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   neither sig_beige sig_white     white
##     77862         1       434       433     82144
```

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

```
##
##   neither   p-value sig_beige sig_white
##    117743     42264       434       433
```

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

```
##
## cluster_not_in_trifid             in_trifid         not_in_trifid
##                 11293                 87024                 62557
```

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

```
##
##    cryptic fantom_cat    gencode     refseq
##      21831       9638     116958      12447
```

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

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

```
nrow(trifid)
```

```
## [1] 160874
```

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

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

```
##             sig
## annotation   beige neither sig_beige sig_white white
##   cryptic    10487       0        24         8 11312
##   fantom_cat  4230       0        21        20  5367
##   gencode    57420       1       371       392 58774
##   refseq      5725       0        18        13  6691
```

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

```
##            in_trifid
## category    cluster_not_in_trifid in_trifid not_in_trifid
##   neither                    7309     64422         46012
##   p-value                    3912     22065         16287
##   sig_beige                    37       257           140
##   sig_white                    35       280           118
```

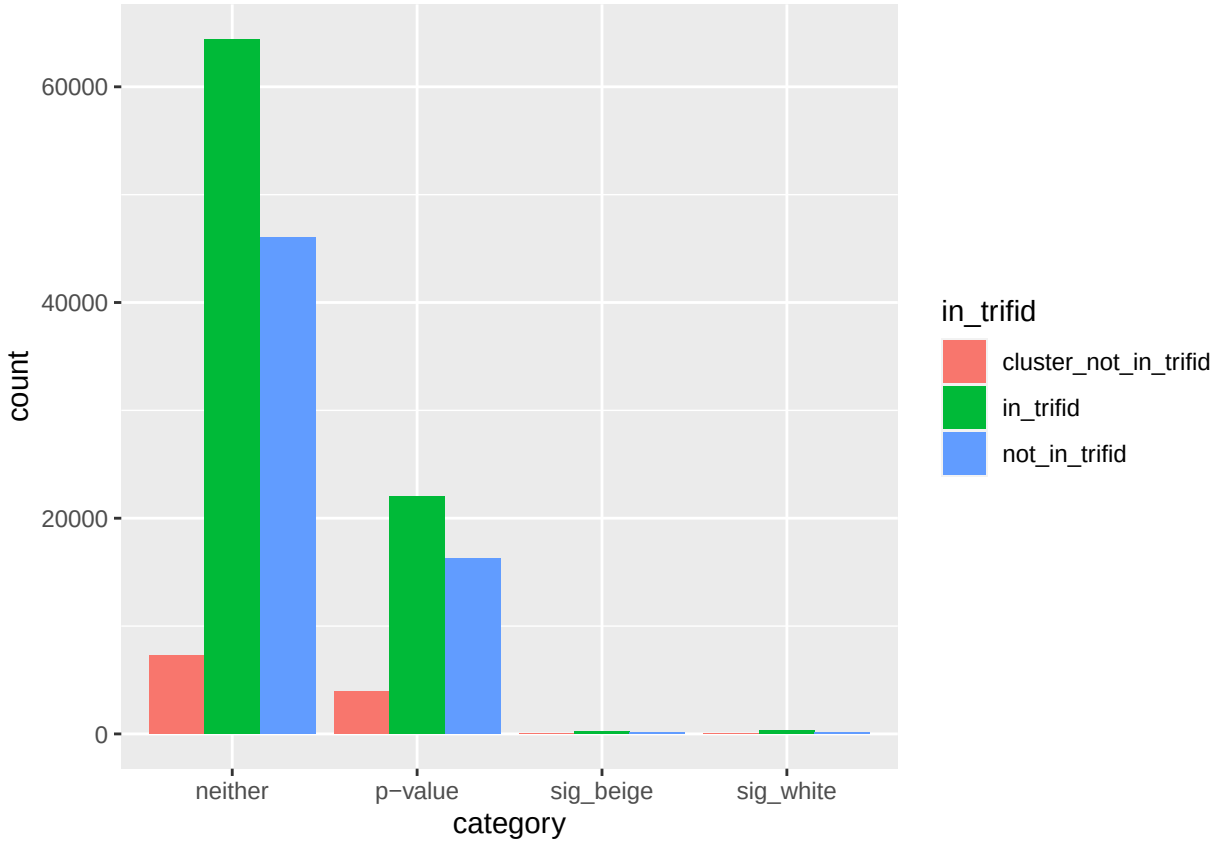
```
trifid[grep("PEMT",trifid$gene),]
```

```
##            in_trifid             intron_coords condition      deltapsi
## 5          in_trifid chr17:17577027:17591531:-     white -0.4096022388
## 80         in_trifid chr17:17577027:17577107:-     beige  0.2084922739
## 2212       in_trifid chr17:17577027:17577414:-     beige  0.0610357482
## 2796       in_trifid chr17:17577027:17582267:-     beige  0.0554167403
## 3950       in_trifid chr17:17577027:17591967:-     beige  0.0476375648
## 18606      in_trifid chr17:17509545:17512509:-     white -0.0189567169
## 30043      in_trifid chr17:17522395:17576920:-     beige  0.0113229995
## 39224      in_trifid chr17:17577027:17591597:-     beige  0.0075591064
## 59616      in_trifid chr17:17505848:17506227:-     beige  0.0026696840
## 62988      in_trifid chr17:17509545:17576920:-     beige  0.0021782712
## 76382      in_trifid chr17:17512654:17522280:-     beige  0.0007176954
## 79409      in_trifid chr17:17506301:17509434:-     white -0.0004726715
## 105315 not_in_trifid chr17:17505848:17506227:-     beige  0.0026696840
## 105317 not_in_trifid chr17:17506301:17509434:-     white -0.0004726715
## 105318 not_in_trifid chr17:17509545:17512509:-     white -0.0189567169
## 105320 not_in_trifid chr17:17512654:17519027:-     beige  0.0034533707
## 105321 not_in_trifid chr17:17512654:17522280:-     beige  0.0007176954
## 105322 not_in_trifid chr17:17512654:17576920:-     beige  0.0001893187
## 105323 not_in_trifid chr17:17522395:17576920:-     beige  0.0113229995
## 105326 not_in_trifid chr17:17577027:17591531:-     white -0.4096022388
## 105327 not_in_trifid chr17:17577027:17591597:-     beige  0.0075591064
##             p.adjust gene  cluster_id annotation mean_trifid_score
## 5      3.184596e-102 PEMT clu_19605_-    gencode         0.1438500
## 80     3.184596e-102 PEMT clu_19605_-     refseq         0.1319000
```

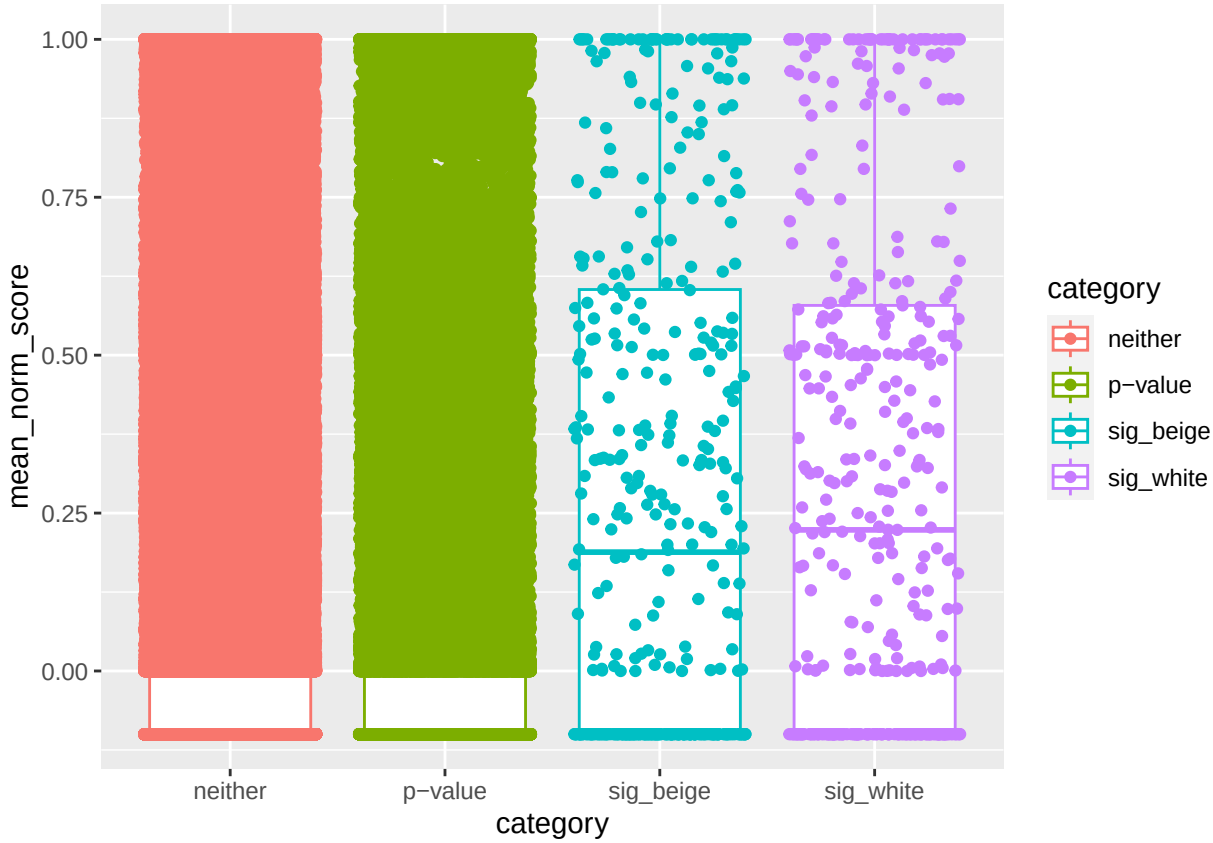
```
## 2212   3.184596e-102 PEMT clu_19605_-    gencode         0.8038000
## 2796   3.184596e-102 PEMT clu_19605_-    gencode         0.8038000
## 3950   3.184596e-102 PEMT clu_19605_-     refseq         0.8739000
## 18606   2.859406e-01 PEMT clu_19604_-    gencode         0.4371200
## 30043   2.859406e-01 PEMT clu_19604_-    gencode         0.4371200
## 39224  3.184596e-102 PEMT clu_19605_-    gencode         0.0040000
## 59616   1.925081e-01 PEMT clu_19603_-    gencode         0.3643500
## 62988   2.859406e-01 PEMT clu_19604_-    gencode         0.0005000
## 76382   2.859406e-01 PEMT clu_19604_-    gencode         0.3644167
## 79409   1.925081e-01 PEMT clu_19603_-    gencode         0.5267750
## 105315  1.925081e-01 PEMT clu_19603_-    gencode        -0.1000000
## 105317  1.925081e-01 PEMT clu_19603_-    gencode        -0.1000000
## 105318  2.859406e-01 PEMT clu_19604_-    gencode        -0.1000000
## 105320  2.859406e-01 PEMT clu_19604_-    gencode        -0.1000000
## 105321  2.859406e-01 PEMT clu_19604_-    gencode        -0.1000000
## 105322  2.859406e-01 PEMT clu_19604_-    gencode        -0.1000000
## 105323  2.859406e-01 PEMT clu_19604_-    gencode        -0.1000000
## 105326 3.184596e-102 PEMT clu_19605_-    gencode        -0.1000000
## 105327 3.184596e-102 PEMT clu_19605_-    gencode        -0.1000000
##        mean_norm_score median_norm_score
## 5             0.179000           0.04725
## 80            0.134600           0.13460
## 2212          1.000000           1.00000
## 2796          1.000000           1.00000
## 3950          0.891400           0.89140
## 18606         0.543840           0.62090
## 30043         0.543840           0.62090
## 39224         0.005000           0.00500
## 59616         0.453300           0.35710
## 62988         0.000600           0.00060
## 76382         0.453400           0.35710
## 79409         0.655375           0.81045
## 105315       -0.100000          -0.10000
## 105317       -0.100000          -0.10000
## 105318       -0.100000          -0.10000
## 105320       -0.100000          -0.10000
## 105321       -0.100000          -0.10000
## 105322       -0.100000          -0.10000
## 105323       -0.100000          -0.10000
## 105326       -0.100000          -0.10000
## 105327       -0.100000          -0.10000
##                                                                                                                transcript
## 5                                                ENST00000255389,ENST00000395781,ENST00000461404,ENST000005801
## 80                                                                                                          XM_006721
## 2212                                                                                                     ENST00000395
## 2796                                                                                                     ENST00000395
## 3950                                                                                                        XM_024505
## 18606                ENST00000395782,ENST00000395783,ENST00000255389,ENST00000395781,ENST000004353
## 30043                ENST00000395782,ENST00000395783,ENST00000255389,ENST00000395781,ENST000004353
## 39224                                                                                                    ENST000004353
## 59616  ENST00000395782,ENST00000395783,ENST00000255389,ENST00000395781,ENST00000435340,ENST000005801
## 62988                                                                                                    ENST000005801
## 76382  ENST00000395782,ENST00000395783,ENST00000255389,ENST00000395781,ENST00000435340,ENST000004614
## 79409                                ENST00000395782,ENST00000395783,ENST00000255389,ENST000005801
```

```
## 105315                                ENST00000582268,ENST00000477595,ENST00000484838,ENST000004903
## 105317                                                                ENST00000490392,ENST000004848
## 105318                                ENST00000472446,ENST00000484838,ENST00000490392,ENST000004210
## 105320                                                                                ENST000004903
## 105321                                                                                ENST000004210
## 105322                                                                                ENST000004724
## 105323                                                                                ENST000004210
## 105326                                                                                ENST000004210
## 105327                                                                                ENST000004724
##                                                transcript_names       sig
## 5                          PEMT-201,PEMT-202,PEMT-207,PEMT-212 sig_white
## 80                                                         <NA> sig_beige
## 2212                                                  PEMT-203     beige
## 2796                                                  PEMT-204     beige
## 3950                                                       <NA>     beige
## 18606           PEMT-203,PEMT-204,PEMT-201,PEMT-202,PEMT-206     white
## 30043           PEMT-203,PEMT-204,PEMT-201,PEMT-202,PEMT-206     beige
## 39224                                                 PEMT-206     beige
## 59616  PEMT-203,PEMT-204,PEMT-201,PEMT-202,PEMT-206,PEMT-212     beige
## 62988                                                 PEMT-212     beige
## 76382  PEMT-203,PEMT-204,PEMT-201,PEMT-202,PEMT-206,PEMT-207     beige
## 79409                    PEMT-203,PEMT-204,PEMT-201,PEMT-212     white
## 105315                                                     <NA>     beige
## 105317                                                     <NA>     white
## 105318                                                     <NA>     white
## 105320                                                     <NA>     beige
## 105321                                                     <NA>     beige
## 105322                                                     <NA>     beige
## 105323                                                     <NA>     beige
## 105326                                                     <NA> sig_white
## 105327                                                     <NA>     beige
##         category
## 5      sig_white
## 80     sig_beige
## 2212     p-value
## 2796     p-value
## 3950     p-value
## 18606    neither
## 30043    neither
## 39224    p-value
## 59616    neither
## 62988    neither
## 76382    neither
## 79409    neither
## 105315   neither
## 105317   neither
## 105318   neither
## 105320   neither
## 105321   neither
## 105322   neither
## 105323   neither
## 105326 sig_white
## 105327   p-value
```

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



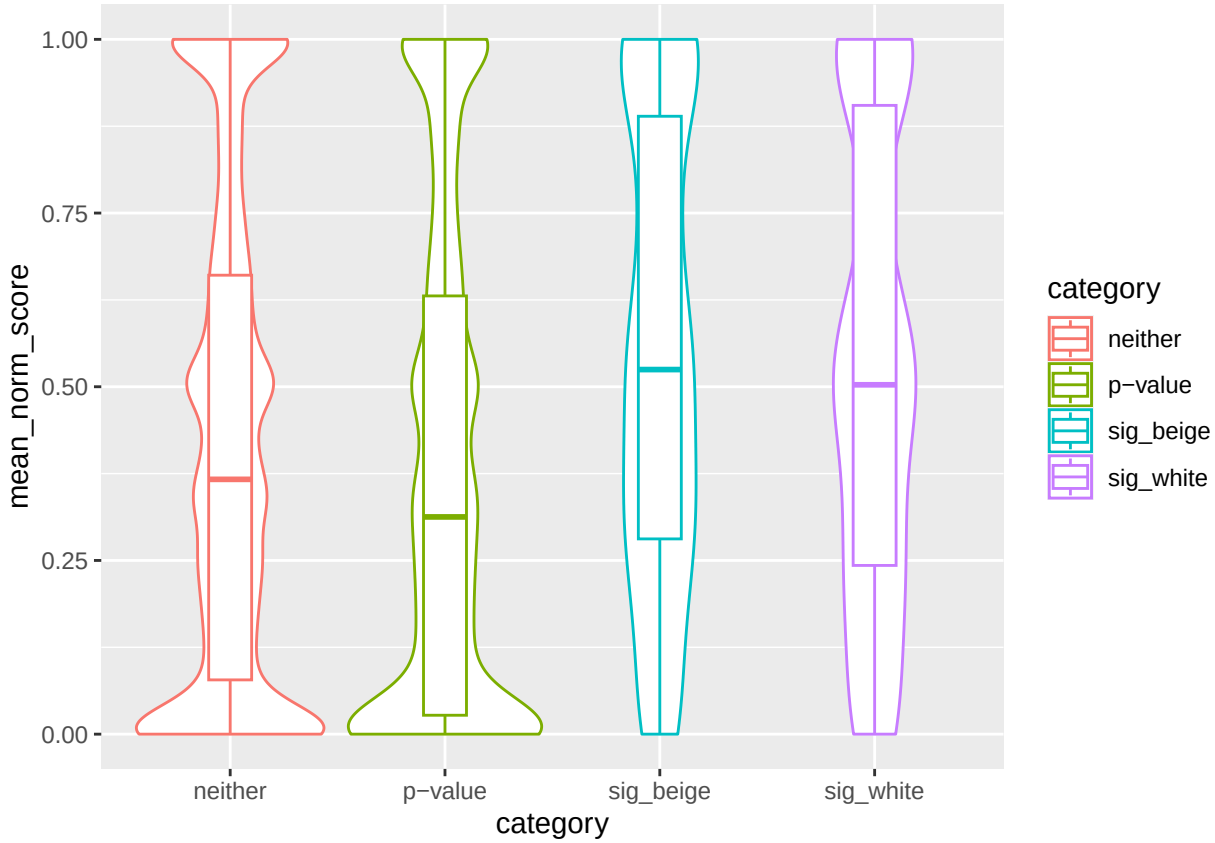

```
ggplot(trifid, aes(x=category, y=mean_norm_score, colour=category)) + geom_boxplot() + geom_jitter()
```



Let's simplify these categories...

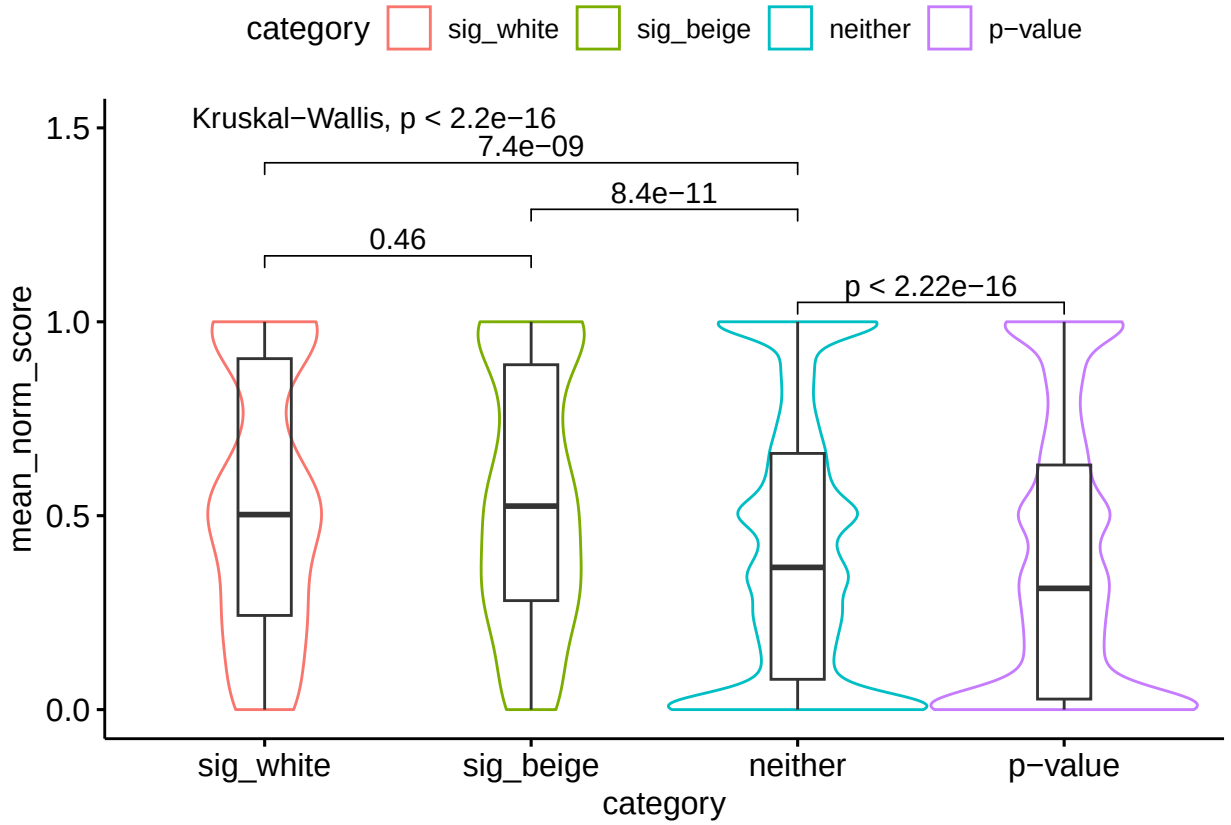
sigbeige, sig white, pvalue, neither

```
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)
```



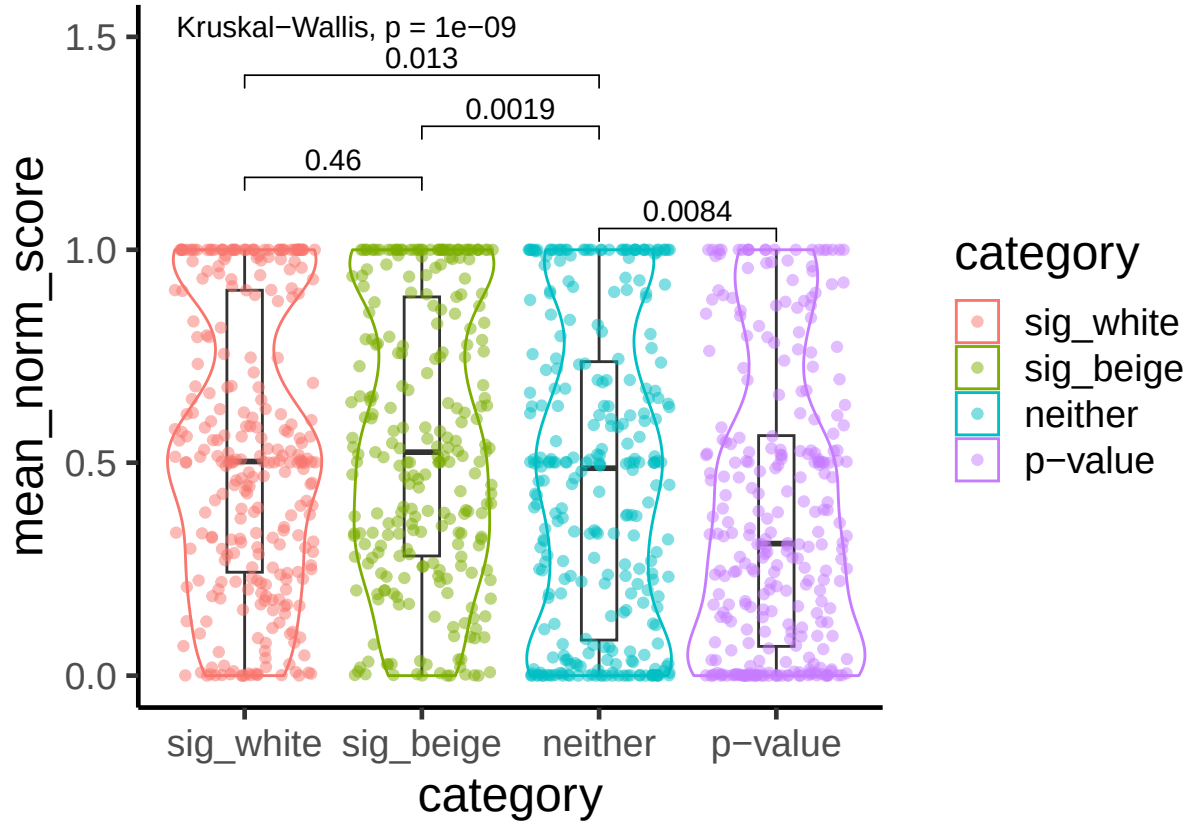
I wonder if the number of introns found for the pvalue and neither category could be subsampled to produce similar numbers and verify the statistics?
The white and beige categories each have ~300 introns in trifid, so lets check.

```
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] 1097
```

```
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)
```



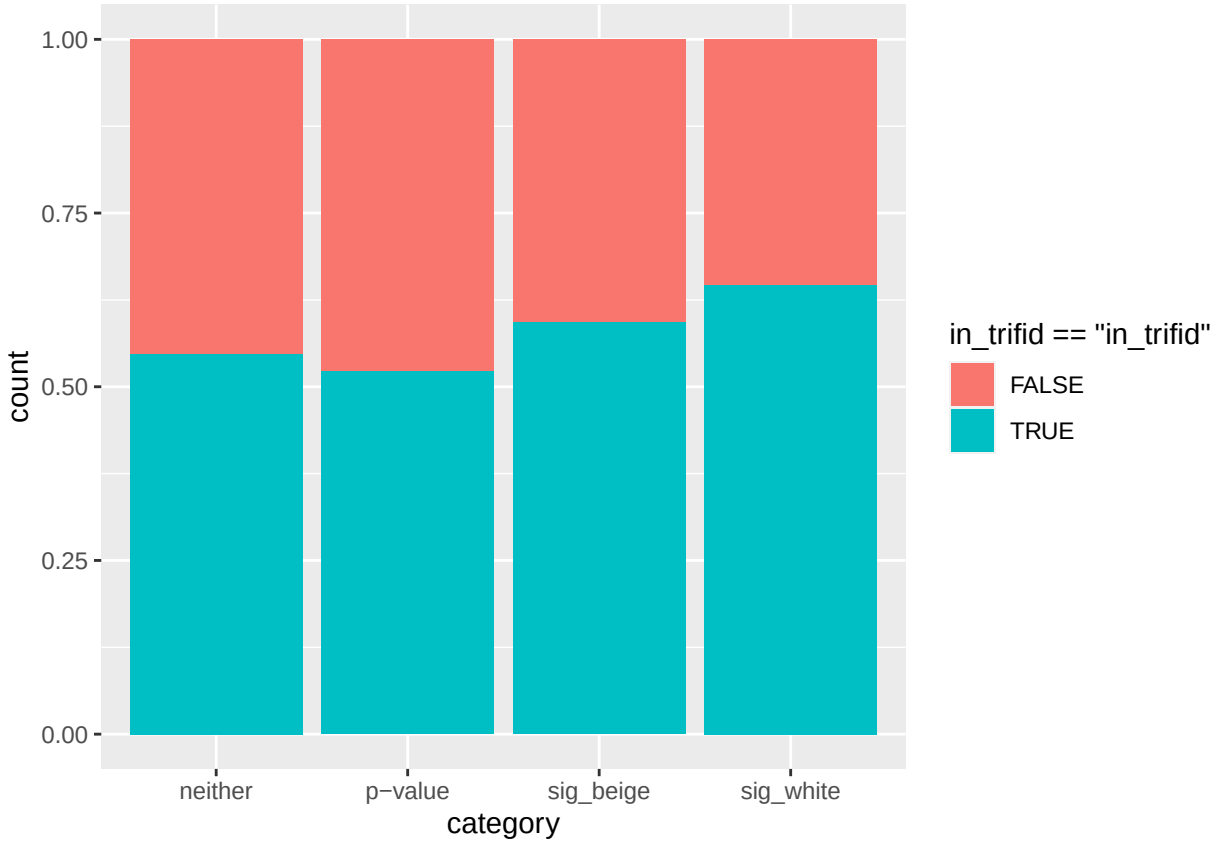

```
ggsave(file=file.path(figs, "trifid_stats.pdf"), width=7, height=4.5)
```

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
```

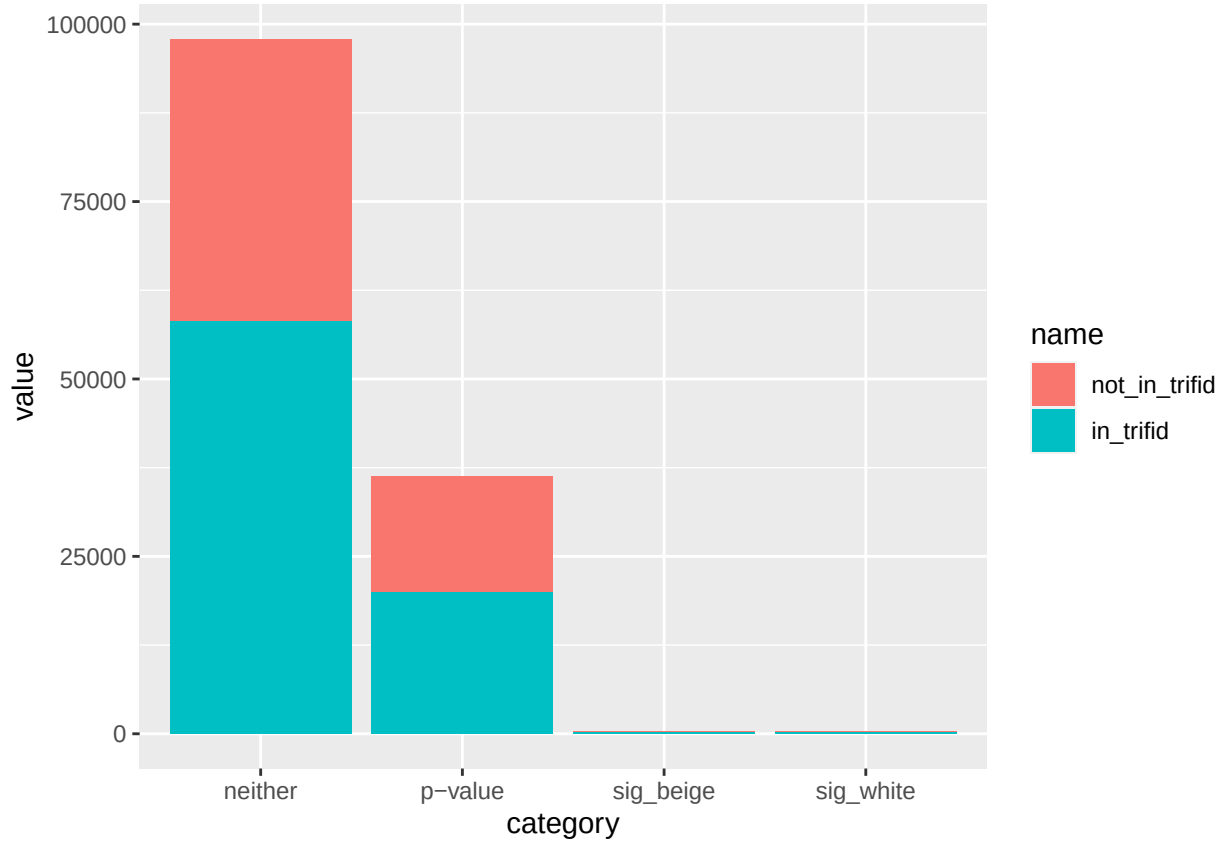
```
## # A tibble: 8 x 3
## # Groups:   category, simple_trifid [8]
##   category  simple_trifid     n
##   <chr>     <chr>         <int>
## 1 neither   in_trifid     64422
## 2 neither   not_in_trifid 53321
## 3 p-value   in_trifid     22065
## 4 p-value   not_in_trifid 20199
## 5 sig_beige in_trifid       257
## 6 sig_beige not_in_trifid   177
## 7 sig_white in_trifid       280
## 8 sig_white not_in_trifid   153
```

```
## 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)
```

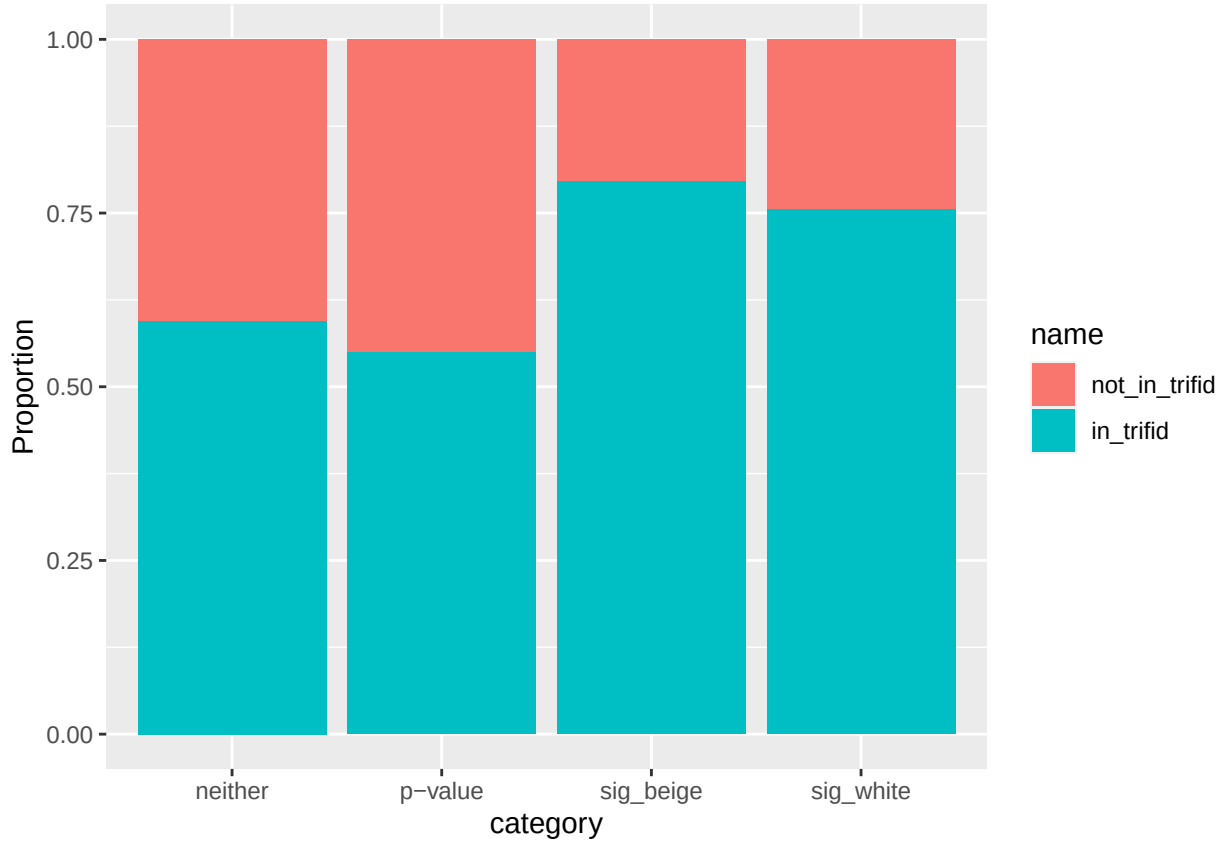
```
##   in_trifid not_in_trifid  category
```

```
## 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%