Skyhawk Therapeutics

Sanjana Gorlla

2022-09-26

Objective *Compare differentially expressed genes from two experiments & assess if the perturbations in the experiments may act in similar pathways*

```
# loading the required packages
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(tidyr)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --

## v tibble  3.1.8     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Tasks & Conclusion *Download the differential expression analysis results from two studies in the EBI expression atlas*

In the first study, cells were treated with a drug, nutlin which activates the P53 pathway

In the second study, the spliceosome gene PRPF8 was silenced with RNAi

Filter the results to only significant changes using the typical cutoffs of pval < 0.05 and abs(log2fc) > 1

Merge the results & prepare a scatter-plot to compare log2 fold-change observed in each gene btwn the experiments

comment on whether these results lead you to think that nutlin is acting on the same pathway as PRPF8 and why

Approach

1. Downloaded the differential expression analysis results from two studies in the EBI expression atlas following the links : (i). *https://www.ebi.ac.uk/gxa/experiments/E-GEOD-53965* (ii). *https://www.ebi.ac.uk/gxa/experiments/E-MTAB-3021/Results*
2. Filtering based on condition:
    - read.table() - load differential expression analysis results
    - subset() - gene id, gene name, log2 fold changes and p-values
    - Indexing + condtion - filtering
3. Merging the results from two experiments:
    - merge() - merged both the experiments based on columns
4. Visualizing & comparing:
    - ggplot() + geom_point - scatterplot of log2fc of nutlin and PRPF8
5. Analyzing the graph and interpretation

Experiment 1 - read.table() - load differential expression analysis results - subset() - gene id, gene name, log2 fold changes and p-values - Indexing + condtion - filtering

```
# loading the experiment1 dataset
exp1<- read.table(file = "/Users/sanjanagorlla/Desktop/skyhawk/E-GEOD-53965-A-AFFY-141-query-results (1)
                  sep = '\t', header = TRUE)

# Extract gene id, gene name, log2 fold changes and p-values of exp1  using subset()
exp1_filter<-subset(exp1, select = c("Gene.ID", "Gene.Name",
                                     "X10.micromolar..Nutlin.3.vs.DMSO.control.foldChange",
                                     "X10.micromolar..Nutlin.3.vs.DMSO.control.pValue"))


# Renaming column names

colnames(exp1_filter) <- c("Gene.ID", "Gene.Name","log2fc","pval" )

# Filter the results to only significant changes
# using the typical cutoffs of pval < 0.05 and abs(log2fc) > 1
# keep rows with p-value<0.05  & |log2FoldChange|>1
exp1_df<- data.frame(exp1_filter[exp1_filter$pval<0.05 &
                                   abs(exp1_filter$log2fc)>1,])

#rank tables high to low log2FoldChange_abs
exp1_final <- exp1_df %>%
  as.data.frame() %>%
  dplyr::arrange(dplyr::desc(log2fc))

# Top 10 genes
head(exp1_final, 10)
```

```
##            Gene.ID Gene.Name log2fc         pval
## 1  ENSG00000144452    ABCA12    3.1 3.193952e-06
## 2  ENSG00000115129    TP53I3    3.0 1.376041e-06
## 3  ENSG00000138271     GPR87    2.9 6.591147e-07
```

```
## 4  ENSG00000164938  TP53INP1    2.6 1.005697e-06
## 5  ENSG00000196562     SULF2    2.5 1.376041e-06
## 6  ENSG00000162643     DNAI3    2.4 5.014760e-06
## 7  ENSG00000124762    CDKN1A    2.3 1.373261e-05
## 8  ENSG00000159388      BTG2    2.3 1.171309e-06
## 9  ENSG00000080546     SESN1    2.2 1.376041e-06
## 10 ENSG00000026103       FAS    2.1 1.439095e-06
```

Experiment 2

```
# loading the experiment2 dataset
exp2<- read.table(file = "/Users/sanjanagorlla/Desktop/skyhawk/E-MTAB-3021-query-results.tsv",
                  sep = '\t', header = TRUE)

#Extract gene id, gene name, log2 fold changes and p-values of exp1  using subset()
exp2_filter <-subset(exp2, select = c("Gene.ID", "Gene.Name",
                                      "PRPF8.vs.control..foldChange",
                                      "PRPF8.vs.control.pValue"))

# Renaming column names
colnames(exp2_filter) <- c("Gene.ID", "Gene.Name","log2fc","pval" )

# Filter the results to only significant changes
# using the typical cutoffs of pval < 0.05 and abs(log2fc) > 1
# keep rows with p-value<0.05  & |log2FoldChange|>1
exp2_df<- data.frame(exp2_filter[exp2_filter$pval<0.05 &
                                   abs(exp2_filter$log2fc)>1,])

#rank tables high to low log2FoldChange_abs
exp2_final <- exp2_df %>%
  as.data.frame() %>%
  dplyr::arrange(dplyr::desc(log2fc))

# Top 10 genes
head(exp2_final, 10)
```

```
##            Gene.ID       Gene.Name log2fc         pval
## 1  ENSG00000153234           NR4A2    5.6 7.720080e-80
## 2  ENSG00000130513           GDF15    5.5 0.000000e+00
## 3  ENSG00000119508           NR4A3    4.8 3.167293e-68
## 4  ENSG00000143217         NECTIN4    3.9 9.212338e-34
## 5  ENSG00000128564             VGF    3.7 5.580199e-20
## 6  ENSG00000235609 ENSG00000235609    3.7 6.107431e-87
## 7  ENSG00000123358           NR4A1    3.6 5.783256e-35
## 8  ENSG00000124762          CDKN1A    3.4 0.000000e+00
## 9  ENSG00000158055           GRHL3    3.4 5.123532e-22
## 10 ENSG00000237437         ASS1P12    3.2 1.360937e-12
```

```
# sub-setting only geneID, genename and log2foldchange for further analysis
exp1df<-as.data.frame(subset(exp1_final, select = c("Gene.ID", "Gene.Name","log2fc")))
exp2df<-as.data.frame(subset(exp2_final, select = c("Gene.ID", "Gene.Name","log2fc")))
```

Merging the results

```r
# Merging the results from both the experiments based on common genes
merge_df12 <- merge(exp1df, exp2df, by=c("Gene.ID", "Gene.Name"))
merge_df12
```

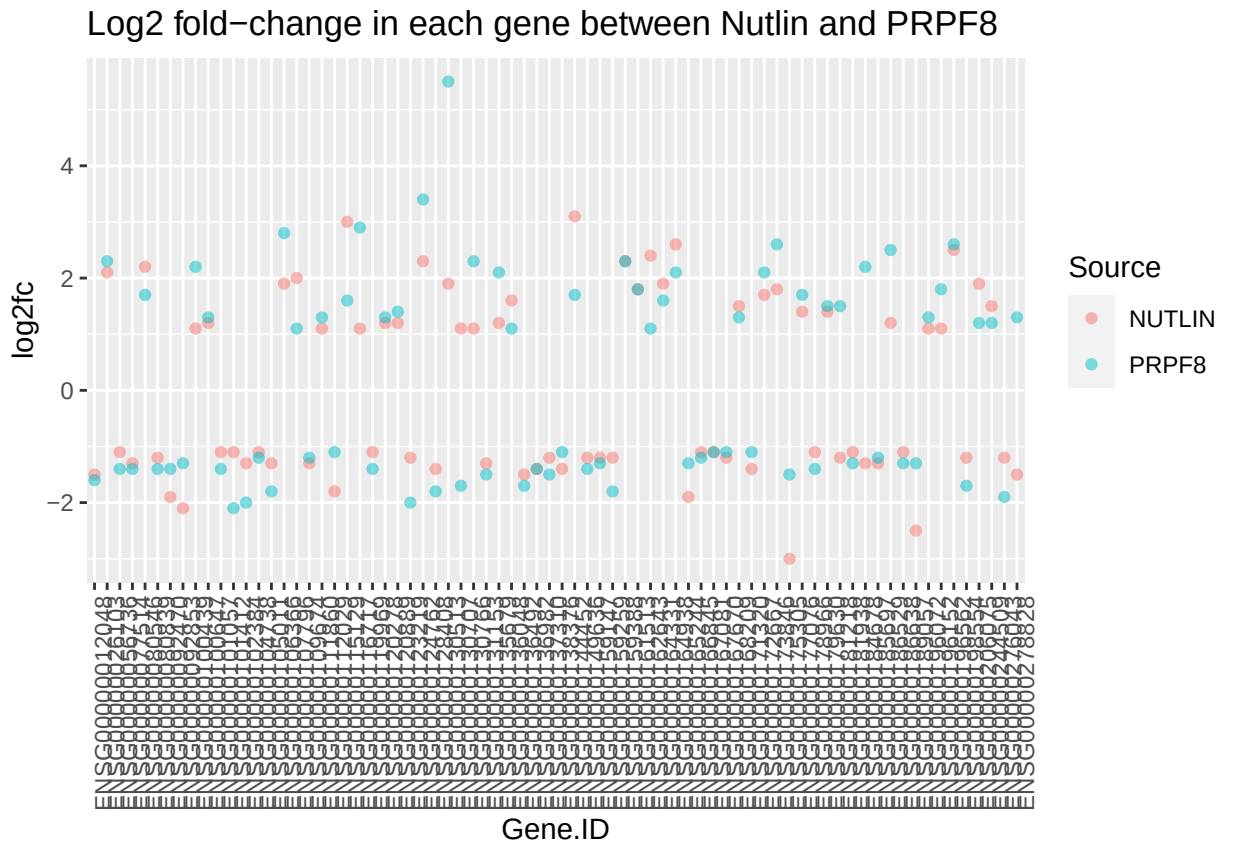
```
##            Gene.ID Gene.Name log2fc.x log2fc.y
## 1  ENSG00000012048     BRCA1     -1.5     -1.6
## 2  ENSG00000026103       FAS      2.1      2.3
## 3  ENSG00000056736    IL17RB     -1.1     -1.4
## 4  ENSG00000077514     POLD3     -1.3     -1.4
## 5  ENSG00000080546     SESN1      2.2      1.7
## 6  ENSG00000080839      RBL1     -1.2     -1.4
## 7  ENSG00000092470     WDR76     -1.9     -1.4
## 8  ENSG00000092853     CLSPN     -2.1     -1.3
## 9  ENSG00000100439     ABHD4      1.1      2.2
## 10 ENSG00000100647     SUSD6      1.2      1.3
## 11 ENSG00000101057     MYBL2     -1.1     -1.4
## 12 ENSG00000101412      E2F1     -1.1     -2.1
## 13 ENSG00000102384     CENPI     -1.3     -2.0
## 14 ENSG00000104738      MCM4     -1.1     -1.2
## 15 ENSG00000105011     ASF1B     -1.3     -1.8
## 16 ENSG00000106366  SERPINE1      1.9      2.8
## 17 ENSG00000107796     ACTA2      2.0      1.1
## 18 ENSG00000109674     NEIL3     -1.3     -1.2
## 19 ENSG00000111860    CEP85L      1.1      1.3
## 20 ENSG00000112029     FBXO5     -1.8     -1.1
## 21 ENSG00000115129    TP53I3      3.0      1.6
## 22 ENSG00000116717   GADD45A      1.1      2.9
## 23 ENSG00000119969     HELLS     -1.1     -1.4
## 24 ENSG00000120278   PLEKHG1      1.2      1.3
## 25 ENSG00000120889 TNFRSF10B      1.2      1.4
## 26 ENSG00000123219     CENPK     -1.2     -2.0
## 27 ENSG00000124762    CDKN1A      2.3      3.4
## 28 ENSG00000128408     RIBC2     -1.4     -1.8
## 29 ENSG00000130513     GDF15      1.9      5.5
## 30 ENSG00000130707      ASS1      1.1     -1.7
## 31 ENSG00000130766     SESN2      1.1      2.3
## 32 ENSG00000131153     GINS2     -1.3     -1.5
## 33 ENSG00000135679      MDM2      1.2      2.1
## 34 ENSG00000136048     DRAM1      1.6      1.1
## 35 ENSG00000136492     BRIP1     -1.5     -1.7
## 36 ENSG00000136982     DSCC1     -1.4     -1.4
## 37 ENSG00000137310     TCF19     -1.2     -1.5
## 38 ENSG00000138376     BARD1     -1.4     -1.1
## 39 ENSG00000144452    ABCA12      3.1      1.7
## 40 ENSG00000149636      DSN1     -1.2     -1.4
## 41 ENSG00000159147    DONSON     -1.2     -1.3
## 42 ENSG00000159259    CHAF1B     -1.2     -1.8
## 43 ENSG00000159388      BTG2      2.3      2.3
## 44 ENSG00000161513      FDXR      1.8      1.8
## 45 ENSG00000162643     DNAI3      2.4      1.1
## 46 ENSG00000164331    ANKRA2      1.9      1.6
## 47 ENSG00000164938  TP53INP1      2.6      2.1
## 48 ENSG00000165244    ZNF367     -1.9     -1.3
```

```
## 49 ENSG00000166845  C18orf54     -1.1     -1.2
## 50 ENSG00000167081      PBX3     -1.1     -1.1
## 51 ENSG00000167670    CHAF1A     -1.2     -1.1
## 52 ENSG00000168209     DDIT4      1.5      1.3
## 53 ENSG00000171320     ESCO2     -1.4     -1.1
## 54 ENSG00000172667     ZMAT3      1.7      2.1
## 55 ENSG00000173846      PLK3      1.8      2.6
## 56 ENSG00000175305     CCNE2     -3.0     -1.5
## 57 ENSG00000177076     ACER2      1.4      1.7
## 58 ENSG00000178966      RMI1     -1.1     -1.4
## 59 ENSG00000179630     LACC1      1.4      1.5
## 60 ENSG00000181218      H2AW     -1.2      1.5
## 61 ENSG00000181938     GINS3     -1.1     -1.3
## 62 ENSG00000184678    H2BC21     -1.3      2.2
## 63 ENSG00000185697     MYBL1     -1.3     -1.2
## 64 ENSG00000186529    CYP4F3      1.2      2.5
## 65 ENSG00000186638     KIF24     -1.1     -1.3
## 66 ENSG00000189057   FAM111B     -2.5     -1.3
## 67 ENSG00000196072   BLOC1S2      1.1      1.3
## 68 ENSG00000196152     ZNF79      1.1      1.8
## 69 ENSG00000196562     SULF2      2.5      2.6
## 70 ENSG00000198554     WDHD1     -1.2     -1.7
## 71 ENSG00000206075  SERPINB5      1.9      1.2
## 72 ENSG00000244509  APOBEC3C      1.5      1.2
## 73 ENSG00000276043     UHRF1     -1.2     -1.9
## 74 ENSG00000278828     H3C10     -1.5      1.3
```

```r
# Final data for analysis
log2fc <- c(merge_df12$log2fc.x, merge_df12$log2fc.y)
Gene.ID <- c(merge_df12$Gene.ID, merge_df12$Gene.ID)
final_data <- data.frame(Gene.ID, log2fc)
final_data$Source <- c(replicate(74,"NUTLIN"),replicate(74,"PRPF8"))
```

Scatterplot for comparision

```r
#scatter-plot to compare log2 fold-change observed in each gene between the experiments
ggplot(final_data, aes(x=Gene.ID, y= log2fc, color=Source))+
  geom_point(alpha = 0.5) + theme(axis.text.x = element_text(angle = 90, hjust = 3)) +
  ggtitle("Log2 fold-change in each gene between Nutlin and PRPF8")
```



Scatterplot for comparision

1. **Cells were treated with a drug, nutlin which activates the P53 pathway**: The connection between MDM2 and p53 is competitively blocked by nutlin-3, which stops p53 from being ubiquitinated and degraded. In numerous model systems, nutlin-3 treatment has been demonstrated to restore p53 activation and the concomitant production of apoptosis, senescence, or reversible cell cycle arrest.
2. **The spliceosome gene PRPF8 was silenced with RNAi** : Alterations in RNA splicing patterns across the human transcriptome that occur in conditions of restricted cellular PRPF8 abundance are defined by the altered splicing of introns with weak 5' splice sites.

Scatterplot : By comparing the genes regulated by both the pathways : we can observe that the genes expressed in P53 pathway due to Nutlin acts very similar to the PRPF8 pathway and is evident with the scatter-plot.

Except the genes PLK3 & KIF24 treated with Nutlin have negative log2fc values which means that the gene is less expressed with nutlin.

All the majority of genes have values >1 which are highly expressed.

Therefore, there is a high possibility that both the experimental perturbations act along similar pathways.

Based on the genes compared nutlin is acting on the same pathway as PRPF8