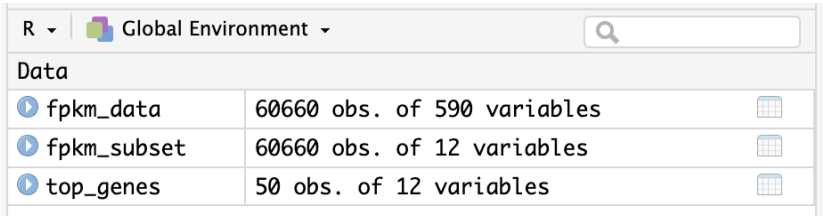
Syed Taqi Hasnain
BIOL 4324

WORKBOOK (R Studio) for LUAD Project

```
> fpkm_data <- read.delim("TCGA-LUAD.star_fpkm.tsv", header = TRUE, sep = "\t")
> fpkm_subset <- fpkm_data[, c(1, 2:11)]
> fpkm_subset$mean_expression <- rowMeans(fpkm_subset[,-1])
> top_genes <- fpkm_subset[order(-fpkm_subset$mean_expression), ][1:50, ]
> write.table(top_genes, "subset_top_genes.tsv", sep="\t", row.names=FALSE)
```

The FPKM data is already processed, but I needed to process this further for the sake of a more manageable dataset and to later identify genes of interest based on mean expression and variance.



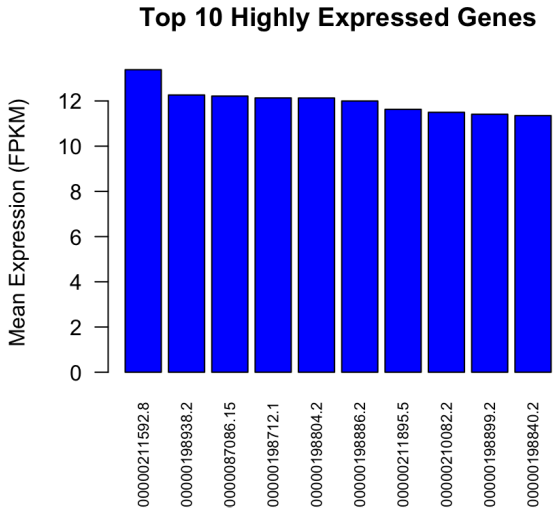

```
> str(top_genes)
'data.frame':   50 obs. of  12 variables:
 $ Ensembl_ID      : chr  "ENSG00000211592.8" "ENSG00000198938.2" "ENSG00000087086.15" "ENSG00000198712.1" ...
 $ TCGA.38.7271.01A: num  14.5 13.4 12.4 13.2 12.9 ...
 $ TCGA.55.7914.01A: num  14.9 11.4 11.4 11.8 11.5 ...
 $ TCGA.95.7043.01A: num  11.7 13.8 15.5 13.2 13.4 ...
 $ TCGA.73.4658.01A: num  12.7 12.1 12.4 11.9 12.2 ...
 $ TCGA.86.8076.01A: num  15.3 12.5 12 12.3 12.7 ...
 $ TCGA.55.7726.01A: num  13.1 13.2 10.8 13.5 13.3 ...
 $ TCGA.44.6147.01A: num  14.1 11.5 10.9 11.4 11.3 ...
 $ TCGA.50.5932.01A: num  11.1 12 10.9 11.6 10.6 ...
 $ TCGA.44.2661.01A: num  15.3 12 13.8 11.6 12.1 ...
 $ TCGA.86.7954.01A: num  11.1 10.7 12 10.8 11.3 ...
 $ mean_expression : num  13.4 12.3 12.2 12.1 12.1 ...
```

```
> head(top_genes)
           Ensembl_ID TCGA.38.7271.01A TCGA.55.7914.01A TCGA.95.7043.01A TCGA.73.4658.01A TCGA.86.8076.01A TCGA.55.7726.01A
21485  ENSG00000211592.8         14.53389         14.90610         11.72588         12.67981         15.30422         13.12776
17956  ENSG00000198938.2         13.42215         11.35158         13.78447         12.13677         12.51372         13.24363
1720  ENSG00000087086.15         12.42268         11.39785         15.51204         12.38996         11.99860         10.80127
17826  ENSG00000198712.1         13.21480         11.84423         13.17352         11.92855         12.28060         13.52371
17875  ENSG00000198804.2         12.93722         11.52541         13.40091         12.16531         12.68033         13.31597
17924  ENSG00000198886.2         13.10835         11.56583         13.55003         11.84267         11.98073         13.55234
      TCGA.44.6147.01A TCGA.50.5932.01A TCGA.44.2661.01A TCGA.86.7954.01A mean_expression
21485         14.05735         11.11281         15.27207         11.09620        13.38161
17956         11.48627         12.04243         11.95360         10.70312        12.26377
1720          10.94672         10.93804         13.76703         11.98552        12.21597
17826         11.37584         11.56307         11.62669         10.80050        12.13315
17875         11.31737         10.55627         12.07371         11.33371        12.13062
17924         10.97152         11.41008         11.23396         10.77159        11.99871
```

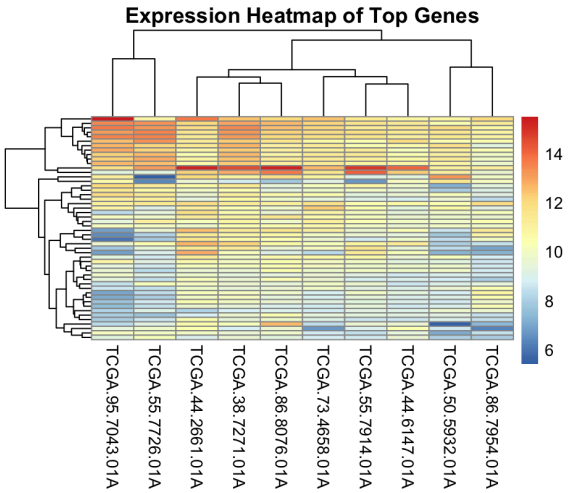
```
> summary(top_genes)
  Ensembl_ID        TCGA.38.7271.01A TCGA.55.7914.01A TCGA.95.7043.01A TCGA.73.4658.01A TCGA.86.8076.01A TCGA.55.7726.01A
 Length:50          Min.   : 8.647   Min.   : 6.644   Min.   : 5.940   Min.   : 6.677   Min.   : 7.772   Min.   : 5.415
 Class :character   1st Qu.: 9.795   1st Qu.: 9.315   1st Qu.: 8.313   1st Qu.: 9.453   1st Qu.: 9.634   1st Qu.: 8.658
 Mode  :character   Median :10.565   Median :10.017   Median : 9.842   Median :10.171   Median :10.503   Median : 9.686
                    Mean   :10.903   Mean   :10.180   Mean   :10.077   Mean   :10.281   Mean   :10.606   Mean   :10.073
                    3rd Qu.:12.268   3rd Qu.:10.917   3rd Qu.:11.850   3rd Qu.:11.222   3rd Qu.:11.271   3rd Qu.:11.289
                    Max.   :14.534   Max.   :14.906   Max.   :15.512   Max.   :12.680   Max.   :15.304   Max.   :13.552
 TCGA.44.6147.01A TCGA.50.5932.01A TCGA.44.2661.01A TCGA.86.7954.01A mean_expression
 Min.   : 8.600   Min.   : 5.466   Min.   : 8.061   Min.   : 6.232   Min.   : 8.919
 1st Qu.: 9.294   1st Qu.: 8.705   1st Qu.: 9.991   1st Qu.: 8.661   1st Qu.: 9.280
 Median : 9.739   Median : 9.755   Median :10.634   Median : 9.733   Median : 9.921
 Mean   :10.004   Mean   : 9.649   Mean   :10.859   Mean   : 9.567   Mean   :10.220
 3rd Qu.:10.623   3rd Qu.:10.884   3rd Qu.:11.576   3rd Qu.:10.496   3rd Qu.:11.254
 Max.   :14.057   Max.   :13.288   Max.   :15.272   Max.   :11.986   Max.   :13.382
```

Once the data was organized, I could begin generating plots

```
> barplot(top_genes$mean_expression[1:10],
+         names.arg = top_genes$Ensembl_ID[1:10], # Use Ensembl_ID for labels
+         las = 2, # Rotate labels for better readability
+         col = "blue",
+         main = "Top 10 Highly Expressed Genes",
+         ylab = "Mean Expression (FPKM)",
+         cex.names = 0.7) # Adjust label size for clarity
```




```
# Subset only expression data
expression_matrix <- as.matrix(top_genes[, 2:11])
rownames(expression_matrix) <- top_genes$Gene
# Create heatmap
pheatmap(expression_matrix,
         cluster_rows = TRUE,
         cluster_cols = TRUE,
         main = "Expression Heatmap of Top Genes")
```

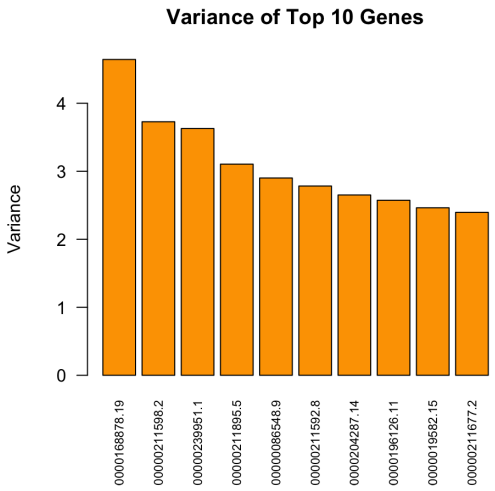


```
> filtered_genes <- top_genes[top_genes$mean_expression > 5, ]
> variances <- apply(top_genes[, 2:11], 1, var)
> top_genes$variance <- variances
> top_genes_high_var <- top_genes[order(-top_genes$variance), ][1:10, ]
```

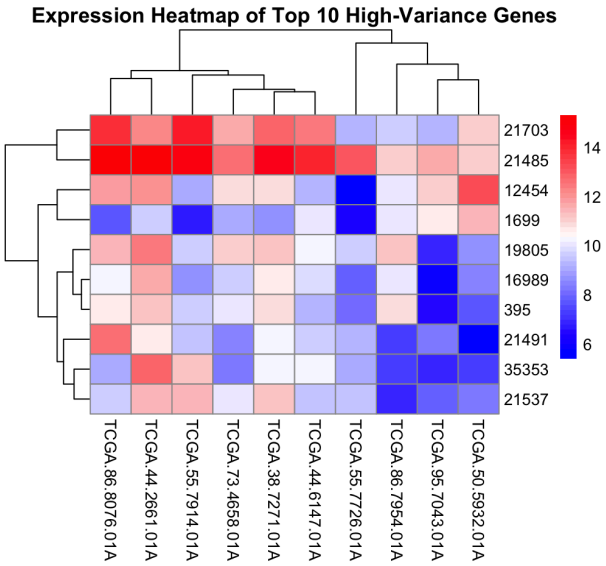
Data	
expression_matr…	num [1:50, 1:10] 14.5 13.4 12.4 13.2 12.…
filtered_genes	50 obs. of 12 variables
fpkm_data	60660 obs. of 590 variables
fpkm_subset	60660 obs. of 12 variables
top_genes	50 obs. of 13 variables
top_genes_high_…	10 obs. of 13 variables
Values	
variances	Named num [1:50] 2.783 0.968 2.151 0.807 0…

```
> print(top_genes_high_var)
             Ensembl_ID TCGA.38.7271.01A TCGA.55.7914.01A TCGA.95.7043.01A TCGA.73.4658.01A TCGA.86.8076.01A TCGA.55.7726.01A
12454 ENSG00000168878.19        10.901415         9.072152        11.081571        10.934702        11.758277         5.414677
21491  ENSG00000211598.2        10.190254         9.555561         8.272099         8.404857        12.585684         9.213056
35353  ENSG00000239951.1        10.219488        11.214584         6.895419         8.305698         9.107754         9.027108
21703  ENSG00000211895.5        12.876462        14.123911         9.327337        11.640165        13.724869         9.311102
1699   ENSG00000086548.9         8.647118         6.644100        10.633260         9.167422         7.771895         6.341890
21485  ENSG00000211592.8        14.533895        14.906104        11.725882        12.679814        15.304218        13.127755
19805 ENSG00000204287.14        11.214218         9.664223         6.922632        11.053895        11.524155         9.603675
16989 ENSG00000196126.11        10.581406         8.728550         5.939509         9.739465        10.265252         7.938647
395   ENSG00000019582.15        10.787223         9.586135         6.452622        10.042574        10.695163         8.118072
21537  ENSG00000211677.2        11.309188        11.474366         7.878378        10.130347         9.577265         9.530984
      TCGA.44.6147.01A TCGA.50.5932.01A TCGA.44.2661.01A TCGA.86.7954.01A mean_expression variance
12454         9.341089        13.287966        12.053102        10.027950        10.387290 4.643844
21491         9.616003         5.465726        10.598625         7.336605         9.123847 3.727952
35353        10.200419         7.253435        12.794326         7.273934         9.229217 3.628611
21703        12.372033        10.989549        12.235449         9.687246        11.628812 3.104780
1699         10.095252        11.520310         9.633642        10.115297         9.057019 2.900952
21485        14.057350        11.112807        15.272074        11.096202        13.381610 2.783077
19805        10.314937         8.594154        12.487310        11.289700        10.266890 2.651267
16989         9.787029         8.427738        11.639217         9.970552         9.301737 2.573241
395           9.366123         7.745210        11.186546        10.783030         9.476270 2.462843
21537         9.500035         8.358307        11.425899         6.955571         9.614034 2.395760
```

```
# Bar plot of variances for the top 10 high-variance genes
barplot(top_genes_high_var$variance,
        names.arg = top_genes_high_var$Ensembl_ID,
        las = 2, col = "orange",
        main = "Variance of Top 10 Genes",
        ylab = "Variance",
        cex.names = 0.7)  # Adjust font size for the labels
```




```
# Generate a heatmap
pheatmap(high_var_matrix,
        cluster_rows = TRUE,
        cluster_cols = TRUE,
        main = "Expression Heatmap of Top 10 High-Variance Genes",
        color = colorRampPalette(c("blue", "white", "red"))(50))
```



```
> top_genes_mean <- top_genes[order(-top_genes$mean_expression), ]
> top_3_by_mean <- top_genes_mean[1:3, ]
> print(top_3_by_mean)
             Ensembl_ID TCGA.38.7271.01A TCGA.55.7914.01A TCGA.95.7043.01A TCGA.73.4658.01A TCGA.86.8076.01A TCGA.55.7726.01A
21485  ENSG00000211592.8         14.53389         14.90610         11.72588         12.67981         15.30422         13.12776
17956  ENSG00000198938.2         13.42215         11.35158         13.78447         12.13677         12.51372         13.24363
1720  ENSG00000087086.15         12.42268         11.39785         15.51204         12.38996         11.99860         10.80127
      TCGA.44.6147.01A TCGA.50.5932.01A TCGA.44.2661.01A TCGA.86.7954.01A mean_expression  variance
21485         14.05735         11.11281         15.27207         11.09620        13.38161 2.7830769
17956         11.48627         12.04243         11.95360         10.70312        12.26377 0.9678355
1720          10.94672         10.93804         13.76703         11.98552        12.21597 2.1508909
> top_genes_variance <- top_genes[order(-top_genes$variance), ]
> top_3_by_variance <- top_genes_variance[1:3, ]
> print(top_3_by_variance)
             Ensembl_ID TCGA.38.7271.01A TCGA.55.7914.01A TCGA.95.7043.01A TCGA.73.4658.01A TCGA.86.8076.01A TCGA.55.7726.01A
12454 ENSG00000168878.19         10.90142         9.072152        11.081571        10.934702        11.758277         5.414677
21491  ENSG00000211598.2         10.19025         9.555561         8.272099         8.404857        12.585684         9.213056
35353  ENSG00000239951.1         10.21949        11.214584         6.895419         8.305698         9.107754         9.027108
      TCGA.44.6147.01A TCGA.50.5932.01A TCGA.44.2661.01A TCGA.86.7954.01A mean_expression variance
12454         9.341089        13.287966         12.05310        10.027950       10.387290 4.643844
21491         9.616003         5.465726         10.59863         7.336605        9.123847 3.727952
35353        10.200419         7.253435         12.79433         7.273934        9.229217 3.628611
```

TOP 3 Genes by Mean Expression

- 21485 ENSG00000211592.8
- 17956 ENSG00000198938.2
- 1720 ENSG00000087086.15

TOP 3 Genes by Variance

- 12454 ENSG00000168878.19
- 21491 ENSG00000211598.2
- 35353 ENSG00000239951.1

Next steps involved doing research on these identified genes and continue adding to project paper