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] 693
```

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

```
## [1] 464
```

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

```
## [1] 660
```

```
head(junctions)
```

```
##    cluster_id annotation   chr     start       end strand   deltapsi
## 1 clu_10181_-    gencode  chr9 128266325 128267458      - -0.1125432
## 2 clu_10209_-    gencode  chr9 129108077 129110483      -  0.1072163
## 3 clu_10638_+    gencode chr12  26195951  26224293      + -0.1295000
## 4 clu_10638_+    gencode chr12  26195531  26224293      +  0.1250475
## 5 clu_10654_+ fantom_cat chr12  27382504  27385481      +  0.1596822
## 6 clu_10654_+    gencode chr12  27380404  27385481      + -0.1596822
##       p.adjust
## 1 9.378884e-15
## 2 2.288239e-08
## 3 1.945157e-02
## 4 1.945157e-02
## 5 1.104588e-07
## 6 1.104588e-07
##
## 1
## 2
## 3
## 4
## 5
## 6 ENST00000311001.9,ENST00000395901.6,ENST00000457040.6,ENST00000266503.9,ENST00000542388.1,ENST0000
##   min_intron_number mode_intron_number               gene               biotype
## 1                 5                  5             GOLGA2        protein_coding
## 2                 1                  1               CRAT        protein_coding
## 3                 1                  1               SSPN        protein_coding
## 4                 1                  1               SSPN protein_coding,lncRNA
## 5                 1                  1 ENSG00000029153.10                  <NA>
## 6                 2                  3             ARNTL2        protein_coding
##   is_first_intron condition  conditions num_introns
## 1           FALSE     white       white           1
## 2            TRUE     beige       beige           1
## 3            TRUE     white beige&white           2
## 4            TRUE     beige beige&white           2
## 5            TRUE     beige beige&white           2
## 6           FALSE     white beige&white           2
```

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

```
##    cluster_id annotation   chr     start       end strand   deltapsi
## 1 clu_10181_-    gencode  chr9 128266325 128267458      - -0.1125432
```

```
## 2 clu_10209_-    gencode  chr9 129108077 129110483      -  0.1072163
## 3 clu_10209_-    gencode  chr9 129108077 129110483      -  0.1072163
## 4 clu_10638_+    gencode chr12  26195951  26224293      + -0.1295000
## 5 clu_10638_+    gencode chr12  26195951  26224293      + -0.1295000
## 6 clu_10638_+    gencode chr12  26195531  26224293      +  0.1250475
##       p.adjust    transcript_ids min_intron_number mode_intron_number   gene
## 1 9.378884e-15 ENST00000458730.2                 5                  5 GOLGA2
## 2 2.288239e-08 ENST00000393384.3                 1                  1   CRAT
## 3 2.288239e-08 ENST00000318080.7                 1                  1   CRAT
## 4 1.945157e-02 ENST00000242729.6                 1                  1   SSPN
## 5 1.945157e-02 ENST00000535504.1                 1                  1   SSPN
## 6 1.945157e-02 ENST00000422622.3                 1                  1   SSPN
##                 biotype is_first_intron condition  conditions num_introns
## 1        protein_coding           FALSE     white       white           1
## 2        protein_coding            TRUE     beige       beige           1
## 3        protein_coding            TRUE     beige       beige           1
## 4        protein_coding            TRUE     white beige&white           2
## 5        protein_coding            TRUE     white beige&white           2
## 6 protein_coding,lncRNA            TRUE     beige beige&white           2
```

```
nrow(lc_trans)
```

```
## [1] 2046
```

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

```
## [1] 1924
```

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

```
## [1] 1833
```

```
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
## 191       ENST00000261439    TBC1D1 ENSG00000065882               TBC1D1-201
## 360       ENST00000182377      FAR2 ENSG00000064763                 FAR2-201
## 497       ENST00000467894      PHC2 ENSG00000134686                 PHC2-206
## 613       ENST00000426335   ARFGAP2 ENSG00000149182              ARFGAP2-202
## 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
## 191 1256.36435     -1.023181 0.2270188 -4.507032 6.57000e-06 2.044759e-03
## 360   88.16492      4.202529 0.8158888  5.150860 2.59000e-07 1.327760e-04
```

```
## 497  107.02797      -2.446358 0.6225960 -3.929287 8.52000e-05 1.550905e-02
## 613   63.34022       9.059762 1.2621878  7.177824 7.08000e-13 1.270000e-09
## 662 2747.99251      -1.368380 0.3645077 -3.754049 1.74001e-04 2.532484e-02
## 744   15.07742      -2.526378 0.6836602 -3.695371 2.19566e-04 2.964458e-02
## 768  313.12911       0.775699 0.1968520  3.940520 8.13000e-05 1.507270e-02
## 784  200.60603       1.022024 0.2740479  3.729362 1.91965e-04 2.727759e-02
```

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

```
##       cluster_id annotation   chr     start       end strand   deltapsi
## 20   clu_10672_+    gencode chr12  29223894  29270412      +  0.2321814
## 169  clu_13672_+    gencode  chr1  45997799  46002805      +  0.2429299
## 179   clu_1389_-    gencode chr11  47172333  47173426      -  0.1031517
## 511  clu_19197_+    gencode  chr3 193618936 193631612      + -0.1419767
## 666  clu_21708_+    gencode  chr8  66452105  66453710      + -0.1056693
## 929  clu_27829_-    gencode  chr1  33334292  33349597      - -0.1461140
## 1162 clu_31216_+    gencode  chr2  48530216  48580587      + -0.1709916
## 1603 clu_39642_+    gencode  chr4  38049898  38054199      + -0.1231900
##          p.adjust  transcript_ids min_intron_number mode_intron_number
## 20   2.283895e-63 ENST00000182377                 1                  1
## 169  1.235699e-15 ENST00000482881                 2                  2
## 179  2.264493e-06 ENST00000426335                 6                  7
## 511  7.776846e-18 ENST00000644959                 1                  5
## 666  3.776492e-03 ENST00000419955                 7                  7
## 929  8.560546e-09 ENST00000467894                 1                  1
## 1162 3.252074e-40 ENST00000404752                 1                  1
## 1603 1.268429e-13 ENST00000261439                 2                  2
##                   gene                                            biotype
## 20                FAR2                                     protein_coding
## 169              MAST2                       protein_coding,retained_intron
## 179            ARFGAP2                 retained_intron,lncRNA,protein_coding
## 511               OPA1 nonsense_mediated_decay,protein_coding,retained_intron
## 666  ADHFE1,AC009879.3                           nonsense_mediated_decay
## 929               PHC2                             lncRNA,protein_coding
## 1162             STON1                                    protein_coding
## 1603            TBC1D1                                    protein_coding
##      is_first_intron condition  conditions num_introns
## 20              TRUE     beige beige&white          2
## 169            FALSE     beige beige&white          2
## 179            FALSE     beige       beige          1
## 511             TRUE     white       white          1
## 666            FALSE     white beige&white          2
## 929             TRUE     white beige&white          2
## 1162            TRUE     white beige&white          3
## 1603           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_10181_-    gencode  chr9 128266325 128267458      - -0.1125432
## 2 clu_10209_-    gencode  chr9 129108077 129110483      -  0.1072163
## 3 clu_10209_-    gencode  chr9 129108077 129110483      -  0.1072163
## 4 clu_10638_+    gencode chr12  26195951  26224293      + -0.1295000
## 5 clu_10638_+    gencode chr12  26195951  26224293      + -0.1295000
## 6 clu_10638_+    gencode chr12  26195531  26224293      +  0.1250475
##       p.adjust  transcript_ids min_intron_number mode_intron_number   gene
## 1 9.378884e-15 ENST00000458730                 5                  5 GOLGA2
## 2 2.288239e-08 ENST00000393384                 1                  1   CRAT
## 3 2.288239e-08 ENST00000318080                 1                  1   CRAT
## 4 1.945157e-02 ENST00000242729                 1                  1   SSPN
## 5 1.945157e-02 ENST00000535504                 1                  1   SSPN
## 6 1.945157e-02 ENST00000422622                 1                  1   SSPN
##                biotype is_first_intron condition  conditions num_introns
## 1        protein_coding           FALSE     white       white           1
## 2        protein_coding            TRUE     beige       beige           1
## 3        protein_coding            TRUE     beige       beige           1
## 4        protein_coding            TRUE     white beige&white           2
## 5        protein_coding            TRUE     white beige&white           2
## 6 protein_coding,lncRNA            TRUE     beige beige&white           2
```

```
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         AAK1 ENST00000623317 -0.1628116        white                      <NA>
## 2        ABCB8 ENST00000358849         NA         <NA>                ABCB8-202
## 3 ABHD14A-ACY1 ENST00000637778         NA         <NA>         ABHD14A-ACY1-228
## 4       ABHD18 ENST00000388795  0.1272483        beige                      <NA>
## 5       ABHD18 ENST00000398965  0.1272483        beige                      <NA>
## 6       ABHD18 ENST00000444616  0.1272483        beige                      <NA>
##   log2FoldChange condition.castella
## 1             NA               <NA>
## 2       1.073463              beige
## 3       4.441791              beige
## 4             NA               <NA>
```

```
## 5             NA               <NA>
## 6             NA               <NA>
```

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

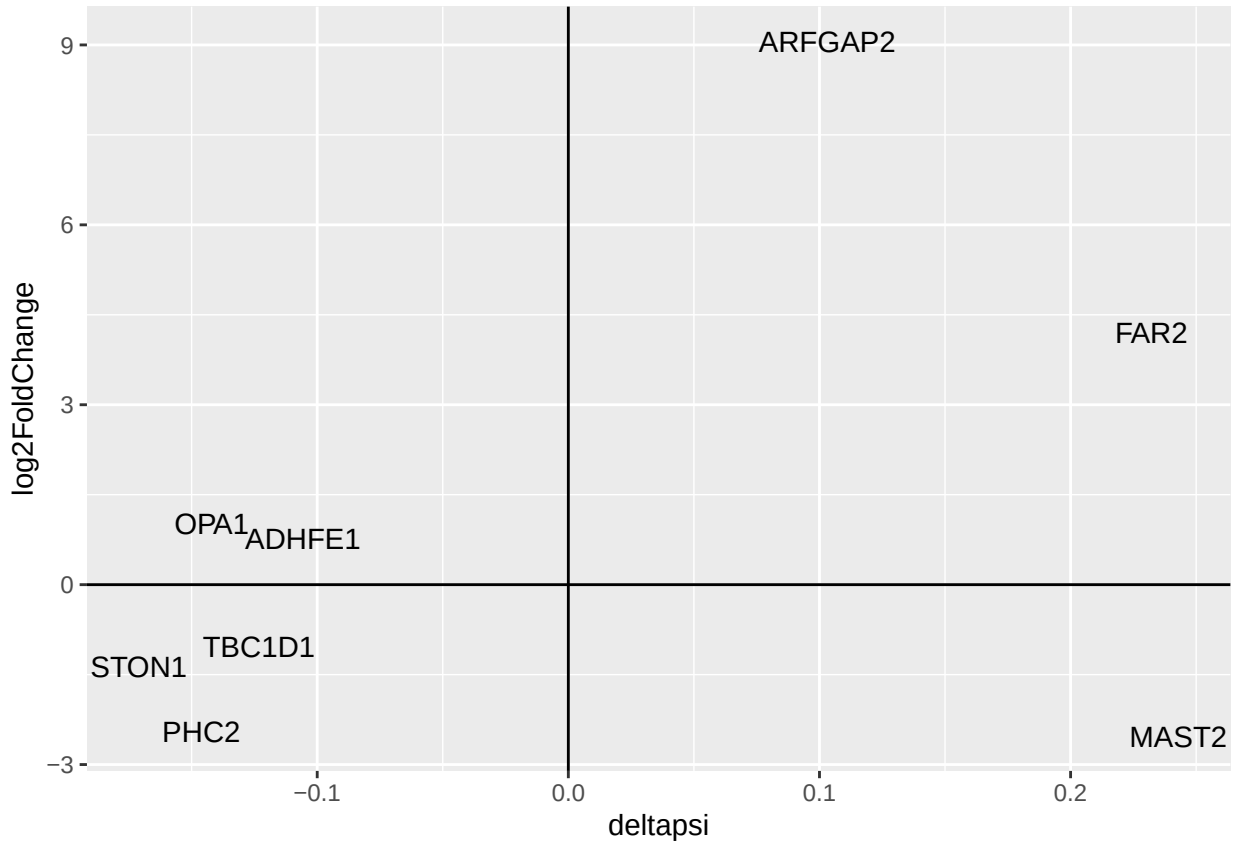
```
##         gene  transcript_ids   deltapsi condition.hp external_transcript_name
## 64    ADHFE1 ENST00000419955 -0.1056693        white               ADHFE1-205
## 146  ARFGAP2 ENST00000426335  0.1031517        beige             ARFGAP2-202
## 901     FAR2 ENST00000182377  0.2321814        beige                FAR2-201
## 1455   MAST2 ENST00000482881  0.2429299        beige               MAST2-207
## 1795    OPA1 ENST00000644959 -0.1419767        white                OPA1-226
## 1884    PHC2 ENST00000467894 -0.1461140        white                PHC2-206
## 2410   STON1 ENST00000404752 -0.1709916        white               STON1-201
## 2450  TBC1D1 ENST00000261439 -0.1231900        white              TBC1D1-201
##      log2FoldChange condition.castella
## 64         0.775699              beige
## 146        9.059762              beige
## 901        4.202529              beige
## 1455      -2.526378              white
## 1795       1.022024              beige
## 1884      -2.446358              white
## 2410      -1.368380              white
## 2450      -1.023181              white
```

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

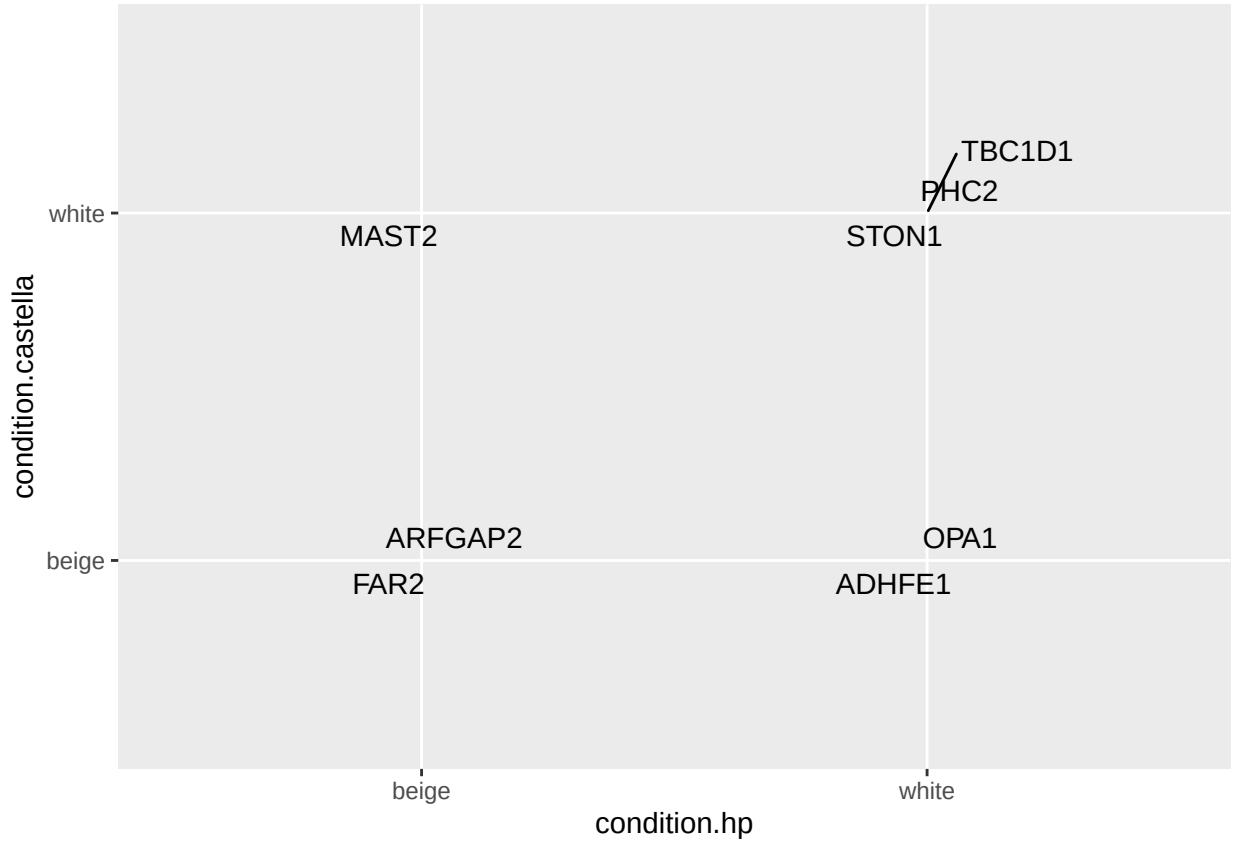
```
## # A tibble: 8 x 3
## # Groups:   condition.hp, condition.castella [8]
##   condition.hp condition.castella     n
##   <chr>        <chr>              <int>
## 1 beige        beige                  2
## 2 beige        white                  1
## 3 beige        <NA>                 977
## 4 white        beige                  2
## 5 white        white                  3
## 6 white        <NA>                1061
## 7 <NA>         beige                343
## 8 <NA>         white                446
```

```
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))
```

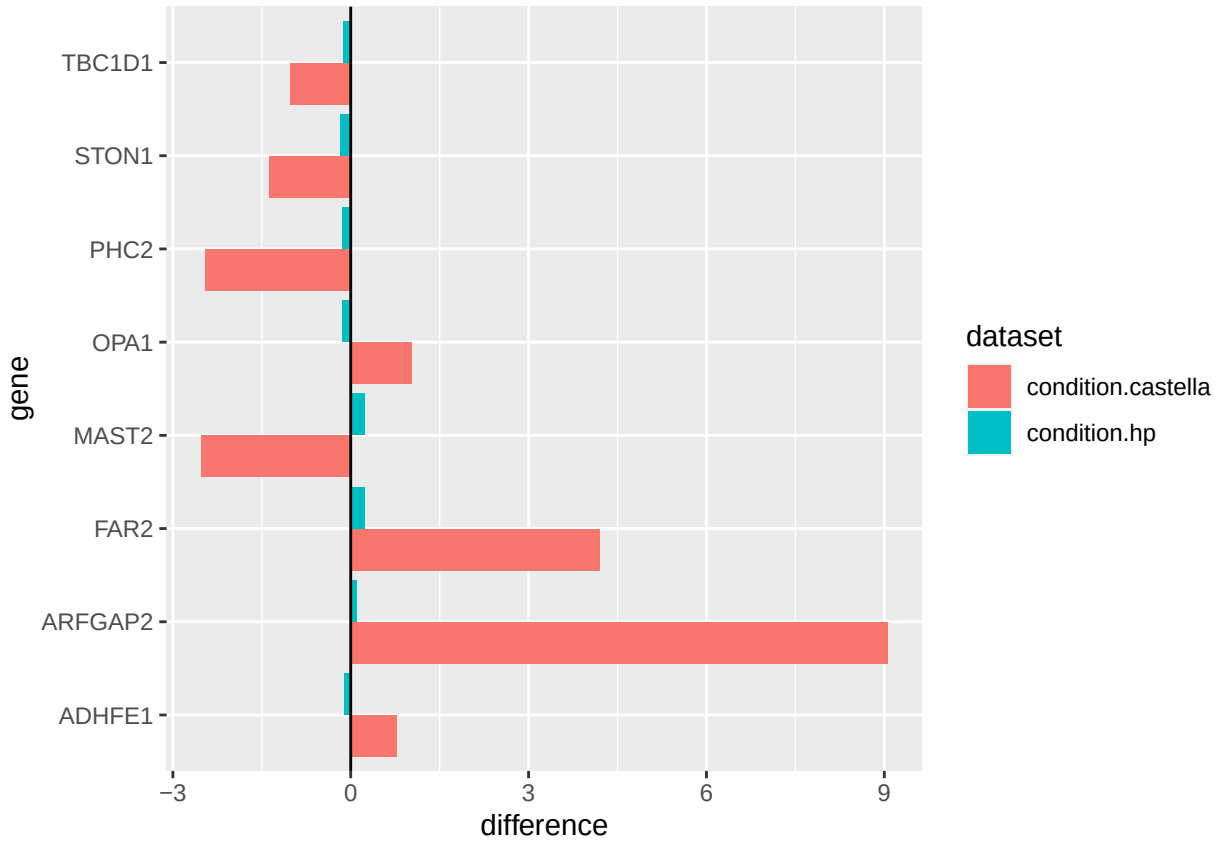



```
nar = pivot_longer(nar, c("condition.hp", "condition.castella"), names_to="dataset", values_to= "conditi
head(nar)
```

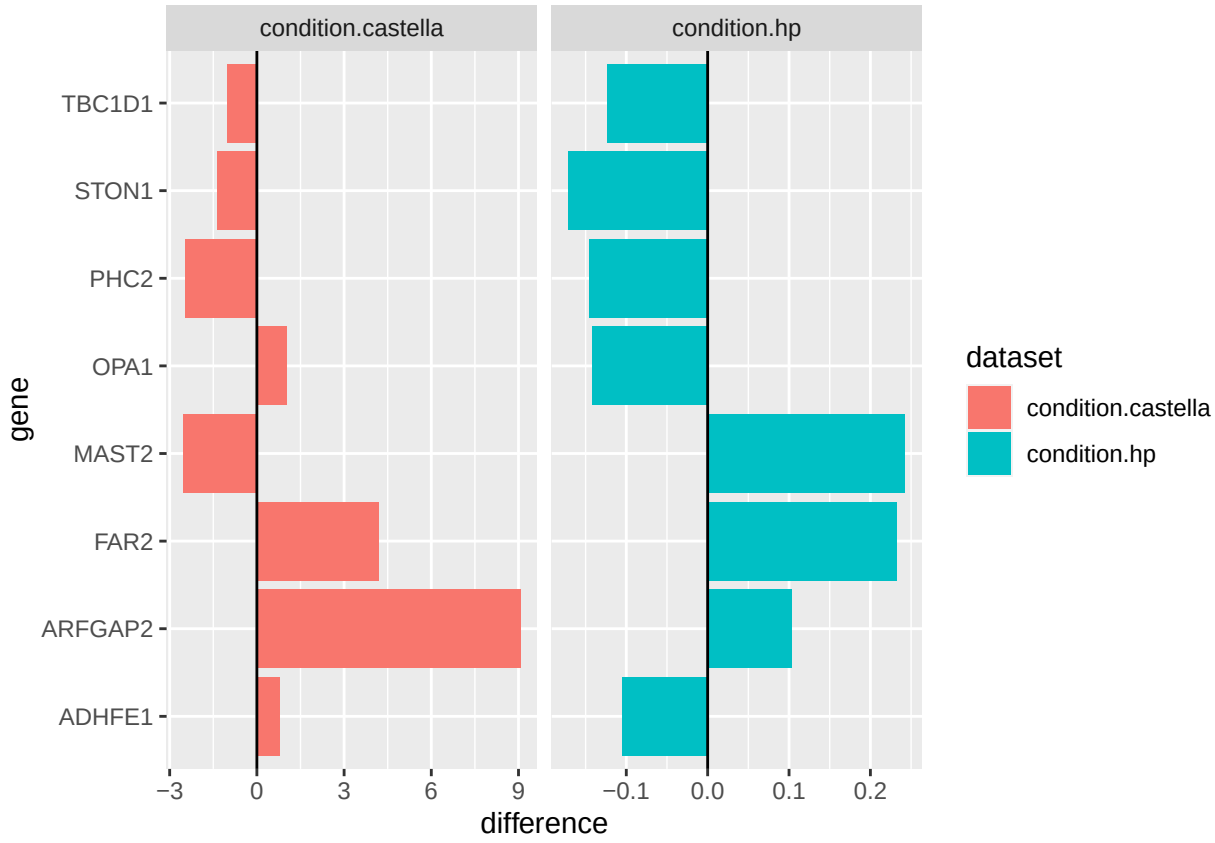
```
## # A tibble: 6 x 7
##   gene    transcript_ids  deltapsi external_transcript_~1 log2FoldChange dataset
##   <chr>   <chr>              <dbl> <chr>                           <dbl> <chr>
## 1 ADHFE1  ENST00000419955   -0.106 ADHFE1-205                      0.776 condit~
## 2 ADHFE1  ENST00000419955   -0.106 ADHFE1-205                      0.776 condit~
## 3 ARFGAP2 ENST00000426335    0.103 ARFGAP2-202                     9.06  condit~
## 4 ARFGAP2 ENST00000426335    0.103 ARFGAP2-202                     9.06  condit~
## 5 FAR2    ENST00000182377    0.232 FAR2-201                        4.20  condit~
## 6 FAR2    ENST00000182377    0.232 FAR2-201                        4.20  condit~
## # i abbreviated name: 1: external_transcript_name
## # 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") + ge
```




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


condition.castella
condition.hp
TBC1D1
STON1
PHC2
OPA1
MAST2
FAR2
ARFGAP2
ADHFE1
gene
−3
0
3
6
9
−0.1
0.0
0.1
0.2
difference
dataset
condition.castella
condition.hp