trifid_stats

2023-12-15

```
library(dplyr)
library(tidyr)
library(ggplot2)
library(ggpubr)
library(here); i_am("R/16_trifid_stats.Rmd")
```

```
figs = here("R/plots")
trifid =  read.delim(here("31_leafcutter", "trifid_all_introns.tsv"))

trifid$category = trifid$sig
trifid$category[trifid$category == "white" & trifid$p.adjust < 0.05] = "p-value"
trifid$category[trifid$category == "beige" & trifid$p.adjust < 0.05] = "p-value"
trifid$category[trifid$p.adjust > 0.05] = "neither"

table(trifid$sig)
```

```
##
##     beige sig_beige sig_white     white
##     63426       385       392     68384
```

```
table(trifid$category)
```

```
##
##   neither   p-value sig_beige sig_white
##     93573     38237       385       392
```

```
table(trifid$in_trifid)
```

```
##
## cluster_not_in_trifid             in_trifid         not_in_trifid
##                 11293                 87024                 34270
```

```
table(trifid$annotation)
```

```
##
##    cryptic fantom_cat    gencode     refseq
##      21831       9629      89495      11632
```

```
length(unique(trifid$intron_coords))
```

```
## [1] 132587
```

```
nrow(trifid)
```

```
## [1] 132587
```

58 clusters (corresponding to 116 two top introns) were not found in trifid.

```
table(trifid[c("annotation", "sig")])
```

```
##             sig
## annotation   beige sig_beige sig_white white
##   cryptic    10449        39        15 11328
##   fantom_cat  4174        21        22  5412
##   gencode    43449       306       341 45399
##   refseq      5354        19        14  6245
```

```
table(trifid[c("category", "in_trifid")])
```

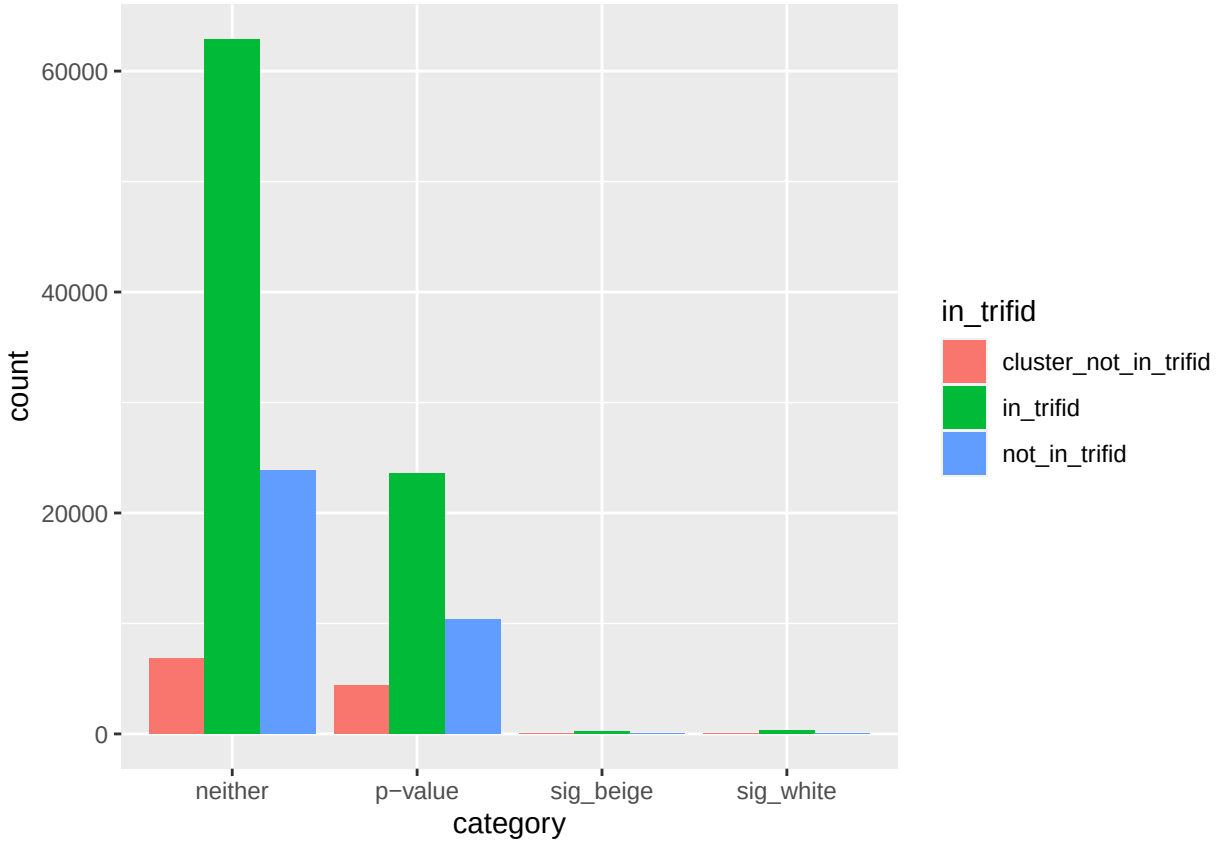
```
##            in_trifid
## category    cluster_not_in_trifid in_trifid not_in_trifid
##   neither                    6822     62894         23857
##   p-value                    4362     23553         10322
##   sig_beige                    54       267            64
##   sig_white                    55       310            27
```

```
trifid[grep("PEMT",trifid$gene),]
```

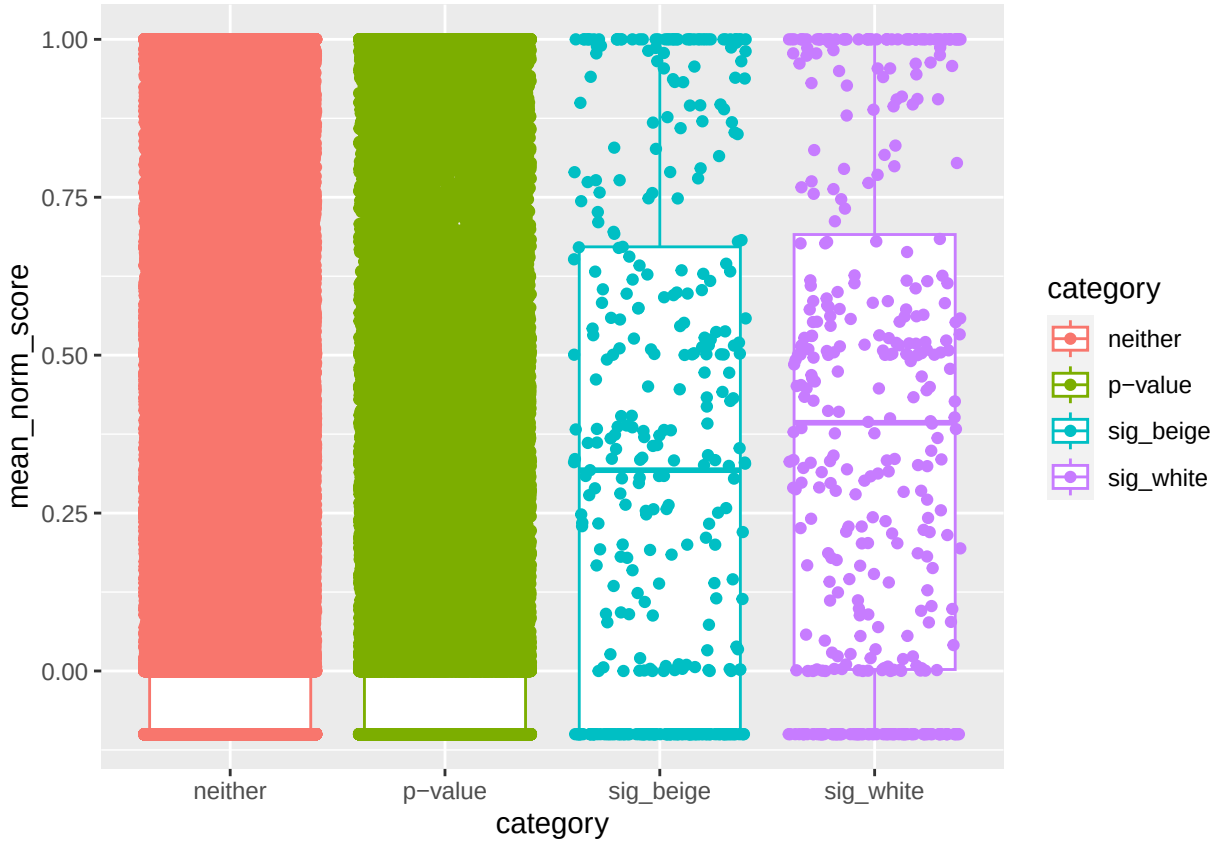
```
##           in_trifid             intron_coords condition      deltapsi
## 2         in_trifid chr17:17577027:17591531:-     white -4.296507e-01
## 105       in_trifid chr17:17577027:17577107:-     beige  1.918093e-01
## 1142      in_trifid chr17:17577027:17582267:-     beige  7.985252e-02
## 1619      in_trifid chr17:17577027:17577414:-     beige  7.101235e-02
## 5446      in_trifid chr17:17577027:17591967:-     beige  4.230864e-02
## 18227     in_trifid chr17:17509545:17512509:-     white -2.027199e-02
## 30418     in_trifid chr17:17522395:17576920:-     beige  1.226553e-02
## 57094     in_trifid chr17:17509545:17576920:-     beige  3.879182e-03
## 57550     in_trifid chr17:17505848:17506227:-     beige  3.798946e-03
## 64846     in_trifid chr17:17577027:17591597:-     beige  2.523690e-03
## 75545     in_trifid chr17:17506301:17509434:-     white -1.104112e-03
## 77323     in_trifid chr17:17512654:17522280:-     white -9.071787e-04
## 96877 not_in_trifid chr17:17512654:17519027:-     beige  2.816277e-03
## 96878 not_in_trifid chr17:17512654:17576920:-     beige  3.072487e-05
##            p.adjust gene  cluster_id annotation mean_trifid_score
## 2     4.360252e-104 PEMT clu_19605_-    gencode         0.1438500
## 105   4.360252e-104 PEMT clu_19605_-     refseq         0.1319000
## 1142  4.360252e-104 PEMT clu_19605_-    gencode         0.8038000
## 1619  4.360252e-104 PEMT clu_19605_-    gencode         0.8038000
## 5446  4.360252e-104 PEMT clu_19605_-     refseq         0.8739000
## 18227  1.914447e-01 PEMT clu_19604_-    gencode         0.4371200
## 30418  1.914447e-01 PEMT clu_19604_-    gencode         0.4371200
## 57094  1.914447e-01 PEMT clu_19604_-    gencode         0.0005000
## 57550  1.846905e-01 PEMT clu_19603_-    gencode         0.3643500
```

```
## 64846 4.360252e-104 PEMT clu_19605_-    gencode         0.0040000
## 75545  1.846905e-01 PEMT clu_19603_-    gencode         0.5267750
## 77323  1.914447e-01 PEMT clu_19604_-    gencode         0.3644167
## 96877  1.914447e-01 PEMT clu_19604_-    gencode        -0.1000000
## 96878  1.914447e-01 PEMT clu_19604_-    gencode        -0.1000000
##       mean_norm_score median_norm_score
## 2            0.179000           0.04725
## 105          0.134600           0.13460
## 1142         1.000000           1.00000
## 1619         1.000000           1.00000
## 5446         0.891400           0.89140
## 18227        0.543840           0.62090
## 30418        0.543840           0.62090
## 57094        0.000600           0.00060
## 57550        0.453300           0.35710
## 64846        0.005000           0.00500
## 75545        0.655375           0.81045
## 77323        0.453400           0.35710
## 96877       -0.100000          -0.10000
## 96878       -0.100000          -0.10000
##                                                                                             transcripts
## 2                                      ENST00000255389,ENST00000395781,ENST00000461404,ENST00000580147
## 105                                                                                            XM_006721418
## 1142                                                                                       ENST00000395783
## 1619                                                                                       ENST00000395782
## 5446                                                                                           XM_024450532
## 18227                  ENST00000395783,ENST00000395782,ENST00000255389,ENST00000395781,ENST00000435340
## 30418                  ENST00000395783,ENST00000395782,ENST00000255389,ENST00000395781,ENST00000435340
## 57094                                                                                      ENST00000580147
## 57550 ENST00000395783,ENST00000395782,ENST00000255389,ENST00000395781,ENST00000435340,ENST00000580147
## 64846                                                                                      ENST00000435340
## 75545                                  ENST00000395783,ENST00000395782,ENST00000255389,ENST00000580147
## 77323 ENST00000395783,ENST00000395782,ENST00000255389,ENST00000395781,ENST00000435340,ENST00000461404
## 96877                                                                                      ENST00000490392
## 96878                                                                                      ENST00000472440
##                                              transcript_names       sig  category
## 2                        PEMT-201,PEMT-202,PEMT-207,PEMT-212 sig_white sig_white
## 105                                                      <NA> sig_beige sig_beige
## 1142                                                 PEMT-204     beige   p-value
## 1619                                                 PEMT-203     beige   p-value
## 5446                                                     <NA>     beige   p-value
## 18227         PEMT-204,PEMT-203,PEMT-201,PEMT-202,PEMT-206     white   neither
## 30418         PEMT-204,PEMT-203,PEMT-201,PEMT-202,PEMT-206     beige   neither
## 57094                                                PEMT-212     beige   neither
## 57550 PEMT-204,PEMT-203,PEMT-201,PEMT-202,PEMT-206,PEMT-212     beige   neither
## 64846                                                PEMT-206     beige   p-value
## 75545                  PEMT-204,PEMT-203,PEMT-201,PEMT-212     white   neither
## 77323 PEMT-204,PEMT-203,PEMT-201,PEMT-202,PEMT-206,PEMT-207     white   neither
## 96877                                                    <NA>     beige   neither
## 96878                                                    <NA>     beige   neither
```

```
ggplot(trifid, aes(x=category, fill=in_trifid)) + geom_bar(position="dodge")
```



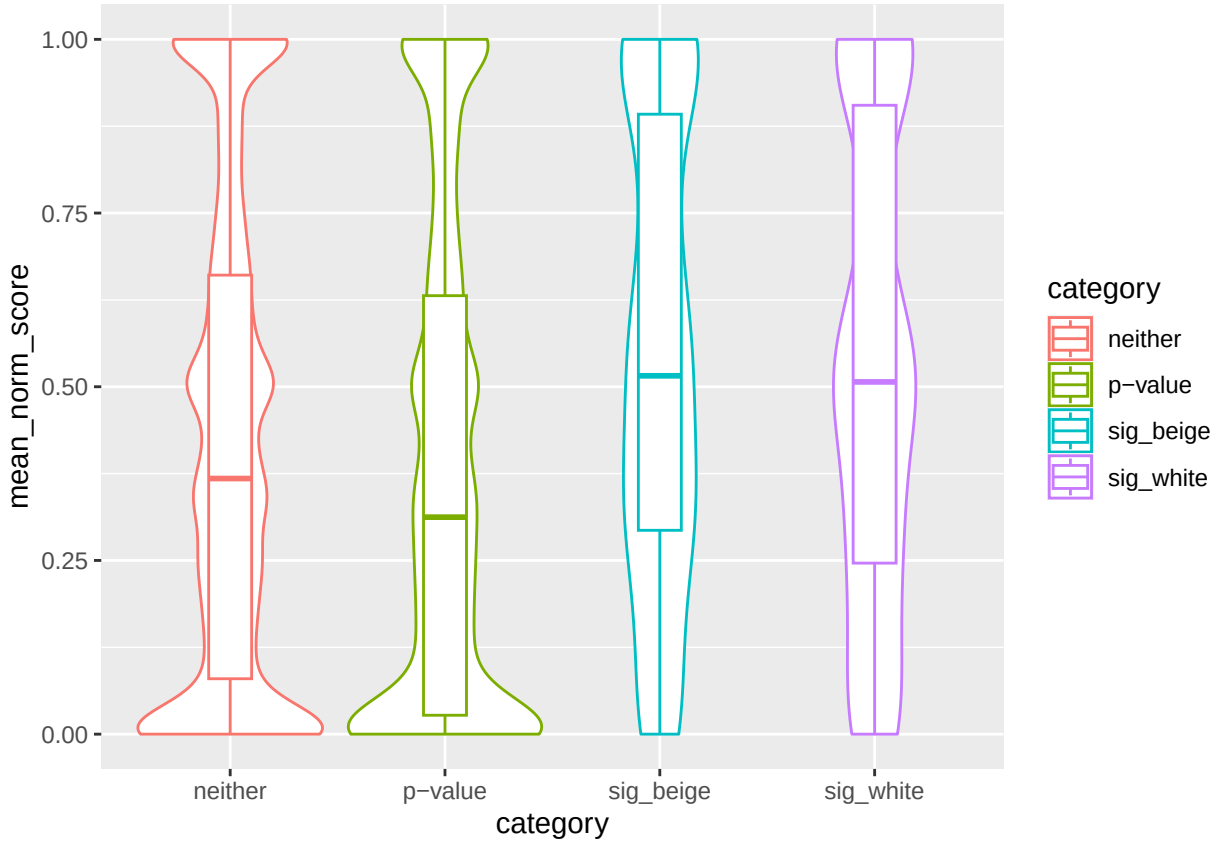

```
ggplot(trifid, aes(x=category, y=mean_norm_score, colour=category)) + geom_boxplot() + geom_jitter()
```



Let's simplify these categories...

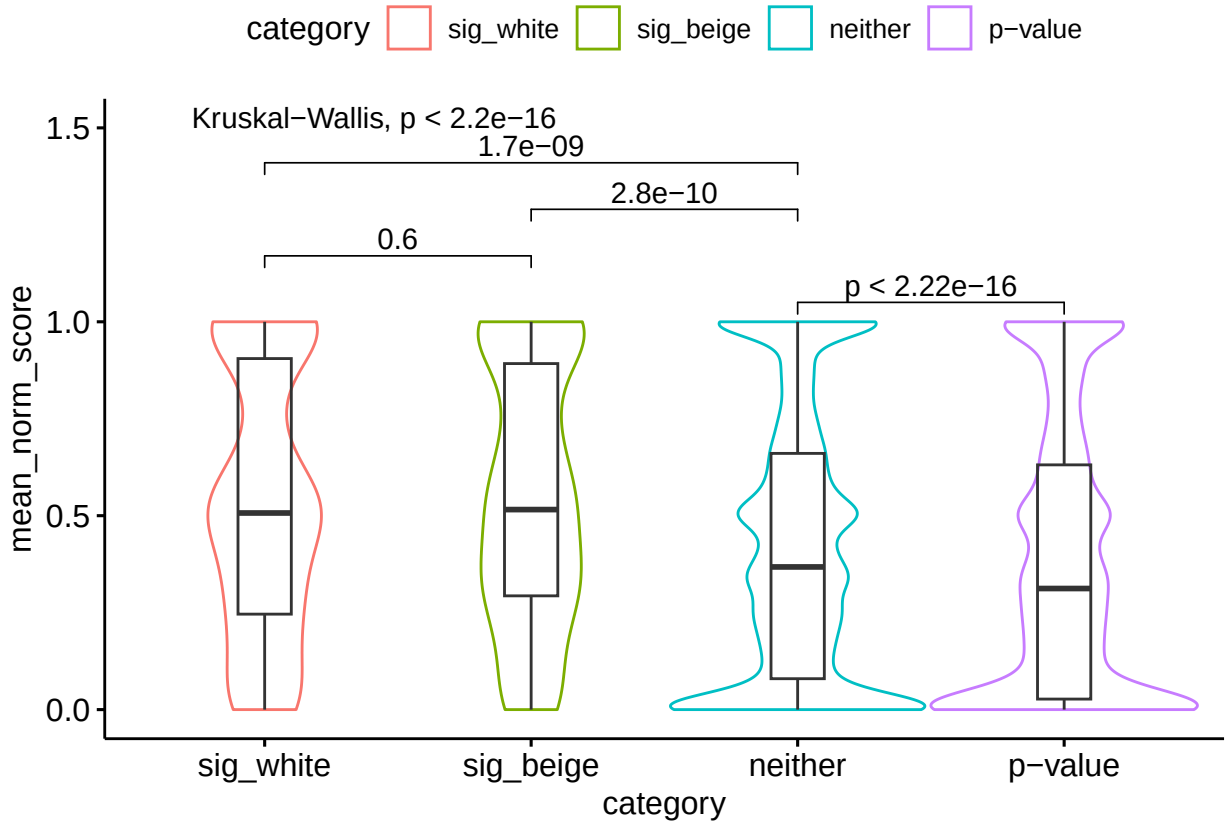
sigbeige, sig white, pvalue, neither

```
ggplot(filter(trifid, in_trifid == "in_trifid"), aes(x=category, y=mean_norm_score, colour=category)) +
  geom_boxplot(width=0.2) #+ geom_jitter()
```



n=257 beige and n=280 white introns are in trifid to be compared like this

```
ggviolin(filter(trifid, in_trifid == "in_trifid"), x="category", y="mean_norm_score", col="category", t
    stat_compare_means(comparisons = list( c("neither","p-value"),
                                           c("sig_white","sig_beige"),
                                             c("sig_beige","neither"),
                                             c("sig_white","neither")) ) +
        stat_compare_means(method = "kruskal.test", label.y=1.5) +
    geom_boxplot(fill=NA, width=0.2)
```



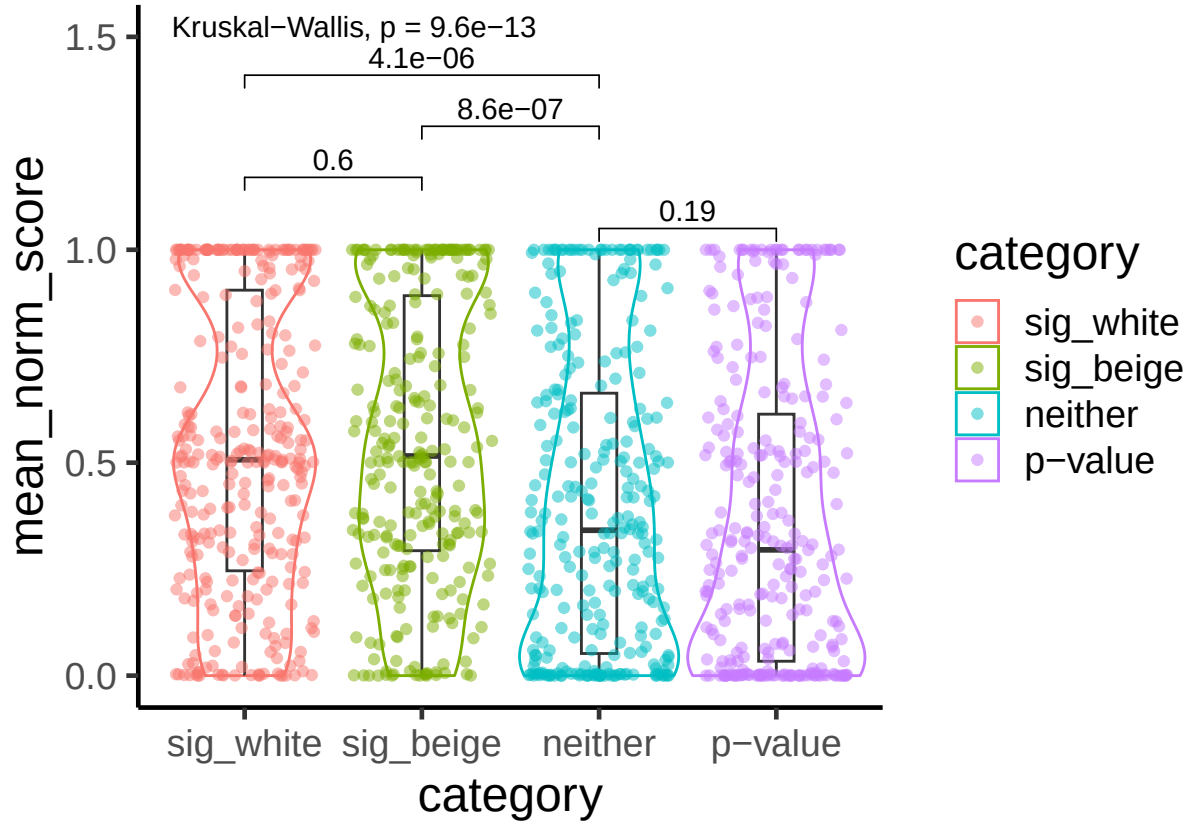
I wonder if the number of introns found for the pvalue and neither category could be subsampled to produce similar numbers and verify the statistics?
The white and beige categories each have ~300 introns in trifid, so lets check.

```
filt = bind_rows(filter(trifid,category %in% c("sig_white","sig_beige") & in_trifid=="in_trifid"),
                 sample_n(filter(trifid,category == "neither" & in_trifid=="in_trifid"), 280),
                 sample_n(filter(trifid,category == "p-value" & in_trifid=="in_trifid"), 280))
nrow(filt)
```

```
## [1] 1137
```

```
write.table(filt, here(figs, "Figure2E_data.tsv"), sep="\t", quote = F, row.names = F)
```

```
ggviolin(filt, x="category", y="mean_norm_score", col="category", trim=T) +
    stat_compare_means(comparisons = list( c("neither","p-value"),
                                           c("sig_white","sig_beige"),
                                             c("sig_beige","neither"),
                                             c("sig_white","neither")) ) +
        stat_compare_means(method = "kruskal.test", label.y=1.5) +
    geom_boxplot(fill=NA, width=0.2)  + geom_jitter(aes(colour=category), alpha=0.5) +
  theme_classic(base_size=18)
```



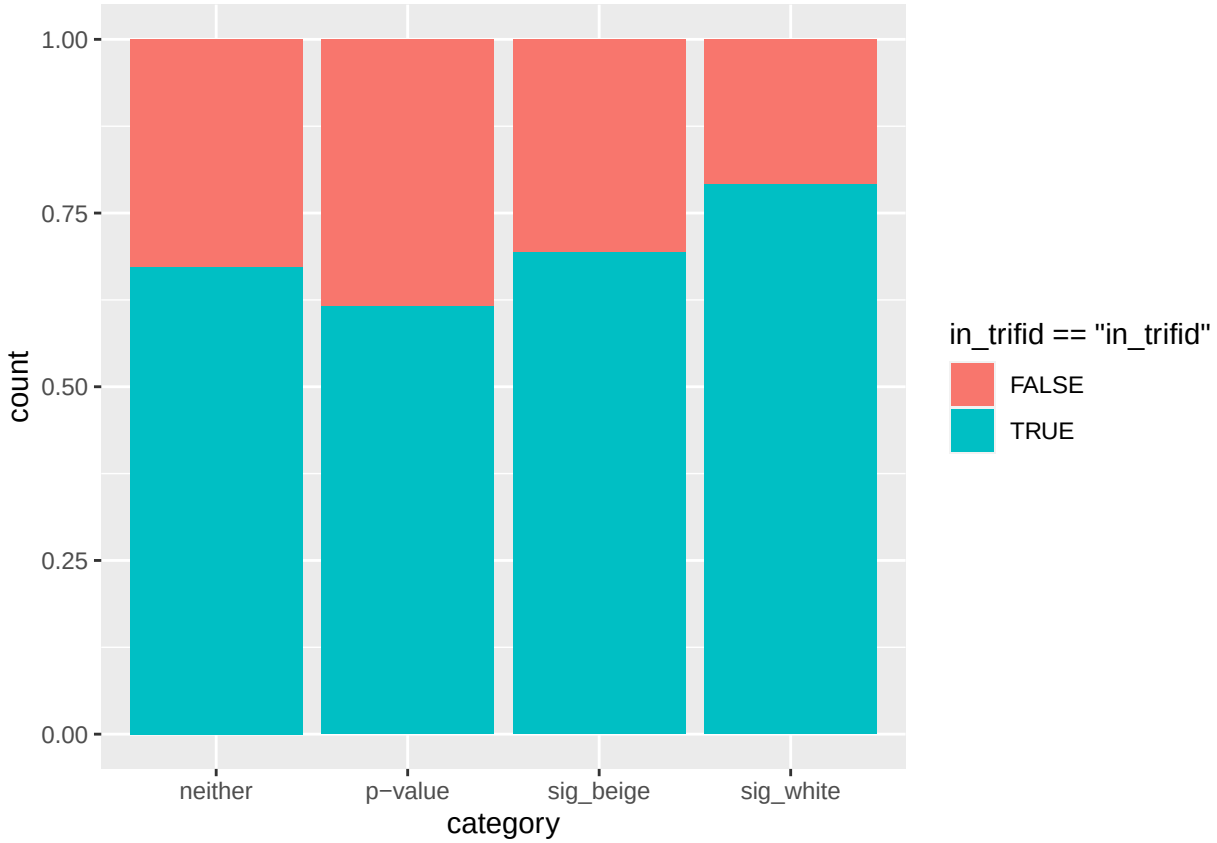

```
ggsave(file=file.path(figs, "trifid_stats.pdf"), width=7, height=4.5)
```

Those p-values seem more reasonable :)

Proportion of trifid not founds

we're going to have to beef up these figures... because I haven't kept all of the unfound introns :)

```
ggplot(trifid, aes(x= category,fill =in_trifid == "in_trifid",)) + geom_bar(position="fill")
```




```
trifid$simple_trifid = trifid$in_trifid
trifid$simple_trifid [trifid$simple_trifid == "cluster_not_in_trifid"] = "not_in_trifid"
prop = group_by(trifid, category, simple_trifid) %>% count()
prop
```

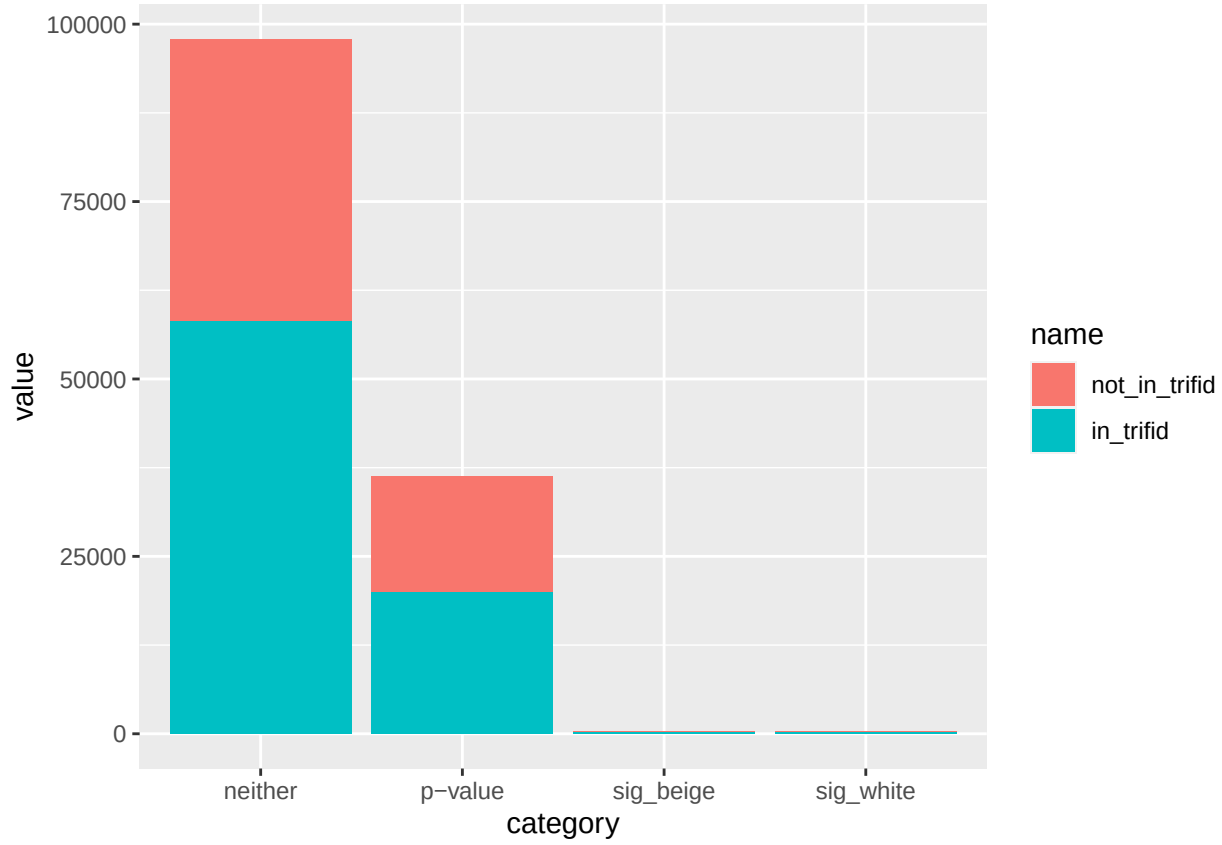
```
## # A tibble: 8 x 3
## # Groups:   category, simple_trifid [8]
##   category  simple_trifid     n
##   <chr>     <chr>         <int>
## 1 neither   in_trifid     62894
## 2 neither   not_in_trifid 30679
## 3 p-value   in_trifid     23553
## 4 p-value   not_in_trifid 14684
## 5 sig_beige in_trifid       267
## 6 sig_beige not_in_trifid   118
## 7 sig_white in_trifid       310
## 8 sig_white not_in_trifid    82
```

```
## we can update the numbers
prop = data.frame(in_trifid = c(58240,20000,280,257),
                  not_in_trifid = c(39673, 16326, 90, 66),
                  category = c("neither","p-value","sig_white","sig_beige"))
head(prop)
```

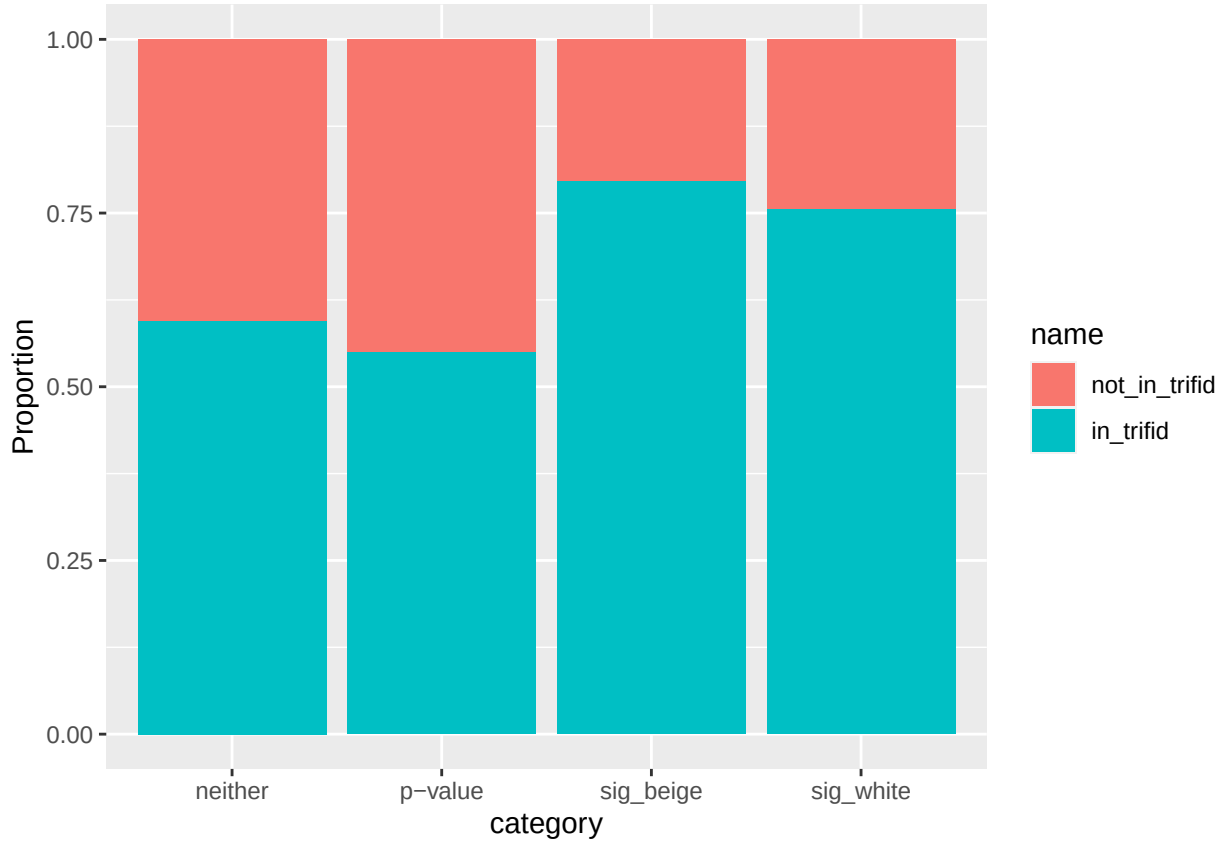
```
##   in_trifid not_in_trifid  category
```

```
## 1     58240         39673   neither
## 2     20000         16326   p-value
## 3       280            90 sig_white
## 4       257            66 sig_beige
```

```
prop = pivot_longer(prop, 1:2)
prop$name = factor(prop$name, levels=c("not_in_trifid", "in_trifid"))
ggplot(prop, aes(x=category, fill= name, y=value)) + geom_bar(stat="identity")
```




```
ggplot(prop, aes(x=category, fill= name, y=value)) + geom_bar(stat="identity", position="fill") + labs(
```



non sig = 58240/ 97913 in trifid/not = 59.4% in sig cluster = 20000/36326 = 55.1% sig white = 280/(280+66) = 81.0% sig_beige = 257/(257+90) = 74.1%