Introduction to Bioinformatics (BIOL 3600) - Project

Omar Moustafa (900222400)

December 11, 2025

Summary of the Original Question and Experiment of the Paper

The original question worked to investigate how gene-expression profiles vary within and among humans and African great apes, with the objective of understanding transcriptional regulatory differences that may contribute to species-specific traits. While genome sequences of humans and different great apes (chimpanzees, bonobos, and gorillas) are well-known, the same can not be said about how their gene-expression patterns differ across comparable cell types. Since gene regulation plays a significant role in shaping complex biological characteristics, the authors worked and aimed to systematically measure these expression differences.

To fulfill this objective, the researchers analyzed genome-wide expression patterns from primary fibroblast cell lines, using 39 samples; 18 human samples and 21 African great ape samples. After comparing these well-controlled fibroblast profiles, they ended up finding that gene-expression patterns could reliably predict the species of the fibroblast donor. They also identified several groups of differentially expressed genes and pathways linked to inherited overgrowth or neurological disorders.

Therefore, the experiment of the paper provides valuable insights into how transcriptional regulation may have played a major role in the rise of species-specific molecular adaptations in humans and African great apes. Building on this, this project also focuses on and thoroughly analyzes a dataset consisting of three human samples and three gorilla fibroblast samples, to identify the true and existing relationships between the two species.

Loading the Necessary Packages

```
library(limma)
library(pagedown)
library(AnnotationDbi)
library(hgu95av2.db)
library(annaffy)
library(pheatmap)
library(gplots)
library(ggplot2)
library(RColorBrewer)
library(edgeR)
library(printr)
```

Loading the Dataset

```
df = read.table("test.tsv", header = T, sep = "\t", row.names = 1, check.names = F)

# First three rows of the dataset
head(df, 3)
```

	H1	H2	H3	G1	G2	G3
100_g_at	6531.0	5562.8	6822.4	7732.1	7191.2	7551.9
1000_at	11486.3	10542.7	10641.4	10408.2	9484.5	7650.2
1001_at	14339.2	13526.1	14444.7	12936.6	13841.7	13285.7

```
# convert the data.frame (table) to a matrix (numeric)
df = as.matrix(df)
```

```
# The dimensions of the dataset
my_dim = dim(df)
number_of_genes = my_dim[1] # The number of genes
number_of_samples = my_dim[2]
cat("Number of Genes:", number_of_genes, "\n")
```

```
## Number of Genes: 4497
```

```
cat("Number of Samples:", number_of_samples, "\n")
```

```
## Number of Samples: 6
```

```
# The ids of the genes are the names of the rows
ids = row.names(df)
head(df)
```

	H1	H2	H3	G1	G2	G3
100_g_at	6531.0	5562.8	6822.4	7732.1	7191.2	7551.9
1000_at	11486.3	10542.7	10641.4	10408.2	9484.5	7650.2
1001_at	14339.2	13526.1	14444.7	12936.6	13841.7	13285.7
1002_f_at	3156.8	2219.5	3264.4	2374.2	2201.8	2525.3
1003_s_at	4002.0	3306.9	3777.0	3760.6	3137.0	2911.5
1004_at	3468.4	3347.4	3332.9	3073.5	3046.0	2914.4

```
sample_names = colnames(df)
print(sample_names)
```

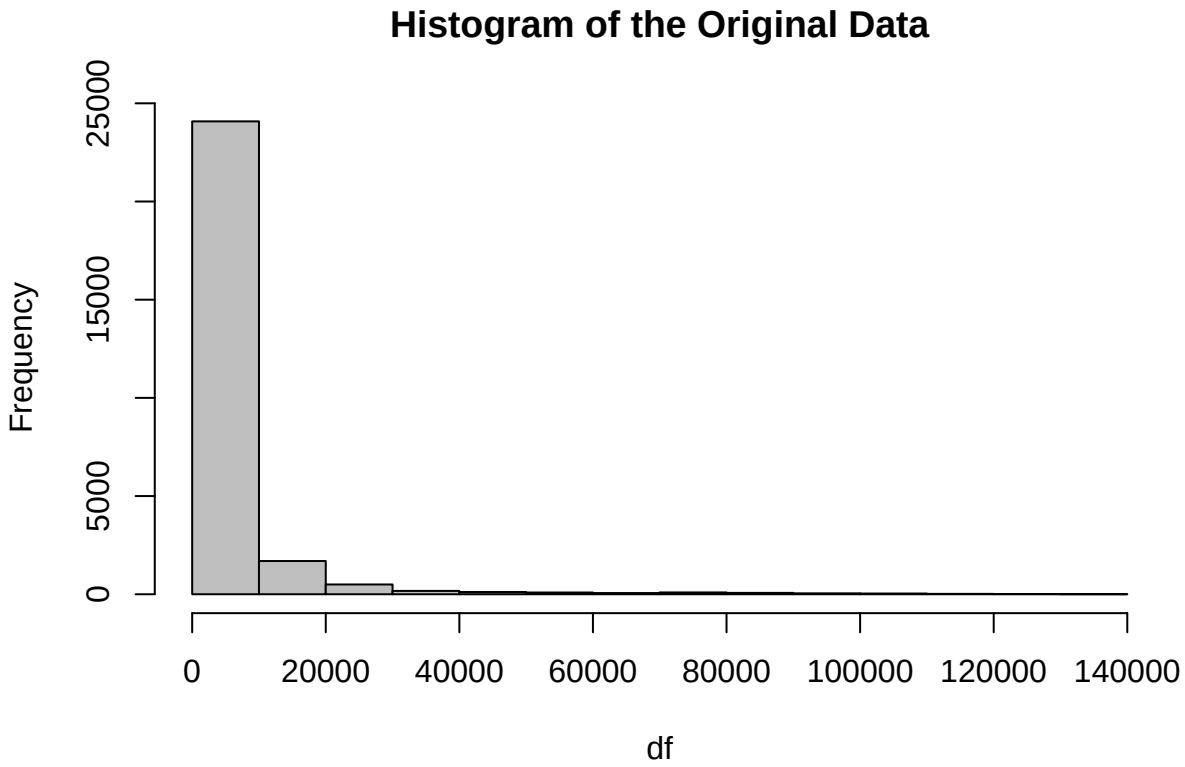
```
## [1] "H1" "H2" "H3" "G1" "G2" "G3"
```

```
group = factor(c("Human", "Human", "Human", "Gorilla", "Gorilla", "Gorilla"))
print(group)
```

```
## [1] Human   Human   Human   Gorilla Gorilla Gorilla
## Levels: Gorilla Human
```

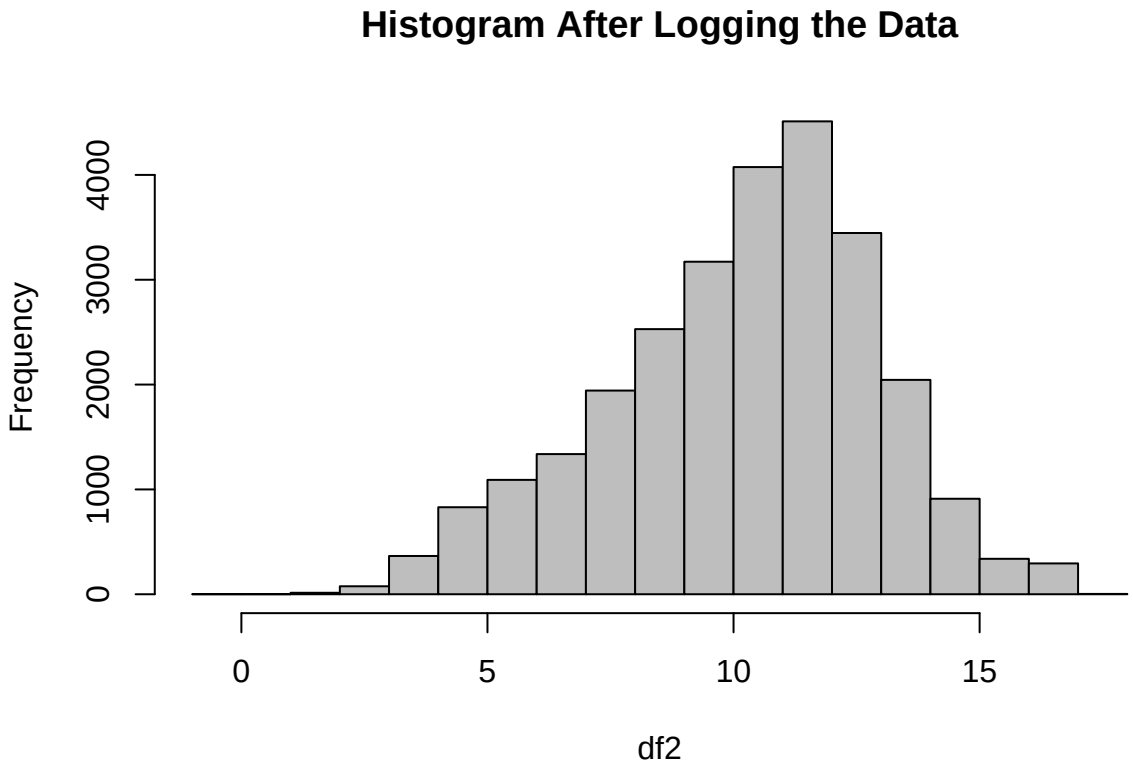
Exploring the Dataset

```
# Histogram
hist(df, col = "gray", main = "Histogram of the Original Data")
```



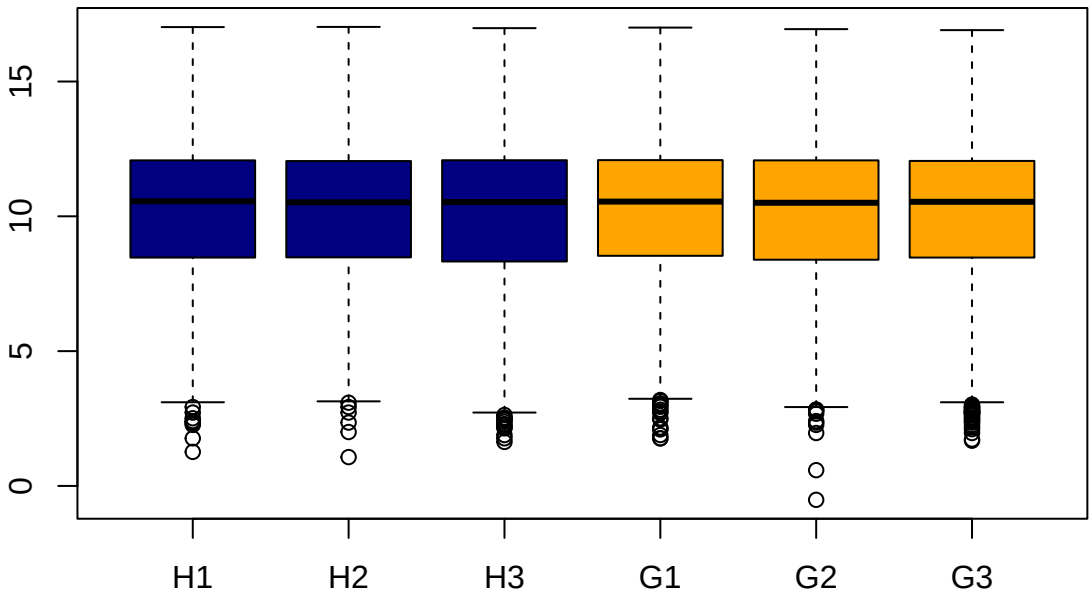
The histogram of the original data shows a highly right-skewed distribution with most gene expression values concentrated at lower intensities and a long, thin tail extending toward the much higher values. This particular skewness is characteristic of microarray data, where a small number of highly expressed genes can dominate the distribution. The presence of extreme values (up to 140,000) makes it quite difficult to visualize the distribution of most genes. Consequently, it is necessary to apply a log transformation to the data to conduct proper statistical analyses.

```
# The histogram above is right-skewed --> Solution: Log the data
df2 = log2(df)
hist(df2, col = "gray", main = "Histogram After Logging the Data")
```



It can be seen that the Log2 transformation successfully normalized the distribution, converting the right-skewed data into an approximately Normal (Gaussian) Distribution centered around 10-12 on the log2 scale. This transformation was essential because most statistical tests assume that the data is normally distributed. Therefore, this newly normalized distribution allows for more reliable statistical inference and insights along with better visualization of gene expression patterns across all values.

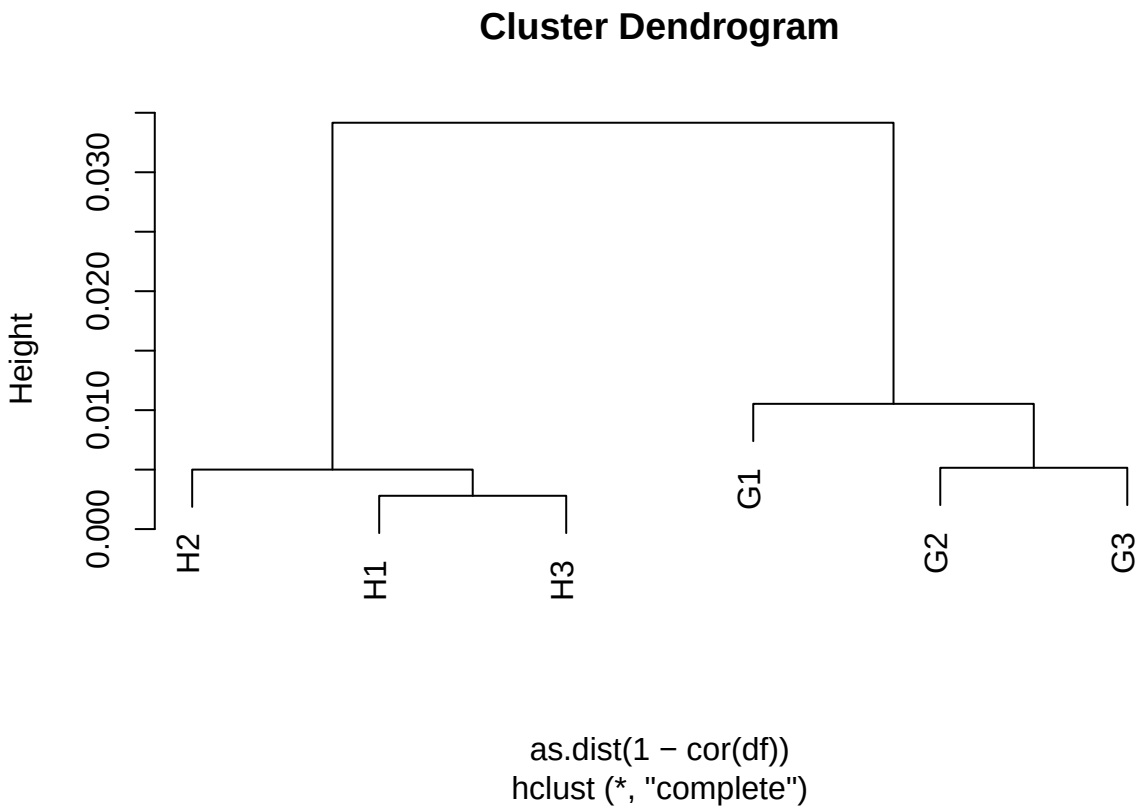
```
# Boxplot
colors = c(rep("navy", 3), rep("orange", 3))
boxplot(df2, col = colors)
```



The boxplots comparing log2-transformed gene expression values across all six samples reveal consistent distributions with similar medians (approximately equal to 10) and lower and upper quartiles. The similar

overall distributions between species indicate that global expression levels are indeed comparable and that differential expression will likely be down to specific genes rather than representing a systematic shift in overall transcription levels. Additionally, it can be seen that there are existing outliers (the points observed below the whiskers) in all six samples, representing highly expressed genes that deviate from the typical expression range.

```
# Hierarchical Clustering
# This is based on the correlation coefficients of the expression values

# Correlation between the similarites
# --> Introducing (1-cor) will give the dissimilarity & then build the tree based on it

hc = hclust(as.dist(1 - cor(df)))
plot(hc)
```



The hierarchical clustering dendrogram above clearly separates the samples into two distinct groups based on their gene expression profiles. The three human samples (H1, H2, H3) cluster together with minimal distance between them, while the three gorilla samples (G1, G2, G3) form a separate cluster with greater, but still quite small distances between them. This clustering pattern demonstrates that gene expression differences between species are more noticeable than the variation within species, confirming that transcriptional profiles can reliably distinguish between human and gorilla fibroblasts.

```
# Splitting Data Matrix into Two - Part 1
ko = df[, 1:3] # KO matrix
head(ko)
```

	H1	H2	H3
100_g_at	6531.0	5562.8	6822.4
1000_at	11486.3	10542.7	10641.4

	H1	H2	H3
1001_at	14339.2	13526.1	14444.7
1002_f_at	3156.8	2219.5	3264.4
1003_s_at	4002.0	3306.9	3777.0
1004_at	3468.4	3347.4	3332.9

```
# Splitting Data Matrix into Two - Part 2
wt = df[, 4:6] # WT matrix
head(wt)
```

	G1	G2	G3
100_g_at	7732.1	7191.2	7551.9
1000_at	10408.2	9484.5	7650.2
1001_at	12936.6	13841.7	13285.7
1002_f_at	2374.2	2201.8	2525.3
1003_s_at	3760.6	3137.0	2911.5
1004_at	3073.5	3046.0	2914.4

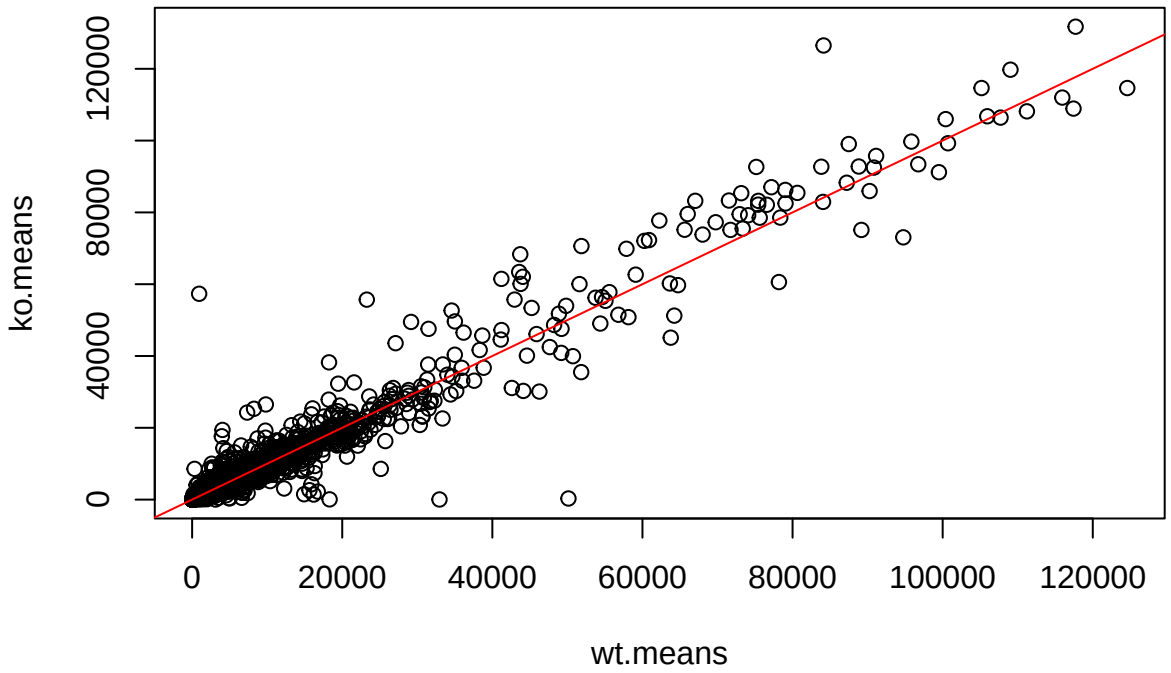
```
# Compute the means of the KO samples
ko.means = rowMeans(ko)
head(ko.means)
```

```
##  100_g_at   1000_at   1001_at 1002_f_at 1003_s_at   1004_at
##  6305.400 10890.133 14103.333  2880.233  3695.300  3382.900
```

```
# Compute the means of the WT samples
wt.means = rowMeans(wt)
head(wt.means)
```

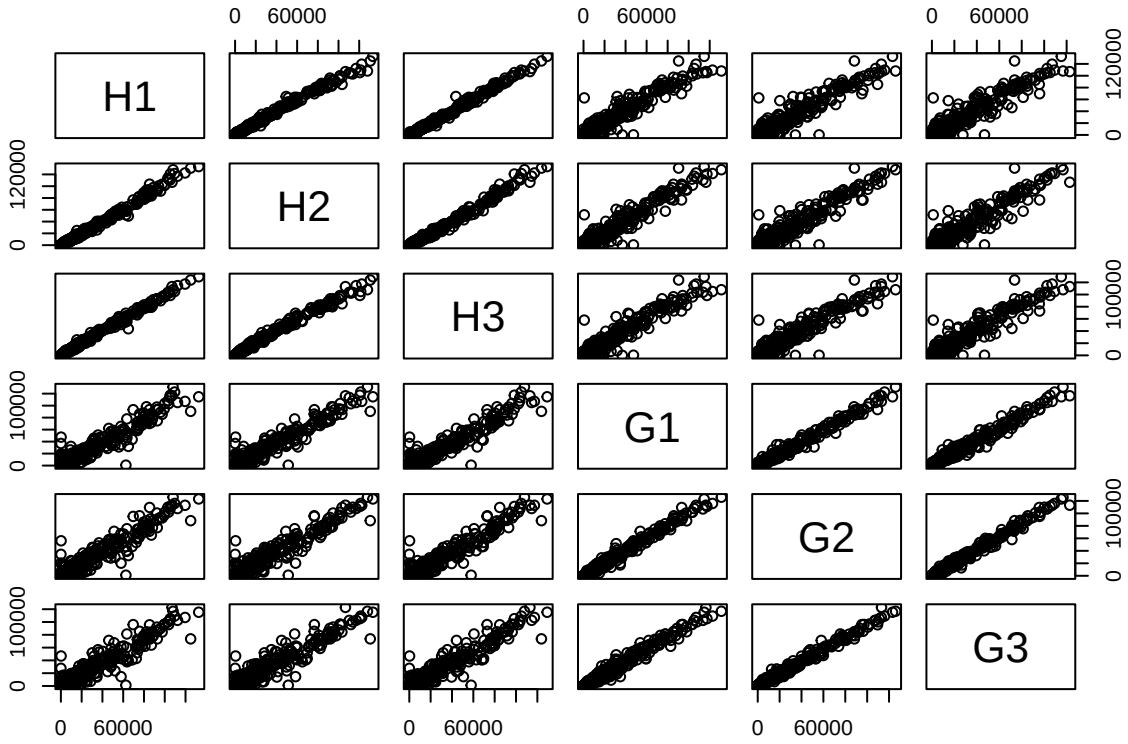
```
##  100_g_at   1000_at   1001_at 1002_f_at 1003_s_at   1004_at
##  7491.733  9180.967 13354.667  2367.100  3269.700  3011.300
```

```
# Scatter Plot
plot(ko.means ~ wt.means) # the '~' correlates the two variables together
abline(0, 1, col = "red") # red diagonal line
```



The scatter plot above, which works to compare the mean expression values between human (ko.means) and gorilla (wt.means) samples, reveals that most genes fall near the diagonal line. This indicates similar expression levels between the two species. However, there are several genes that deviate from this line, representing candidates for differentially expressed genes (DEGs). These deviations suggest species-specific transcriptional regulation patterns that may contribute to phenotypic differences between humans and gorillas.

```
# Checking to see if any of the samples are outliers
pairs(df) # All pairwise comparisons
```



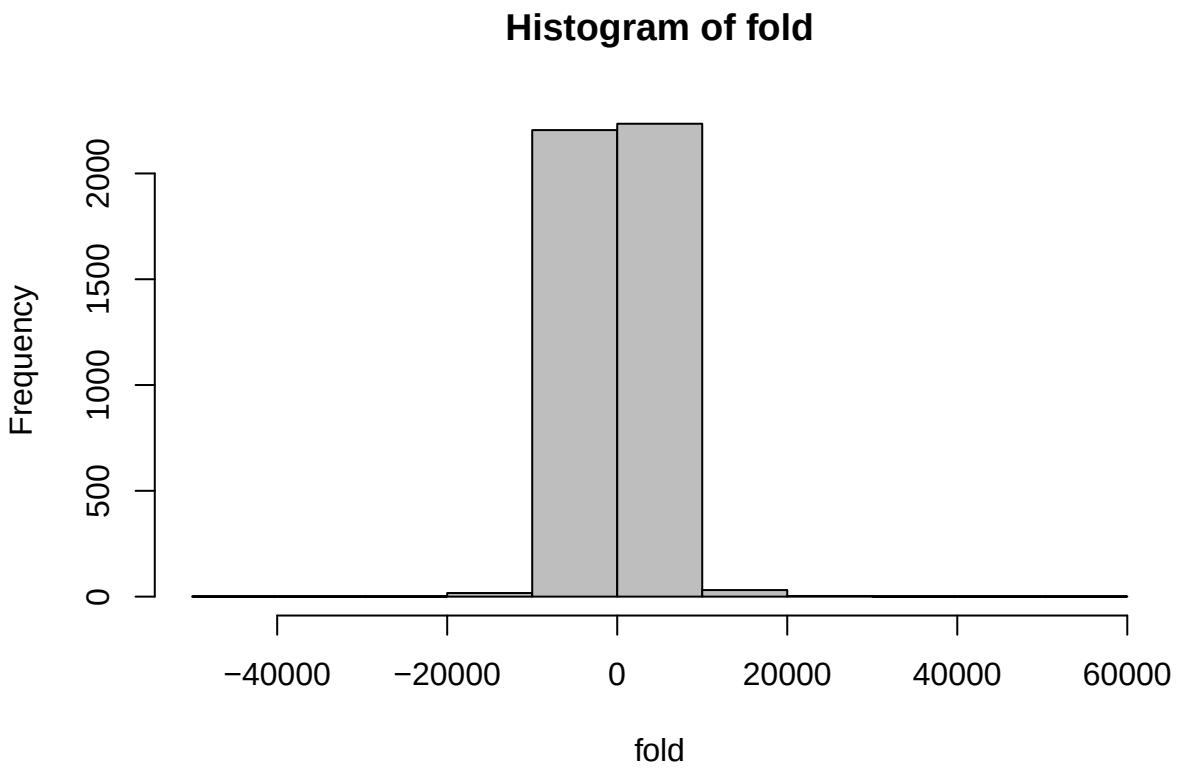
The pairwise comparison plot above shows strong positive correlations between all of the samples, with

especially tight correlations within each species group. From this plot, there are no obvious or extreme outlier samples detected, which indicates that all samples are of good quality and indeed suitable for differential expression analysis. The consistent patterns within species groups further validate the biological replicates used within this study.

```
# The universal p-value is 0.05:
# Anything below that is statistically significant and anything above it is not
# If p-value < 0.05: Reject null hypothesis due to confidence that what happened was NOT by chance

# Biological significance (fold-change):
fold = ko.means - wt.means # Difference between the means
head(fold)
```

```
##   100_g_at    1000_at    1001_at  1002_f_at  1003_s_at    1004_at
## -1186.3333  1709.1667   748.6667   513.1333   425.6000   371.6000
```

```
# Histogram of the fold
hist(fold, col = "gray")
```



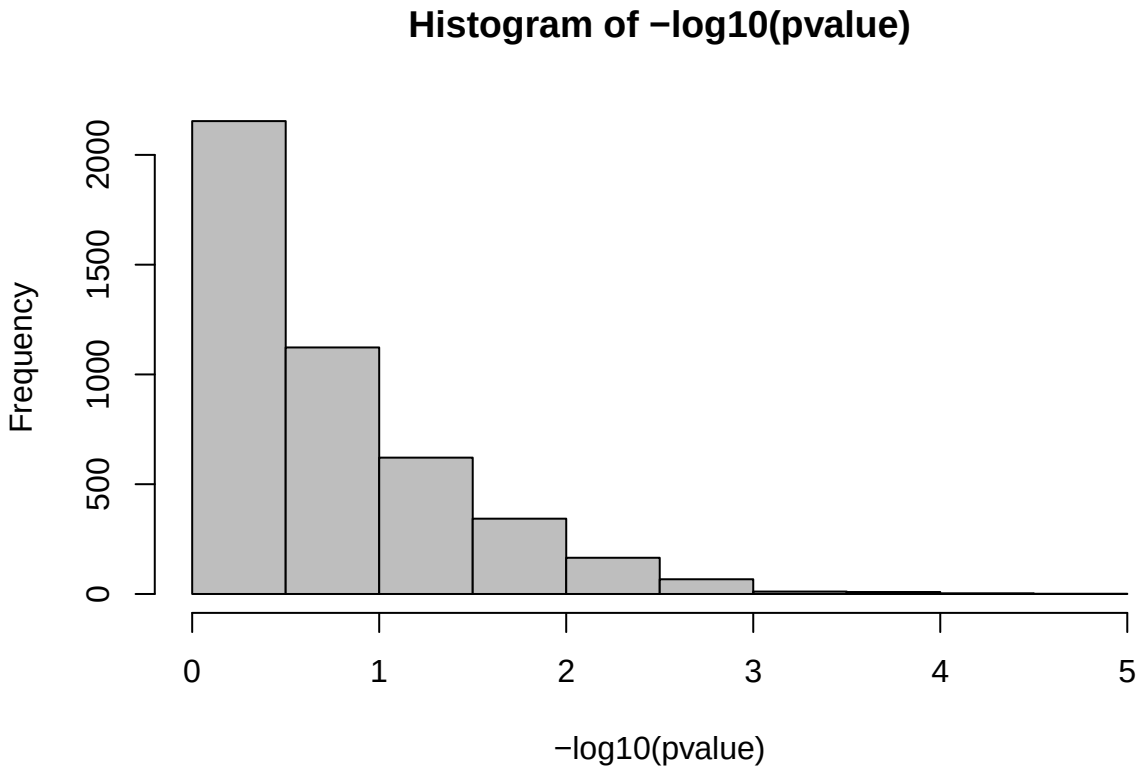
The fold change histogram displays a roughly symmetric distribution centered around zero, which indicate that most of the genes show similar expression levels between human and gorilla samples. The shape of the distribution reveals that large expression differences are quite rare, with the vast majority of genes falling within +/- 10,000. With that being said, the presence of genes with extreme fold changes (extending to +/- 20,000) demonstrates that substantial species-specific expression differences do exist for a subset of genes. These specific outliers represent the most promising candidates for genes contributing to phenotypic differences between the two species.

```
# We use the t-test when we're comparing 2 samples

pvalue = NULL # Empty list for the p-values

for(i in 1 : number_of_genes) { # for each gene from to the number of genes
```

```
  x = wt[i, ] # wt values of gene number i
  y = ko[i, ] # ko values of gene number i
  t = t.test(x, y) # t-test between the two conditions
  pvalue[i] = t$p.value # Store p-value number i into the list of p-values
}

head(pvalue)
```

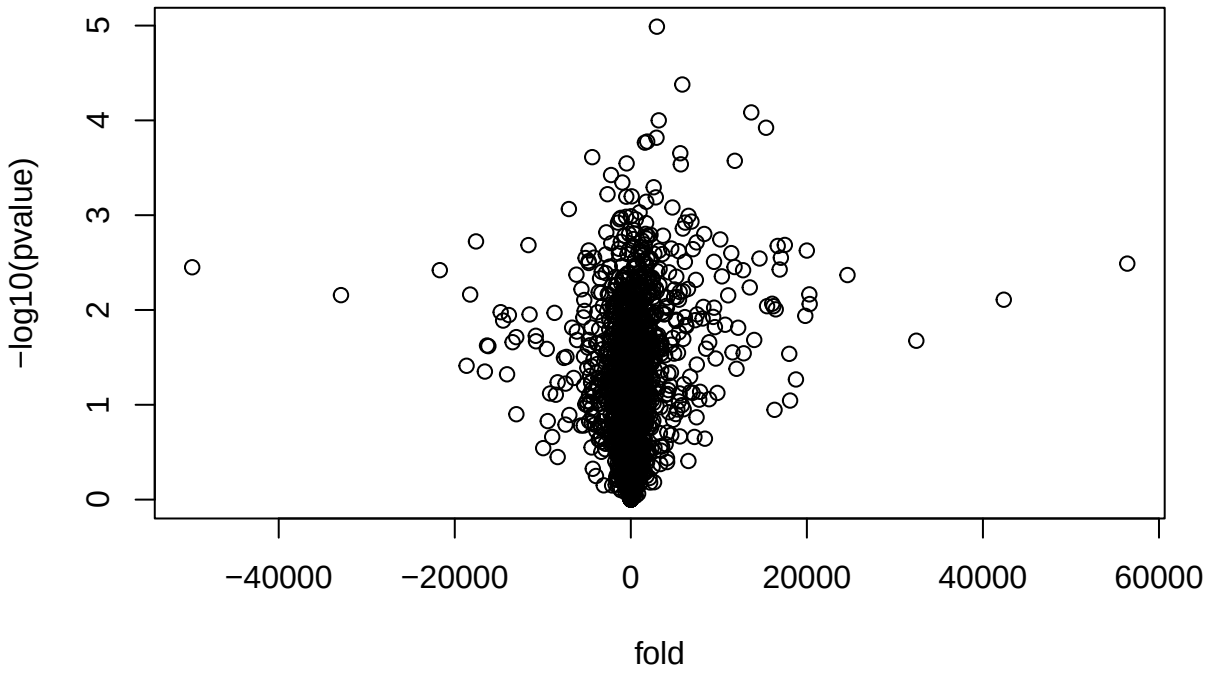
```
## [1] 0.073026163 0.158884955 0.129441802 0.258540610 0.264908834 0.004947113
```

```
# Histogram of p-values (-log10)
hist(-log10(pvalue), col = "gray")
```



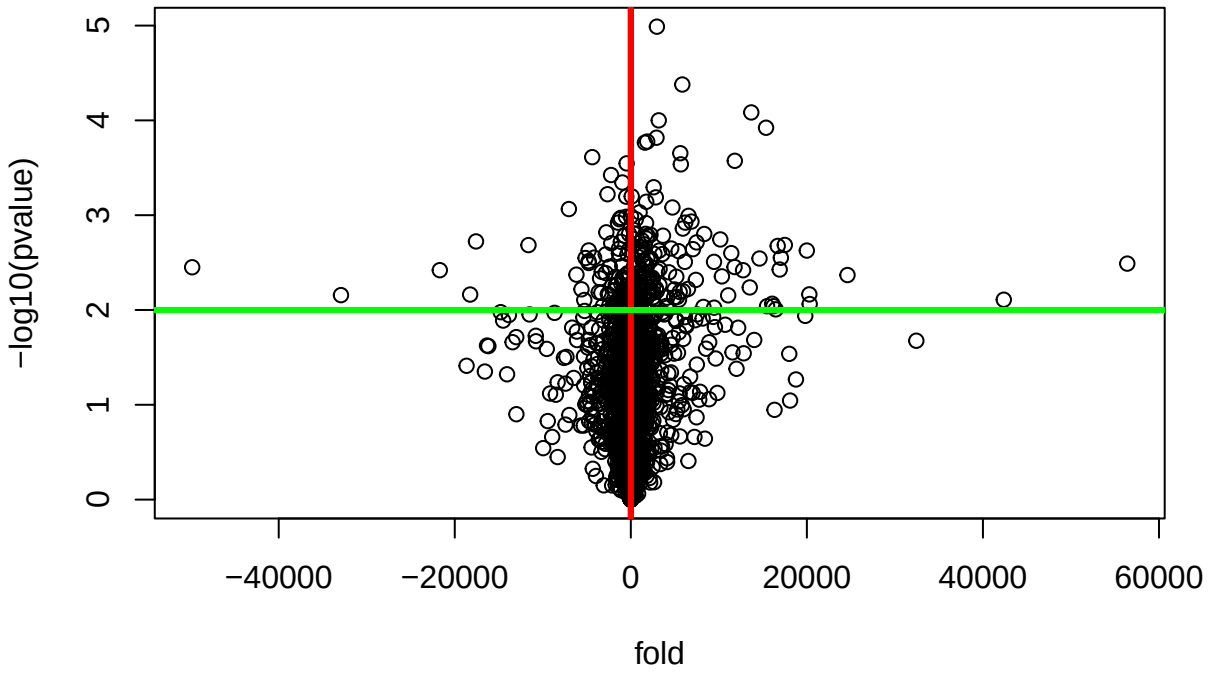
The histogram of -log10(p-values) above shows that most genes have low significance values as most of the genes cluster around 0-1 on the x-axis (corresponding to p-values of 1.0 to 0.1), indicating that their expression differences are likely due to random variation rather than true biological differences. On the other hand, there is still a smaller subset of genes shows with high -log10(p-value) scores, ranging from 2 to 5 (corresponding to p-values of 0.01 to 0.00001). Building on this, the shape of the distribution highlights that most of the genes are not differentially expressed, but a subset within them does show clear, reproducible differences between conditions.

```
# Volcano Plot
plot(-log10(pvalue) ~ fold)
```



This first volcano plot visualizes both biological significance (fold change, x-axis) and statistical significance (-log10 p-value, y-axis) simultaneously. The majority of the genes cluster near zero fold change with low significance, while DEGs appear in the upper left and right regions of the figure, which represents genes with both large expression changes and high statistical confidence.

```
# Volcano Plot with filtering
# filter according to the fold change & the p-value
fold_cutoff = 2 # double higher or double lower

pvalue_cutoff = 0.01 # instead of 0.05 to be stringent & obtain the highly significant

plot(-log10(pvalue) ~ fold)

abline(v = fold_cutoff, col = "blue", lwd = 3)

abline(v = -fold_cutoff, col = "red", lwd = 3)

abline(h = -log10(pvalue_cutoff), col = "green", lwd = 3)
```



Buildin on the previous and initial volcano plot, the filtering thresholds are now visualized as colored lines on the volcano plot: vertical red lines mark the fold change cutoff, while the horizontal green line indicates the p-value threshold (0.01, shown as -log10(0.01) = 2). These stringent criteria define four quadrants with the genes in the two upper quadrants (high significance and large fold change) being the ones selected as the DEGs. This approach ensures that identified genes are both biologically meaningful (large expression change) and statistically reliable (low probability of false discovery), therefore, lowering the risk of including genes that display changes by chance or genes that are statistically significant but biologically meaningless.

```
# Filtering for DEGs - 1

filter_by_fold = abs(fold) >= fold_cutoff # Biological
sum(filter_by_fold) # Number of genes satisfy the condition
```

```
## [1] 4445
```

```
cat("")

filter_by_pvalue = pvalue <= pvalue_cutoff # Statistical
sum(filter_by_pvalue)
```

```
## [1] 256
```

```
cat("")

# Merged to give those that are BOTH biological & statistical
filter_combined = filter_by_fold & filter_by_pvalue # Combined
sum(filter_combined)
```

```
## [1] 256
```

```
cat("")
```

```
# Filtering for DEGs - 2
filtered = df[filter_combined, ]

dim(filtered)
```

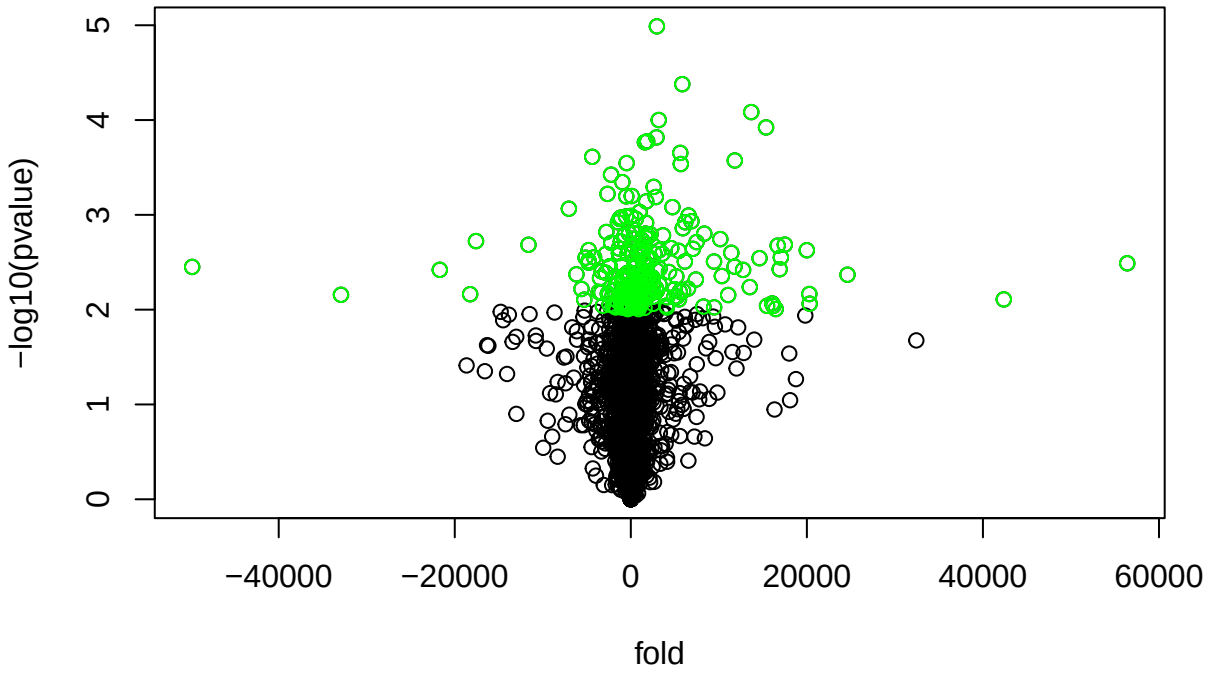
```
## [1] 256   6
```

```
head(filtered)
```

	H1	H2	H3	G1	G2	G3
1004_at	3468.4	3347.4	3332.9	3073.5	3046.0	2914.4
1011_s_at	9218.3	9631.5	9316.8	8560.7	8389.7	8277.3
1018_at	8388.7	8396.1	8936.6	7213.7	6786.1	7406.0
1071_at	9841.4	9339.6	9860.2	7779.5	7961.7	7603.4
1114_at	17391.0	16400.9	17315.7	14184.5	13678.3	13583.3
1116_at	2192.6	2297.2	1909.8	1035.0	813.4	1010.4

After using stringent filtering criteria (fold change >= 2 and p-value =< 0.01), 256 DEGs between human and gorilla fibroblasts were identified. Of these, 165 genes (64.5%) were up-regulated in humans compared to gorillas, with the remaining 91 genes (35.5%) being down-regulated. This asymmetry in regulation suggests that certainbiological pathways may be preferentially activated or repressed in human cells relative to gorilla cells.

```
# Filtering for DEGs - 3

plot(-log10(pvalue)~ fold)

# Changing the color of the genes that are beyond the threshold
points(-log10(pvalue[filter_combined])~ fold[filter_combined], col = "green")
```



With the DEGs highlighted in green, the volcano plot above displays that true DEGs occupy the regions of high statistical significance (upper portion of the plot) combined with substantial fold changes (far left and right). The relatively symmetric distribution of green points on both sides of zero indicates a balanced number of up-regulated and down-regulated genes. Therefore, this visual representation effectively confirms the presence of a robust set of genes that are both highly statistically significant and biologically meaningful.

```
# Up-Regulated Genes
# Screen for the up-regulated genes (positive fold)
```

```r
filter_up = filter_combined & fold > 0
head(filter_up)
```

```
##  100_g_at   1000_at   1001_at 1002_f_at 1003_s_at   1004_at
##     FALSE     FALSE     FALSE     FALSE     FALSE      TRUE
```

```r
# Number of filtered genes
sum(filter_up)
```

```
## [1] 165
```

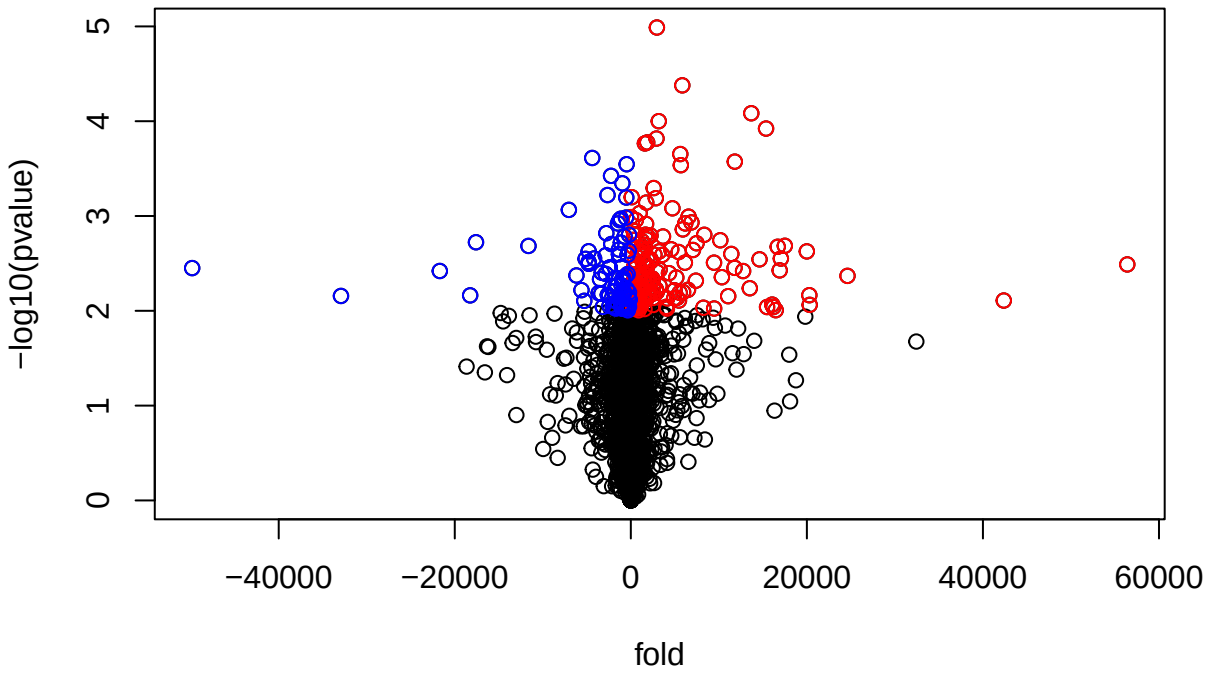
```r
# Down-Regulated Genes
# Screen for the down-regulated genes (negative fold)
filter_down = filter_combined & fold < 0
head(filter_down)
```

```
##  100_g_at   1000_at   1001_at 1002_f_at 1003_s_at   1004_at
##     FALSE     FALSE     FALSE     FALSE     FALSE     FALSE
```

```r
# Number of filtered genes
sum(filter_down)
```

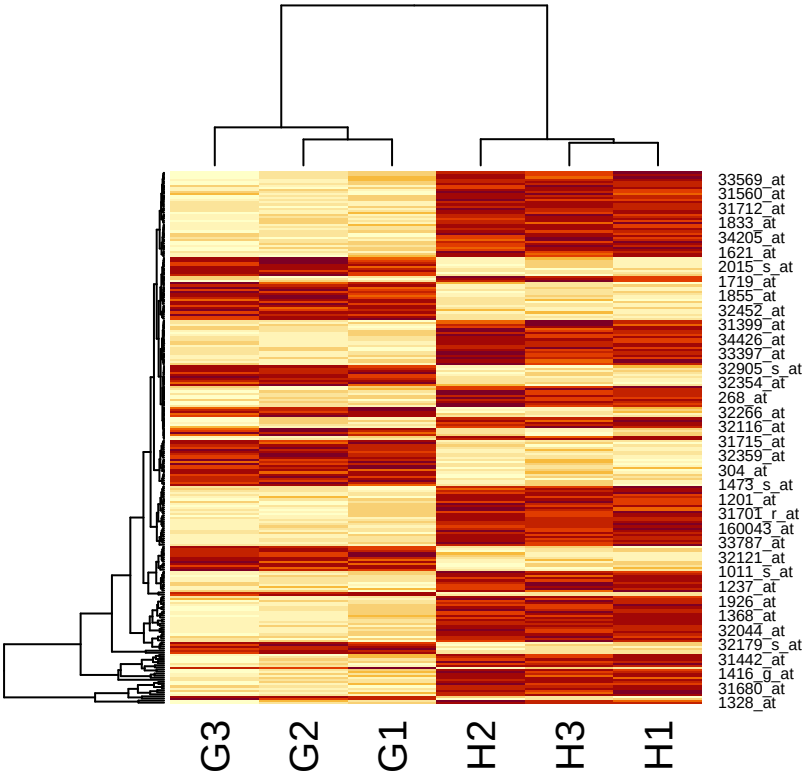
```
## [1] 91
```

```r
# Giving the positive and negative filtered genes different colors & visualizing it
plot(-log10(pvalue) ~ fold)
points(-log10(pvalue[filter_up]) ~ fold[filter_up], col = "red")
points(-log10(pvalue[filter_down]) ~ fold[filter_down], col = "blue")
```



The color-coded volcano above plot clearly distinguishes up-regulated genes (red) from down-regulated genes (blue). The up-regulated genes show a broader distribution of fold changes, with some genes exhibiting very high expression differences ($>$ 20,000 fold change). This suggests that these may represent human-specific gene expression programs. As for the down-regulated genes, these show a more moderate range of fold changes, indicating more subtle regulatory differences for genes with higher expression in gorillas than humans.

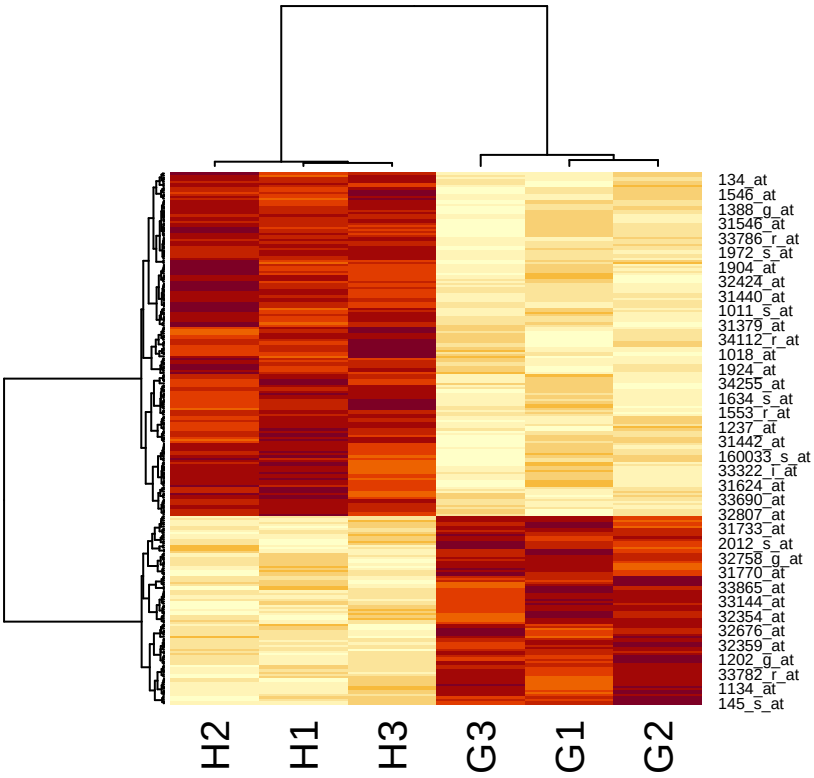
```
heatmap(filtered)
```



This first and basic heatmap displays the 256 DEGs across all six samples, with color intensity representing expression levels. The heatmap shows clear vertical separation between human (H1, H2, H3) and gorilla (G1, G2, G3) samples, with the horizontal bands indicating groups of genes with coordinated expression patterns. The yellow regions represent high expression while the red regions indicate low expression. This color scheme within each species group demonstrate that these genes can function as reliable molecular signatures to accurately distinguish between human and gorilla fibroblasts.

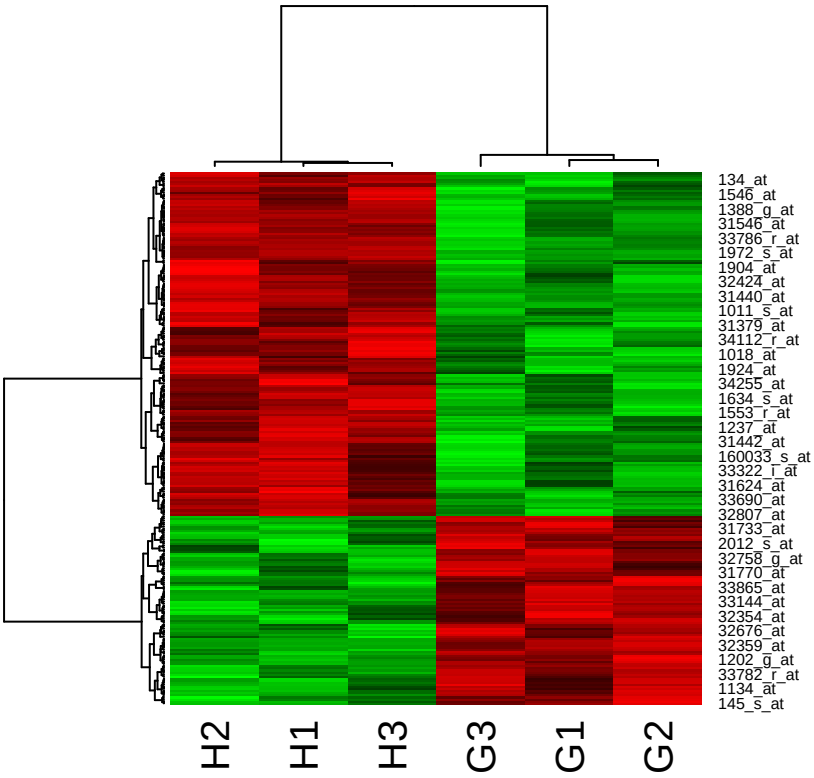
```
# Clustering of the columns (samples)
col_dendrogram = as.dendrogram(hclust(as.dist(1-cor(filtered))))

# Clustering of the rows (genes)
row_dendrogram = as.dendrogram(hclust(as.dist(1-cor(t(filtered)))))
```

```
# Heatmap with the rows and columns clustered by correlation coefficients
heatmap(filtered, Rowv=row_dendrogram, Colv = col_dendrogram)
```



This optimized heatmap uses correlation-based hierarchical clustering for both rows (genes) and columns (samples), giving a clearer display of the data's underlying structure. Unlike the original heatmap, the column dendrogram here perfectly separates human samples (H2, H1, H3) from gorilla samples (G3, G1, G2), with minimal within-species distances. As for the row dendrogram, it organizes the genes into distinct clusters based on their expression patterns, identifying groups of co-expressed genes. Thus, these gene clusters represent functionally related genes which provides valuable insights into the coordinated biological processes that differ between the two species.

```
heatmap(filtered, Rowv = row_dendrogram, Colv = col_dendrogram, col = rev(redgreen(1024)))
```



In the this heatmap of the bunch, the red-green color scheme reveals coordinated patterns of gene expression, with blocks of genes showing consistently higher expression in one species versus the other. These particular patterns suggest the presence of co-regulated gene modules that may be controlled by shared transcription factors which then contributes to species-specific cellular phenotypes.

Collectively, the heatmaps provide an overall visual representation of the 256 DEGs across all six samples. The hierarchical clustering successfully separates human and gorilla samples into distinct groups, with the dendrogram showing clear species-based clustering at both the column (sample) and row (gene) levels.

```
# Annotations
# To obtain the functional annotation of the differentially expressed genes:
# we're 1st going to extract their probe ids:

filterd_ids = row.names(filtered) # ids of the filtered DE genes
length(filterd_ids)
```

```
## [1] 256
```

```
head(filterd_ids)
```

```
## [1] "1004_at"   "1011_s_at" "1018_at"   "1071_at"   "1114_at"   "1116_at"
```

```
annotation_table = aafTableAnn(filterd_ids, "hgu95av2.db")
saveHTML(annotation_table, file = "project_filtered.html")
browseURL("project_filtered.html")
# Convert the HTML file to PDF
chrome_print(
  input = "project_filtered.html",
  output = "project_filtered.pdf"
)
```

```r
write.table(filterd_ids, "filterd_ids.txt", quote = FALSE, row.names = FALSE, col.names = FALSE)
```

```r
up_ids = ids[filter_up] # IDs of up-regulated genes
length(up_ids)
```

```
## [1] 165
```

```r
head(up_ids)
```

```
## [1] "1004_at"   "1011_s_at" "1018_at"   "1071_at"   "1114_at"   "1116_at"
```

```r
# Obtaining the annotation of the up- and down- regulated genes into tables
up_table = aafTableAnn(up_ids, "hgu95av2.db")
saveHTML(up_table, file = "project_up_table.html")
browseURL("project_up_table.html")
# Convert the HTML file to PDF
chrome_print(
  input = "project_up_table.html",
  output = "project_up_table.pdf"
)
```

```r
write.table(up_ids, "up_ids.txt", quote = FALSE, row.names = FALSE, col.names = FALSE)
```

```r
down_ids = ids[filter_down] # IDs of down-regulared genes
length(down_ids)
```

```
## [1] 91
```

```r
head(down_ids)
```

```
## [1] "1134_at"   "1202_g_at" "1213_at"   "1236_s_at" "1239_s_at" "1254_at"
```

```r
down_table = aafTableAnn(down_ids, "hgu95av2.db")
saveHTML(down_table, file = "project_down_table.html")
browseURL("project_down_table.html")
# Convert the HTML file to PDF
chrome_print(
  input = "project_down_table.html",
  output = "project_down_table.pdf"
)
```

```r
write.table(down_ids, "down_ids.txt", quote = FALSE, row.names = FALSE, col.names = FALSE)
```

Conclusion

In conclusion, this differential expression analysis of human and gorilla fibroblast samples successfully identified 256 genes with significant expression differences between the two species. The analysis thoroughly revealed that gene expression profiles can be used to reliably distinguish between human and gorilla samples, which is consistent with the findings of the original paper.

The key findings from this conducted analysis include the following:

1. Clear separation of samples by species in hierarchical clustering and heatmap analyses, demonstrating robust species-specific expression patterns.
2. Identification of 165 up-regulated and 91 down-regulated genes in humans compared to gorillas, using stringent statistical criteria (fold change >= 2 and p-value <= 0.01).
3. The asymmetric distribution of up- versus down-regulated genes suggests differential activation of biological pathways between species.

4. High-quality data with no clear or extreme outlier samples, ensuring reliable differential expression results.

The annotated gene lists generated from this analysis provide a foundation for further investigation into the functional pathways and biological processes that distinguish between the cellular physiologies of humans and gorillas. These transcriptional differences likely contribute to the species-specific traits observed at the organismic level. This supports the hypothesis that gene regulation plays a truly critical role in areas such as primate evolution and species divergence.

Future work and directions could include pathway enrichment analysis to identify which biological processes are most significantly affected by these expression differences, as well as investigation of the regulatory mechanisms, such as transcription factors, driving these species-specific expression patterns.

Overall, this study confirms that the interdisciplinary field of Bioinformatics, specifically gene expression analysis, is a truly powerful and essential tool for understanding the molecular basis of species differences between humans and their closest evolutionary relatives.