MetaStudy

Moritz Madern

2023-01-14

This script compiles differential expression data from files of different experiments (each provided in a standardized format & containing gene-wise FC vs p-values) and performs meta-analysis on specified genes.

Specify parameters

```
## specify file path to datasets to read in (all datasets should be located in the same folder)
filepath_datasets = "./Datasets"


## specify file path to table that defines the conditions within those datasets of interest
filepath_tableConditions = "Condition_datasets_11012023.csv"


## specify genes to be analyzed (note that they will be concverted to capital letters to make pattern s
vector_geneNames = c("Cdc45", "eIF2A","Rpi7","Gapdh","Ncor1","Stat1", "Gata3", "Stat4")


## specify subset of those genes (from vector_geneNames) to be illustrated in a Heatmap.
vector_geneNames_heatmap = c("Cdc45","Ncor1","Stat1", "Stat4", "Gata3")


## specify heatmap color scheme. Suppurted colorschemes are "viridis" and "redblue"
heatmap_colorscheme = "redblue"
```

Read in datasets

Read in conditions table:

```
##   dataset condition.number
## 1       1       condition1
## 2       2       condition3
## 3       3       condition1
## 4       4       condition2
## 5       5       condition1
## 6       7       condition2
## 7       8       condition2
## 8       9       condition3
## 9      11       condition1
```

Read in all datasets as specified by filepath_datasets:

```
## [1] "./Datasets/Dataset1  E-GEOD-57945-query-results-.tsv"
## [1] "./Datasets/Dataset11 E-GEOD-42768-A-AFFY-130-query-results.tsv"
## [1] "./Datasets/Dataset2 E-MTAB-7860-query-results-2.tsv"
## [1] "./Datasets/Dataset3 E-GEOD-2461-A-AFFY-33-query-results.tsv"
## [1] "./Datasets/Dataset4 E-GEOD-65114-A-GEOD-16686-query-results.tsv"
## [1] "./Datasets/Dataset5 E-GEOD-6731-A-AFFY-1-query-results.tsv"
## [1] "./Datasets/Dataset7 E-MTAB-9850-query-results-2.tsv"
## [1] "./Datasets/Dataset8 E-GEOD-20621-A-AFFY-45-query-results-2.tsv"
## [1] "./Datasets/Dataset9 E-GEOD-27302-A-AFFY-45-query-results.tsv"
```

Note: Input data needs required standard formatting! Here is an example:

```
##              Gene.ID Gene.Name Design.Element condition1.foldChange
## 1 ENSMUSG00000000003      Pbsn     1449320_at                    NA
## 2 ENSMUSG00000000028     Cdc45     1416575_at                   0.3
## 3 ENSMUSG00000000031       H19   1448194_a_at                    NA
## 4 ENSMUSG00000000037     Scml2   1421514_a_at                    NA
## 5 ENSMUSG00000000049      Apoh     1416677_at                   0.9
## 6 ENSMUSG00000000056      Narf     1451678_at                    NA
##   condition1.pValue X2.week.vs.0.week.tStat condition2.foldChange
## 1                NA                     NA                   1.5
## 2        0.02063145                4.654057                   0.7
## 3                NA                     NA                   0.4
## 4                NA                     NA                  -0.1
## 5        0.45467262                1.351620                    NA
## 6                NA                     NA                  -0.5
##   condition2.pValue X4.week.vs.0.week.tStat condition3.foldChange
## 1      0.3376295888                1.221822                    NA
## 2      0.0003306578                7.593565                   0.7
## 3      0.1316269670                1.961724                   0.5
## 4      0.2162355323               -1.585398                   0.1
## 5                NA                     NA                   0.2
## 6      0.0005123569               -7.010167                  -0.2
##   condition3.pValue X6.week.vs.0.week.tStat
## 1                NA                     NA
## 2      0.0003571848               7.1609105
## 3      0.0339536615               2.9269671
## 4      0.6257187039               0.6249972
## 5      0.1797140710               1.7151928
## 6      0.0438151488              -2.7407916
```

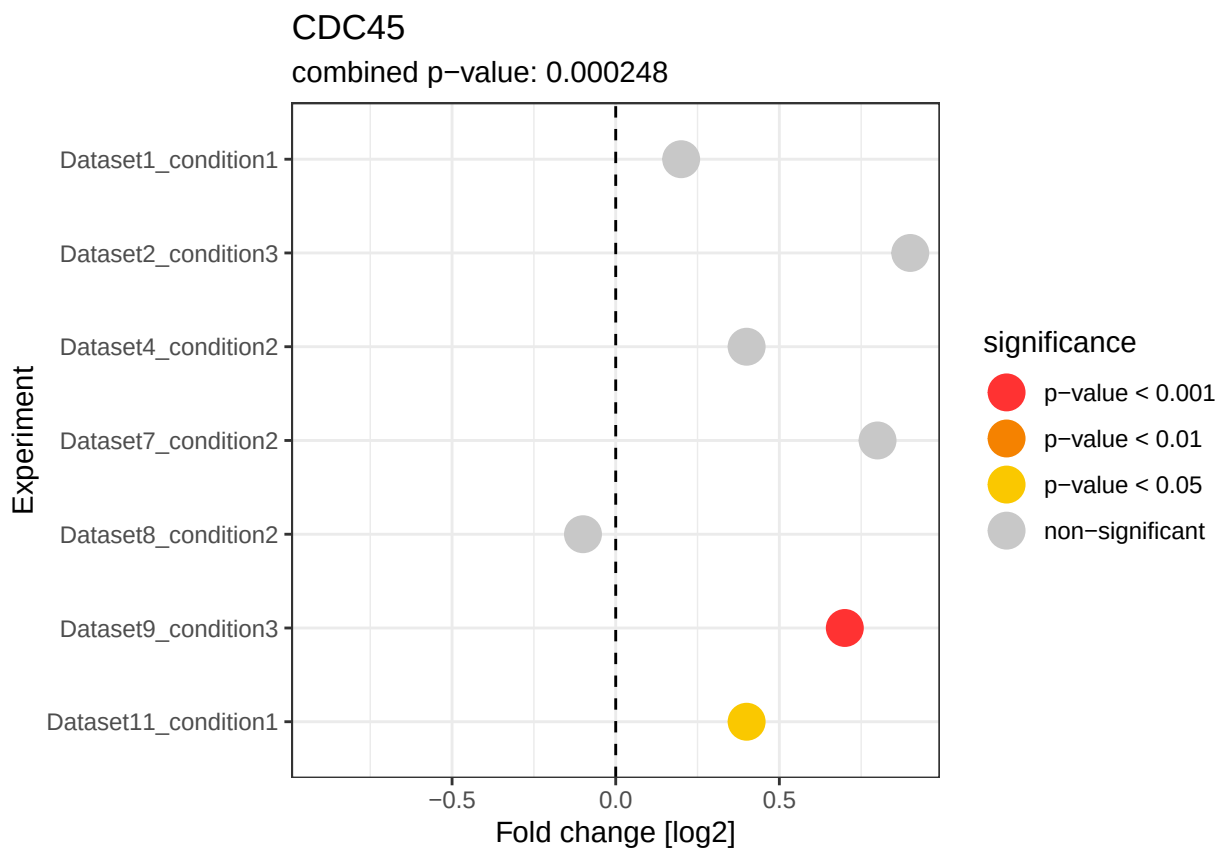
Note: For Dataset 5, formatting of the .tsv file is somehow different which causes somehing weird to happen when reading into R. I found an automated workaround (i.e. without adapting the script), but this resulted in the loss of the very first row (Gene TSPAN6) for this dataset.

Plot data for each specified gene

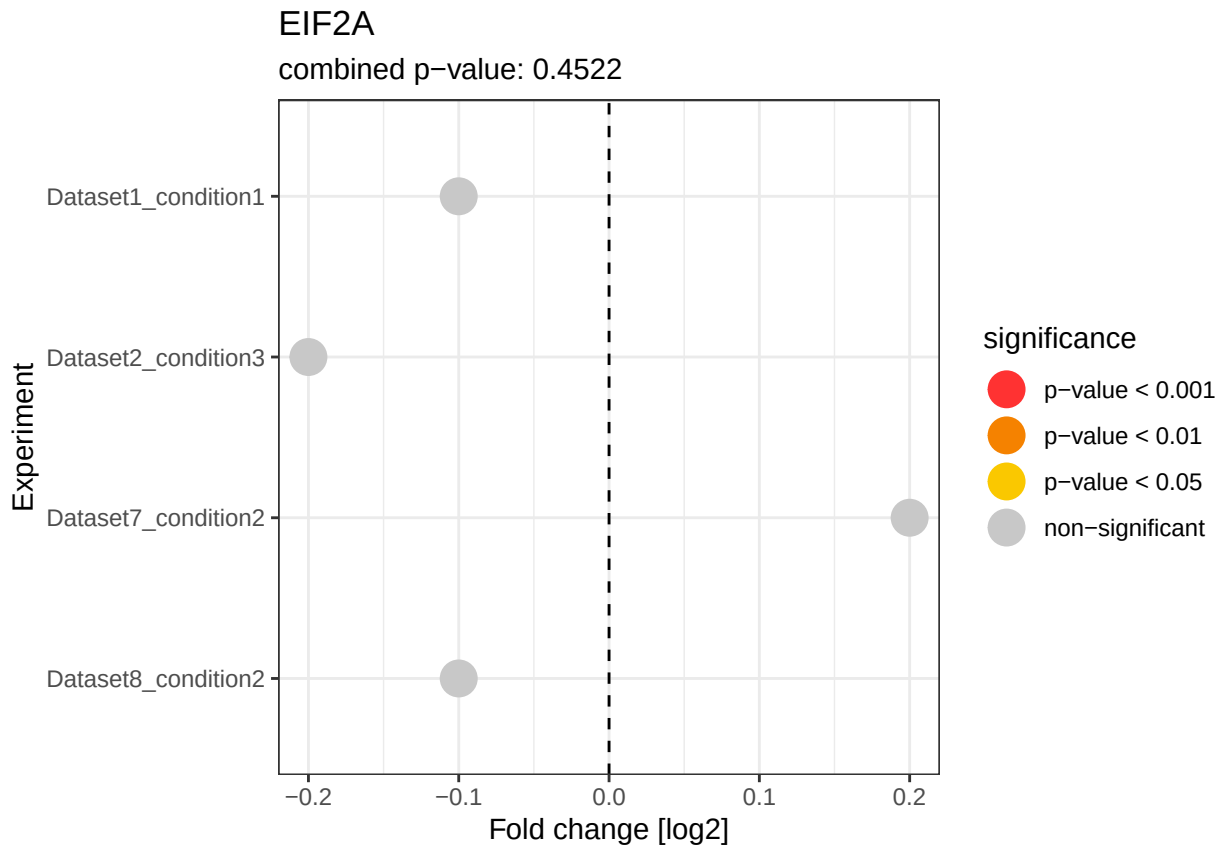
Plots show foldChange vs experiment, points are colored by statistical significance. Note that experiments in which the respective gene was not quantified are not shown. Furthermore, for each gene, the individual p-values were combined to an overall p-value via Fisher's method (https://en.wikipedia.org/wiki/Fisher%

27s_method), rounded to 6 digits. This method assumes that the individual p-values to be pooled are independent but test for the same Null hypothesis. If these assumptions are not met, ignore the combined p-value.

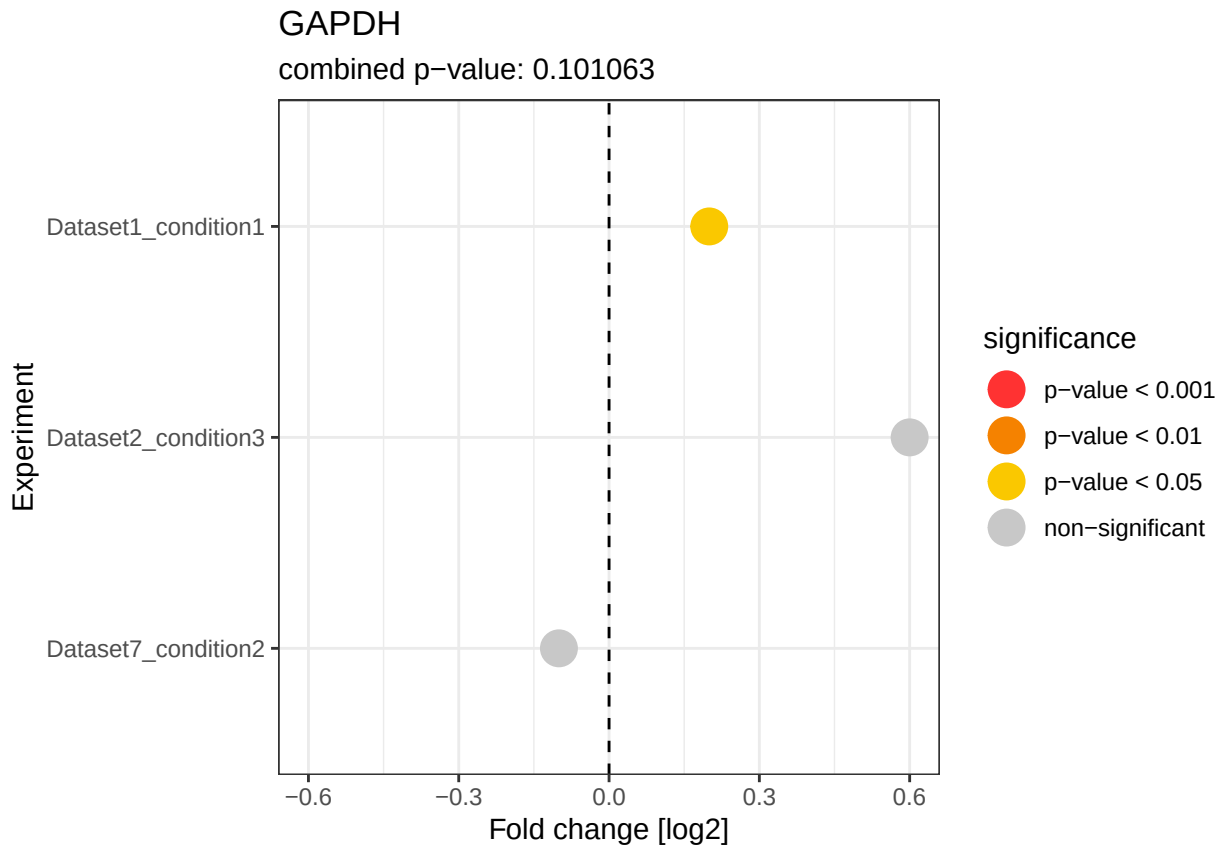
```
## [[1]]
```



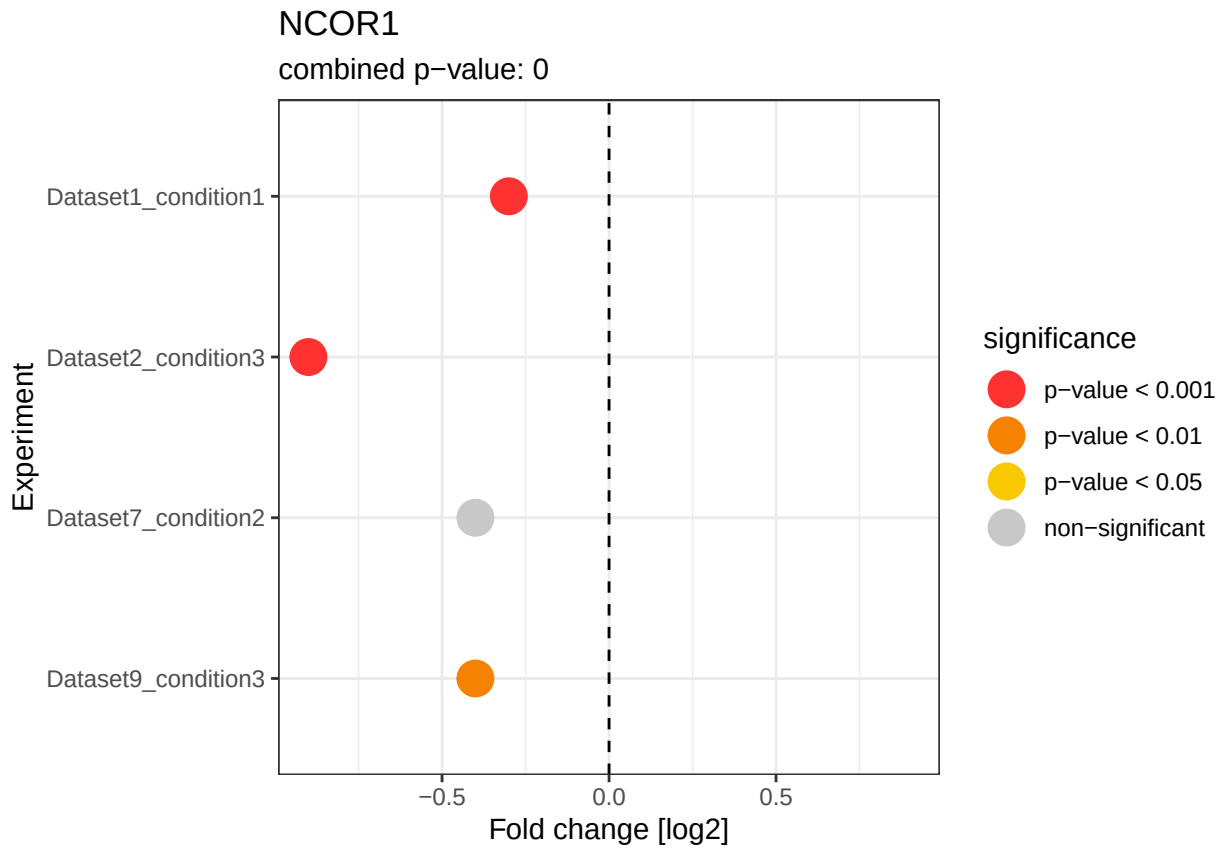

```
##
## [[2]]
```


EIF2A
combined p−value: 0.4522
Dataset1_condition1
Dataset2_condition3
Dataset7_condition2
Dataset8_condition2
Experiment
−0.2
−0.1
0.0
0.1
0.2
Fold change [log2]
significance
p−value < 0.001
p−value < 0.01
p−value < 0.05
non−significant

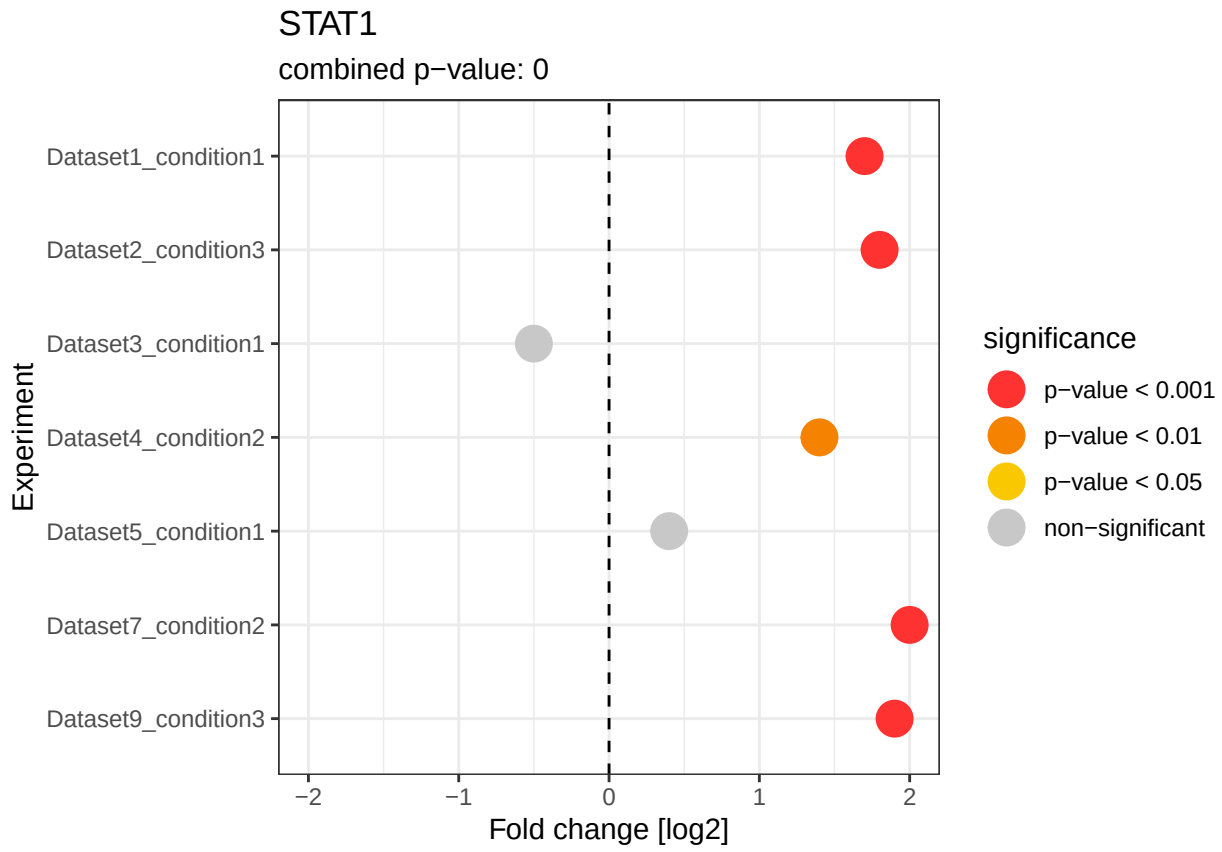
```
##
## [[3]]
```


GAPDH
combined p−value: 0.101063
Dataset1_condition1
Dataset2_condition3
Dataset7_condition2
Experiment
−0.6
−0.3
0.0
0.3
0.6
Fold change [log2]
significance
p−value < 0.001
p−value < 0.01
p−value < 0.05
non−significant

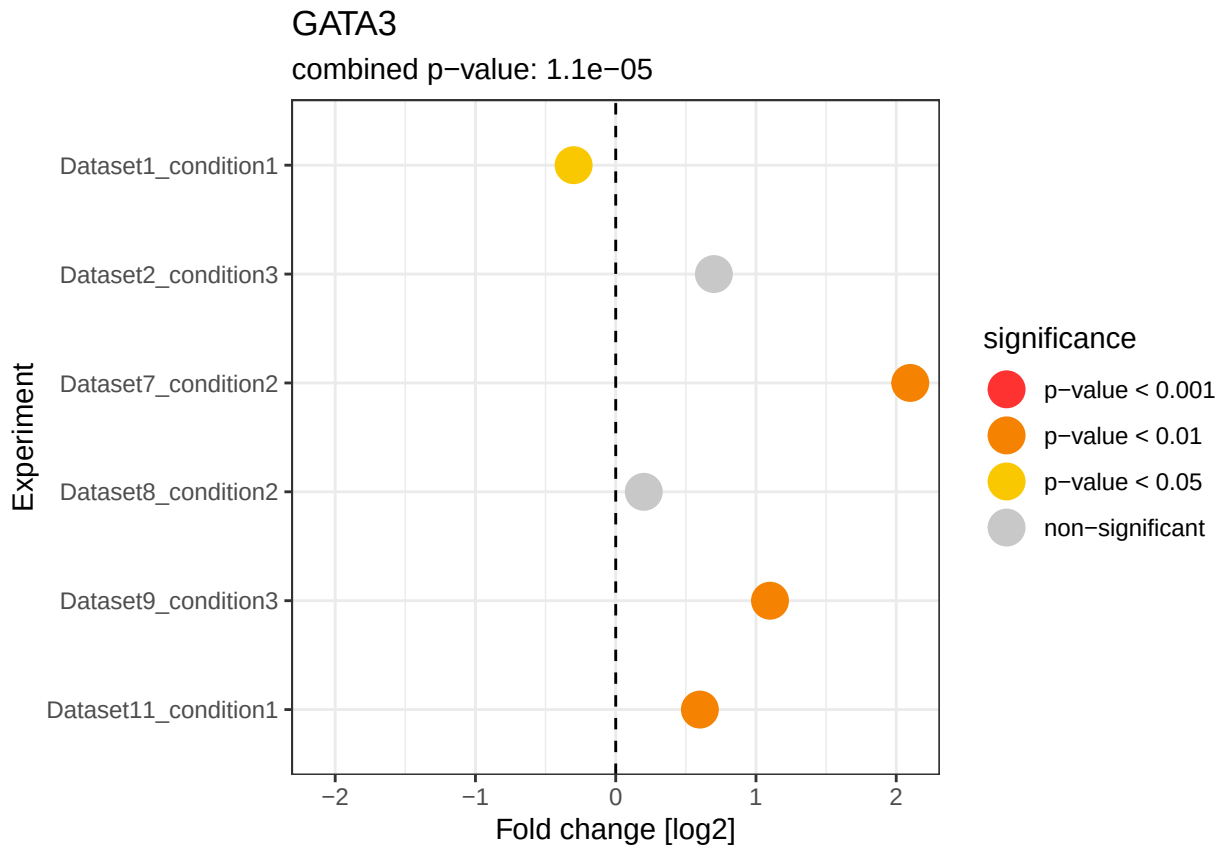
```
##
## [[4]]
```


NCOR1
combined p−value: 0
Dataset1_condition1
Dataset2_condition3
Dataset7_condition2
Dataset9_condition3
Experiment
−0.5
0.0
0.5
Fold change [log2]
significance
p−value < 0.001
p−value < 0.01
p−value < 0.05
non−significant

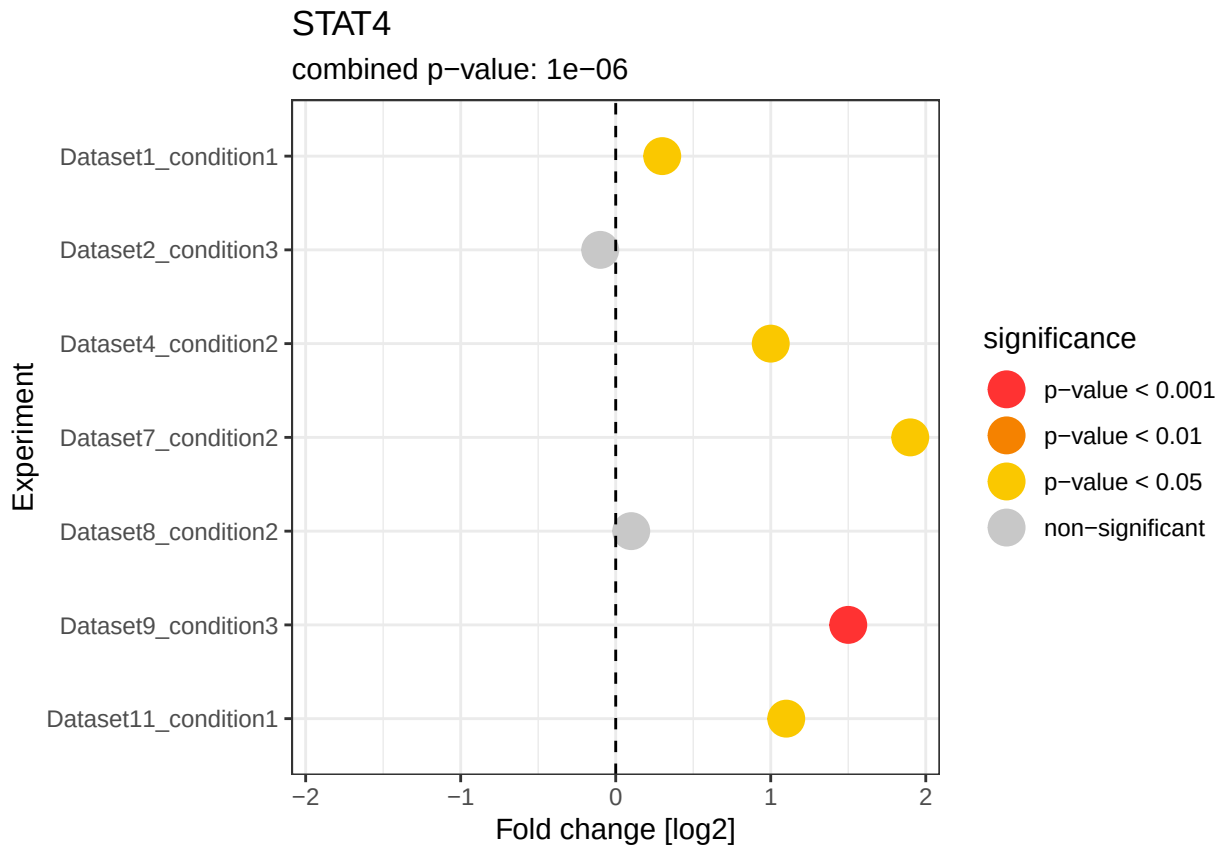
```
##
## [[5]]
```


STAT1
combined p−value: 0
Dataset1_condition1
Dataset2_condition3
Dataset3_condition1
Dataset4_condition2
Dataset5_condition1
Dataset7_condition2
Dataset9_condition3
Experiment
−2
−1
0
1
2
Fold change [log2]
significance
p−value < 0.001
p−value < 0.01
p−value < 0.05
non−significant

```
##
## [[6]]
```


GATA3
combined p−value: 1.1e−05
Dataset1_condition1
Dataset2_condition3
Dataset7_condition2
Dataset8_condition2
Dataset9_condition3
Dataset11_condition1
Experiment
−2
−1
0
1
2
Fold change [log2]
significance
p−value < 0.001
p−value < 0.01
p−value < 0.05
non−significant

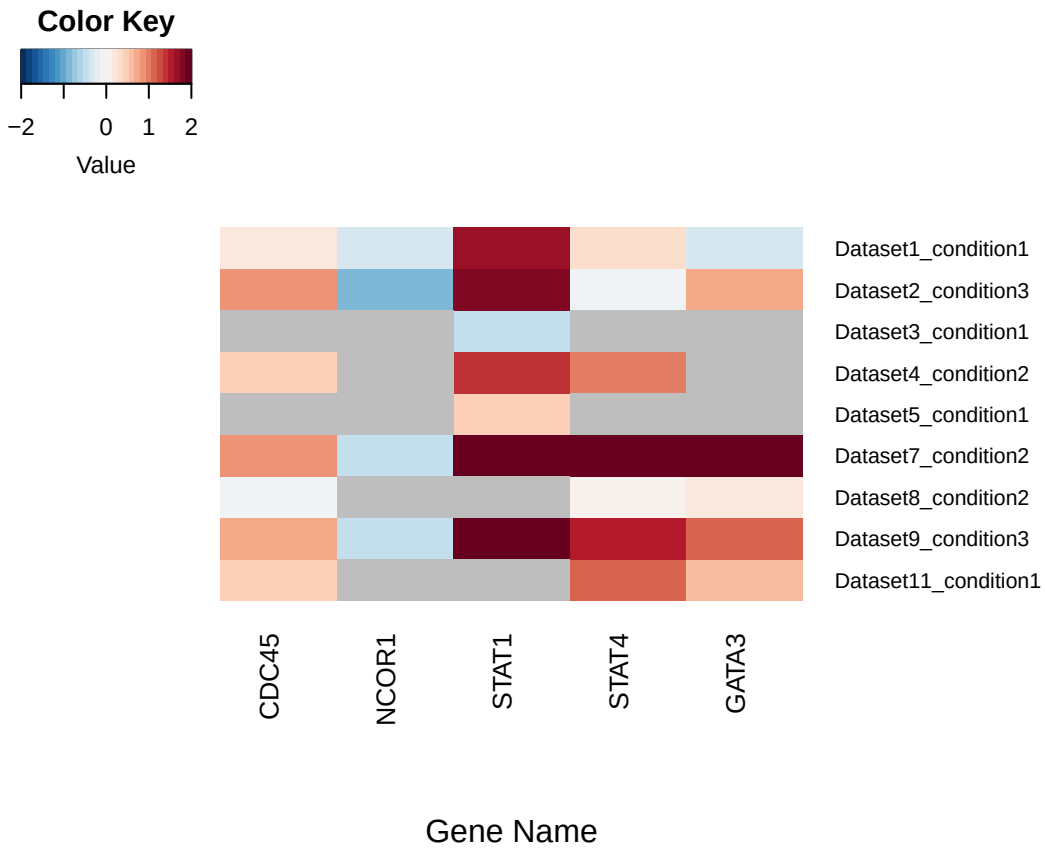
```
##
## [[7]]
```



Note: Upon compiling the script to pdf, each plot is saved as separate pdf and png file in the newly generated folder called "figures" in your working directory.

Plot Heatmap

For selected genes, a heatmap is plotted that depicts fold changes. Missing values (i.e. NAs) are colored as grey.


Color Key
−2
0
1
2
Value
Dataset1_condition1
Dataset2_condition3
Dataset3_condition1
Dataset4_condition2
Dataset5_condition1
Dataset7_condition2
Dataset8_condition2
Dataset9_condition3
Dataset11_condition1
CDC45
NCOR1
STAT1
STAT4
GATA3
Gene Name