TRIFID_using_extra_introns

Makes Supplementary Table 5.

Trifid_stats - Figure 2 Go_networks_plus_trifid.Rmd - Figure 2

Theres a refseq TRIFID table, so for introns with only a refseq id I use the TRIFID score for refseq transcripts.

```
library(biomaRt)
library(tidyr)
library(dplyr)
library(ggplot2)
library(ggrepel)

library(EnhancedVolcano)
library(here)
i_am("R/15_TRIFID_using_extra_introns.Rmd")
```

```
lc = read.delim(here("31_leafcutter", "three_database_info_all_junctions.tsv"))
lc = unite(lc, "intron_coords", chr, start, end, strand, sep = ":")
lc[grep("PEMT",lc$gene),]
```

```
##       annotation             intron_coords  cluster_id      deltapsi
## 21       gencode chr17:17577027:17577414:- clu_19605_-  0.0610357482
## 22       gencode chr17:17577027:17582267:- clu_19605_-  0.0554167403
## 23       gencode chr17:17577027:17591531:- clu_19605_- -0.4096022388
## 24       gencode chr17:17577027:17591597:- clu_19605_-  0.0075591064
## 36900    gencode chr17:17505848:17506227:- clu_19603_-  0.0026696840
## 36901    gencode chr17:17506301:17509434:- clu_19603_- -0.0004726715
## 43646    gencode chr17:17509545:17512509:- clu_19604_- -0.0189567169
## 43647    gencode chr17:17509545:17576920:- clu_19604_-  0.0021782712
## 43648    gencode chr17:17512654:17519027:- clu_19604_-  0.0034533707
## 43649    gencode chr17:17512654:17522280:- clu_19604_-  0.0007176954
## 43650    gencode chr17:17512654:17576920:- clu_19604_-  0.0001893187
## 43651    gencode chr17:17522395:17576920:- clu_19604_-  0.0113229995
## 94116     refseq chr17:17577027:17577107:- clu_19605_-  0.2084922739
## 94117     refseq chr17:17577027:17591967:- clu_19605_-  0.0476375648
##             p.adjust
## 21    3.184596e-102
## 22    3.184596e-102
## 23    3.184596e-102
## 24    3.184596e-102
## 36900  1.925081e-01
## 36901  1.925081e-01
## 43646  2.859406e-01
## 43647  2.859406e-01
## 43648  2.859406e-01
## 43649  2.859406e-01
```

```
## 43650  2.859406e-01
## 43651  2.859406e-01
## 94116 3.184596e-102
## 94117 3.184596e-102
##
## 21
## 22
## 23                                                                                         ENST
## 24
## 36900 ENST00000395781.6,ENST00000435340.6,ENST00000580147.5,ENST00000582268.5,ENST00000477595.5,ENST
## 36901                                                                              ENST00000395783.5,ENST
## 43646                ENST00000255389.10,ENST00000435340.6,ENST00000472446.1,ENST00000395781.6,ENS
## 43647
## 43648
## 43649                                                             ENST00000421096.5,ENST00000255389.10,ENS
## 43650
## 43651                                                                              ENST00000395781.6,ENST
## 94116
## 94117
##       min_intron_number mode_intron_number gene
## 21                    1                  1 PEMT
## 22                    1                  1 PEMT
## 23                    1                  1 PEMT
## 24                    1                  1 PEMT
## 36900                 1                  4 PEMT
## 36901                 3                  3 PEMT
## 43646                 2                  4 PEMT
## 43647                 2                  2 PEMT
## 43648                 1                  1 PEMT
## 43649                 3                  3 PEMT
## 43650                 2                  2 PEMT
## 43651                 2                  2 PEMT
## 94116                 1                  1 PEMT
## 94117                 1                  1 PEMT
##                                          biotype is_first_intron
## 21                                protein_coding            TRUE
## 22                                protein_coding            TRUE
## 23    lncRNA,nonsense_mediated_decay,protein_coding            TRUE
## 24                         protein_coding,lncRNA            TRUE
## 36900 protein_coding,nonsense_mediated_decay,lncRNA            TRUE
## 36901 protein_coding,nonsense_mediated_decay,lncRNA           FALSE
## 43646                      protein_coding,lncRNA           FALSE
## 43647                    nonsense_mediated_decay           FALSE
## 43648                                     lncRNA            TRUE
## 43649 lncRNA,protein_coding,nonsense_mediated_decay           FALSE
## 43650                                     lncRNA           FALSE
## 43651                      protein_coding,lncRNA           FALSE
## 94116                                       <NA>            TRUE
## 94117                                       <NA>            TRUE
```

```
head(lc)
```

```
##   annotation            intron_coords  cluster_id      deltapsi      p.adjust
## 1    gencode chr7:43648652:43650493:- clu_35616_- -1.189192e-02 1.007194e-106
```

```
## 2    gencode chr7:43648652:43650612:- clu_35616_- -2.327379e-02 1.007194e-106
## 3    gencode chr7:43648652:43665658:- clu_35616_-  6.207827e-03 1.007194e-106
## 4    gencode chr7:43648652:43711400:- clu_35616_- -5.406415e-06 1.007194e-106
## 5    gencode chr7:43648652:43729429:- clu_35616_- -4.443950e-02 1.007194e-106
## 6    gencode chr7:43650712:43656033:- clu_35616_- -7.926449e-03 1.007194e-106
##                                                                                                                  trans
## 1                                                                                                              ENST000003
## 2 ENST00000446564.5,ENST00000431651.5,ENST00000418140.5,ENST00000448704.5,ENST00000451651.5,ENST00000
## 3                                                                                                               ENST00000
## 4                                                                                                               ENST00000
## 5                  ENST00000223336.10,ENST00000415076.6,ENST00000420441.1,ENST00000446330.6,ENST00000
## 6                                                                                                               ENST00000
##   min_intron_number mode_intron_number gene
## 1                 2                  2 COA1
## 2                 2                  2 COA1
## 3                 2                  2 COA1
## 4                 2                  2 COA1
## 5                 1                  1 COA1
## 6                 4                  4 COA1
##                                    biotype is_first_intron
## 1                           protein_coding           FALSE
## 2 nonsense_mediated_decay,protein_coding,lncRNA           FALSE
## 3                     nonsense_mediated_decay           FALSE
## 4                           protein_coding           FALSE
## 5       protein_coding,nonsense_mediated_decay            TRUE
## 6                           protein_coding           FALSE
```

```
dim(lc) # some duplications based on gene names/ antisense transcripts.
```

```
## [1] 132587     11
```

```
trifid = read.delim(here("annotations/trifid/gencode37_trifid_predictions.tsv"))
head(trifid)
```

```
##           gene_id gene_name   transcript_id  translation_id          flags
## 1 ENSG00000187010       RHD ENST00000342055 ENSP00000339577 protein_coding
## 2 ENSG00000187010       RHD ENST00000328664 ENSP00000331871 protein_coding
## 3 ENSG00000187010       RHD ENST00000417538 ENSP00000396420 protein_coding
## 4 ENSG00000187010       RHD ENST00000423810 ENSP00000399640 protein_coding
## 5 ENSG00000187010       RHD ENST00000622561 ENSP00000478087 protein_coding
## 6 ENSG00000187010       RHD ENST00000454452 ENSP00000413849 protein_coding
##        ccdsid      appris               ann_type length trifid_score
## 1 CCDS60028.1       MINOR            Alternative    493       0.2105
## 2   CCDS262.1 PRINCIPAL:1              Principal    417       0.4021
## 3 CCDS60031.1       MINOR            Alternative    378       0.0572
## 4 CCDS60027.1       MINOR Alternative Duplication    431       0.0415
## 5 CCDS60027.1       MINOR            Alternative    431       0.0223
## 6 CCDS53285.1       MINOR            Alternative    321       0.0652
##   norm_trifid_score
## 1            0.4210
## 2            0.8043
## 3            0.1145
## 4            0.0830
```

```
## 5            0.0447
## 6            0.1303
```

```
length(unique(trifid$transcript_id))#104 688
```

```
## [1] 104688
```

```
trefseq = read.delim(here("annotations/trifid/refseq110_trifid_predictions.tsv"))
nrow(trefseq)
```

```
## [1] 129456
```

```
head(trefseq)
```

```
##   gene_id gene_name transcript_id translation_id flags      ccdsid      appris
## 1    9997      SCO2  NM_001169111   NP_001162582  mRNA CCDS14095.1 PRINCIPAL:1
## 2    9997      SCO2  NM_001169110   NP_001162581  mRNA CCDS14095.1 PRINCIPAL:1
## 3    9997      SCO2  NM_001169109   NP_001162580  mRNA CCDS14095.1 PRINCIPAL:1
## 4    9997      SCO2     NM_005138      NP_005129  mRNA CCDS14095.1 PRINCIPAL:1
## 5    9994  CASP8AP2  NM_001137667   NP_001131139  mRNA           - PRINCIPAL:1
## 6    9994  CASP8AP2     NM_012115      NP_036247  mRNA           - PRINCIPAL:1
##   ann_type length trifid_score norm_trifid_score
## 1        -    266       0.5779            1.0000
## 2        -    266       0.5779            1.0000
## 3        -    266       0.5779            1.0000
## 4        -    266       0.5779            1.0000
## 5        -   1982       0.2909            0.5818
## 6        -   1982       0.2909            0.5818
```

```
trefseq$gene_id = as.character(trefseq$gene_id)
trifid = bind_rows(trifid, trefseq)
nrow(trifid)
```

```
## [1] 234144
```

The reason trifid has so few transcripts is only those with translated sequences are included.

Convert lc introns to transcript table

```
transcripts = separate_longer_delim(lc, transcript_ids, delim=",")
transcripts$transcript_ids = gsub("\\.[0-9]*","",gsub("rna-","",transcripts$transcript_ids))
head(transcripts)
```

```
##   annotation            intron_coords  cluster_id    deltapsi      p.adjust
## 1    gencode chr7:43648652:43650493:- clu_35616_- -0.01189192 1.007194e-106
## 2    gencode chr7:43648652:43650612:- clu_35616_- -0.02327379 1.007194e-106
## 3    gencode chr7:43648652:43650612:- clu_35616_- -0.02327379 1.007194e-106
## 4    gencode chr7:43648652:43650612:- clu_35616_- -0.02327379 1.007194e-106
```

```
## 5    gencode chr7:43648652:43650612:- clu_35616_- -0.02327379 1.007194e-106
## 6    gencode chr7:43648652:43650612:- clu_35616_- -0.02327379 1.007194e-106
##    transcript_ids min_intron_number mode_intron_number gene
## 1 ENST00000310564                 2                   2 COA1
## 2 ENST00000446564                 2                   2 COA1
## 3 ENST00000431651                 2                   2 COA1
## 4 ENST00000418140                 2                   2 COA1
## 5 ENST00000448704                 2                   2 COA1
## 6 ENST00000451651                 2                   2 COA1
##                                         biotype is_first_intron
## 1                                 protein_coding           FALSE
## 2 nonsense_mediated_decay,protein_coding,lncRNA           FALSE
## 3 nonsense_mediated_decay,protein_coding,lncRNA           FALSE
## 4 nonsense_mediated_decay,protein_coding,lncRNA           FALSE
## 5 nonsense_mediated_decay,protein_coding,lncRNA           FALSE
## 6 nonsense_mediated_decay,protein_coding,lncRNA           FALSE
```

```
dim(transcripts)
```

```
## [1] 379523     11
```

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

```
##        annotation             intron_coords  cluster_id      deltapsi
## 59        gencode chr17:17577027:17577414:- clu_19605_-  0.0610357482
## 60        gencode chr17:17577027:17582267:- clu_19605_-  0.0554167403
## 61        gencode chr17:17577027:17591531:- clu_19605_- -0.4096022388
## 62        gencode chr17:17577027:17591531:- clu_19605_- -0.4096022388
## 63        gencode chr17:17577027:17591531:- clu_19605_- -0.4096022388
## 64        gencode chr17:17577027:17591531:- clu_19605_- -0.4096022388
## 65        gencode chr17:17577027:17591531:- clu_19605_- -0.4096022388
## 66        gencode chr17:17577027:17591597:- clu_19605_-  0.0075591064
## 67        gencode chr17:17577027:17591597:- clu_19605_-  0.0075591064
## 123864    gencode chr17:17505848:17506227:- clu_19603_-  0.0026696840
## 123865    gencode chr17:17505848:17506227:- clu_19603_-  0.0026696840
## 123866    gencode chr17:17505848:17506227:- clu_19603_-  0.0026696840
## 123867    gencode chr17:17505848:17506227:- clu_19603_-  0.0026696840
## 123868    gencode chr17:17505848:17506227:- clu_19603_-  0.0026696840
## 123869    gencode chr17:17505848:17506227:- clu_19603_-  0.0026696840
## 123870    gencode chr17:17505848:17506227:- clu_19603_-  0.0026696840
## 123871    gencode chr17:17505848:17506227:- clu_19603_-  0.0026696840
## 123872    gencode chr17:17505848:17506227:- clu_19603_-  0.0026696840
## 123873    gencode chr17:17505848:17506227:- clu_19603_-  0.0026696840
## 123874    gencode chr17:17506301:17509434:- clu_19603_- -0.0004726715
## 123875    gencode chr17:17506301:17509434:- clu_19603_- -0.0004726715
## 123876    gencode chr17:17506301:17509434:- clu_19603_- -0.0004726715
## 123877    gencode chr17:17506301:17509434:- clu_19603_- -0.0004726715
## 123878    gencode chr17:17506301:17509434:- clu_19603_- -0.0004726715
## 123879    gencode chr17:17506301:17509434:- clu_19603_- -0.0004726715
## 146525    gencode chr17:17509545:17512509:- clu_19604_- -0.0189567169
## 146526    gencode chr17:17509545:17512509:- clu_19604_- -0.0189567169
## 146527    gencode chr17:17509545:17512509:- clu_19604_- -0.0189567169
## 146528    gencode chr17:17509545:17512509:- clu_19604_- -0.0189567169
```

```
## 146529    gencode chr17:17509545:17512509:- clu_19604_- -0.0189567169
## 146530    gencode chr17:17509545:17512509:- clu_19604_- -0.0189567169
## 146531    gencode chr17:17509545:17512509:- clu_19604_- -0.0189567169
## 146532    gencode chr17:17509545:17512509:- clu_19604_- -0.0189567169
## 146533    gencode chr17:17509545:17512509:- clu_19604_- -0.0189567169
## 146534    gencode chr17:17509545:17576920:- clu_19604_-  0.0021782712
## 146535    gencode chr17:17512654:17519027:- clu_19604_-  0.0034533707
## 146536    gencode chr17:17512654:17522280:- clu_19604_-  0.0007176954
## 146537    gencode chr17:17512654:17522280:- clu_19604_-  0.0007176954
## 146538    gencode chr17:17512654:17522280:- clu_19604_-  0.0007176954
## 146539    gencode chr17:17512654:17522280:- clu_19604_-  0.0007176954
## 146540    gencode chr17:17512654:17522280:- clu_19604_-  0.0007176954
## 146541    gencode chr17:17512654:17522280:- clu_19604_-  0.0007176954
## 146542    gencode chr17:17512654:17522280:- clu_19604_-  0.0007176954
## 146543    gencode chr17:17512654:17576920:- clu_19604_-  0.0001893187
## 146544    gencode chr17:17522395:17576920:- clu_19604_-  0.0113229995
## 146545    gencode chr17:17522395:17576920:- clu_19604_-  0.0113229995
## 146546    gencode chr17:17522395:17576920:- clu_19604_-  0.0113229995
## 146547    gencode chr17:17522395:17576920:- clu_19604_-  0.0113229995
## 146548    gencode chr17:17522395:17576920:- clu_19604_-  0.0113229995
## 146549    gencode chr17:17522395:17576920:- clu_19604_-  0.0113229995
## 320814     refseq chr17:17577027:17577107:- clu_19605_-  0.2084922739
## 320815     refseq chr17:17577027:17591967:- clu_19605_-  0.0476375648
##              p.adjust  transcript_ids min_intron_number mode_intron_number gene
## 59     3.184596e-102 ENST00000395782                 1                   1 PEMT
## 60     3.184596e-102 ENST00000395783                 1                   1 PEMT
## 61     3.184596e-102 ENST00000421096                 1                   1 PEMT
## 62     3.184596e-102 ENST00000580147                 1                   1 PEMT
## 63     3.184596e-102 ENST00000461404                 1                   1 PEMT
## 64     3.184596e-102 ENST00000255389                 1                   1 PEMT
## 65     3.184596e-102 ENST00000395781                 1                   1 PEMT
## 66     3.184596e-102 ENST00000435340                 1                   1 PEMT
## 67     3.184596e-102 ENST00000472446                 1                   1 PEMT
## 123864  1.925081e-01 ENST00000395781                 1                   4 PEMT
## 123865  1.925081e-01 ENST00000435340                 1                   4 PEMT
## 123866  1.925081e-01 ENST00000580147                 1                   4 PEMT
## 123867  1.925081e-01 ENST00000582268                 1                   4 PEMT
## 123868  1.925081e-01 ENST00000477595                 1                   4 PEMT
## 123869  1.925081e-01 ENST00000395783                 1                   4 PEMT
## 123870  1.925081e-01 ENST00000255389                 1                   4 PEMT
## 123871  1.925081e-01 ENST00000395782                 1                   4 PEMT
## 123872  1.925081e-01 ENST00000484838                 1                   4 PEMT
## 123873  1.925081e-01 ENST00000490392                 1                   4 PEMT
## 123874  1.925081e-01 ENST00000395783                 3                   3 PEMT
## 123875  1.925081e-01 ENST00000580147                 3                   3 PEMT
## 123876  1.925081e-01 ENST00000490392                 3                   3 PEMT
## 123877  1.925081e-01 ENST00000395782                 3                   3 PEMT
## 123878  1.925081e-01 ENST00000484838                 3                   3 PEMT
## 123879  1.925081e-01 ENST00000255389                 3                   3 PEMT
## 146525  2.859406e-01 ENST00000255389                 2                   4 PEMT
## 146526  2.859406e-01 ENST00000435340                 2                   4 PEMT
## 146527  2.859406e-01 ENST00000472446                 2                   4 PEMT
## 146528  2.859406e-01 ENST00000395781                 2                   4 PEMT
## 146529  2.859406e-01 ENST00000484838                 2                   4 PEMT
```

```
## 146530  2.859406e-01 ENST00000490392                2                  4 PEMT
## 146531  2.859406e-01 ENST00000395782                2                  4 PEMT
## 146532  2.859406e-01 ENST00000421096                2                  4 PEMT
## 146533  2.859406e-01 ENST00000395783                2                  4 PEMT
## 146534  2.859406e-01 ENST00000580147                2                  2 PEMT
## 146535  2.859406e-01 ENST00000490392                1                  1 PEMT
## 146536  2.859406e-01 ENST00000421096                3                  3 PEMT
## 146537  2.859406e-01 ENST00000255389                3                  3 PEMT
## 146538  2.859406e-01 ENST00000395781                3                  3 PEMT
## 146539  2.859406e-01 ENST00000395783                3                  3 PEMT
## 146540  2.859406e-01 ENST00000461404                3                  3 PEMT
## 146541  2.859406e-01 ENST00000435340                3                  3 PEMT
## 146542  2.859406e-01 ENST00000395782                3                  3 PEMT
## 146543  2.859406e-01 ENST00000472446                2                  2 PEMT
## 146544  2.859406e-01 ENST00000395781                2                  2 PEMT
## 146545  2.859406e-01 ENST00000395783                2                  2 PEMT
## 146546  2.859406e-01 ENST00000395782                2                  2 PEMT
## 146547  2.859406e-01 ENST00000255389                2                  2 PEMT
## 146548  2.859406e-01 ENST00000421096                2                  2 PEMT
## 146549  2.859406e-01 ENST00000435340                2                  2 PEMT
## 320814 3.184596e-102    XM_006721418                1                  1 PEMT
## 320815 3.184596e-102    XM_024450532                1                  1 PEMT
##                                                 biotype is_first_intron
## 59                                       protein_coding            TRUE
## 60                                       protein_coding            TRUE
## 61     lncRNA,nonsense_mediated_decay,protein_coding            TRUE
## 62     lncRNA,nonsense_mediated_decay,protein_coding            TRUE
## 63     lncRNA,nonsense_mediated_decay,protein_coding            TRUE
## 64     lncRNA,nonsense_mediated_decay,protein_coding            TRUE
## 65     lncRNA,nonsense_mediated_decay,protein_coding            TRUE
## 66                               protein_coding,lncRNA            TRUE
## 67                               protein_coding,lncRNA            TRUE
## 123864 protein_coding,nonsense_mediated_decay,lncRNA            TRUE
## 123865 protein_coding,nonsense_mediated_decay,lncRNA            TRUE
## 123866 protein_coding,nonsense_mediated_decay,lncRNA            TRUE
## 123867 protein_coding,nonsense_mediated_decay,lncRNA            TRUE
## 123868 protein_coding,nonsense_mediated_decay,lncRNA            TRUE
## 123869 protein_coding,nonsense_mediated_decay,lncRNA            TRUE
## 123870 protein_coding,nonsense_mediated_decay,lncRNA            TRUE
## 123871 protein_coding,nonsense_mediated_decay,lncRNA            TRUE
## 123872 protein_coding,nonsense_mediated_decay,lncRNA            TRUE
## 123873 protein_coding,nonsense_mediated_decay,lncRNA            TRUE
## 123874 protein_coding,nonsense_mediated_decay,lncRNA           FALSE
## 123875 protein_coding,nonsense_mediated_decay,lncRNA           FALSE
## 123876 protein_coding,nonsense_mediated_decay,lncRNA           FALSE
## 123877 protein_coding,nonsense_mediated_decay,lncRNA           FALSE
## 123878 protein_coding,nonsense_mediated_decay,lncRNA           FALSE
## 123879 protein_coding,nonsense_mediated_decay,lncRNA           FALSE
## 146525                           protein_coding,lncRNA           FALSE
## 146526                           protein_coding,lncRNA           FALSE
## 146527                           protein_coding,lncRNA           FALSE
## 146528                           protein_coding,lncRNA           FALSE
## 146529                           protein_coding,lncRNA           FALSE
## 146530                           protein_coding,lncRNA           FALSE
```

```
## 146531                          protein_coding,lncRNA          FALSE
## 146532                          protein_coding,lncRNA          FALSE
## 146533                          protein_coding,lncRNA          FALSE
## 146534                       nonsense_mediated_decay          FALSE
## 146535                                       lncRNA           TRUE
## 146536 lncRNA,protein_coding,nonsense_mediated_decay          FALSE
## 146537 lncRNA,protein_coding,nonsense_mediated_decay          FALSE
## 146538 lncRNA,protein_coding,nonsense_mediated_decay          FALSE
## 146539 lncRNA,protein_coding,nonsense_mediated_decay          FALSE
## 146540 lncRNA,protein_coding,nonsense_mediated_decay          FALSE
## 146541 lncRNA,protein_coding,nonsense_mediated_decay          FALSE
## 146542 lncRNA,protein_coding,nonsense_mediated_decay          FALSE
## 146543                                       lncRNA          FALSE
## 146544                          protein_coding,lncRNA          FALSE
## 146545                          protein_coding,lncRNA          FALSE
## 146546                          protein_coding,lncRNA          FALSE
## 146547                          protein_coding,lncRNA          FALSE
## 146548                          protein_coding,lncRNA          FALSE
## 146549                          protein_coding,lncRNA          FALSE
## 320814                                          <NA>           TRUE
## 320815                                          <NA>           TRUE
```

some numbers on this

```
sig_trans = filter(transcripts, p.adjust < 0.05 & abs(deltapsi) >= 0.1)
sprintf("Leafcutter significant introns (%d) correspond to %d unique transcripts (represented in %d rows
        length(unique(sig_trans$intron_coords)),
        length(unique(sig_trans$transcript_ids)),
        length(sig_trans$transcript_ids))
```

```
## [1] "Leafcutter significant introns (693) correspond to 1924 unique transcripts (represented in 2046
```

```
sprintf("%d of these transcripts are represented in the trifid database.\n",
        sum(unique(sig_trans$transcript_ids) %in% trifid$transcript_id))
```

```
## [1] "1383 of these transcripts are represented in the trifid database.\n"
```

```
transcripts$transcript_ids[grep("PEMT",transcripts$gene,)] %in% trifid$transcript_id
```

```
##  [1]  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE
## [13] FALSE FALSE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE FALSE  TRUE FALSE
## [25]  TRUE  TRUE  TRUE FALSE  TRUE FALSE FALSE  TRUE FALSE  TRUE  TRUE FALSE
## [37] FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE
## [49] FALSE  TRUE  TRUE  TRUE
```

```
cat("Unrepresented transcripts include those from genes like: \n")
```

```
## Unrepresented transcripts include those from genes like:
```

```
cat(head(unique(sig_trans$gene[!sig_trans$transcript_ids %in% trifid$transcript_id])))
```

```
## PEMT CA5BP1 CA5B PPARG CD44 FAR2
```

```
missing = transcripts[!transcripts$transcript_ids %in% trifid$transcript_id,] #annotate later
table(missing$annotation)
```

```
##
##    cryptic fantom_cat    gencode     refseq
##      21831      18773      75171       4124
```

```
missing[grep( "PEMT", missing$gene),] #Just 1 out of 6 significant transcripts are missing for PEMT
```

```
##        annotation             intron_coords  cluster_id      deltapsi
## 61        gencode chr17:17577027:17591531:- clu_19605_- -0.4096022388
## 67        gencode chr17:17577027:17591597:- clu_19605_-  0.0075591064
## 123867    gencode chr17:17505848:17506227:- clu_19603_-  0.0026696840
## 123868    gencode chr17:17505848:17506227:- clu_19603_-  0.0026696840
## 123872    gencode chr17:17505848:17506227:- clu_19603_-  0.0026696840
## 123873    gencode chr17:17505848:17506227:- clu_19603_-  0.0026696840
## 123876    gencode chr17:17506301:17509434:- clu_19603_- -0.0004726715
## 123878    gencode chr17:17506301:17509434:- clu_19603_- -0.0004726715
## 146527    gencode chr17:17509545:17512509:- clu_19604_- -0.0189567169
## 146529    gencode chr17:17509545:17512509:- clu_19604_- -0.0189567169
## 146530    gencode chr17:17509545:17512509:- clu_19604_- -0.0189567169
## 146532    gencode chr17:17509545:17512509:- clu_19604_- -0.0189567169
## 146535    gencode chr17:17512654:17519027:- clu_19604_-  0.0034533707
## 146536    gencode chr17:17512654:17522280:- clu_19604_-  0.0007176954
## 146543    gencode chr17:17512654:17576920:- clu_19604_-  0.0001893187
## 146548    gencode chr17:17522395:17576920:- clu_19604_-  0.0113229995
##             p.adjust  transcript_ids min_intron_number mode_intron_number gene
## 61     3.184596e-102 ENST00000421096                 1                  1 PEMT
## 67     3.184596e-102 ENST00000472446                 1                  1 PEMT
## 123867  1.925081e-01 ENST00000582268                 1                  4 PEMT
## 123868  1.925081e-01 ENST00000477595                 1                  4 PEMT
## 123872  1.925081e-01 ENST00000484838                 1                  4 PEMT
## 123873  1.925081e-01 ENST00000490392                 1                  4 PEMT
## 123876  1.925081e-01 ENST00000490392                 3                  3 PEMT
## 123878  1.925081e-01 ENST00000484838                 3                  3 PEMT
## 146527  2.859406e-01 ENST00000472446                 2                  4 PEMT
## 146529  2.859406e-01 ENST00000484838                 2                  4 PEMT
## 146530  2.859406e-01 ENST00000490392                 2                  4 PEMT
## 146532  2.859406e-01 ENST00000421096                 2                  4 PEMT
## 146535  2.859406e-01 ENST00000490392                 1                  1 PEMT
## 146536  2.859406e-01 ENST00000421096                 3                  3 PEMT
## 146543  2.859406e-01 ENST00000472446                 2                  2 PEMT
## 146548  2.859406e-01 ENST00000421096                 2                  2 PEMT
##                                           biotype is_first_intron
## 61     lncRNA,nonsense_mediated_decay,protein_coding            TRUE
## 67                             protein_coding,lncRNA            TRUE
## 123867 protein_coding,nonsense_mediated_decay,lncRNA            TRUE
```

```
## 123868 protein_coding,nonsense_mediated_decay,lncRNA            TRUE
## 123872 protein_coding,nonsense_mediated_decay,lncRNA            TRUE
## 123873 protein_coding,nonsense_mediated_decay,lncRNA            TRUE
## 123876 protein_coding,nonsense_mediated_decay,lncRNA           FALSE
## 123878 protein_coding,nonsense_mediated_decay,lncRNA           FALSE
## 146527                         protein_coding,lncRNA           FALSE
## 146529                         protein_coding,lncRNA           FALSE
## 146530                         protein_coding,lncRNA           FALSE
## 146532                         protein_coding,lncRNA           FALSE
## 146535                                      lncRNA            TRUE
## 146536 lncRNA,protein_coding,nonsense_mediated_decay           FALSE
## 146543                                      lncRNA           FALSE
## 146548                         protein_coding,lncRNA           FALSE
```

Of course fantom and cryptic transcripts cannot be annotated by this trifid (which is based on the ensembl + refseq annotation); but also many transcripts from gencode cannot be found there either - perhaps they're new transcripts, or not protein coding versions. The genocde annotation is always changing.

Merge trifid and leafcutter

```
mtrans = merge(transcripts, trifid, by.x ="transcript_ids",by.y="transcript_id")
mtrans = mutate(mtrans, condition = if_else(deltapsi > 0, "beige", if_else(deltapsi < 0, "white", "none
                sig = p.adjust < 0.05 & abs(deltapsi) >= 0.1)

length(unique(mtrans$transcript_ids)) #82 326 transcripts in total
```

```
## [1] 82326
```

```
length(unique(filter(mtrans, sig) %>% pull(transcript_ids))) #1301 + 82 from significant introns > 0.1
```

```
## [1] 1383
```

```
mtrans = arrange(mtrans, desc(trifid_score), desc(norm_trifid_score), p.adjust, desc(abs(deltapsi)))
head(mtrans[c("gene_name", "trifid_score","norm_trifid_score","deltapsi","p.adjust")], n=20)
```

```
##    gene_name trifid_score norm_trifid_score      deltapsi     p.adjust
## 1     LRRC28            1                 1 -0.1351265692 2.720861e-14
## 2      COPS5            1                 1  0.0594822450 1.763160e-08
## 3      STK38            1                 1  0.0251638798 2.609567e-07
## 4        C1D            1                 1  0.0098738938 3.061780e-07
## 5      AAMDC            1                 1 -0.0008931935 7.197419e-07
## 6      AAMDC            1                 1 -0.0006013790 7.197419e-07
## 7      ACTR6            1                 1  0.0444343616 8.870709e-06
## 8     ARFIP1            1                 1  0.0038268440 2.680114e-05
## 9     ARFIP1            1                 1  0.0037346590 2.680114e-05
## 10    RPRD1A            1                 1 -0.0201042561 6.950028e-05
## 11    RPRD1A            1                 1 -0.0182828468 6.950028e-05
## 12    RPRD1A            1                 1  0.0077775865 6.950028e-05
## 13     GNAI2            1                 1 -0.0235736966 7.140425e-05
## 14      GNB4            1                 1  0.0222376961 2.309501e-04
```

```
## 15      GDI2            1                 1 -0.0325581934 5.852116e-04
## 16      GDI2            1                 1  0.0300832463 5.852116e-04
## 17    ACTR10            1                 1  0.0278916145 2.554726e-03
## 18    ACTR10            1                 1 -0.0268661873 2.554726e-03
## 19  ATP6V1C1            1                 1  0.0237742392 4.773516e-03
## 20  ATP6V1C1            1                 1 -0.0148102227 4.773516e-03
```

```
head(mtrans[mtrans$gene == "PEMT",])
```

```
##        transcript_ids annotation             intron_coords  cluster_id
## 39508    XM_024450532     refseq chr17:17577027:17591967:- clu_19605_-
## 49409 ENST00000395782    gencode chr17:17577027:17577414:- clu_19605_-
## 49410 ENST00000395783    gencode chr17:17577027:17582267:- clu_19605_-
## 49422 ENST00000395782    gencode chr17:17505848:17506227:- clu_19603_-
## 49423 ENST00000395783    gencode chr17:17505848:17506227:- clu_19603_-
## 49424 ENST00000395782    gencode chr17:17506301:17509434:- clu_19603_-
##            deltapsi      p.adjust min_intron_number mode_intron_number gene
## 39508  0.0476375648 3.184596e-102                 1                  1 PEMT
## 49409  0.0610357482 3.184596e-102                 1                  1 PEMT
## 49410  0.0554167403 3.184596e-102                 1                  1 PEMT
## 49422  0.0026696840  1.925081e-01                 1                  4 PEMT
## 49423  0.0026696840  1.925081e-01                 1                  4 PEMT
## 49424 -0.0004726715  1.925081e-01                 3                  3 PEMT
##                                           biotype is_first_intron
## 39508                                       <NA>            TRUE
## 49409                             protein_coding            TRUE
## 49410                             protein_coding            TRUE
## 49422 protein_coding,nonsense_mediated_decay,lncRNA            TRUE
## 49423 protein_coding,nonsense_mediated_decay,lncRNA            TRUE
## 49424 protein_coding,nonsense_mediated_decay,lncRNA           FALSE
##               gene_id gene_name  translation_id          flags      ccdsid
## 39508           10400      PEMT    XP_024306300           mRNA           -
## 49409 ENSG00000133027      PEMT ENSP00000379128 protein_coding CCDS11187.1
## 49410 ENSG00000133027      PEMT ENSP00000379129 protein_coding CCDS11187.1
## 49422 ENSG00000133027      PEMT ENSP00000379128 protein_coding CCDS11187.1
## 49423 ENSG00000133027      PEMT ENSP00000379129 protein_coding CCDS11187.1
## 49424 ENSG00000133027      PEMT ENSP00000379128 protein_coding CCDS11187.1
##            appris              ann_type length trifid_score norm_trifid_score
## 39508 PRINCIPAL:1                     -    199       0.8739            0.8914
## 49409 PRINCIPAL:1             Principal    199       0.8038            1.0000
## 49410 PRINCIPAL:1 Principal Duplication    199       0.8038            1.0000
## 49422 PRINCIPAL:1             Principal    199       0.8038            1.0000
## 49423 PRINCIPAL:1 Principal Duplication    199       0.8038            1.0000
## 49424 PRINCIPAL:1             Principal    199       0.8038            1.0000
##       condition   sig
## 39508     beige FALSE
## 49409     beige FALSE
## 49410     beige FALSE
## 49422     beige FALSE
## 49423     beige FALSE
## 49424     white FALSE
```

summarise transcripts with introns in multiple directions

```
table(mtrans$num_introns)
```

```
## < table of extent 0 >
```

```
table(mtrans$condition)
```

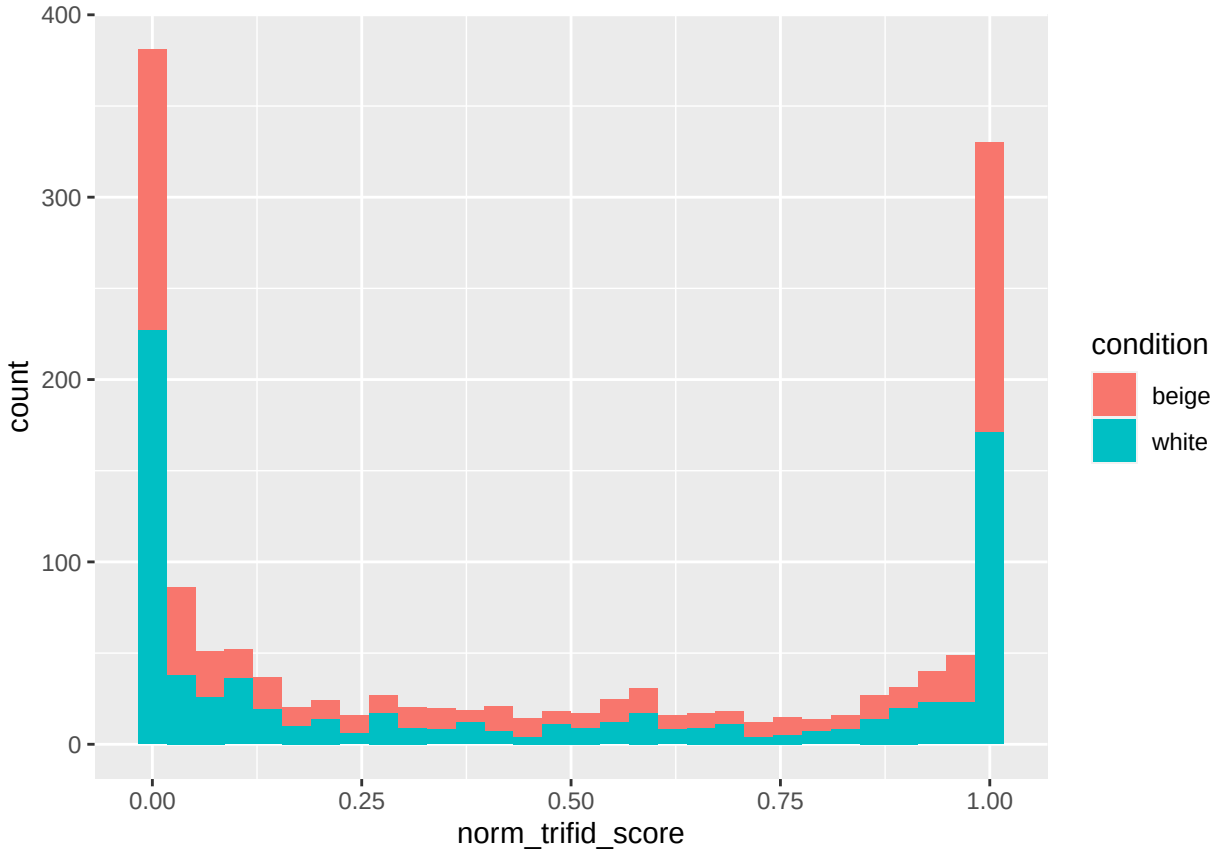
```
##
##  beige   none  white
## 133239      2 126383
```

```
table(paste(mtrans$sig, mtrans$condition))
```

```
##
## FALSE beige  FALSE none FALSE white  TRUE beige  TRUE white
##      132560           2      125598         679         785
```

Histogram

```
ggplot(filter(mtrans, sig)) + geom_histogram(aes(x=norm_trifid_score, fill=condition), bins=30)
```



Flags

```
table(mtrans$flags)
```

```
##
##                       mRNA             non_stop_decay
##                      25475                        146
##    nonsense_mediated_decay nonsense_mediated_decay,RT
##                      41896                       1836
##             protein_coding          protein_coding,RT
##                     189018                       1251
##                  TR_C_gene
##                          2
```

```
table(filter(mtrans, sig) %>% pull(flags))
```

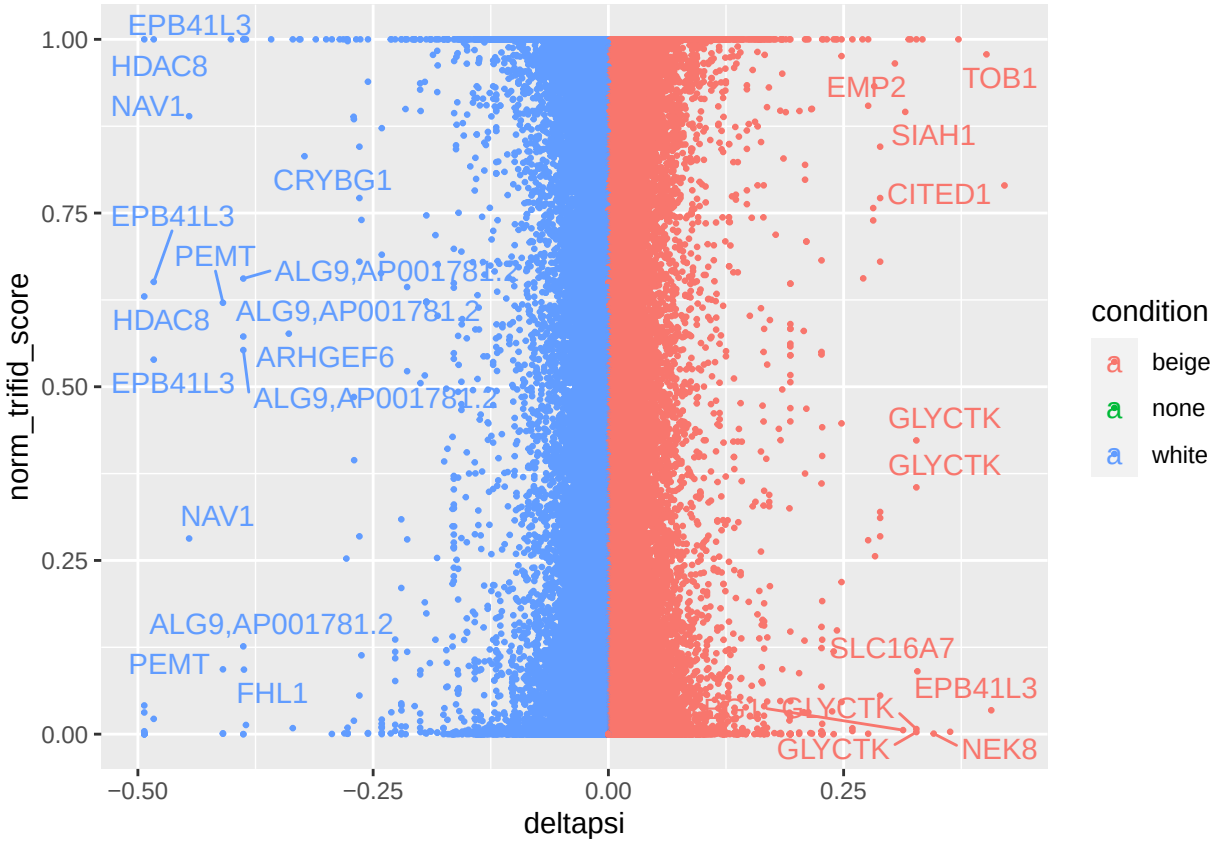
```
##
##                       mRNA             non_stop_decay
##                         82                          1
##    nonsense_mediated_decay nonsense_mediated_decay,RT
##                        185                         10
##             protein_coding          protein_coding,RT
##                       1178                          8
```

Correlation between dPSI and TRIFID score?

There is actually, if you look within the same cluster if we have a intron with a higher PSI it tends to have higher functionality - or maybe it just has more transcripts?

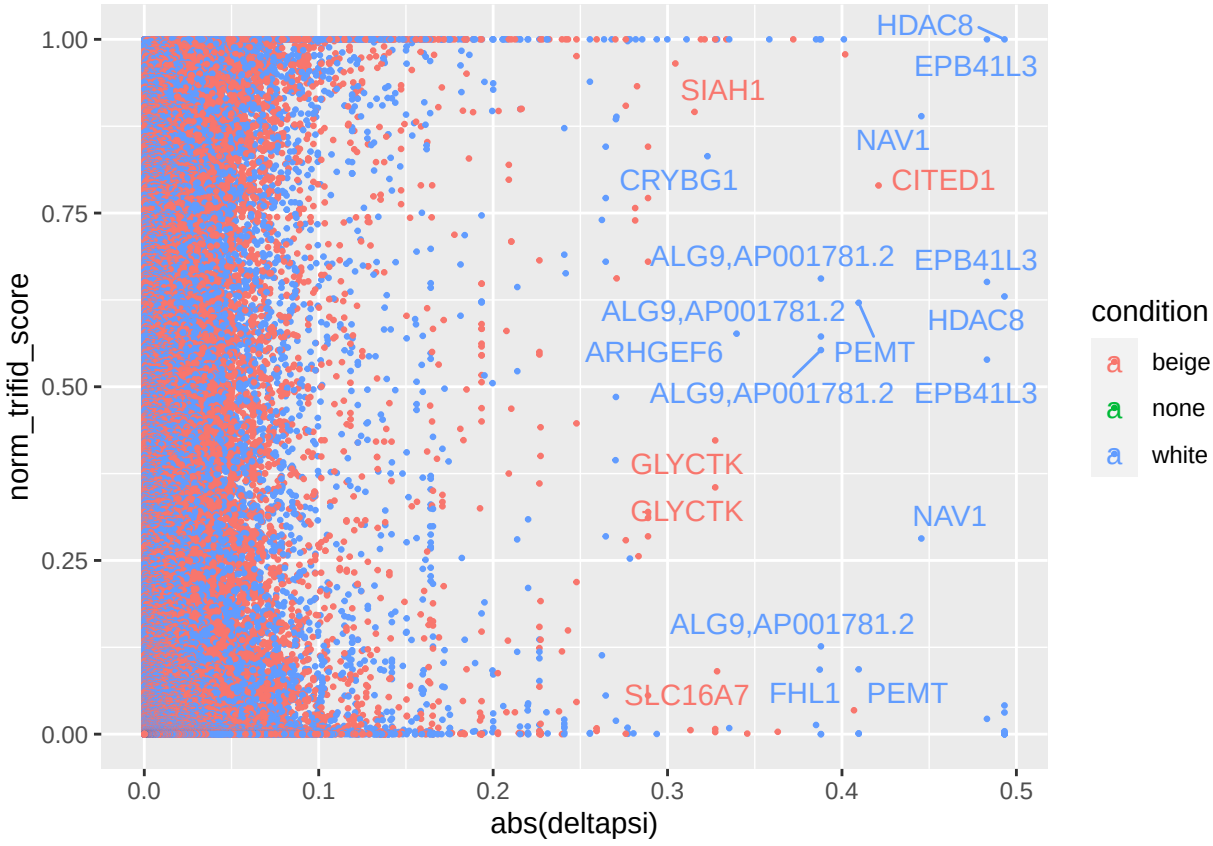
```
ggplot(mtrans, aes(x=deltapsi, y=norm_trifid_score, colour=condition)) + geom_point(size=0.5) +
    geom_text_repel(data= filter(mtrans, abs(deltapsi) >0.3), aes(label=gene))
```

```
## Warning: ggrepel: 33 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```




```
ggplot(mtrans, aes(x=abs(deltapsi), y=norm_trifid_score, colour=condition)) + geom_point(size=0.5) +
    geom_text_repel(data= filter(mtrans, abs(deltapsi) >0.3), aes(label=gene))
```

```
## Warning: ggrepel: 40 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



Annotate transcript names

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

biomaRt::searchAttributes(mart = mart, pattern = "transcript.*name")
```

```
##                                name                  description         page
## 24        external_transcript_name               Transcript name feature_page
## 25 external_transcript_source_name     Source of transcript name feature_page
## 58           entrezgene_trans_name EntrezGene transcript name ID feature_page
## 77              mirbase_trans_name    miRBase transcript name ID feature_page
## 92                 rfam_trans_name       RFAM transcript name ID feature_page
```

```
annot = getBM(c("external_transcript_name","ensembl_gene_id", "ensembl_transcript_id"),
              filters = "ensembl_transcript_id",
              values =  mtrans$transcript_ids,
              mart = mart, useCache = F)

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

```
##   external_transcript_name ensembl_gene_id ensembl_transcript_id
```

```
## 1                VTI1B-204 ENSG00000100568       ENST00000554659
## 2                VPS36-201 ENSG00000136100       ENST00000378060
```

```
## [1] 69152     3
```

```
#Add gene names to filt_series
mtrans = merge(mtrans,annot, by.x= "transcript_ids",
                        by.y = "ensembl_transcript_id", sort=FALSE,
               all.x = T)
#head(mtrans)
remove(annot)
```

Averaging score across isoforms at each intron

```
summary(mtrans$trifid_score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0009  0.0777  0.2988  0.6362  1.0000
```

```
head(mtrans)
```

```
##    transcript_ids annotation             intron_coords  cluster_id     deltapsi
## 1 ENST00000301981    gencode chr15:99276616:99287257:+ clu_30898_+  0.010503776
## 2 ENST00000301981    gencode chr15:99334129:99352369:+ clu_30899_+ -0.041238134
## 3 ENST00000301981    gencode chr15:99251541:99255898:+ clu_30897_+ -0.135126569
## 4 ENST00000301981    gencode chr15:99256125:99276576:+ clu_30898_+  0.011506472
## 5 ENST00000301981    gencode chr15:99352471:99361336:+ clu_30900_+  0.004463087
## 6 ENST00000301981    gencode chr15:99361511:99363106:+ clu_30901_+  0.002213603
##       p.adjust min_intron_number mode_intron_number   gene
## 1 1.748305e-01                 3                  3 LRRC28
## 2 3.326508e-02                 1                  1 LRRC28
## 3 2.720861e-14                 1                  1 LRRC28
## 4 1.748305e-01                 2                  2 LRRC28
## 5 8.785136e-01                 2                  4 LRRC28
## 6 5.514856e-01                 1                  8 LRRC28
##                                                      biotype is_first_intron
## 1                                        lncRNA,protein_coding           FALSE
## 2 protein_coding,retained_intron,lncRNA,nonsense_mediated_decay            TRUE
## 3 nonsense_mediated_decay,protein_coding,retained_intron,lncRNA            TRUE
## 4                 lncRNA,protein_coding,nonsense_mediated_decay           FALSE
## 5        retained_intron,protein_coding,nonsense_mediated_decay           FALSE
## 6        retained_intron,protein_coding,nonsense_mediated_decay            TRUE
##           gene_id gene_name  translation_id          flags      ccdsid
## 1 ENSG00000168904    LRRC28 ENSP00000304923 protein_coding CCDS10380.1
## 2 ENSG00000168904    LRRC28 ENSP00000304923 protein_coding CCDS10380.1
## 3 ENSG00000168904    LRRC28 ENSP00000304923 protein_coding CCDS10380.1
## 4 ENSG00000168904    LRRC28 ENSP00000304923 protein_coding CCDS10380.1
## 5 ENSG00000168904    LRRC28 ENSP00000304923 protein_coding CCDS10380.1
## 6 ENSG00000168904    LRRC28 ENSP00000304923 protein_coding CCDS10380.1
##        appris  ann_type length trifid_score norm_trifid_score condition   sig
```

```
## 1 PRINCIPAL:1 Principal    367            1                 1     beige FALSE
## 2 PRINCIPAL:1 Principal    367            1                 1     white FALSE
## 3 PRINCIPAL:1 Principal    367            1                 1     white  TRUE
## 4 PRINCIPAL:1 Principal    367            1                 1     beige FALSE
## 5 PRINCIPAL:1 Principal    367            1                 1     beige FALSE
## 6 PRINCIPAL:1 Principal    367            1                 1     beige FALSE
##   external_transcript_name ensembl_gene_id
## 1               LRRC28-201 ENSG00000168904
## 2               LRRC28-201 ENSG00000168904
## 3               LRRC28-201 ENSG00000168904
## 4               LRRC28-201 ENSG00000168904
## 5               LRRC28-201 ENSG00000168904
## 6               LRRC28-201 ENSG00000168904
```

```
dim(mtrans)
```

```
## [1] 289037     25
```

```
#mtrans = filter(mtrans, sig)

#unknown gene names get given a ".", change to NA
mtrans = mutate(mtrans, gene = if_else(gene==".", cluster_id, gene))

#this step  averages the score if multiple transcripts are implicated at an intron
introns = group_by(mtrans, intron_coords,condition, deltapsi, p.adjust, gene, cluster_id, annotation) %
                                                                    mean_norm_score = mean(norm_trifid_sco
                                                                     median_norm_score = median(norm_trifi
                                                                    transcripts = paste(transcript_ids, se
                                                                    transcript_names = paste(external_trans
```

```
## `summarise()` has grouped output by 'intron_coords', 'condition', 'deltapsi',
## 'p.adjust', 'gene', 'cluster_id'. You can override using the `.groups`
## argument.
```

```
introns = arrange(introns, desc(abs(deltapsi)), desc(mean_trifid_score), p.adjust,)
dim(introns) #78756 introns with annotations
```

```
## [1] 87024    12
```

```
head(introns)
```

```
## # A tibble: 6 x 12
## # Groups:   intron_coords, condition, deltapsi, p.adjust, gene, cluster_id [6]
##   intron_coords         condition deltapsi  p.adjust gene  cluster_id annotation
##   <chr>                 <chr>        <dbl>     <dbl> <chr> <chr>      <chr>
## 1 chrX:72330076:723517~ white       -0.493 1.86e- 81 HDAC8 clu_291_-  gencode
## 2 chr18:5489194:554391~ white       -0.483 4.27e- 13 EPB4~ clu_21093~ gencode
## 3 chr1:201718755:20178~ white       -0.446 5.73e- 43 NAV1  clu_14762~ gencode
## 4 chrX:72302934:723070~ beige        0.421 1.86e- 81 CITE~ clu_291_-  gencode
## 5 chr17:17577027:17591~ white       -0.410 3.18e-102 PEMT  clu_19605~ gencode
## 6 chr18:5489161:563037~ beige        0.407 4.27e- 13 EPB4~ clu_21093~ gencode
## # i 5 more variables: mean_trifid_score <dbl>, mean_norm_score <dbl>,
## #   median_norm_score <dbl>, transcripts <chr>, transcript_names <chr>
```

```
length(unique(introns$intron_coords))
```

```
## [1] 87024
```

If an intron is not in trifid; report as -0.1

```
sig_clusters = unique(lc$cluster_id[lc$p.adjust < 0.05 & abs(lc$deltapsi) > 0.1])
summary(sig_clusters %in% introns$cluster_id)#379 clusters have at least 1 trifid annotation; 58 do n
```

```
##    Mode   FALSE    TRUE
## logical      51     386
```

```
#must have another annotated intron in the cluster, but not have a score already for that intron
to_add = missing#[missing$cluster_id %in% sig_clusters &
                 #    !missing$intron_coords %in% introns$intron_coords,]
nrow(to_add) #160 of the unannotated trifid transcripts, have an annotated cluster and are for an unsco
```

```
## [1] 119899
```

```
to_add = mutate(to_add, gene = if_else(gene==".", cluster_id, gene))
to_add = group_by(to_add, intron_coords, deltapsi, p.adjust, gene, cluster_id, annotation) %>%
    summarise(transcripts = paste(transcript_ids,collapse=","),
              )
```

```
## `summarise()` has grouped output by 'intron_coords', 'deltapsi', 'p.adjust',
## 'gene', 'cluster_id'. You can override using the `.groups` argument.
```

```
to_add = mutate(to_add, mean_trifid_score = -0.1,
                        mean_norm_score = -0.1,
                        median_norm_score = -0.1,
                condition=  if_else(deltapsi > 0, "beige", "white"))
head(to_add)
```

```
## # A tibble: 6 x 11
## # Groups:   intron_coords, deltapsi, p.adjust, gene, cluster_id [6]
##   intron_coords        deltapsi p.adjust gene  cluster_id annotation transcripts
##   <chr>                   <dbl>    <dbl> <chr> <chr>      <chr>      <chr>
## 1 chr10:100006342:100~  7.87e-4    0.454 ENSG~ clu_37954~ fantom_cat FTMT237000~
## 2 chr10:1001013:10058~  2.47e-2    0.108 GTPB~ clu_29373~ gencode    ENST000004~
## 3 chr10:1001013:10070~ -9.44e-5    0.108 <NA>  clu_29373~ cryptic    Unknown
## 4 chr10:100176070:100~  3.00e-5    0.296 ERLI~ clu_37957~ refseq     NR_144757
## 5 chr10:100185637:100~  8.04e-4    0.349 ERLI~ clu_37958~ refseq     NR_144760
## 6 chr10:100190968:100~  2.97e-2    0.153 CHUK  clu_37959~ gencode    ENST000005~
## # i 4 more variables: mean_trifid_score <dbl>, mean_norm_score <dbl>,
## #   median_norm_score <dbl>, condition <chr>
```

```
nrow(to_add)
```

```
## [1] 73850
```

```
length(unique(to_add$intron_coords))
```

```
## [1] 73850
```

```
to_add= distinct(to_add) #managed to duplicate
nrow(to_add)
```

```
## [1] 73850
```

```
all_introns = bind_rows(in_trifid=introns, not_in_trifid=to_add[to_add$cluster_id %in% introns$cluster_
                        cluster_not_in_trifid= to_add[!to_add$cluster_id %in% introns$cluster_id,],
                        .id = "in_trifid" )
nrow(all_introns)
```

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

```
length(unique(all_introns$intron_coords))
```

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

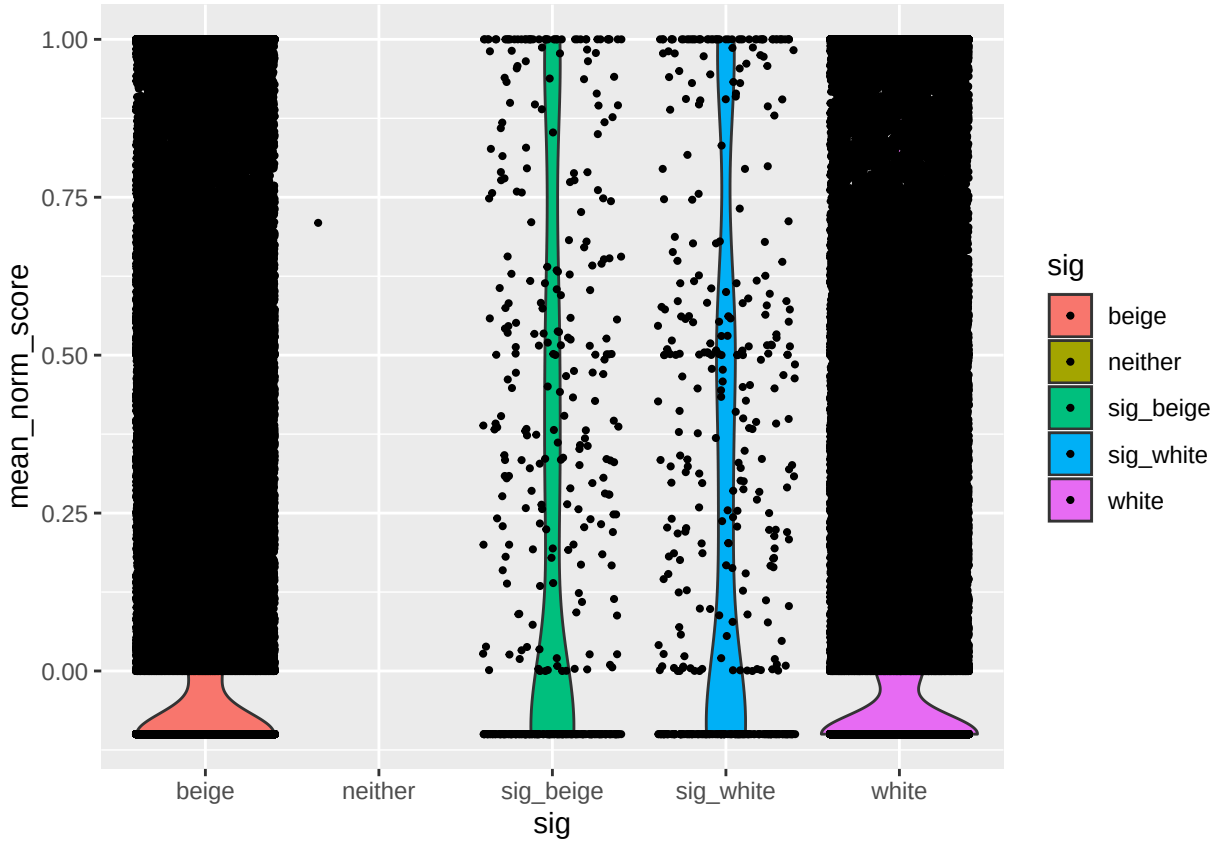
Violin plots on *average* TRIFID score

```
all_introns =mutate(all_introns, sig = if_else(p.adjust < 0.05 & deltapsi > 0.1, "sig_beige",
                                        if_else(p.adjust < 0.05 & deltapsi < -0.1, "sig_white",
                                                if_else(deltapsi > 0, "beige",
                                                        if_else(deltapsi < -0, "white", "neither
table(all_introns$sig)
```

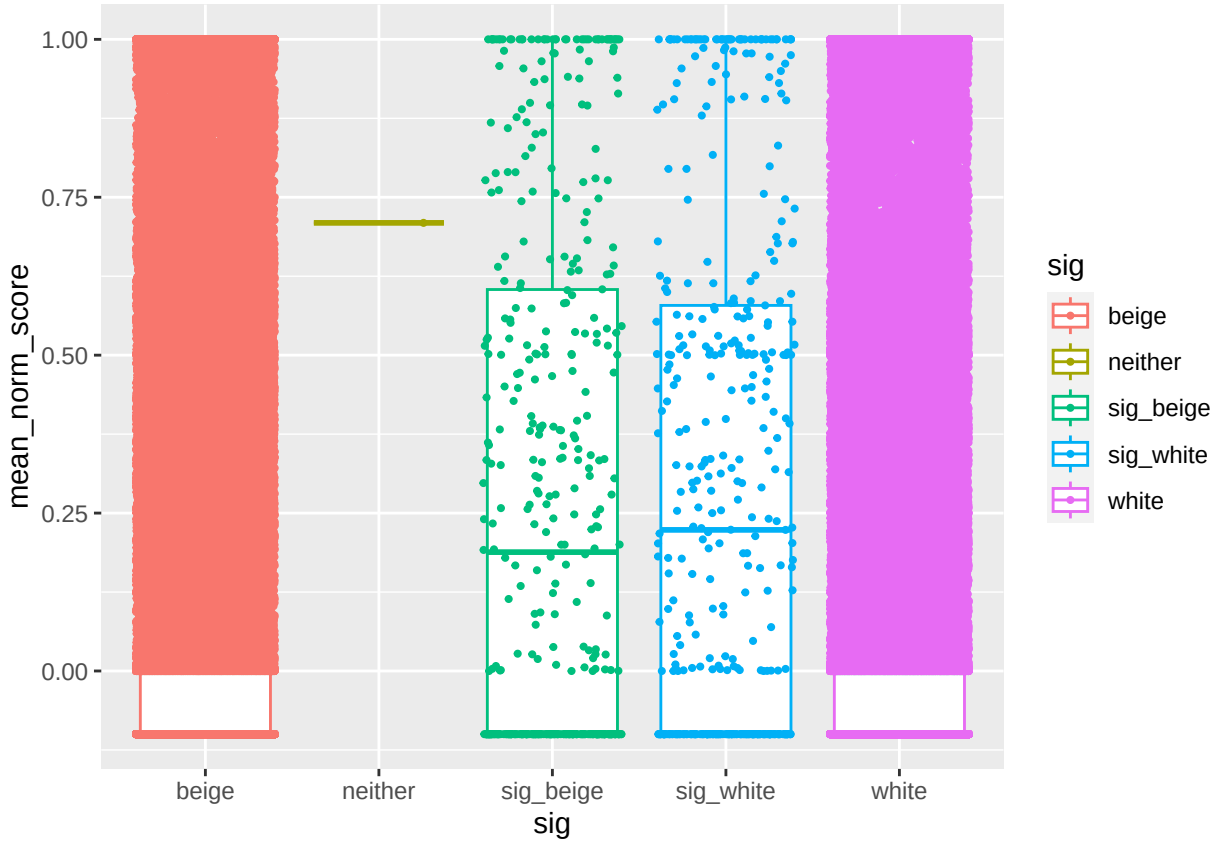
```
##
##     beige   neither sig_beige sig_white     white
##     77862         1       434       433     82144
```

```
ggplot(all_introns, aes(x=sig, y=mean_norm_score, fill=sig)) + geom_violin() +
    geom_jitter(size=0.75)
```

```
## Warning: Groups with fewer than two data points have been dropped.
```




```
ggplot(all_introns, aes(x= sig,  y=mean_norm_score,colour=sig)) + geom_boxplot() +
    geom_jitter(size=0.75)
```



The non-averaged score would be appropriate if the number of transcripts per intron was similar between the groups. Trifid scores (more approprate if we're comparing across gene instead of pairwise between within each gene); has the average score HIGHER in white than beige. also a nice validation of the psi threshold.

```
write.table(all_introns, here("31_leafcutter/trifid_all_introns.tsv"), sep="\t",quote=F, row.names = F)
```

Select significant introns + alt introns

```
alt_introns = read.delim(here("31_leafcutter/alt_introns_195.tsv"))
alt_introns = unite(alt_introns, "intron_coords", chr, start, end, strand, sep = ":")
sig_junctions = filter(all_introns, (p.adjust < 0.05 & abs(deltapsi) > 0.1 ) |
                                    intron_coords %in% alt_introns$intron_coords)
nrow(sig_junctions)
```

```
## [1] 1100
```

```
table(sig_junctions$sig)
```

```
##
##     beige sig_beige sig_white     white
##       117       434       433       116
```

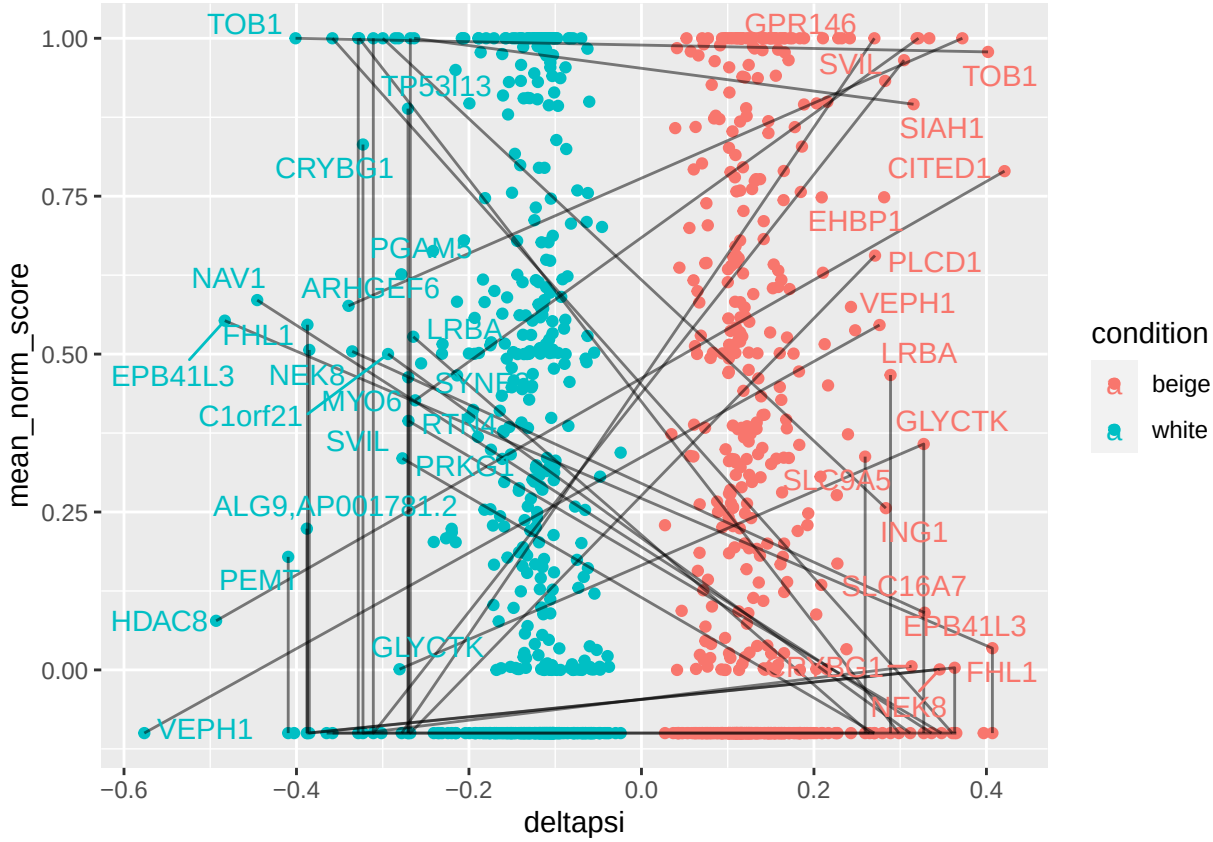
```
write.table(sig_junctions, here("31_leafcutter/trifid_with_alt_introns.tsv"), sep="\t",quote=F, row.name
```

Plot the TRIFID difference

```
ggplot(sig_junctions, aes(colour=condition, y=mean_norm_score, x=deltapsi)) + geom_jitter() +
    geom_line(data=filter(sig_junctions, abs(deltapsi) > 0.25), aes(group=cluster_id), colour="black", a
    geom_text_repel(data=filter(sig_junctions, abs(deltapsi) > 0.25), aes(label=gene))
```

```
## Warning: Removed 4 rows containing missing values (‘geom_text_repel()‘).
```

```
## Warning: ggrepel: 42 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



Calculate the TRIFID difference

```
#if three introns are signifciant for a cluster
#average TRIFID score between the two introns in the same direction

paste_uq = function(x){
    return = paste(unique(x), collapse=",")
```

```
}
mean_of_3s = filter(sig_junctions, in_trifid != "cluster_not_in_trifid") %>%
    group_by( cluster_id, condition, p.adjust) %>% summarise(mean_norm_score = mean(mean_norm_score),
                                                             deltapsi = mean(deltapsi),
                                                             gene = paste(unique(gene), collapse=","
                                                             transcript_names= paste(unique(transcri
```

```
## `summarise()` has grouped output by 'cluster_id', 'condition'. You can override
## using the `.groups` argument.
```

```
head(mean_of_3s)
```

```
## # A tibble: 6 x 7
## # Groups:   cluster_id, condition [6]
##   cluster_id  condition p.adjust mean_norm_score deltapsi gene  transcript_names
##   <chr>       <chr>        <dbl>           <dbl>    <dbl> <chr> <chr>
## 1 clu_10181_- beige     9.38e-15          0.206    0.0620 GOLG~ GOLGA2-214,GOLG~
## 2 clu_10181_- white     9.38e-15          0       -0.113  GOLG~ GOLGA2-203
## 3 clu_10209_- beige     2.29e- 8          0.502    0.107  CRAT  CRAT-201,CRAT-2~
## 4 clu_10209_- white     2.29e- 8          0.0049  -0.0484 CRAT  CRAT-207
## 5 clu_10638_+ beige     1.95e- 2          0.226    0.125  SSPN  SSPN-202,NA
## 6 clu_10638_+ white     1.95e- 2          0.505   -0.130  SSPN  SSPN-201,SSPN-2~
```

```
filter(mean_of_3s, is.na(deltapsi))
```

```
## # A tibble: 0 x 7
## # Groups:   cluster_id, condition [0]
## # i 7 variables: cluster_id <chr>, condition <chr>, p.adjust <dbl>,
## #   mean_norm_score <dbl>, deltapsi <dbl>, gene <chr>, transcript_names <chr>
```

```
trifid_diff = pivot_wider(mean_of_3s, names_from = condition, values_from = c(deltapsi, mean_norm_score
                          id_cols=c("cluster_id", "p.adjust"))
head(trifid_diff)
```

```
## # A tibble: 6 x 8
## # Groups:   cluster_id [6]
##   cluster_id  p.adjust deltapsi_beige deltapsi_white mean_norm_score_beige
##   <chr>          <dbl>          <dbl>          <dbl>                 <dbl>
## 1 clu_10181_- 9.38e-15         0.0620        -0.113                 0.206
## 2 clu_10209_- 2.29e- 8         0.107         -0.0484                0.502
## 3 clu_10638_+ 1.95e- 2         0.125         -0.130                 0.226
## 4 clu_10654_+ 1.10e- 7         0.160         -0.160                -0.1
## 5 clu_10672_+ 2.28e-63         0.232         -0.328                 1
## 6 clu_10690_+ 1.03e- 6         0.114         -0.121                 0.0621
## # i 3 more variables: mean_norm_score_white <dbl>, gene_beige <chr>,
## #   gene_white <chr>
```

```
## If your other intron is not in trifid speak now
## deltapsi we should have from the other table, just the trifid score we could set to -1
filter(trifid_diff, is.na(deltapsi_beige ) | is.na(deltapsi_white )) # just 28 with an intron pair miss
```

```
## # A tibble: 2 x 8
## # Groups:   cluster_id [2]
##   cluster_id  p.adjust deltapsi_beige deltapsi_white mean_norm_score_beige
##   <chr>          <dbl>          <dbl>          <dbl>                 <dbl>
## 1 clu_18299_+ 5.75e-29         NA             -0.101                 NA
## 2 clu_30508_+ 3.91e-34          0.116         NA                     -0.1
## # i 3 more variables: mean_norm_score_white <dbl>, gene_beige <chr>,
## #   gene_white <chr>
```

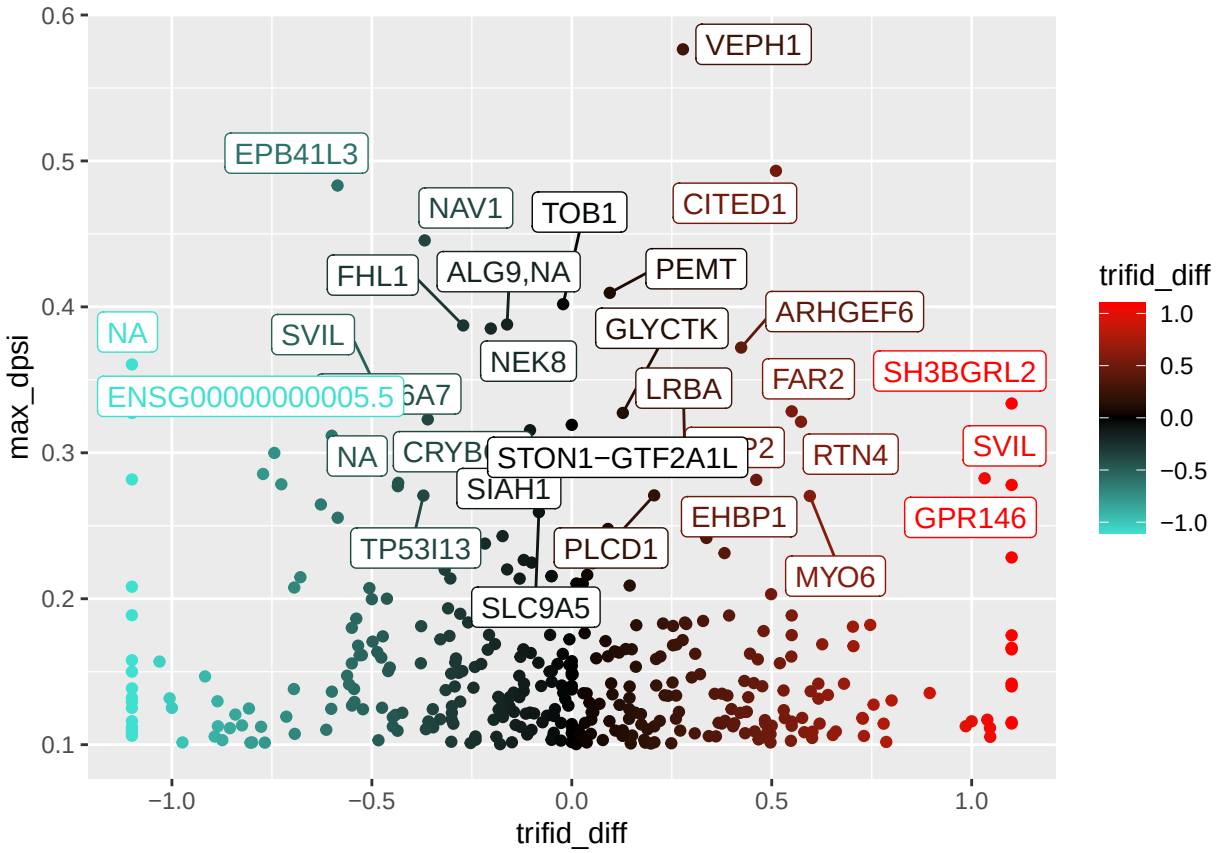
```
trifid_diff = mutate(trifid_diff, trifid_diff = mean_norm_score_beige-mean_norm_score_white, max_dpsi =
head(trifid_diff)
```

```
## # A tibble: 6 x 10
## # Groups:   cluster_id [6]
##   cluster_id  p.adjust deltapsi_beige deltapsi_white mean_norm_score_beige
##   <chr>          <dbl>          <dbl>          <dbl>                 <dbl>
## 1 clu_10181_- 9.38e-15         0.0620        -0.113                 0.206
## 2 clu_10209_- 2.29e- 8         0.107         -0.0484                0.502
## 3 clu_10638_+ 1.95e- 2         0.125         -0.130                 0.226
## 4 clu_10654_+ 1.10e- 7         0.160         -0.160                -0.1
## 5 clu_10672_+ 2.28e-63         0.232         -0.328                 1
## 6 clu_10690_+ 1.03e- 6         0.114         -0.121                 0.0621
## # i 5 more variables: mean_norm_score_white <dbl>, gene_beige <chr>,
## #   gene_white <chr>, trifid_diff <dbl>, max_dpsi <dbl>
```

```
figs = here("R/plots")
ggplot(trifid_diff, aes(y=max_dpsi, x=trifid_diff, colour=trifid_diff)) + geom_point() +
    geom_label_repel(data=filter(trifid_diff, max_dpsi > 0.25), aes(label=gene_beige)) +
    scale_color_gradient2(low="turquoise", mid="black", high="red")
```

```
## Warning: Removed 2 rows containing missing values (‘geom_point()‘).
```

```
## Warning: ggrepel: 8 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```




```
ggsave(file.path(figs, "trifid_difference_v_dpsi.pdf"))
```

```
## Saving 6.5 x 4.5 in image
```

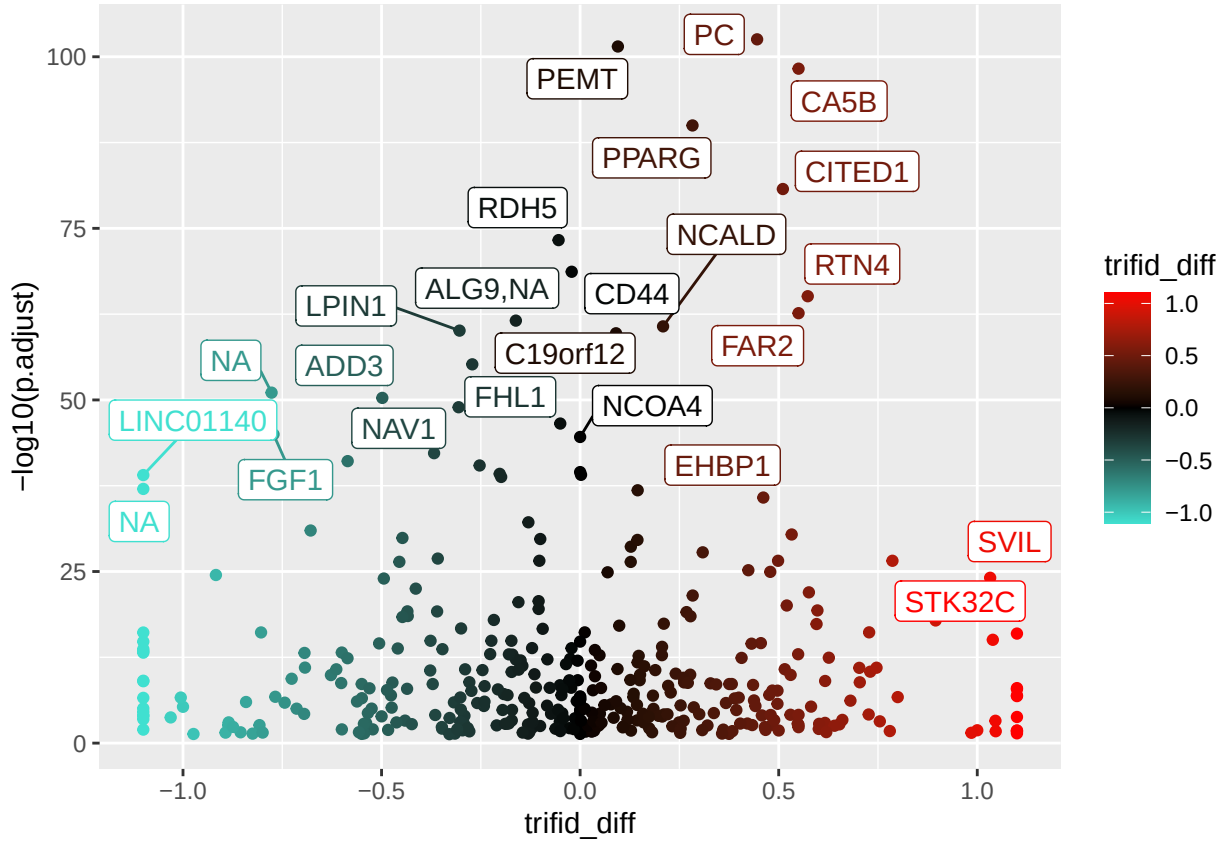
```
## Warning: Removed 2 rows containing missing values (‘geom_point()‘).
## ggrepel: 8 unlabeled data points (too many overlaps). Consider increasing max.overlaps
```

```
ggplot(trifid_diff, aes(y=-log10(p.adjust), x=trifid_diff, colour=trifid_diff)) + geom_point() +
    geom_label_repel(data=filter(trifid_diff, p.adjust < 0.01), aes(label=gene_beige)) +
    scale_color_gradient2(low="turquoise", mid="black", high="red")
```

```
## Warning: Removed 2 rows containing missing values (‘geom_point()‘).
```

```
## Warning: Removed 2 rows containing missing values (‘geom_label_repel()‘).
```

```
## Warning: ggrepel: 317 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```




```
ggsave(file.path(figs, "trifid_difference_v_pvalue.pdf"))
```

```
## Saving 6.5 x 4.5 in image
```

```
## Warning: Removed 2 rows containing missing values (‘geom_point()‘).
```

```
## Warning: Removed 2 rows containing missing values (‘geom_label_repel()‘).
```

```
## Warning: ggrepel: 317 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```

Hmm.... CITED1 really shouldn't be here because we're comparing across genes which doesn't necessarily make sense.

```
write.table(trifid_diff, here("31_leafcutter/trifid_DIFFERENCE_with_Alt_introns.tsv"), sep="\t",quote=F
```

Actin changes

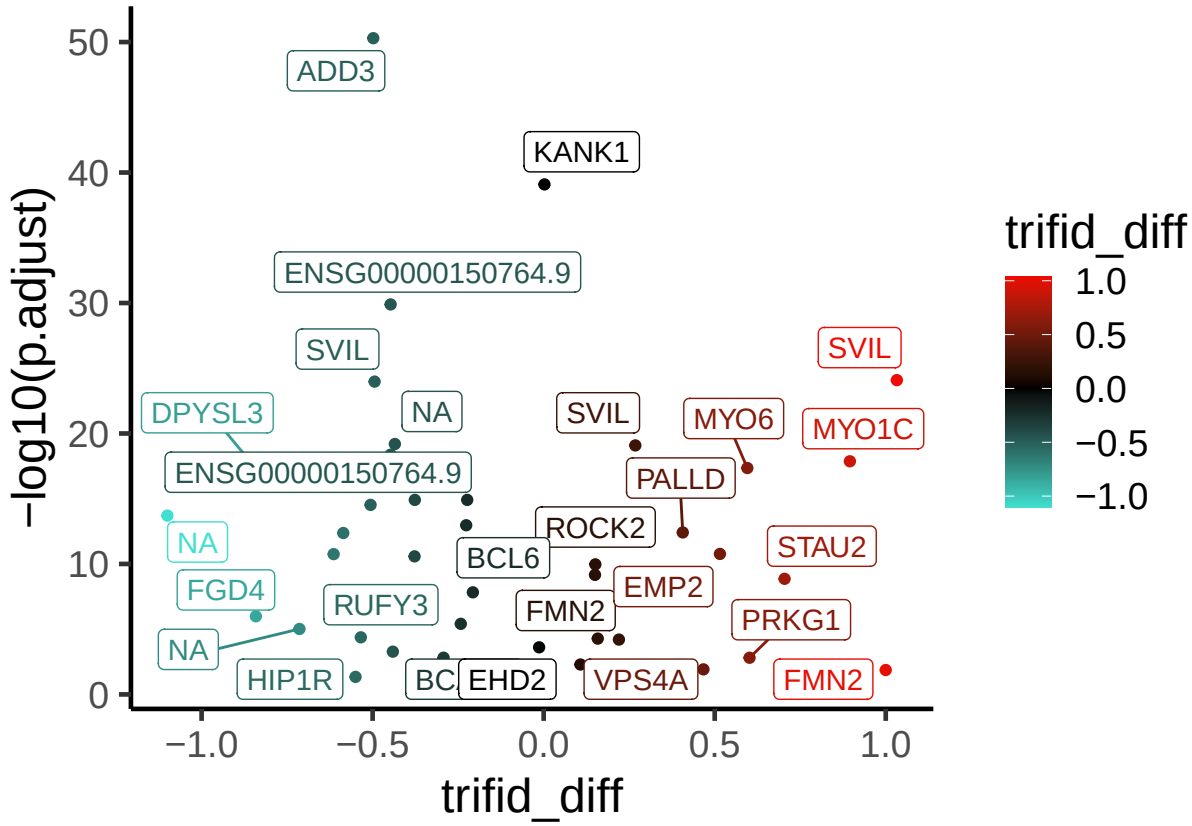
```
from_go_term = "ADD3/KANK1/DIXDC1/SVIL/PRKG1/MYO1C/INPP5K/MYO6/DPYSL3/SH3KBP1/ARHGEF10L/ARHGEF2/IQSEC1/

actin = stringr::str_split_1(from_go_term, "/")
str(actin)
```

```
##  chr [1:38] "ADD3" "KANK1" "DIXDC1" "SVIL" "PRKG1" "MYO1C" "INPP5K" "MYO6" ...
```

```
ggplot(filter(trifid_diff, gene_beige %in% actin | gene_white %in% actin),
       aes(y=-log10(p.adjust), x=trifid_diff, colour=trifid_diff)) + geom_point() +
    geom_label_repel( aes(label=gene_beige)) +
    scale_color_gradient2(low="turquoise", mid="black", high="red") + theme_classic(base_size=18)
```

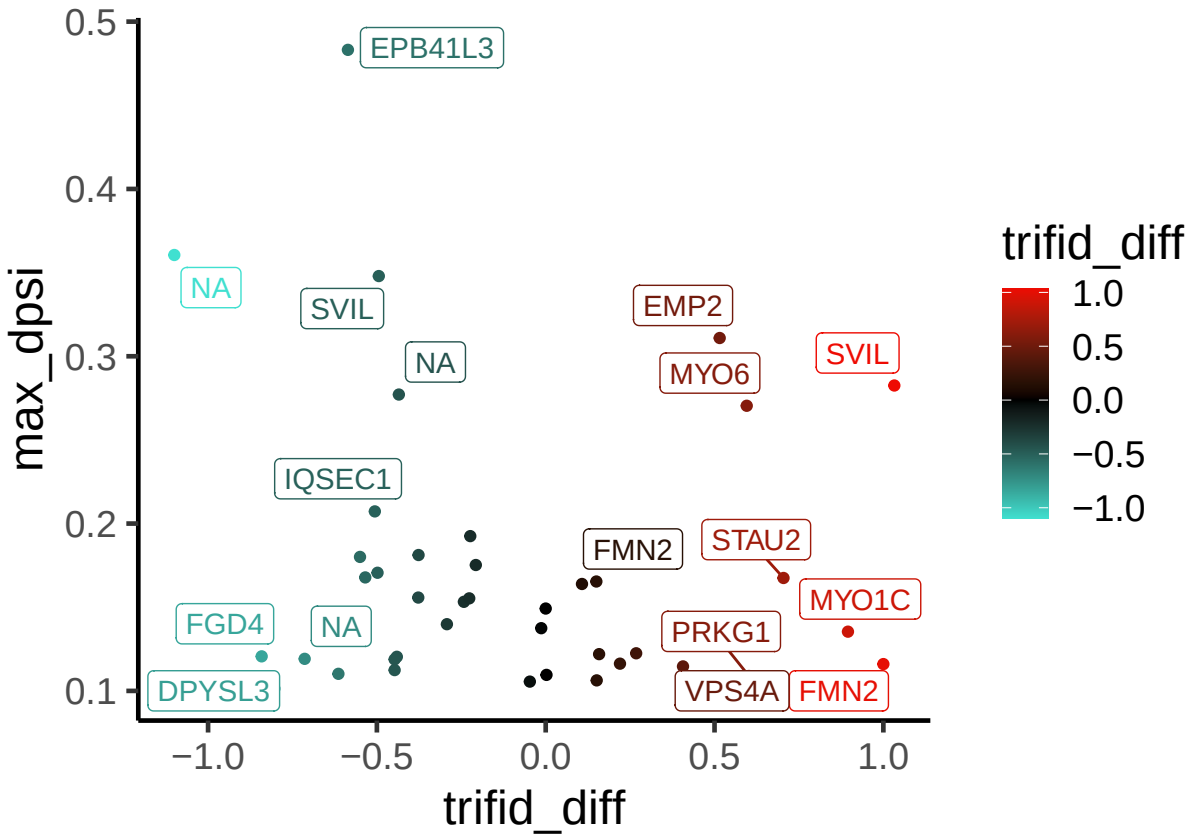
```
## Warning: ggrepel: 14 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```




```
ggsave(file.path(figs, "actin_trifid_scores.pdf"), width = 10, height = 10)

ggplot(filter(trifid_diff, gene_beige %in% actin | gene_white %in% actin),
       aes(y=max_dpsi, x=trifid_diff, colour=trifid_diff)) + geom_point() +
    geom_label_repel( aes(label=gene_beige)) +
    scale_color_gradient2(low="turquoise", mid="black", high="red") + theme_classic(base_size=18)
```

```
## Warning: ggrepel: 24 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```




```
ggsave(file.path(figs, "actin_trifid_scores_dpsi.pdf"), width = 10, height = 10)
```