castella_overlap_lc

2024-01-19

Supplementary Figure 4

```
library(readxl)
library(tidyr)
library(dplyr)
library(biomaRt)
library(here); i_am("R/18_castella_overlap.Rmd")
```

```
castella = read_excel(here("annotations/Castella2023/Castella_etal_2023_Supplementary_Table_1_DiffTrans
                           sheet="isoforms-de")
#castella = read.delim(file.path(other_data, "comparison_to_castella_etal_2023.txt"))
head(castella)
```

```
## # A tibble: 6 x 7
##   TranscriptID    baseMean log2FoldChange lfcSE  stat   pvalue     padj
##   <chr>              <dbl>          <dbl> <dbl> <dbl>    <dbl>    <dbl>
## 1 ENST00000416363     897.           8.64 0.512  16.9 9.54e-64 4.01e-59
## 2 ENST00000400394   13676.          16.8  1.32   12.7 3.59e-37 6.03e-33
## 3 ENST00000419277    1783.          13.9  1.24   11.2 3.60e-29 4.31e-25
## 4 ENST00000611398    1043.          13.1  1.26   10.4 3.06e-25 3.21e-21
## 5 ENST00000425571     761.          12.6  1.23   10.3 1.10e-24 1.03e-20
## 6 ENST00000309539     261.          -2.73 0.272 -10.0 9.37e-24 7.87e-20
```

```
str(castella$TranscriptID)
```

```
##  chr [1:803] "ENST00000416363" "ENST00000400394" "ENST00000419277" ...
```

```
length(unique(castella$TranscriptID))
```

```
## [1] 803
```

loading leafcutter ensembl transcripts

```
junctions = read.delim(here("31_leafcutter/three_database_info_sig_junctions.tsv"))

nrow(junctions)
```

```
## [1] 777
```

```r
length(unique(junctions$gene)) #missing ~8 genes with unknown gene names
```

```
## [1] 516
```

```r
length(unique(junctions$transcript_ids))
```

```
## [1] 720
```

```r
head(junctions)
```

```
##    cluster_id annotation   chr     start       end strand   deltapsi
## 1 clu_10104_-    gencode  chr9 123401912 123403402      - -0.1076770
## 2 clu_10181_-    gencode  chr9 128266325 128267458      - -0.1130214
## 3 clu_10209_-    gencode  chr9 129108077 129110483      -  0.1074815
## 4 clu_10638_+    gencode chr12  26195951  26224293      + -0.1222677
## 5 clu_10638_+    gencode chr12  26195531  26224293      +  0.1056313
## 6 clu_10654_+    gencode chr12  27380404  27385481      + -0.1918059
##       p.adjust
## 1 1.437183e-02
## 2 8.760649e-15
## 3 2.104379e-08
## 4 1.870231e-02
## 5 1.870231e-02
## 6 1.084347e-07
##
## 1                                                                                            ENST0000(
## 2
## 3
## 4
## 5
## 6 ENST00000395901.6,ENST00000542388.1,ENST00000311001.9,ENST00000261178.9,ENST00000457040.6,ENST0000(
##   min_intron_number mode_intron_number    gene                biotype
## 1                 1                  1 DENND1A protein_coding,lncRNA
## 2                 5                  5  GOLGA2         protein_coding
## 3                 1                  1    CRAT         protein_coding
## 4                 1                  1    SSPN         protein_coding
## 5                 1                  1    SSPN protein_coding,lncRNA
## 6                 2                  3  ARNTL2         protein_coding
##   genes_in_cluster is_first_intron condition  conditions num_introns
## 1          DENND1A            TRUE     white       white           1
## 2           GOLGA2           FALSE     white       white           1
## 3             CRAT            TRUE     beige       beige           1
## 4             SSPN            TRUE     white beige&white           2
## 5             SSPN            TRUE     beige beige&white           2
## 6           ARNTL2           FALSE     white beige&white           2
```

```r
lc_trans = separate_longer_delim(junctions, transcript_ids, delim = ",")
head(lc_trans)
```

```
##    cluster_id annotation  chr     start       end strand   deltapsi
## 1 clu_10104_-    gencode chr9 123401912 123403402      - -0.1076770
```

```
## 2 clu_10104_-    gencode chr9 123401912 123403402      - -0.1076770
## 3 clu_10104_-    gencode chr9 123401912 123403402      - -0.1076770
## 4 clu_10181_-    gencode chr9 128266325 128267458      - -0.1130214
## 5 clu_10209_-    gencode chr9 129108077 129110483      -  0.1074815
## 6 clu_10209_-    gencode chr9 129108077 129110483      -  0.1074815
##       p.adjust    transcript_ids min_intron_number mode_intron_number    gene
## 1 1.437183e-02 ENST00000394215.6                 1                   1 DENND1A
## 2 1.437183e-02 ENST00000373620.7                 1                   1 DENND1A
## 3 1.437183e-02 ENST00000475421.1                 1                   1 DENND1A
## 4 8.760649e-15 ENST00000458730.2                 5                   5  GOLGA2
## 5 2.104379e-08 ENST00000393384.3                 1                   1    CRAT
## 6 2.104379e-08 ENST00000318080.7                 1                   1    CRAT
##                biotype genes_in_cluster is_first_intron condition conditions
## 1 protein_coding,lncRNA          DENND1A            TRUE     white      white
## 2 protein_coding,lncRNA          DENND1A            TRUE     white      white
## 3 protein_coding,lncRNA          DENND1A            TRUE     white      white
## 4        protein_coding           GOLGA2           FALSE     white      white
## 5        protein_coding             CRAT            TRUE     beige      beige
## 6        protein_coding             CRAT            TRUE     beige      beige
##   num_introns
## 1           1
## 2           1
## 3           1
## 4           1
## 5           1
## 6           1
```

```
nrow(lc_trans)
```

```
## [1] 2233
```

```
length(unique(lc_trans$transcript_ids))
```

```
## [1] 2094
```

```
length(unique(grep("^ENS", lc_trans$transcript_ids)))
```

```
## [1] 1987
```

```
lc_trans$transcript_ids = gsub("\\..*","", lc_trans$transcript_ids)
```

```
library(biomaRt)
mart <- useMart(biomart = "ensembl",
  dataset = "hsapiens_gene_ensembl",
   host = "https://sep2019.archive.ensembl.org")

annot = getBM(c("external_gene_name", "ensembl_gene_id", "ensembl_transcript_id", "external_transcript_
              filters = "ensembl_transcript_id",
              values =  castella$TranscriptID,
              mart = mart, useCache = F)

head(annot, n=2); dim(annot)
```

```
##   external_gene_name ensembl_gene_id ensembl_transcript_id
## 1         AC004556.3 ENSG00000276345       ENST00000612848
## 2               RPS9 ENSG00000278081       ENST00000630852
##   external_transcript_name
## 1           AC004556.3-201
## 2                RPS9-255
```

```
## [1] 797   4
```

```
#Tidying up the annot table
colnames(annot)[1] = "gene_name"

#Add gene names to filt_series
castella = merge(annot, castella, by.y= "TranscriptID",
                        by.x = "ensembl_transcript_id", sort=FALSE)
#head(tpm)
remove(annot)
```

```
nrow(castella) #6 transcripts cannot be found by ensembl
```

```
## [1] 797
```

```
head(castella)
```

```
##   ensembl_transcript_id  gene_name ensembl_gene_id external_transcript_name
## 1       ENST00000612848 AC004556.3 ENSG00000276345           AC004556.3-201
## 2       ENST00000630852       RPS9 ENSG00000278081                RPS9-255
## 3       ENST00000613328 AL662796.1 ENSG00000277263           AL662796.1-201
## 4       ENST00000621600       CCL4 ENSG00000277943                CCL4-208
## 5       ENST00000613036       CCL4 ENSG00000277943                CCL4-207
## 6       ENST00000485428   ALDH18A1 ENSG00000059573             ALDH18A1-204
##     baseMean log2FoldChange     lfcSE      stat   pvalue     padj
## 1  331.32067      6.6186874 0.8053731  8.218163 2.07e-16 7.55e-13
## 2 1797.46232      0.9024061 0.1580807  5.708515 1.14e-08 8.78e-06
## 3 1063.11497    -10.0918623 1.4815721 -6.811590 9.65e-12 1.45e-08
## 4 1646.58519     -3.1975006 0.5153564 -6.204445 5.49e-10 5.69e-07
## 5   84.03103     -3.6358717 0.6830706 -5.322835 1.02e-07 5.88e-05
## 6   34.02005      8.0962554 1.2477614  6.488625 8.66e-11 1.09e-07
```

```
summary(castella$ensembl_transcript_id %in% lc_trans$transcript_ids)
```

```
##    Mode   FALSE    TRUE
## logical     789       8
```

```
castella[castella$ensembl_transcript_id %in% lc_trans$transcript_ids,]
```

```
##     ensembl_transcript_id gene_name ensembl_gene_id external_transcript_name
## 140       ENST00000400706    WASH8P ENSG00000226210               WASH8P-201
## 191       ENST00000261439    TBC1D1 ENSG00000065882               TBC1D1-201
## 360       ENST00000182377      FAR2 ENSG00000064763                 FAR2-201
```

```
## 497       ENST00000467894      PHC2 ENSG00000134686                PHC2-206
## 662       ENST00000404752     STON1 ENSG00000243244               STON1-201
## 744       ENST00000482881     MAST2 ENSG00000086015               MAST2-207
## 768       ENST00000419955    ADHFE1 ENSG00000147576              ADHFE1-205
## 784       ENST00000644959      OPA1 ENSG00000198836                OPA1-226
##       baseMean log2FoldChange     lfcSE      stat      pvalue        padj
## 140  985.97774      -1.144270 0.3194296 -3.582229 0.000340674 0.039790683
## 191 1256.36435      -1.023181 0.2270188 -4.507032 0.000006570 0.002044759
## 360   88.16492       4.202529 0.8158888  5.150860 0.000000259 0.000132776
## 497  107.02797      -2.446358 0.6225960 -3.929287 0.000085200 0.015509048
## 662 2747.99251      -1.368380 0.3645077 -3.754049 0.000174001 0.025324837
## 744   15.07742      -2.526378 0.6836602 -3.695371 0.000219566 0.029644579
## 768  313.12911       0.775699 0.1968520  3.940520 0.000081300 0.015072702
## 784  200.60603       1.022024 0.2740479  3.729362 0.000191965 0.027277586
```

```
lc_trans[lc_trans$transcript_ids %in% castella$ensembl_transcript_id,]
```

```
##       cluster_id annotation   chr     start       end strand   deltapsi
## 23   clu_10672_+    gencode chr12  29223894  29270412      +  0.2412781
## 154  clu_13672_+    gencode  chr1  45997799  46002805      +  0.2439310
## 561  clu_19197_+    gencode  chr3 193618936 193631612      + -0.1381253
## 730  clu_21708_+    gencode  chr8  66452105  66453710      + -0.1024494
## 895  clu_24930_-    gencode chr12     17859     18037      -  0.1123566
## 1004 clu_27829_-    gencode  chr1  33334292  33349597      - -0.1459678
## 1238 clu_31216_+    gencode  chr2  48530216  48580587      + -0.2204830
## 1759 clu_39642_+    gencode  chr4  38049898  38054199      + -0.1464586
##          p.adjust  transcript_ids min_intron_number mode_intron_number
## 23   1.387532e-62 ENST00000182377                 1                  1
## 154  1.191401e-15 ENST00000482881                 2                  2
## 561  6.342723e-18 ENST00000644959                 1                  5
## 730  3.604441e-03 ENST00000419955                 7                  7
## 895  8.291420e-03 ENST00000400706                 4                  4
## 1004 8.402584e-09 ENST00000467894                 1                  1
## 1238 3.171417e-40 ENST00000404752                 1                  1
## 1759 1.216652e-13 ENST00000261439                 2                  2
##                   gene                                              biotype
## 23                FAR2                                       protein_coding
## 154              MAST2                          protein_coding,retained_intron
## 561               OPA1 nonsense_mediated_decay,protein_coding,retained_intron
## 730  AC009879.3,ADHFE1                              nonsense_mediated_decay
## 895             WASH8P                               unprocessed_pseudogene
## 1004              PHC2                                protein_coding,lncRNA
## 1238             STON1                                       protein_coding
## 1759            TBC1D1                                       protein_coding
##         genes_in_cluster is_first_intron condition  conditions num_introns
## 23                  FAR2            TRUE     beige beige&white           2
## 154                MAST2           FALSE     beige beige&white           2
## 561                 OPA1            TRUE     white       white           1
## 730    AC009879.3,ADHFE1           FALSE     white beige&white           2
## 895                WASH8P           FALSE     beige       beige           1
## 1004                PHC2            TRUE     white beige&white           2
## 1238 STON1,STON1-GTF2A1L            TRUE     white beige&white           3
## 1759              TBC1D1           FALSE     white       white           1
```

```
lc_trans$gene = gsub(",.*","", lc_trans$gene)
```

castella ensembl_transcript_id log2FoldChange direction (white|beige)

```
castella = mutate(castella, condition = if_else(log2FoldChange > 0, "beige",  "white"))
castella[grep("CKMT", castella$gene_name), ] #double checking logfc direction check
```

```
##     ensembl_transcript_id gene_name ensembl_gene_id external_transcript_name
## 138       ENST00000515615     CKMT2 ENSG00000131730                CKMT2-212
## 139       ENST00000437669     CKMT2 ENSG00000131730                CKMT2-203
## 390       ENST00000437534    CKMT1B ENSG00000237289               CKMT1B-205
##       baseMean log2FoldChange    lfcSE     stat   pvalue        padj condition
## 138   21.95421       7.446402 1.635463 4.553084 5.29e-06 0.001707523     beige
## 139 1025.94310       4.183403 1.064604 3.929538 8.51e-05 0.015509048     beige
## 390  538.75403       4.801066 1.157486 4.147838 3.36e-05 0.007576838     beige
```

lc_trans transcript_ids deltapsi condition (white|beige)

```
head(lc_trans)
```

```
##    cluster_id annotation  chr     start       end strand   deltapsi
## 1 clu_10104_-    gencode chr9 123401912 123403402      - -0.1076770
## 2 clu_10104_-    gencode chr9 123401912 123403402      - -0.1076770
## 3 clu_10104_-    gencode chr9 123401912 123403402      - -0.1076770
## 4 clu_10181_-    gencode chr9 128266325 128267458      - -0.1130214
## 5 clu_10209_-    gencode chr9 129108077 129110483      -  0.1074815
## 6 clu_10209_-    gencode chr9 129108077 129110483      -  0.1074815
##       p.adjust  transcript_ids min_intron_number mode_intron_number    gene
## 1 1.437183e-02 ENST00000394215                 1                   1 DENND1A
## 2 1.437183e-02 ENST00000373620                 1                   1 DENND1A
## 3 1.437183e-02 ENST00000475421                 1                   1 DENND1A
## 4 8.760649e-15 ENST00000458730                 5                   5  GOLGA2
## 5 2.104379e-08 ENST00000393384                 1                   1    CRAT
## 6 2.104379e-08 ENST00000318080                 1                   1    CRAT
##                 biotype genes_in_cluster is_first_intron condition conditions
## 1 protein_coding,lncRNA          DENND1A            TRUE     white      white
## 2 protein_coding,lncRNA          DENND1A            TRUE     white      white
## 3 protein_coding,lncRNA          DENND1A            TRUE     white      white
## 4        protein_coding           GOLGA2           FALSE     white      white
## 5        protein_coding             CRAT            TRUE     beige      beige
## 6        protein_coding             CRAT            TRUE     beige      beige
##   num_introns
## 1           1
## 2           1
## 3           1
## 4           1
## 5           1
## 6           1
```

```
both = merge(lc_trans[c("gene","transcript_ids", "deltapsi","condition")],
             castella[c("gene_name","ensembl_transcript_id", "external_transcript_name","log2FoldChange
```

```
            by.x=c("gene","transcript_ids"), by.y= c("gene_name","ensembl_transcript_id"),
      all = T, suffixes = c(".hp",".castella"))
head(both)
```

```
##           gene  transcript_ids   deltapsi condition.hp external_transcript_name
## 1         ABCB8 ENST00000358849         NA         <NA>                ABCB8-202
## 2 ABHD14A-ACY1 ENST00000637778         NA         <NA>         ABHD14A-ACY1-228
## 3         ABI1 ENST00000346832         NA         <NA>                 ABI1-201
## 4   AC002074.1 ENST00000642601 -0.1231060        white                     <NA>
## 5   AC002467.1 ENST00000609979  0.1294696        beige                     <NA>
## 6   AC002467.1 ENST00000653575  0.1294696        beige                     <NA>
##   log2FoldChange condition.castella
## 1       1.073463              beige
## 2       4.441791              beige
## 3       1.379262              beige
## 4             NA               <NA>
## 5             NA               <NA>
## 6             NA               <NA>
```

```
nar = both[!is.na(both$condition.hp) & !is.na(both$condition.castella),]
nar
```

```
##        gene  transcript_ids   deltapsi condition.hp external_transcript_name
## 924    FAR2 ENST00000182377  0.2412781        beige                FAR2-201
## 1502  MAST2 ENST00000482881  0.2439310        beige               MAST2-207
## 1864   OPA1 ENST00000644959 -0.1381253        white                OPA1-226
## 1974   PHC2 ENST00000467894 -0.1459678        white                PHC2-206
## 2520  STON1 ENST00000404752 -0.2204830        white               STON1-201
## 2564 TBC1D1 ENST00000261439 -0.1464586        white              TBC1D1-201
## 2829 WASH8P ENST00000400706  0.1123566        beige              WASH8P-201
##      log2FoldChange condition.castella
## 924        4.202529              beige
## 1502      -2.526378              white
## 1864       1.022024              beige
## 1974      -2.446358              white
## 2520      -1.368380              white
## 2564      -1.023181              white
## 2829      -1.144270              white
```

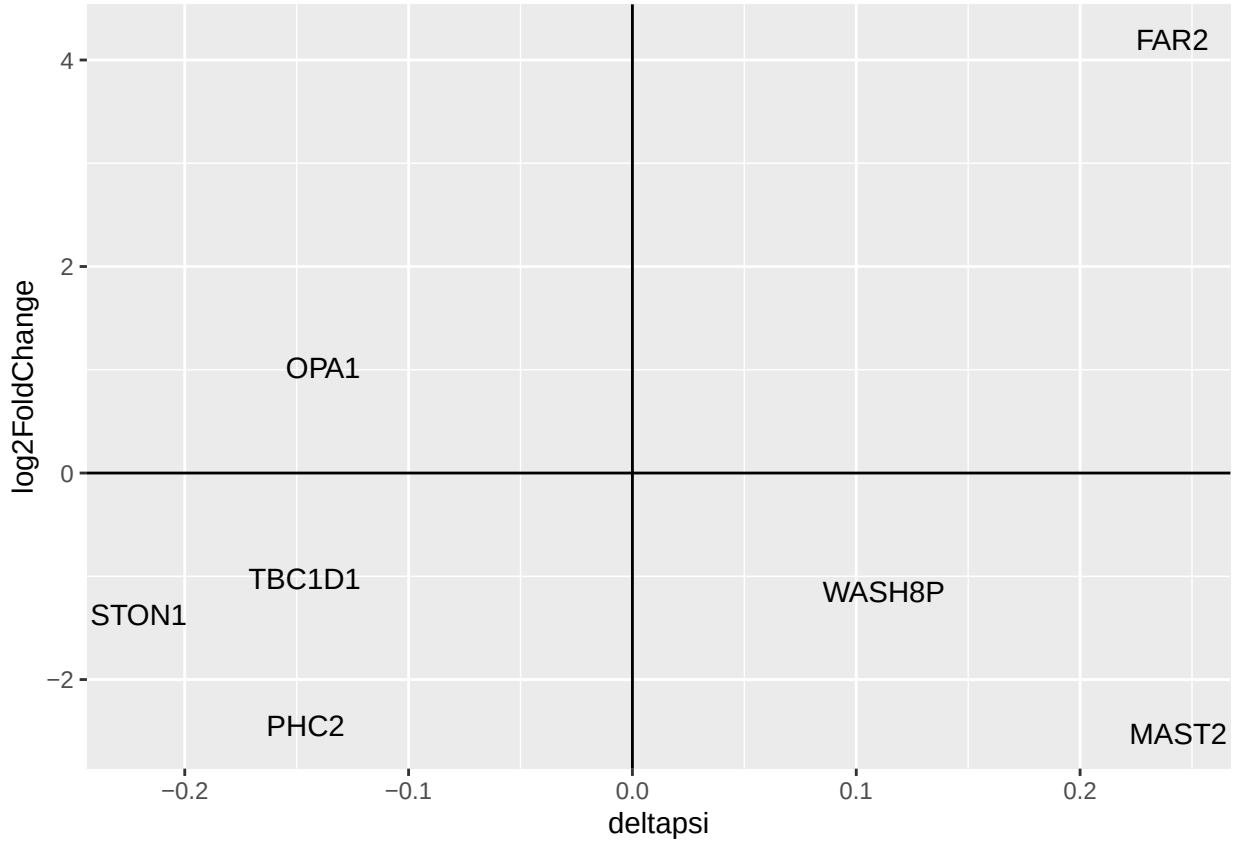
```
freq = group_by(both, condition.hp, condition.castella) %>% count()
freq
```

```
## # A tibble: 8 x 3
## # Groups:   condition.hp, condition.castella [8]
##   condition.hp condition.castella     n
##   <chr>        <chr>              <int>
## 1 beige        beige                  1
## 2 beige        white                  2
## 3 beige        <NA>                1052
## 4 white        beige                  1
## 5 white        white                  3
## 6 white        <NA>                1174
```

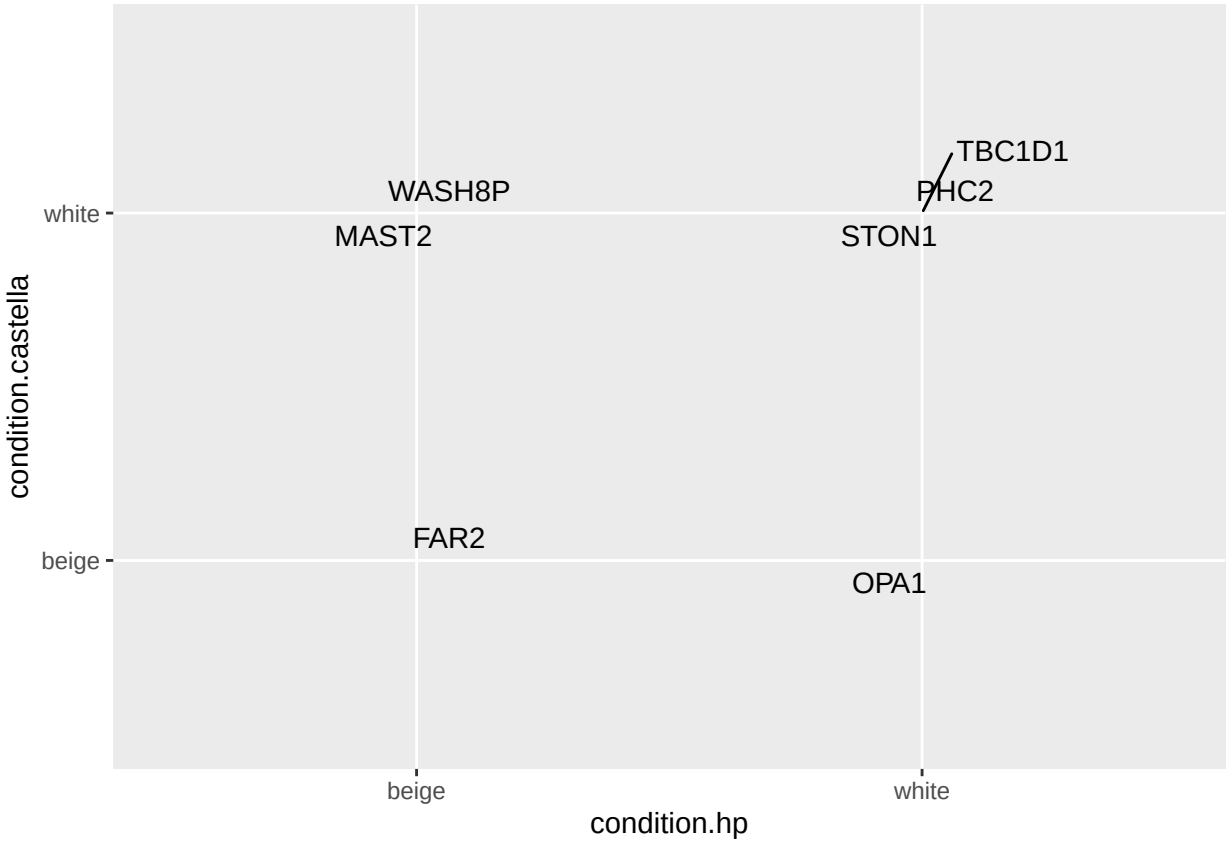
```
## 7 <NA>          beige                 345
## 8 <NA>          white                 445
```

```
library(ggplot2)
library(ggrepel)

ggplot(nar) + geom_text(aes(x=deltapsi, y=log2FoldChange, label=gene)) + geom_hline(aes(yintercept=0)) +
```




```
ggplot(nar) + geom_text_repel(aes(x=condition.hp, y=condition.castella, label=gene))
```

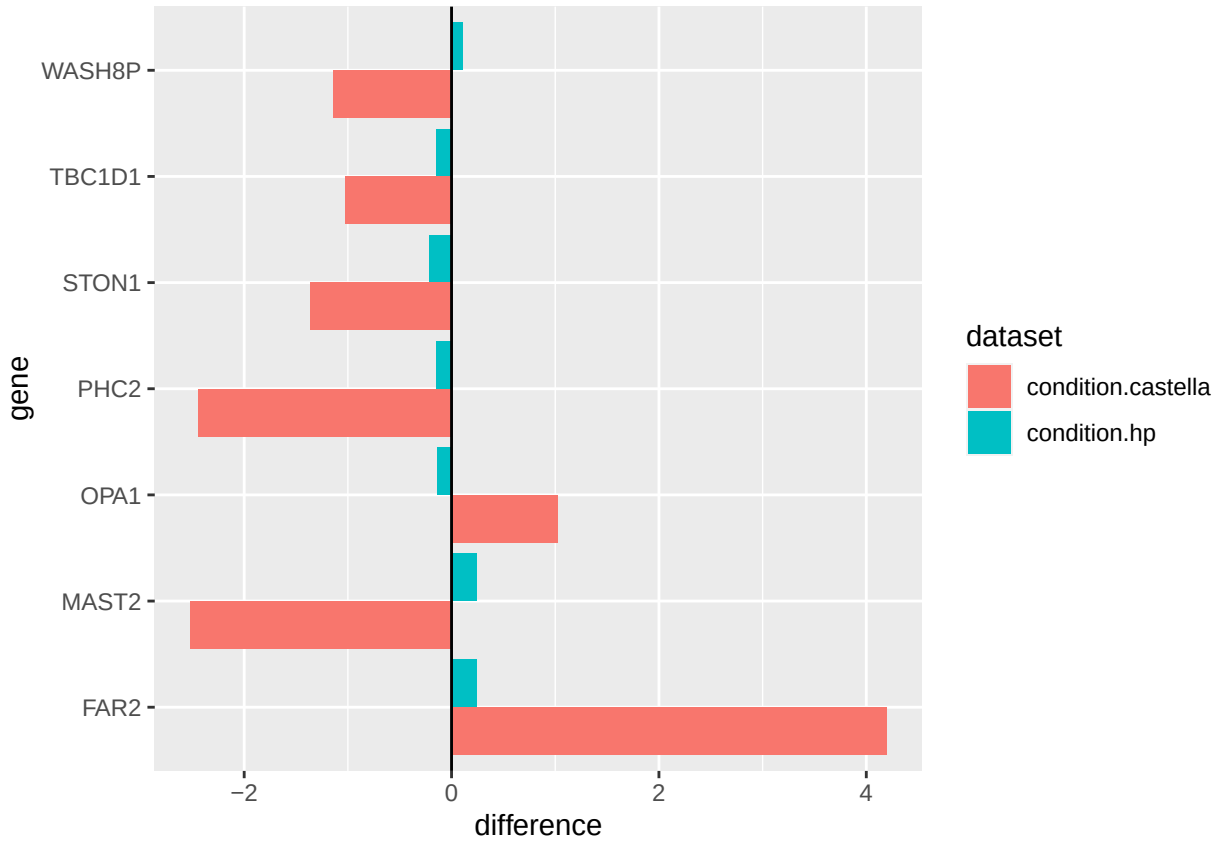



```
nar = pivot_longer(nar, c("condition.hp", "condition.castella"), names_to="dataset", values_to= "conditi
head(nar)
```

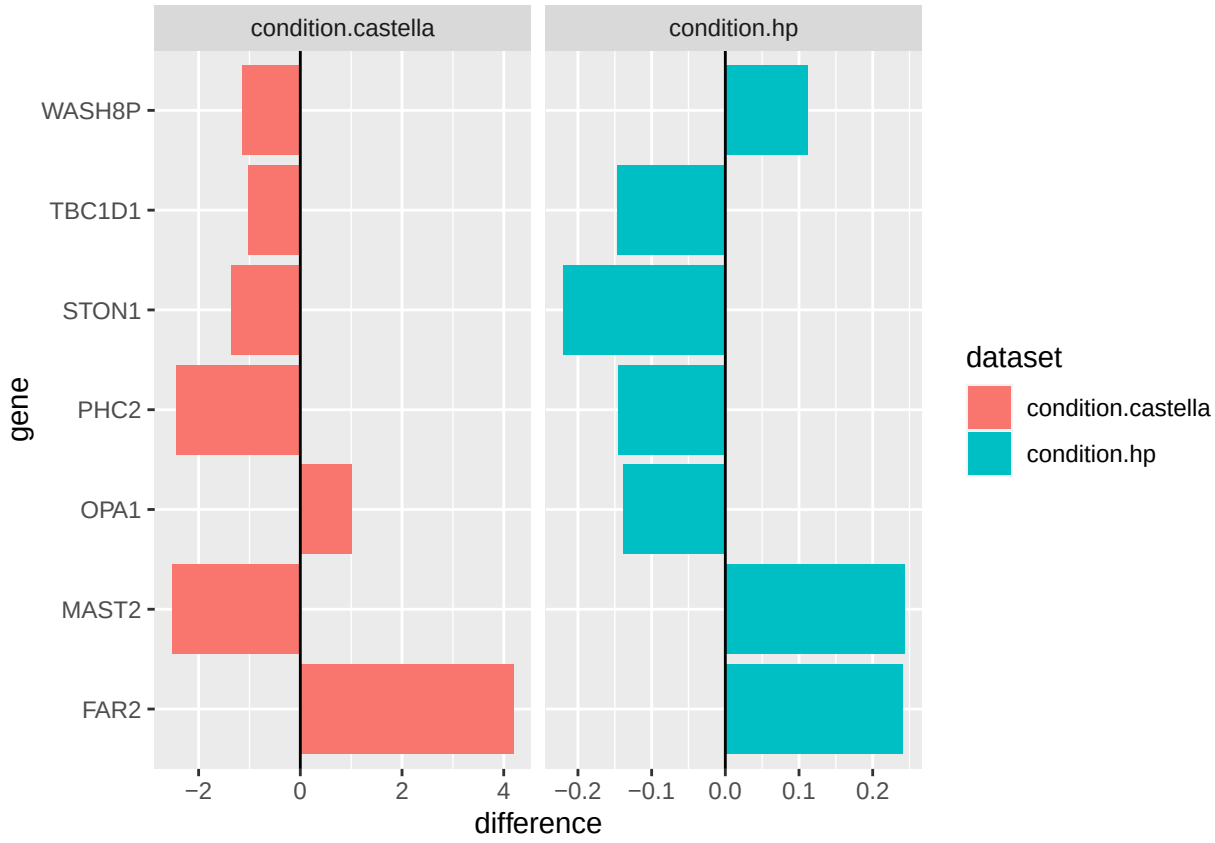
```
## # A tibble: 6 x 7
##   gene  transcript_ids  deltapsi external_transcript_name log2FoldChange dataset
##   <chr> <chr>              <dbl> <chr>                              <dbl> <chr>
## 1 FAR2  ENST00000182377    0.241 FAR2-201                            4.20 condit~
## 2 FAR2  ENST00000182377    0.241 FAR2-201                            4.20 condit~
## 3 MAST2 ENST00000482881    0.244 MAST2-207                          -2.53 condit~
## 4 MAST2 ENST00000482881    0.244 MAST2-207                          -2.53 condit~
## 5 OPA1  ENST00000644959   -0.138 OPA1-226                            1.02 condit~
## 6 OPA1  ENST00000644959   -0.138 OPA1-226                            1.02 condit~
## # i 1 more variable: condition <chr>
```

```
nar = mutate(nar, difference = if_else(dataset == "condition.hp", deltapsi, log2FoldChange))

ggplot(nar, aes(x=difference, fill=dataset, y=gene)) + geom_bar(stat='identity', position="dodge") + geo
```




```
ggplot(nar, aes(x=difference, fill=dataset, y=gene)) + geom_bar(stat='identity', position="dodge") + ge
```


condition.castella
condition.hp
WASH8P
TBC1D1
STON1
PHC2
OPA1
MAST2
FAR2
gene
−2
0
2
4
−0.2
−0.1
0.0
0.1
0.2
difference
dataset
condition.castella
condition.hp