Reporte

Jordi García Garcés

2025-02-01

Contents

Analysis of the transcriptome of people with Hutchinson-Gilford Progeria Syndrome

This data was recovered from the study "Predicting age from the transcriptome of human dermal fibroblasts" in recount3.

```
# Libraries
library("edgeR")
```

```
## Loading required package: limma
```

```
library("ggplot2")
library("pheatmap")
library("RColorBrewer")
library("recount3")
```

```
## Loading required package: SummarizedExperiment
```

```
## Loading required package: MatrixGenerics
```

```
## Loading required package: matrixStats
```

```
##
## Attaching package: 'MatrixGenerics'
```

```
## The following objects are masked from 'package:matrixStats':
##
##     colAlls, colAnyNAs, colAnys, colAvgsPerRowSet, colCollapse,
##     colCounts, colCummaxs, colCummins, colCumprods, colCumsums,
##     colDiffs, colIQRDiffs, colIQRs, colLogSumExps, colMadDiffs,
##     colMads, colMaxs, colMeans2, colMedians, colMins, colOrderStats,
##     colProds, colQuantiles, colRanges, colRanks, colSdDiffs, colSds,
##     colSums2, colTabulates, colVarDiffs, colVars, colWeightedMads,
##     colWeightedMeans, colWeightedMedians, colWeightedSds,
##     colWeightedVars, rowAlls, rowAnyNAs, rowAnys, rowAvgsPerColSet,
##     rowCollapse, rowCounts, rowCummaxs, rowCummins, rowCumprods,
##     rowCumsums, rowDiffs, rowIQRDiffs, rowIQRs, rowLogSumExps,
##     rowMadDiffs, rowMads, rowMaxs, rowMeans2, rowMedians, rowMins,
##     rowOrderStats, rowProds, rowQuantiles, rowRanges, rowRanks,
##     rowSdDiffs, rowSds, rowSums2, rowTabulates, rowVarDiffs, rowVars,
##     rowWeightedMads, rowWeightedMeans, rowWeightedMedians,
##     rowWeightedSds, rowWeightedVars

## Loading required package: GenomicRanges

## Loading required package: stats4

## Loading required package: BiocGenerics

##
## Attaching package: 'BiocGenerics'

## The following object is masked from 'package:limma':
##
##     plotMA

## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs

## The following objects are masked from 'package:base':
##
##     anyDuplicated, aperm, append, as.data.frame, basename, cbind,
##     colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,
##     get, grep, grepl, intersect, is.unsorted, lapply, Map, mapply,
##     match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
##     Position, rank, rbind, Reduce, rownames, sapply, saveRDS, setdiff,
##     table, tapply, union, unique, unsplit, which.max, which.min

## Loading required package: S4Vectors

##
## Attaching package: 'S4Vectors'

## The following object is masked from 'package:utils':
##
##     findMatches
```

```
## The following objects are masked from 'package:base':
##
##     expand.grid, I, unname

## Loading required package: IRanges

## Loading required package: GenomeInfoDb

## Loading required package: Biobase

## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.

##
## Attaching package: 'Biobase'

## The following object is masked from 'package:MatrixGenerics':
##
##     rowMedians

## The following objects are masked from 'package:matrixStats':
##
##     anyMissing, rowMedians
```

```
# Download the dataset recovered from recount3
rse_gene_SRP144355 <- recount3::create_rse_manual(
    project = "SRP144355",
    project_home = "data_sources/sra",
    organism = "human",
    annotation = "gencode_v26",
    type = "gene"
)
```

```
## 2025-02-02 12:40:13.119139 downloading and reading the metadata.

## 2025-02-02 12:40:13.77651 caching file sra.sra.SRP144355.MD.gz.

## 2025-02-02 12:40:14.53072 caching file sra.recount_project.SRP144355.MD.gz.

## 2025-02-02 12:40:15.194309 caching file sra.recount_qc.SRP144355.MD.gz.

## 2025-02-02 12:40:15.788201 caching file sra.recount_seq_qc.SRP144355.MD.gz.

## 2025-02-02 12:40:16.419359 caching file sra.recount_pred.SRP144355.MD.gz.

## 2025-02-02 12:40:16.756216 downloading and reading the feature information.
```

```
## 2025-02-02 12:40:17.17472 caching file human.gene_sums.G026.gtf.gz.

## 2025-02-02 12:40:18.545456 downloading and reading the counts: 143 samples across 63856 features.

## 2025-02-02 12:40:18.975304 caching file sra.gene_sums.SRP144355.G026.gz.

## 2025-02-02 12:40:22.955305 constructing the RangedSummarizedExperiment (rse) object.
```

```
# Analysis of the reads of our data set
assay(rse_gene_SRP144355, "counts") <- compute_read_counts(rse_gene_SRP144355)
# Attributes of Sequence Read Archive
rse_gene_SRP144355$sra.sample_attributes[1:3]
```

```
## [1] "age;;66|cell id;;GM03529|disease;;Normal|ethnicity;;Black|Sex;;male|source_name;;Skin; Thigh"
## [2] "age;;8yr|cell id;;PRF167|disease;;HGPS|ethnicity;;Unknown|Sex;;male|source_name;;Unknown"
## [3] "age;;1|cell id;;AG08498|disease;;Normal|ethnicity;;Asian|Sex;;male|source_name;;Skin; Foreskin"
```

```
# Access to the metadata of sra
rse_gene_SRP144355 <- expand_sra_attributes(rse_gene_SRP144355)
colData(rse_gene_SRP144355)[
    ,
    grepl("^sra_attribute", colnames(colData(rse_gene_SRP144355)))
]
```

```
## DataFrame with 143 rows and 6 columns
##             sra_attribute.age sra_attribute.cell_id sra_attribute.disease
##                   <character>           <character>           <character>
## SRR7093938                 66               GM03529                Normal
## SRR7093943               8yr                PRF167                  HGPS
## SRR7093809                 1               AG08498                Normal
## SRR7093810                12               AG16409                Normal
## SRR7093811                24               AG11732                Normal
## ...                      ...                   ...                   ...
## SRR7093947           8yr6mos             HGADFN169                  HGPS
## SRR7093948          6yr11mos             HGADFN178                  HGPS
## SRR7093949           5yr0mos             HGADFN122                  HGPS
## SRR7093950          8yr10mos             HGADFN143                  HGPS
## SRR7093951           3yr0mos             HGADFN367                  HGPS
##             sra_attribute.ethnicity sra_attribute.Sex sra_attribute.source_name
##                         <character>       <character>               <character>
## SRR7093938                   Black              male               Skin; Thigh
## SRR7093943                 Unknown              male                   Unknown
## SRR7093809                   Asian              male            Skin; Foreskin
## SRR7093810                Caucasian              male         Skin; Unspecified
## SRR7093811                Caucasian            female                 Skin; Arm
## ...                             ...               ...                       ...
## SRR7093947                 Unknown              male                   Unknown
## SRR7093948                 Unknown            female                   Unknown
## SRR7093949                 Unknown            female                   Unknown
## SRR7093950                 Unknown              male                   Unknown
## SRR7093951                 Unknown            female                   Unknown
```

```
# Go from character to numeric or factor
rse_gene_SRP144355$sra_attribute.age <- as.numeric(rse_gene_SRP144355$sra_attribute.age)
rse_gene_SRP144355$sra_attribute.disease <-
  factor(tolower(rse_gene_SRP144355$sra_attribute.disease))
rse_gene_SRP144355$sra_attribute.Sex <- factor(rse_gene_SRP144355$sra_attribute.Sex)
```

```
# Summary of the attributes of interest
summary(as.data.frame(colData(rse_gene_SRP144355)[
    ,
    grepl("^sra_attribute\\.(age|disease|Sex)", colnames(colData(rse_gene_SRP144355)))
]))
```

```
##  sra_attribute.age sra_attribute.disease sra_attribute.Sex
##  Min.   : 1.00     hgps  : 10            female: 41
##  1st Qu.:25.00     normal:133            male  :102
##  Median :46.00
##  Mean   :48.84
##  3rd Qu.:78.00
##  Max.   :96.00
##  NA's   :10
```

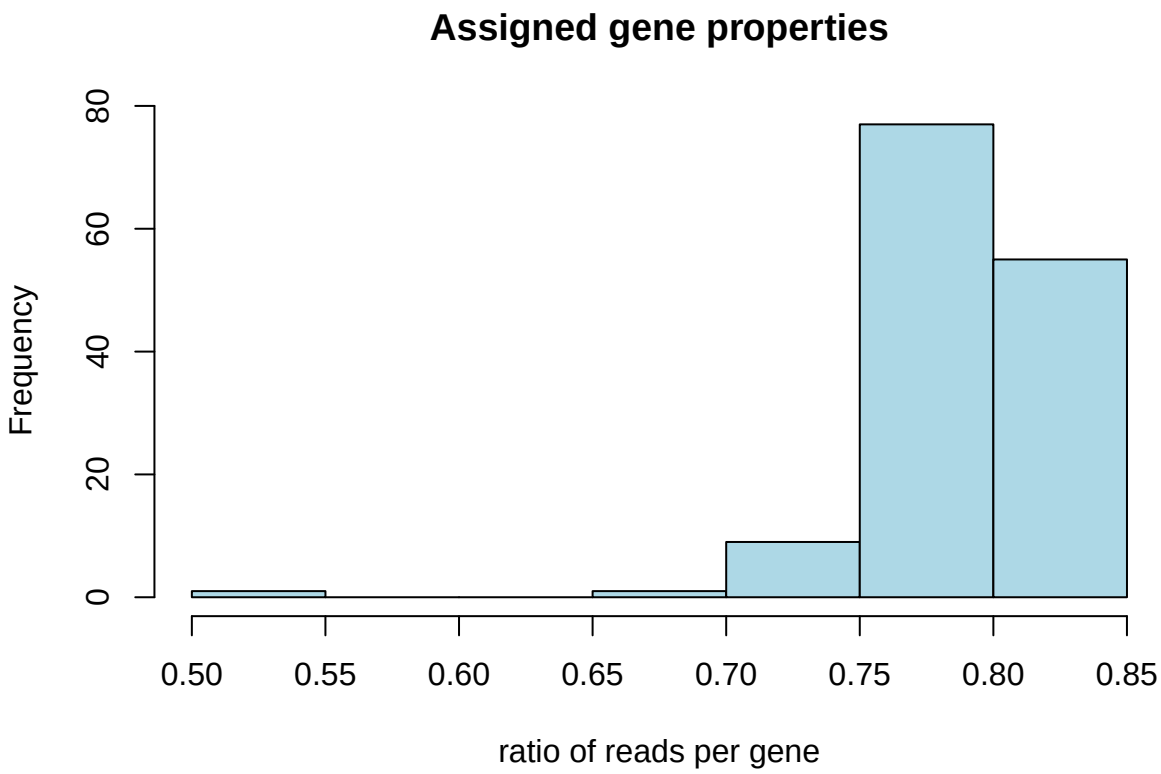
```
# Quality check
rse_gene_SRP144355$assigned_gene_prop <-
  rse_gene_SRP144355$recount_qc.gene_fc_count_all.assigned /
  rse_gene_SRP144355$recount_qc.gene_fc_count_all.total

summary(rse_gene_SRP144355$assigned_gene_prop)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.5359  0.7866  0.7973  0.7901  0.8041  0.8183
```

```
# Quality of assigned_gene_prop and attribute disease
with(colData(rse_gene_SRP144355), tapply(assigned_gene_prop, sra_attribute.disease, summary))
```

```
## $hgps
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.5359  0.7904  0.8082  0.7760  0.8137  0.8183
##
## $normal
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.6794  0.7860  0.7970  0.7911  0.8026  0.8181
```

```
# Visualization with histogram of the quality
hist(rse_gene_SRP144355$assigned_gene_prop, col = "lightblue",
     main = "Assigned gene properties", xlab = "ratio of reads per gene")
```




```
table(rse_gene_SRP144355$assigned_gene_prop < 0.5)
```

```
##
## FALSE
##   143
```

```
# All have good quality
```

Data normalization

```
# Filtering and normalization with edger
library(edgeR)

# object dgelist used by edger
dge <- DGEList(
    counts = assay(rse_gene_SRP144355, "raw_counts"),
    genes = rowData(rse_gene_SRP144355)
)
# Apply filterByExpr to remove low expression genes
dge$samples$group <- factor(rse_gene_SRP144355$sra_attribute.disease )
keep_genes <- filterByExpr(dge)
#Filters genes in the DGEList object
dge <- dge[keep_genes, , keep.lib.sizes=FALSE]
```

```
# Normalize data
dge <- calcNormFactors(dge)

# Dimensions before and after filtering
dim(rse_gene_SRP144355)  # Before
```

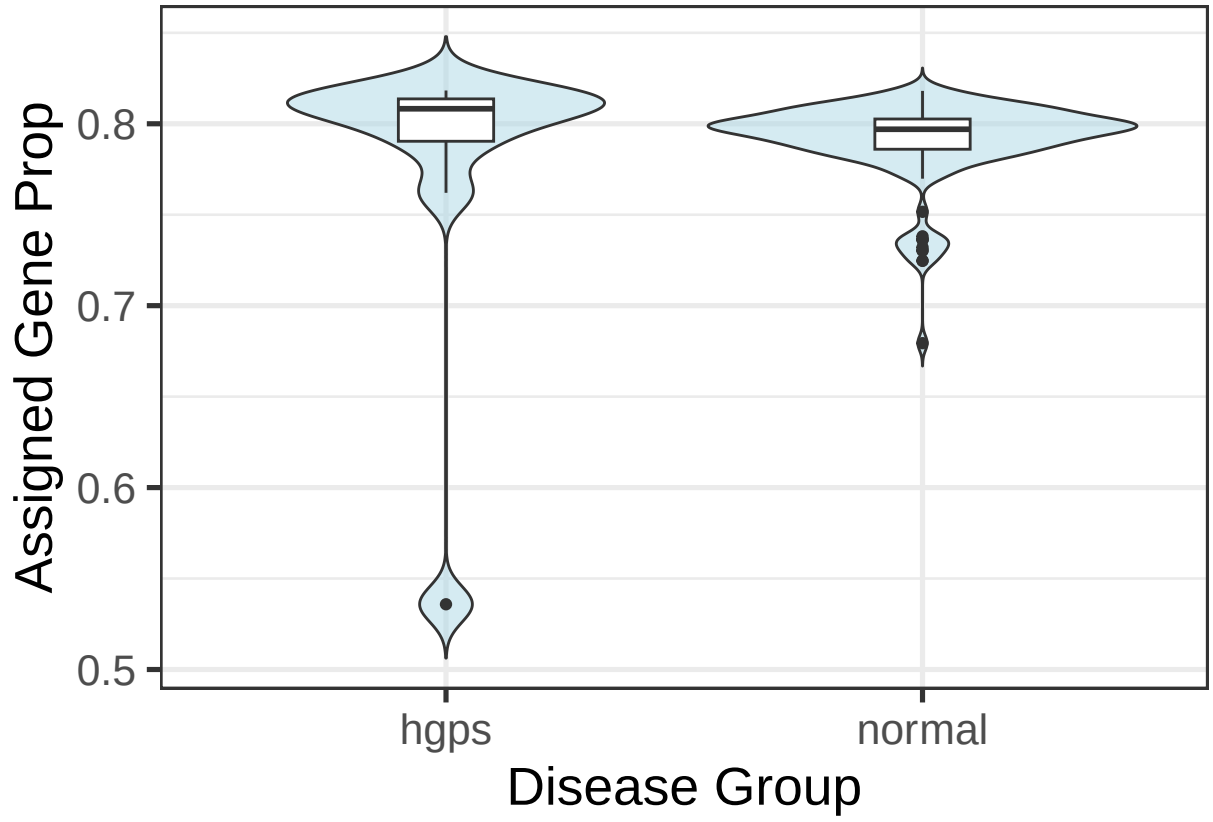
```
## [1] 63856   143
```

```
dim(dge)                 # After
```

```
## [1] 36476   143
```

Expression analysis

```
# Boxplots and violin plots of normal and hgps
ggplot(as.data.frame(colData(rse_gene_SRP144355)), aes(y = assigned_gene_prop,
                                                       x = sra_attribute.disease)) +
    geom_violin(trim = FALSE, fill = "lightblue", alpha = 0.5) +
    geom_boxplot(width = 0.2) +
    theme_bw(base_size = 20) +
    ylab("Assigned Gene Prop") +
    xlab("Disease Group")
```



```
mod <- model.matrix(~ rse_gene_SRP144355$sra_attribute.disease + sra_attribute.Sex + assigned_gene_prop
    data = colData(rse_gene_SRP144355)
)
colnames(mod)
```

```
## [1] "(Intercept)"
## [2] "rse_gene_SRP144355$sra_attribute.diseasenormal"
## [3] "sra_attribute.Sexmale"
## [4] "assigned_gene_prop"
```

```
library("limma")
vGene <- voom(dge, mod, plot = FALSE)
```

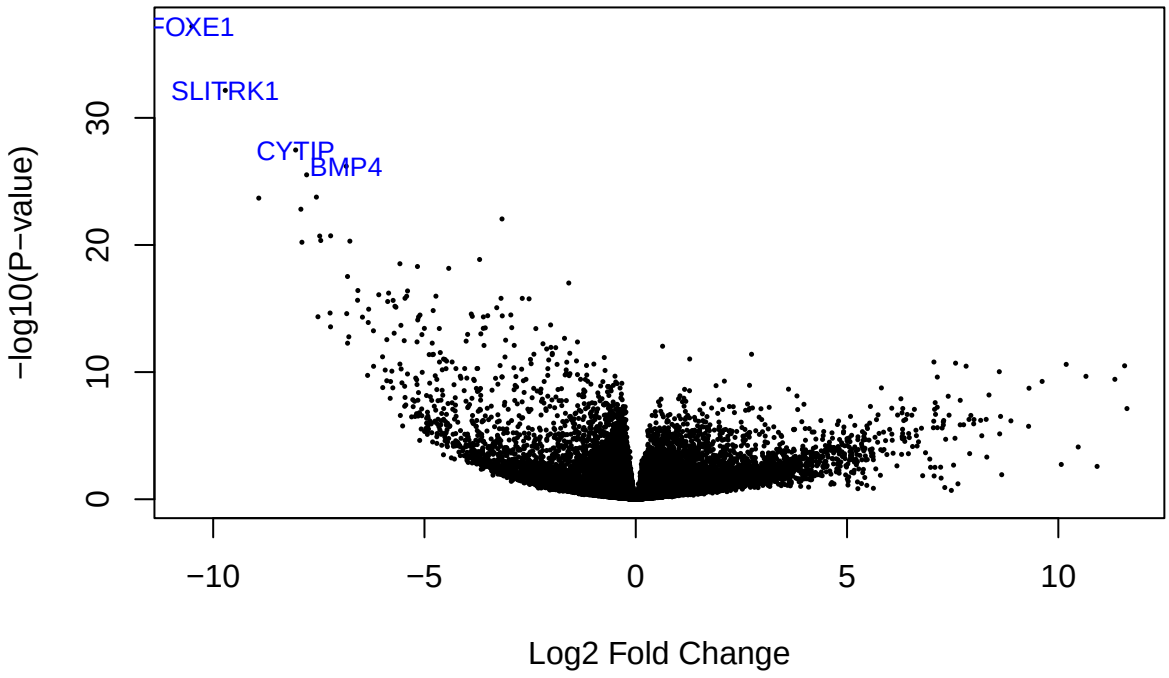
```
eb_results <- eBayes(lmFit(vGene))

de_results <- topTable(
    eb_results,
    coef = 2,
    number = nrow(rse_gene_SRP144355),
    sort.by = "none"
)
dim(de_results)
```

```
## [1] 36476    16
```

```
table(de_results$adj.P.Val < 0.05)
```

```
##
## FALSE  TRUE
## 31799  4677
```

```
# Volcano plot "normal" respect "hgps"
volcanoplot(eb_results, coef = 2, highlight = 4, names = de_results$gene_name)
```

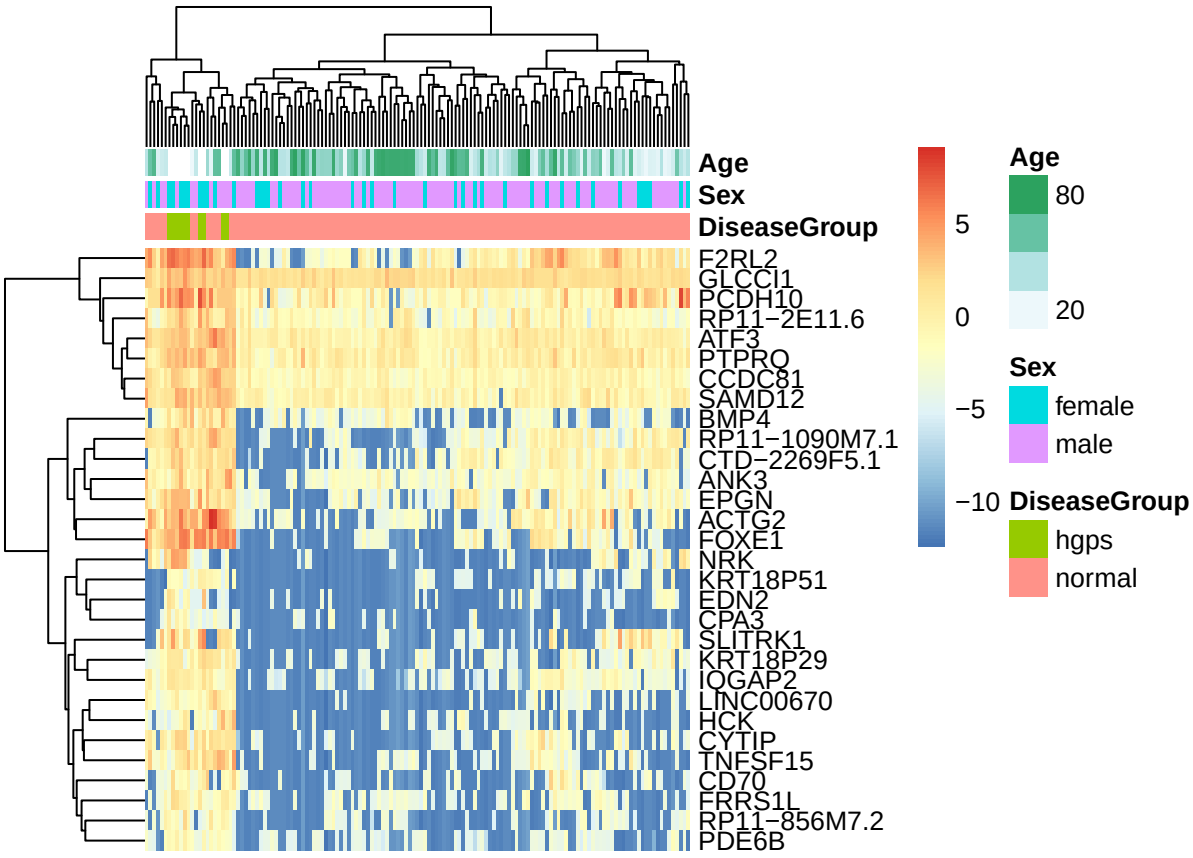



```
# Genes with highest P-value
de_results[de_results$gene_name %in% c("FOXE1", "SLITRK1", "CYTIP", "BMP4"), ]
```

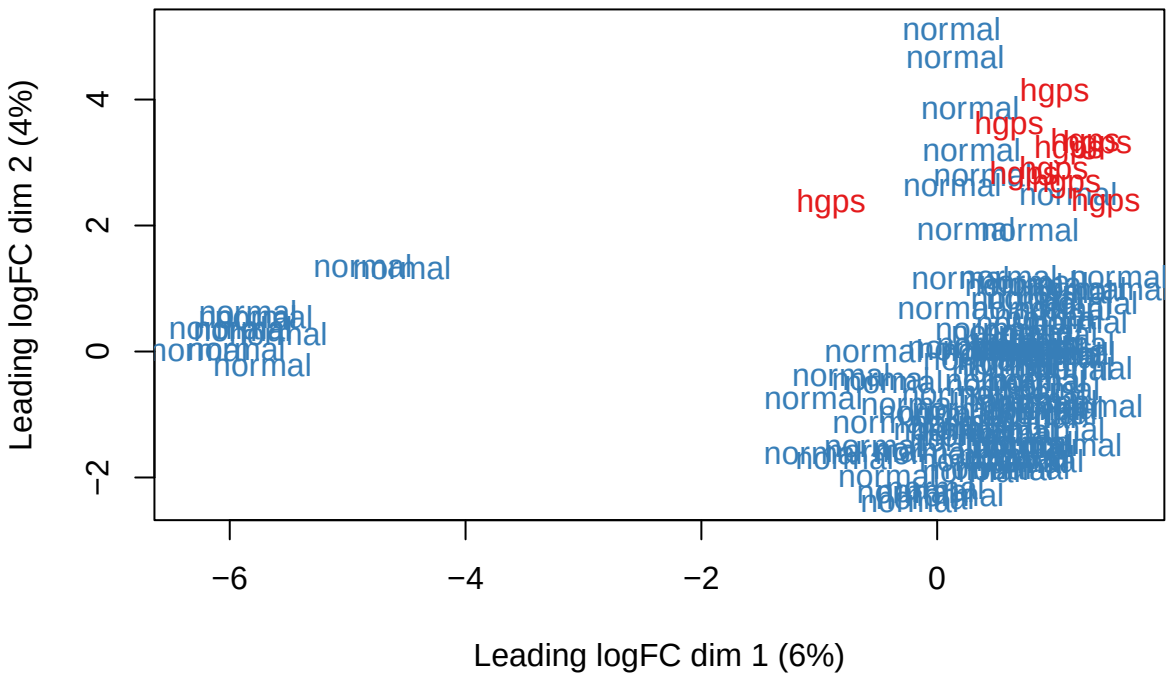
```
##                     source type bp_length phase            gene_id
## ENSG00000178235.7  HAVANA gene      5189    NA  ENSG00000178235.7
## ENSG00000125378.15 HAVANA gene      3082    NA ENSG00000125378.15
## ENSG00000115165.9  HAVANA gene      3428    NA  ENSG00000115165.9
## ENSG00000178919.8  HAVANA gene      3462    NA  ENSG00000178919.8
##                          gene_type gene_name level          havana_gene
## ENSG00000178235.7  protein_coding   SLITRK1     2 OTTHUMG00000017149.1
## ENSG00000125378.15 protein_coding      BMP4     1 OTTHUMG00000140303.4
## ENSG00000115165.9  protein_coding     CYTIP     1 OTTHUMG00000154551.6
## ENSG00000178919.8  protein_coding     FOXE1     2 OTTHUMG00000020333.1
##                           tag      logFC   AveExpr         t      P.Value
## ENSG00000178235.7        <NA>  -9.713210 -7.061852 -15.69420 6.897473e-33
## ENSG00000125378.15 retrogene  -6.851335 -4.697394 -13.35926 6.407172e-27
## ENSG00000115165.9        <NA>  -8.048797 -6.428105 -13.85358 3.373544e-28
## ENSG00000178919.8        <NA> -10.505225 -5.009710 -17.73915 6.336154e-38
##                        adj.P.Val        B
## ENSG00000178235.7  1.257961e-28 63.21938
## ENSG00000125378.15 5.842700e-23 50.08250
## ENSG00000115165.9  4.101780e-24 52.65615
## ENSG00000178919.8  2.311175e-33 74.73089
```

```
## Extract values from the genes of interest
exprs_heatmap <- vGene$E[rank(de_results$adj.P.Val) <= 30, ]

## And with friendlier column names
df <- as.data.frame(colData(rse_gene_SRP144355)[, c("sra_attribute.disease",
                                                    "sra_attribute.Sex",
                                                    "sra_attribute.age")])
colnames(df) <- c("DiseaseGroup", "Sex", "Age")
```

```
## We save the IDs of our 30 genes
nombres_originales <- rownames(exprs_heatmap)

rownames(exprs_heatmap) <- rowRanges(rse_gene_SRP144355)$gene_name[
    match(rownames(exprs_heatmap),
    rowRanges(rse_gene_SRP144355)$gene_id)
]

## heatmap with the gene names
pheatmap(
    exprs_heatmap,
    cluster_rows = TRUE,
    cluster_cols = TRUE,
    show_rownames = TRUE,
    show_colnames = FALSE,
    annotation_col = df
)
```




```r
# library RColorBrewer
col.group <- df$DiseaseGroup
levels(col.group) <- brewer.pal(nlevels(col.group), "Set1")
col.group <- as.character(col.group)
# MDS by groups of age
plotMDS(vGene$E, labels = df$DiseaseGroup, col = col.group)
```



Biological Analysis

The Hutchinson-Gilford Progeria Syndrome (HGPS) is a premature aging disorder caused by a mutation in the "LMNA" gene(1). In this study I aimed to explore wheter this mutation or syndrome affects individuals not only physical level, but also at the transciptomic level.

Upon analyzing the overexpressed and underexpressed genes, no absolute relationship was observed between individuals with HGPS. Howeever, we can conclude that there are certain genes that are notably overexpressed in individuals with HGPS compared to the majority of individuals without the syndrome.

Interestingly, the results show that age does not seem to create a distinct cluster, despite the fact that HGPS is a premature aging disorder.

In conclusion, while detecting a comprehensive transcriptomic relationship among individuals with HGPS is challenging, it is evident that some genes exhibit significant differences in expression. Therefore, further analysis of these genes, particularly in relation to the LMNA gene or the syndrome itself, could provide valuable insights into the underlying mechanisms of HGPS.

References

1. Cisneros B, García-Aguirre I, De Ita M, Arrieta-Cruz I, Rosas-Vargas H. Hutchinson-Gilford Progeria Syndrome: Cellular Mechanisms and Therapeutic Perspectives. Arch Med Res. 2023 Jul;54(5):102837. doi: 10.1016/j.arcmed.2023.06.002. Epub 2023 Jun 28. PMID: 37390702.