get_splice_sites

2023-10-31

Supplementary Figure 2

For introns not found in 3 db (gencode, refseq and fantom_cat), classify cryptic splice sites.

Requires: leafviz processed object which contains classifiations under the column "verdict"

```
library(tidyr)
library(dplyr)
library(ggplot2)
library(here); i_am("R/19_get_cryptic_splice_sites.Rmd")
bed_cols = c("chr","start","end","genes","deltapsi","strand")
```

```
junctions= read.delim(here("31_leafcutter","three_database_info_sig_junctions.tsv"), header=T)
nrow(junctions)
```

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

```
head(junctions)
```

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

```
## 3                 1                  1    CRAT        protein_coding
## 4                 1                  1    SSPN        protein_coding
## 5                 1                  1    SSPN protein_coding,lncRNA
## 6                 2                  3  ARNTL2        protein_coding
##   genes_in_cluster is_first_intron condition  conditions num_introns
## 1          DENND1A            TRUE     white       white           1
## 2           GOLGA2           FALSE     white       white           1
## 3             CRAT            TRUE     beige       beige           1
## 4             SSPN            TRUE     white beige&white           2
## 5             SSPN            TRUE     beige beige&white           2
## 6           ARNTL2           FALSE     white beige&white           2
```

Clip splice site regions

```
load(here("31_leafcutter/leafviz.RData"))
head(introns) #<- we only need this table
```

```
##          clusterID gene          ensemblID   chr    start      end
## 112697 clu_35616_-    . ENSG00000133027.18  chr7 43749288 43750147
## 39234  clu_19605_- PEMT ENSG00000133027.18 chr17 17577027 17577107
## 39239  clu_19605_- PEMT ENSG00000173599.15 chr17 17577027 17591531
## 7517    clu_1700_-   PC ENSG00000169239.13 chr11 66872159 66907821
## 128912 clu_15162_+ CA5B ENSG00000169239.13  chrX 15675778 15688661
## 128916 clu_15162_+ CA5B ENSG00000132170.21  chrX 15738352 15749971
##                  verdict deltapsi
## 112697    unknown_strand    0.117
## 39234  cryptic_fiveprime    0.192
## 39239          annotated   -0.430
## 7517           annotated    0.159
## 128912         annotated   -0.147
## 128916         annotated    0.127
##                                                                                                                transcripts
## 112697                                                                                                                  NA
## 39234                                                                                                                    .
## 39239  ENST00000255389.10+ENST00000395781.6+ENST00000421096.5+ENST00000461404.1+ENST00000580147.5
## 7517                                                                                                  ENST00000393955.6
## 128912                                                                                 ENST00000380333.5+ENST00000448692.5
## 128916                                  ENST00000318636.8+ENST00000380319.2+ENST00000474624.5+ENST00000478923.1
```

```
nrow(introns)# prefiltered for p.adjust and deltapsi
```

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

```
table(junctions$annotation)
```

```
##
##    cryptic fantom_cat    gencode     refseq
##         54         43        647         33
```

```
with_verdict = merge(junctions, introns[c("clusterID","chr","start","end","verdict")],
                     by.x=c("cluster_id","chr","start","end"), by.y=c("clusterID","chr","start","end"))
table(with_verdict$verdict)
```

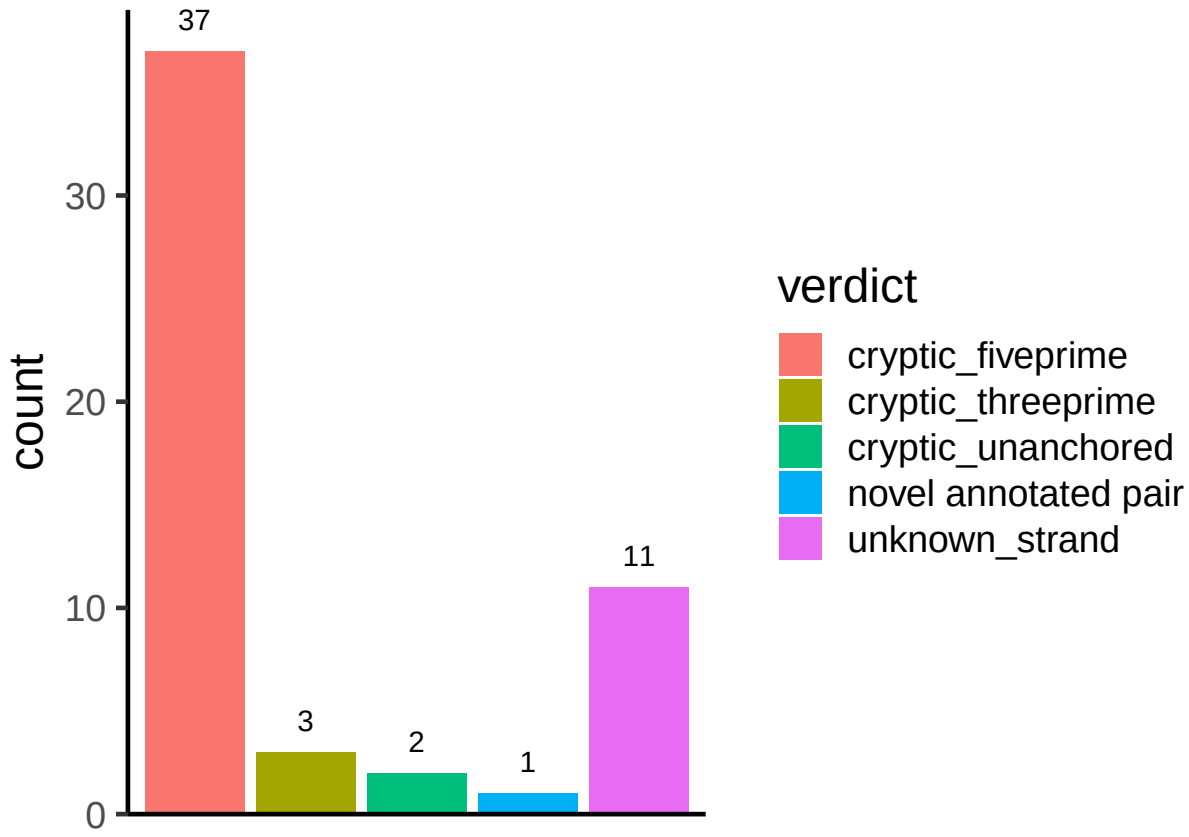
```
##
##            annotated    cryptic_fiveprime   cryptic_threeprime
##                  647                   82                   12
##   cryptic_unanchored novel annotated pair       unknown_strand
##                    4                   11                   21
```

```
table(with_verdict$verdict, with_verdict$annotation)
```

```
##
##                        cryptic fantom_cat gencode refseq
##   annotated                  0          0     647      0
##   cryptic_fiveprime         37         25       0     20
##   cryptic_threeprime         3          5       0      4
##   cryptic_unanchored         2          2       0      0
##   novel annotated pair       1          5       0      5
##   unknown_strand            11          6       0      4
```

```
ggplot(filter(with_verdict,annotation=="cryptic"), aes(x=verdict, fill=verdict)) + geom_bar()+
     theme_classic(base_size=18) + theme(axis.text.x = element_blank(), axis.ticks.x = element_blank(),
                                         panel.border = element_blank(), axis.title.x = element_blank())
  geom_text(aes(label = after_stat(count), group = annotation),
    stat="count", vjust = -1) + scale_y_continuous(expand = c(0, 0), limits = c(0, 39))
```

```
## Warning: The following aesthetics were dropped during statistical transformation: fill
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a ‘group‘ aesthetic or to convert a numerical
##   variable into a factor?
```




```
ggsave(here("R/plots/cryptic_splice_site_annotation.pdf"))
```

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

```
## Warning: The following aesthetics were dropped during statistical transformation: fill
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a 'group' aesthetic or to convert a numerical
##   variable into a factor?
```

Make intron coord bed

not exon to exon coords (as leafcutter provides), but interior intron coords. Also need to convert from gtf (1-based) to bed (0-based) , which means subtracting 1 from the upstream position.

```
sig_introns = select(with_verdict, chr, start, end, gene, deltapsi, strand, p.adjust, annotation, transc
  mutate( start = start, end = end - 1)

head(sig_introns)
```

```
##     chr     start       end    gene   deltapsi strand     p.adjust annotation
## 1  chr9 123401912 123403401 DENND1A -0.1076770      - 1.437183e-02    gencode
## 2  chr9 128266325 128267457  GOLGA2 -0.1130214      - 8.760649e-15    gencode
```

```
## 3  chr9 129108077 129110482    CRAT  0.1074815      - 2.104379e-08    gencode
## 4 chr12  26195531  26224292    SSPN  0.1056313      + 1.870231e-02    gencode
## 5 chr12  26195951  26224292    SSPN -0.1222677      + 1.870231e-02    gencode
## 6 chr12  27380404  27385480  ARNTL2 -0.1918059      + 1.084347e-07    gencode
##
## 1                                                                                          ENST0000
## 2
## 3
## 4
## 5
## 6 ENST00000395901.6,ENST00000542388.1,ENST00000311001.9,ENST00000261178.9,ENST00000457040.6,ENST0000
##     verdict
## 1 annotated
## 2 annotated
## 3 annotated
## 4 annotated
## 5 annotated
## 6 annotated
```

```r
nrow(sig_introns)
```

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

```r
write.table(sig_introns, here("31_leafcutter/sig_leafcutter_interior_intron_coords.bed"),
            row.names = F, col.names = F , quote=F, sep="\t")
```

```r
cryptic_introns = filter(sig_introns, annotation == "cryptic")
nrow(cryptic_introns)
```

```
## [1] 54
```

```r
five_prime_flanks = mutate(cryptic_introns, start = if_else(strand == "+", start, end - 5),
                                                        end= if_else(strand=="+", start+5, end),
                           annotation = if_else(verdict %in% c("cryptic_fiveprime", "cryptic_unanchored
                                                  "cryptic_5'",
                                                  "5'"),
                           gene = paste(annotation, gene, sep="_"))
nrow(five_prime_flanks)
```

```
## [1] 54
```

```r
#head(five_prime_flanks)

three_prime_flanks = mutate(cryptic_introns, start = if_else(strand == "+", end-3, start),
                                                        end= if_else(strand=="+", end, start +3),
                            annotation = if_else(verdict %in% c("cryptic_threeprime", "cryptic_unanchor
                                                  "cryptic_3'",
                                                  "3'"),
                            gene = paste(annotation, gene, sep="_"))
#head(three_prime_flanks)

flanks = rbind(five_prime_flanks, three_prime_flanks)
head(flanks)
```

```
##     chr     start       end          gene   deltapsi strand     p.adjust
## 1 chr12 132710160 132710165 cryptic_5'_NA  0.1875376      + 4.068023e-10
## 2  chr1  14856060  14856065 cryptic_5'_NA  0.1805027      + 5.387364e-13
## 3  chr1 120233347 120233352 cryptic_5'_NA  0.1619583      + 4.224799e-03
## 4  chr1 144720690 144720695 cryptic_5'_NA  0.1064807      + 1.826322e-02
## 5  chr1 144721138 144721143 cryptic_5'_NA -0.1185484      + 1.826322e-02
## 6  chr1 240404659 240404664 cryptic_5'_NA  0.1274724      + 1.025979e-10
##   annotation transcript_ids           verdict
## 1 cryptic_5'        Unknown  cryptic_fiveprime
## 2 cryptic_5'        Unknown cryptic_unanchored
## 3 cryptic_5'        Unknown  cryptic_fiveprime
## 4 cryptic_5'        Unknown     unknown_strand
## 5 cryptic_5'        Unknown     unknown_strand
## 6 cryptic_5'        Unknown  cryptic_fiveprime
```

```
flanks = filter(flanks, grepl("cryptic", annotation))
nrow(flanks)
```

```
## [1] 68
```

```
table(flanks$verdict)
```

```
##
##    cryptic_fiveprime   cryptic_threeprime   cryptic_unanchored
##                   37                    3                    4
## novel annotated pair       unknown_strand
##                    2                   22
```

```
write.table(flanks, here("31_leafcutter/leafcutter_cryptic_splice_sites_5bp_3bp.bed"), sep="\t", quote =
```