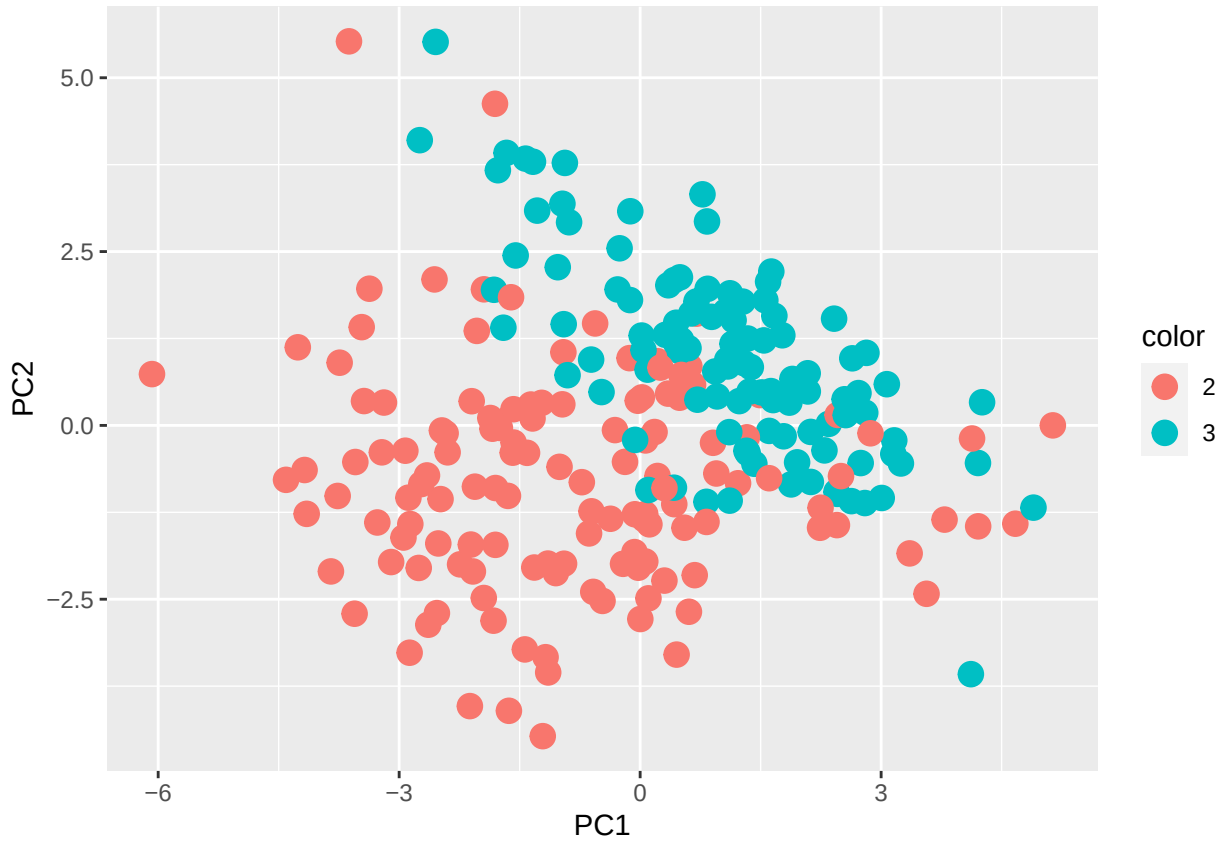
Plot Counts

```
plotCounts(dds, gene = which.min(res$padj), intgroup = "Class")
```



PCA Plot

```
# run on both ntd and vsd
pca_res <- prcomp(t(assay(ntd)[combined_genes, ]), scale. = TRUE)
score <- pca_res$x

score = as.data.frame(score)
score$color <- as.factor(colDataProject[, 1])


ggplot(score, aes(x = PC1, y = PC2, color = color)) + geom_point(size = 4)
```



Pathway Analysis

```
library(AnnotationDbi)
```

```
## Warning: package 'AnnotationDbi' was built under R version 4.3.2
```

```
##
## Attaching package: 'AnnotationDbi'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(org.Hs.eg.db)
```

```
##
```

```
ensembl_ids <- sub("\\..*$", "", row.names(res))
```

```
res$symbol = mapIds(org.Hs.eg.db, keys = ensembl_ids, column = "SYMBOL", keytype = "ENSEMBL",
    multiVals = "first")
```

```
## 'select()' returned 1:many mapping between keys and columns
```

```
# Now use mapIds with the modified Ensembl IDs
res$entrez <- mapIds(org.Hs.eg.db, keys = ensembl_ids, column = "ENTREZID", keytype = "ENSEMBL",
    multiVals = "first")
```

```
## 'select()' returned 1:many mapping between keys and columns
```

```
res$name = mapIds(org.Hs.eg.db, keys = ensembl_ids, column = "GENENAME", keytype = "ENSEMBL",
    multiVals = "first")
```

```
## 'select()' returned 1:many mapping between keys and columns
```

```
head(res, 10)
```

```
## log2 fold change (MLE): Class 3 vs 2
## Wald test p-value: Class 3 vs 2
## DataFrame with 10 rows and 9 columns
##                     baseMean log2FoldChange     lfcSE      stat       pvalue
##                    <numeric>      <numeric> <numeric> <numeric>    <numeric>
## ENSG00000143452.16   383.113       -7.09849  0.315844  -22.4747 7.33798e-112
## ENSG00000156219.17   582.833       -6.58493  0.299589  -21.9799 4.48416e-107
## ENSG00000203688.6    383.918       -6.58580  0.306868  -21.4613 3.57848e-102
## ENSG00000102854.16  2267.676       -6.46215  0.316825  -20.3966  1.79156e-92
## ENSG00000135069.14  1646.752       -3.95761  0.200985  -19.6910  2.57323e-86
## ENSG00000166535.20  1797.479       -5.85571  0.300930  -19.4587  2.45904e-84
## ENSG00000107159.13   551.394       -5.59023  0.294623  -18.9742  2.78713e-80
## ENSG00000189001.11   383.310       -6.61228  0.350331  -18.8744  1.85181e-79
## ENSG00000105173.14   732.624       -3.10242  0.167565  -18.5147  1.57142e-76
## ENSG00000163064.7   1437.923       -4.42679  0.242924  -18.2229  3.39608e-74
##                           padj      symbol      entrez                  name
##                      <numeric> <character> <character>           <character>
## ENSG00000143452.16 3.13618e-107     HORMAD1       84072 HORMA domain contain..
## ENSG00000156219.17 9.58243e-103        ART3         419 ADP-ribosyltransfera..
## ENSG00000203688.6   5.09802e-98   LINC02487      441178 long intergenic non-..
## ENSG00000102854.16  1.91424e-88        MSLN       10232             mesothelin
## ENSG00000135069.14  2.19954e-82       PSAT1       29968 phosphoserine aminot..
## ENSG00000166535.20  1.75162e-80       A2ML1      144568 alpha-2-macroglobuli..
## ENSG00000107159.13  1.70170e-76         CA9         768   carbonic anhydrase 9
## ENSG00000189001.11  9.89306e-76        SBSN      374897             suprabasin
## ENSG00000105173.14  7.46233e-73       CCNE1         898              cyclin E1
## ENSG00000163064.7   1.45145e-70         EN1        2019   engrailed homeobox 1
```

```
library(pathview)
```

```
## ##############################################################################
## Pathview is an open source software package distributed under GNU General
## Public License version 3 (GPLv3). Details of GPLv3 is available at
## http://www.gnu.org/licenses/gpl-3.0.html. Particullary, users are required to
## formally cite the original Pathview paper (not just mention it) in publications
## or products. For details, do citation("pathview") within R.
```

```
##
## The pathview downloads and uses KEGG data. Non-academic uses may require a KEGG
## license agreement (details at http://www.kegg.jp/kegg/legal.html).
## #############################################################################
```

```
library("gage")
```

```
##
```

```
library(gageData)


data(kegg.sets.hs)
data(sigmet.idx.hs)

# Focus on signaling and metabolic pathways only
kegg.sets.hs = kegg.sets.hs[sigmet.idx.hs]


foldchanges = res$log2FoldChange
names(foldchanges) = res$entrez
head(foldchanges)
```

```
##     84072       419    441178     10232     29968    144568
## -7.098494 -6.584935 -6.585796 -6.462153 -3.957608 -5.855715
```

```
keggres = gage(foldchanges, gsets = kegg.sets.hs)


attributes(keggres)
```

```
## $names
## [1] "greater" "less"    "stats"
```

```
head(keggres$less)
```

```
##                                                    p.geomean stat.mean
## hsa04110 Cell cycle                              3.808599e-06 -4.573557
## hsa04612 Antigen processing and presentation     2.056645e-03 -2.927925
## hsa04650 Natural killer cell mediated cytotoxicity 2.830189e-03 -2.790122
## hsa03050 Proteasome                              2.830269e-03 -2.863305
## hsa03008 Ribosome biogenesis in eukaryotes       3.084227e-03 -2.797932
## hsa03030 DNA replication                         3.622777e-03 -2.784680
##                                                          p.val        q.val
## hsa04110 Cell cycle                              3.808599e-06 0.0006246102
## hsa04612 Antigen processing and presentation     2.056645e-03 0.0990225669
## hsa04650 Natural killer cell mediated cytotoxicity 2.830189e-03 0.0990225669
## hsa03050 Proteasome                              2.830269e-03 0.0990225669
## hsa03008 Ribosome biogenesis in eukaryotes       3.084227e-03 0.0990225669
## hsa03030 DNA replication                         3.622777e-03 0.0990225669
##                                                  set.size         exp1
```

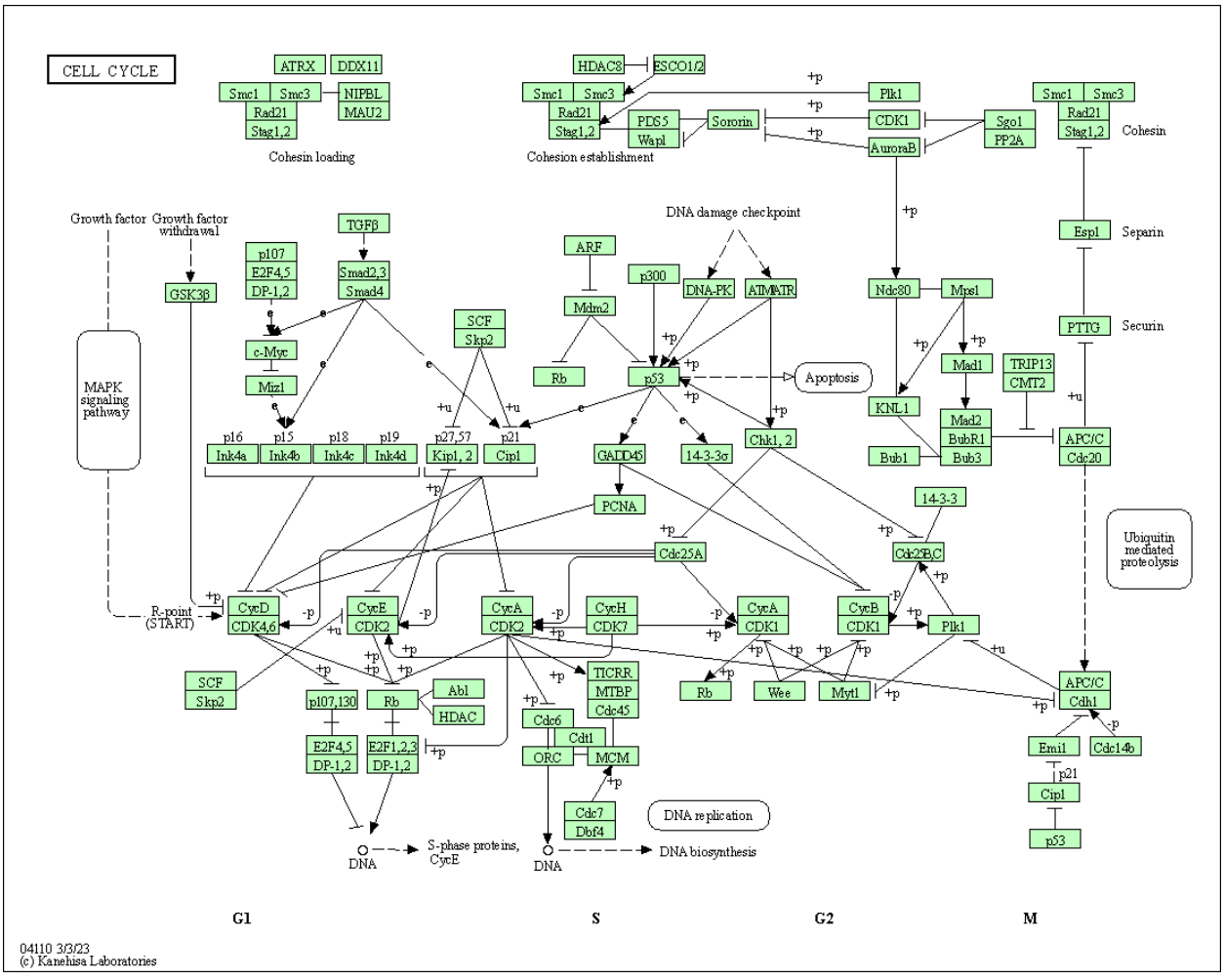
```
## hsa04110 Cell cycle                                          124 3.808599e-06
## hsa04612 Antigen processing and presentation                  68 2.056645e-03
## hsa04650 Natural killer cell mediated cytotoxicity           131 2.830189e-03
## hsa03050 Proteasome                                           44 2.830269e-03
## hsa03008 Ribosome biogenesis in eukaryotes                    73 3.084227e-03
## hsa03030 DNA replication                                      36 3.622777e-03
```

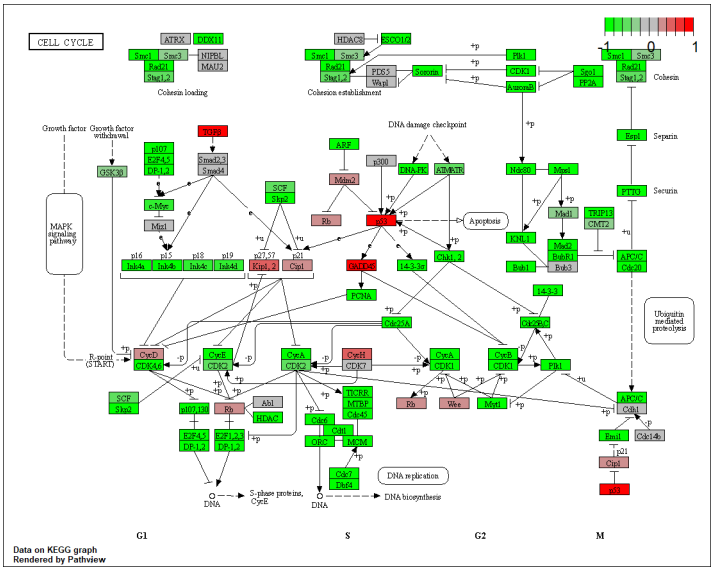
```
pathview(gene.data = foldchanges, pathway.id = "hsa04110")
```

```
## 'select()' returned 1:1 mapping between keys and columns
```

```
## Info: Working in directory C:/Users/justi/OneDrive/Documents/R_Projects/Project
```

```
## Info: Writing image file hsa04110.pathview.png
```





Code Contributions

Evan: coded data formatting & differential expression, coded gene annotation, coded pathway analysis, wrote methods sections for those portions of code, wrote results section for figures generated in those code sections, made slides for said portions

Justin: coded MAF summary, coded oncoplot generation, coded survival analysis KM plots, created pipeline summary diagram in draw.io, wrote methods sections for said code portions, wrote results for figures from said portions, made slides for said portions

Andrew: