lc_introns_to_TSSs_three_database

Caveat: we can only assign 1st introns if the they are annotated to a transcript - here we add refseq and fantom cat annotations to the ensembl one.
Also adding in ALL introns to get an accurate background

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(ggplot2)
library(readr)
library(here)
```

```
## here() starts at /projects/imb-pkbphil/sp/rnaseq/six_donor_trans/splicing_paper
```

```
here::i_am("R/13_lc_introns_to_TSSs_three_database.Rmd")
```

```
## here() starts at /projects/imb-pkbphil/sp/rnaseq/six_donor_trans/splicing_paper
```

```
LEAF_INDEX = here('annotations')
leaf_junctions = read.delim(here("31_leafcutter","three_database_info_all_junctions.tsv"))
dim(leaf_junctions)
```

```
## [1] 132587     15
```

```
head(leaf_junctions)
```

```
##   annotation  chr    start      end strand  cluster_id      deltapsi
## 1    gencode chr7 43648652 43650493      - clu_35616_- -0.0141028170
## 2    gencode chr7 43648652 43650612      - clu_35616_- -0.0408035987
## 3    gencode chr7 43648652 43665658      - clu_35616_- -0.0009734716
## 4    gencode chr7 43648652 43711400      - clu_35616_- -0.0003668545
```

```
## 5    gencode chr7 43648652 43729429      - clu_35616_- -0.0681581659
## 6    gencode chr7 43650712 43656033      - clu_35616_- -0.0034620531
##        p.adjust
## 1 2.192287e-123
## 2 2.192287e-123
## 3 2.192287e-123
## 4 2.192287e-123
## 5 2.192287e-123
## 6 2.192287e-123
##                                                                                                 trans
## 1                                                                                           ENST00000
## 2 ENST00000446564.5,ENST00000448704.5,ENST00000451651.5,ENST00000418140.5,ENST00000431651.5,ENST0000
## 3                                                                                            ENST0000
## 4                                                                                            ENST0000
## 5                 ENST00000457939.1,ENST00000420441.1,ENST00000415076.6,ENST00000223336.10,ENST0000
## 6                                                                                            ENST0000
##   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 genes_in_cluster
## 1                              protein_coding             COA1
## 2 nonsense_mediated_decay,protein_coding,lncRNA             COA1
## 3                     nonsense_mediated_decay             COA1
## 4                              protein_coding             COA1
## 5        nonsense_mediated_decay,protein_coding             COA1
## 6                              protein_coding             COA1
##   is_first_intron
## 1           FALSE
## 2           FALSE
## 3           FALSE
## 4           FALSE
## 5            TRUE
## 6           FALSE
```

```
leaf_junctions = mutate(leaf_junctions, pot_first_intron = min_intron_number == 1, intron_id = paste(ch
leaf_junctions$pot_first_intron = factor(leaf_junctions$pot_first_intron, levels=c("TRUE","FALSE"))
leaf_junctions$annotation = factor(leaf_junctions$annotation, levels=c("gencode","refseq","fantom_cat")

length(unique(leaf_junctions$intron_id))
```

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

```
sig = filter(leaf_junctions, p.adjust < 0.05 & abs(deltapsi)> 0.1)
nrow(sig); #length(unique(sig$intron_id))
```

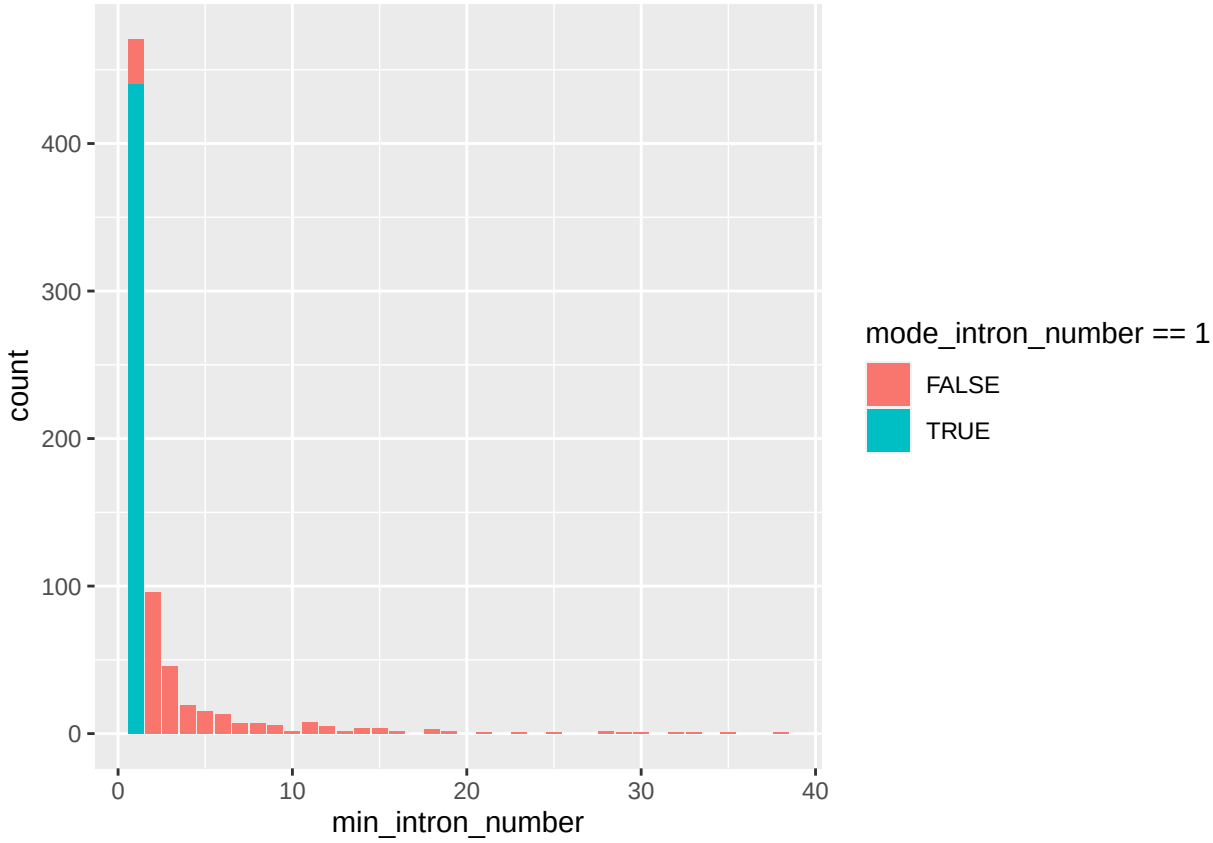
```
## [1] 777
```

```
summary(sig$min_intron_number)
```

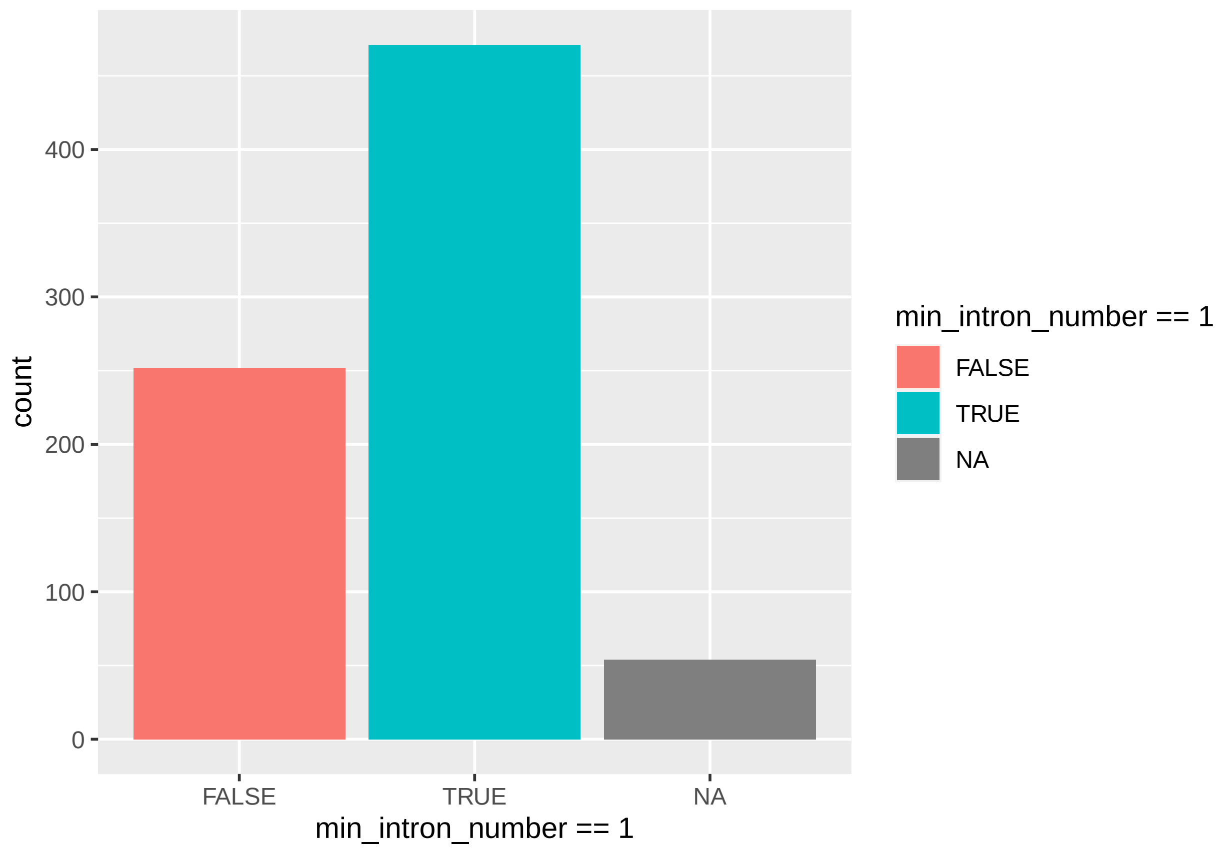
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00    1.00    1.00    2.69    2.00   38.00      54
```

```
ggplot(sig, aes(x=min_intron_number, fill=mode_intron_number==1)) + geom_bar()
```

```
## Warning: Removed 54 rows containing non-finite values (`stat_count()`).
```




```
ggplot(sig, aes(x=min_intron_number==1, fill=min_intron_number==1)) + geom_bar()
```




```
head(leaf_junctions)
```

```
##   annotation  chr    start      end strand  cluster_id      deltapsi
## 1    gencode chr7 43648652 43650493      - clu_35616_- -0.0141028170
## 2    gencode chr7 43648652 43650612      - clu_35616_- -0.0408035987
## 3    gencode chr7 43648652 43665658      - clu_35616_- -0.0009734716
## 4    gencode chr7 43648652 43711400      - clu_35616_- -0.0003668545
## 5    gencode chr7 43648652 43729429      - clu_35616_- -0.0681581659
## 6    gencode chr7 43650712 43656033      - clu_35616_- -0.0034620531
##        p.adjust
## 1 2.192287e-123
## 2 2.192287e-123
## 3 2.192287e-123
## 4 2.192287e-123
## 5 2.192287e-123
## 6 2.192287e-123
##                                                                                                 trans
## 1                                                                                             ENST00000
## 2 ENST00000446564.5,ENST00000448704.5,ENST00000451651.5,ENST00000418140.5,ENST00000431651.5,ENST0000
## 3                                                                                              ENST0000
## 4                                                                                              ENST0000
## 5                 ENST00000457939.1,ENST00000420441.1,ENST00000415076.6,ENST00000223336.10,ENST0000
## 6                                                                                              ENST0000
##   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 genes_in_cluster
## 1                            protein_coding             COA1
## 2 nonsense_mediated_decay,protein_coding,lncRNA             COA1
## 3                   nonsense_mediated_decay             COA1
## 4                            protein_coding             COA1
## 5        nonsense_mediated_decay,protein_coding             COA1
## 6                            protein_coding             COA1
##   is_first_intron pot_first_intron                          intron_id
## 1           FALSE            FALSE chr7:43648652:43650493:clu_35616_-
## 2           FALSE            FALSE chr7:43648652:43650612:clu_35616_-
## 3           FALSE            FALSE chr7:43648652:43665658:clu_35616_-
## 4           FALSE            FALSE chr7:43648652:43711400:clu_35616_-
## 5            TRUE             TRUE chr7:43648652:43729429:clu_35616_-
## 6           FALSE            FALSE chr7:43650712:43656033:clu_35616_-
```

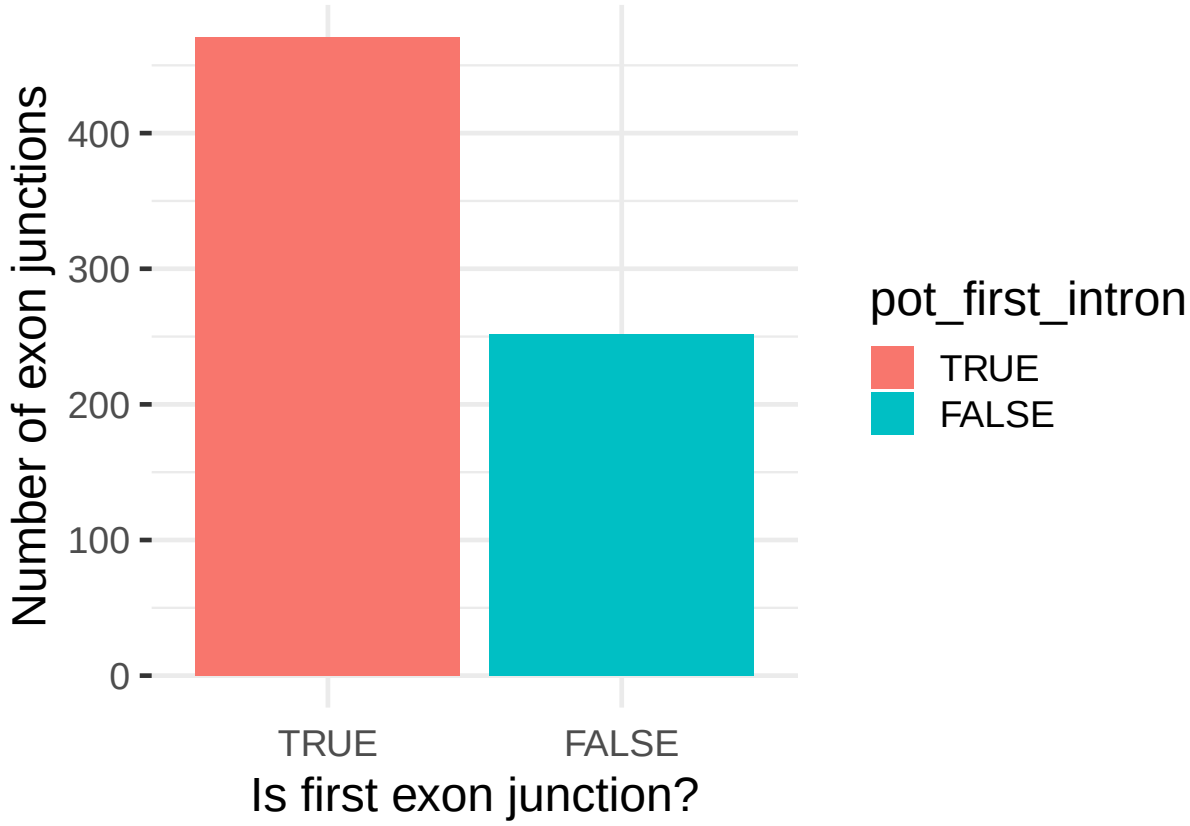
```
summary(sig$pot_first_intron)
```

```
##  TRUE FALSE  NA's
##   471   252    54
```

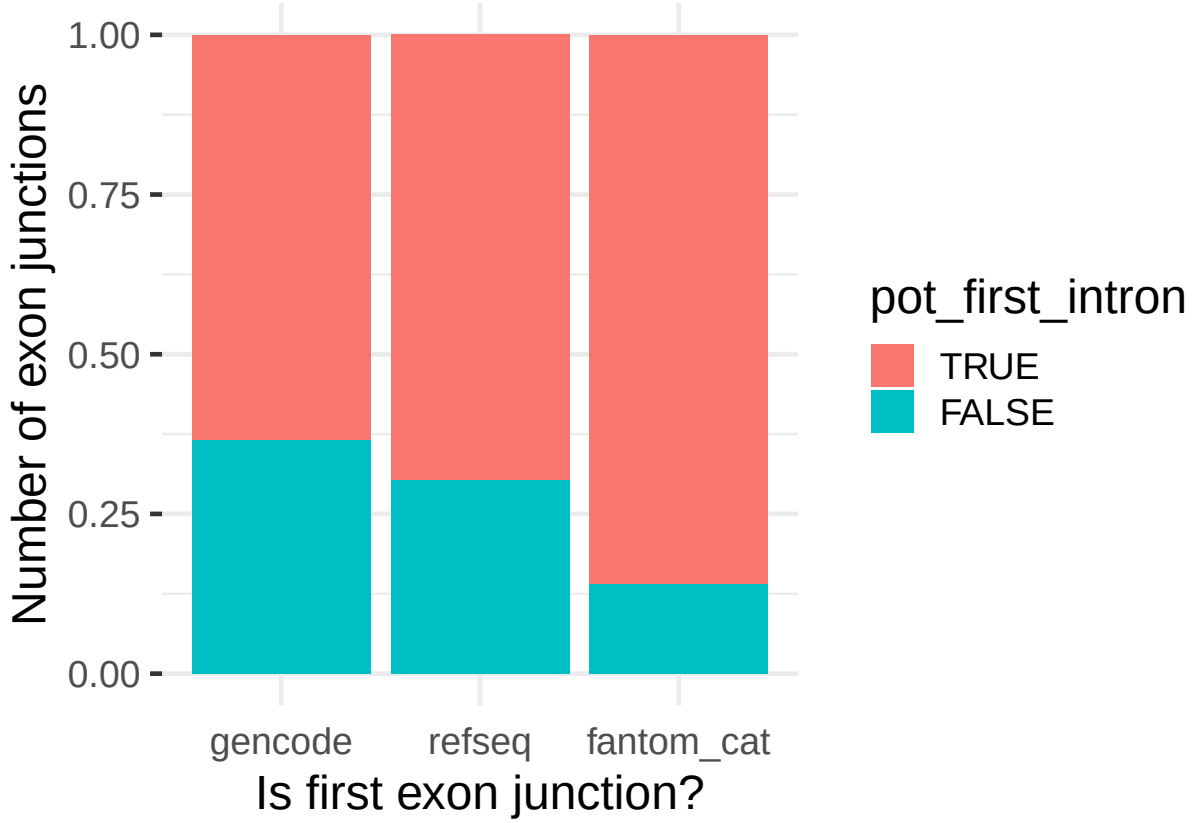
```
table(sig$pot_first_intron, sig$annotation)
```

```
##
##         gencode refseq fantom_cat
##   TRUE      411     23         37
##   FALSE     236     10          6
```

```
ggplot(filter(sig, !is.na(pot_first_intron) ),aes(x=pot_first_intron, fill=pot_first_intron)) + geom_ba
 labs(x="Is first exon junction?", y="Number of exon junctions") +
        theme_bw(base_size=18) + theme(axis.ticks.x = element_blank(), panel.border = element_blank())
```




```
ggplot(filter(sig, !is.na(pot_first_intron) ),aes(x=annotation, fill=pot_first_intron)) + geom_bar(posi
  labs(x="Is first exon junction?", y="Number of exon junctions") +
         theme_bw(base_size=18) + theme(axis.ticks.x = element_blank(), panel.border = element_blank())
```




```
ggsave(here("R/plots", "annotated_first_intron_barplot.pdf"))
```

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

Get 1st exon coords for 1st introns

per intron, could be several 1st exons/TSSs. but thats okay, just log them here

```
first_introns = filter(leaf_junctions, pot_first_intron==TRUE) %>% rename(intron_start = start, intron_
nrow(first_introns)
```

```
## [1] 42486
```

```
nrow(filter(first_introns, p.adjust < 0.05 & abs(deltapsi)>0.1))#471 1st introns
```

```
## [1] 471
```

```
head(first_introns)
```

```
##   annotation   chr intron_start intron_end strand  cluster_id     deltapsi
## 1    gencode  chr7     43648652   43729429      - clu_35616_- -0.068158166
## 2    gencode  chr7     43650712   43729429      - clu_35616_- -0.072381672
```

```
## 3    gencode  chr7     43665722   43729429      - clu_35616_- -0.040087563
## 4    gencode  chr7     43711449   43729429      - clu_35616_- -0.001227714
## 5    gencode chr17     17577027   17577414      - clu_19605_-  0.071012350
## 6    gencode chr17     17577027   17582267      - clu_19605_-  0.079852522
##        p.adjust
## 1 2.192287e-123
## 2 2.192287e-123
## 3 2.192287e-123
## 4 2.192287e-123
## 5 4.360252e-104
## 6 4.360252e-104
##                                                                                                 trans
## 1                  ENST00000457939.1,ENST00000420441.1,ENST00000415076.6,ENST00000223336.10,ENST0000
## 2 ENST00000451651.5,ENST00000448704.5,ENST00000446564.5,ENST00000310564.10,ENST00000418140.5,ENST0000
## 3                                                                           ENST00000431651.5,ENST0000
## 4                                                                                             ENST0000
## 5                                                                                             ENST0000
## 6                                                                                             ENST0000
##   min_intron_number mode_intron_number gene
## 1                 1                  1 COA1
## 2                 1                  1 COA1
## 3                 1                  1 COA1
## 4                 1                  1 COA1
## 5                 1                  1 PEMT
## 6                 1                  1 PEMT
##                                        biotype genes_in_cluster
## 1        nonsense_mediated_decay,protein_coding             COA1
## 2 nonsense_mediated_decay,protein_coding,lncRNA             COA1
## 3        protein_coding,nonsense_mediated_decay             COA1
## 4                               protein_coding             COA1
## 5                               protein_coding             PEMT
## 6                               protein_coding             PEMT
##   is_first_intron pot_first_intron                           intron_id
## 1            TRUE             TRUE  chr7:43648652:43729429:clu_35616_-
## 2            TRUE             TRUE  chr7:43650712:43729429:clu_35616_-
## 3            TRUE             TRUE  chr7:43665722:43729429:clu_35616_-
## 4            TRUE             TRUE  chr7:43711449:43729429:clu_35616_-
## 5            TRUE             TRUE chr17:17577027:17577414:clu_19605_-
## 6            TRUE             TRUE chr17:17577027:17582267:clu_19605_-
```

```
table(first_introns$strand)
```

```
##
##     -     +
## 21269 21217
```

```
prefixes = c("gencode/gencode.v32", "refseq/refseq_2023","fantom/fantom_cat.lv3_robust")
ref_names= c("gencode","refseq","fantom")
indexes = paste0(prefixes, "_all_exons.txt.gz")
TSSes = paste0(prefixes, ".TSS.txt")
names(indexes) = ref_names
names(TSSes) = ref_names
```

```
load_index = function(filename, path){
  index = read.delim(file.path(path, filename),
                          col.names=c("chr","exon_start","exon_end","strand","gene"))
  index = distinct(index)
  print(nrow(index))
  return(index)
}

load_TSS = function(filename, path){
  tss = read.delim(file.path(path, filename), col.names = c("chr","TSS", "strand"))
  tss = distinct(tss)
  print(nrow(tss))
  return(tss)
}
#gencode = load_index(indexes[1], LEAF_INDEX)
#gTSS = load_TSS(TSSes[1], LEAF_INDEX)
```

```
intron_to_TSS = function(first_introns, index_fn, tss_fn, path=LEAF_INDEX){
  exon_index = load_index(index_fn, path)
  tss = load_TSS(tss_fn, path)


  first_introns$gene = NULL
  pos_upexons = merge(exon_index, filter(first_introns,strand =="+"),
                  by.x=c("chr", "strand", "exon_end"),
                  by.y=c("chr", "strand", "intron_start"))

  neg_upexons = merge(exon_index, filter(first_introns,strand =="-"),
                  by.x=c("chr", "strand", "exon_start"),
                  by.y=c("chr", "strand", "intron_end"))


  first_exons = rbind(filter(pos_upexons,exon_start %in% tss$TSS),
                    filter(neg_upexons, exon_end %in% tss$TSS) %>% rename(intron_end=intron_start))

  print(sprintf("all exons: %d\n only TSSes: %d",nrow(pos_upexons) + nrow(neg_upexons), nrow(first_exons
  return(first_exons)
}
```

```
gencode_TSSes = intron_to_TSS(filter(first_introns, annotation =="gencode"),
                              indexes[["gencode"]], TSSes[["gencode"]], path=LEAF_INDEX)
```

```
## [1] 639644
## [1] 204091
## [1] "all exons: 114253\n only TSSes: 98661"
```

```
nrow(filter(gencode_TSSes, p.adjust < 0.05 & abs(deltapsi) > 0.1))
```

```
## [1] 1027
```

```
filter(gencode_TSSes, p.adjust < 0.05 & abs(deltapsi) > 0.1) %>% pull(intron_id) %>% n_distinct()
```

```
## [1] 411
```

```
refseq_TSSes = intron_to_TSS(filter(first_introns, annotation =="refseq"),
                             indexes[["refseq"]], TSSes[["refseq"]], path=LEAF_INDEX)
```

```
## [1] 382847
## [1] 83607
## [1] "all exons: 5801\n only TSSes: 5120"
```

```
filter(refseq_TSSes,gene == "PEMT")
```

```
##     chr strand exon_end exon_start gene annotation intron_end  cluster_id
## 1 chr17      - 17577496   17577107 PEMT     refseq   17577027 clu_19605_-
## 2 chr17      - 17592142   17591967 PEMT     refseq   17577027 clu_19605_-
##     deltapsi      p.adjust     transcript_ids min_intron_number
## 1 0.19180926 4.360252e-104 rna-XM_006721418.5                 1
## 2 0.04230864 4.360252e-104 rna-XM_024450532.2                 1
##   mode_intron_number biotype genes_in_cluster is_first_intron pot_first_intron
## 1                  1    <NA>             PEMT            TRUE             TRUE
## 2                  1    <NA>             PEMT            TRUE             TRUE
##                           intron_id
## 1 chr17:17577027:17577107:clu_19605_-
## 2 chr17:17577027:17591967:clu_19605_-
```

```
filter(refseq_TSSes, p.adjust < 0.05 & abs(deltapsi) > 0.1) %>% pull(intron_id) %>% n_distinct()
```

```
## [1] 23
```

```
fantom_TSSes = intron_to_TSS(filter(first_introns, annotation =="fantom_cat"),
                             indexes[["fantom"]], TSSes[["fantom"]], path=LEAF_INDEX)
```

```
## [1] 945340
## [1] 196565
## [1] "all exons: 7227\n only TSSes: 5215"
```

```
filter(fantom_TSSes, p.adjust < 0.05 & abs(deltapsi) > 0.1) %>% pull(intron_id) %>% n_distinct()
```

```
## [1] 37
```

```
all_first = bind_rows(gencode_TSSes, refseq_TSSes, fantom_TSSes)
head(all_first)
```

```
##    chr strand  exon_end exon_start    gene annotation intron_end  cluster_id
## 1 chr1      + 100038316  100038095 MFSD14A    gencode  100049909 clu_13961_+
## 2 chr1      + 100133315  100133165  TRMT13    gencode  100136882 clu_13963_+
## 3 chr1      + 100133315  100133215  TRMT13    gencode  100136882 clu_13963_+
```

```
## 4 chr1      + 100133315  100133163  TRMT13    gencode  100136882 clu_13963_+
## 5 chr1      + 100133315  100133150  TRMT13    gencode  100136882 clu_13963_+
## 6 chr1      +  10033694   10032958   UBE4B    gencode   10072028 clu_13109_+
##        deltapsi     p.adjust
## 1  0.0632683516 1.077793e-05
## 2 -0.0005606123 2.972521e-01
## 3 -0.0005606123 2.972521e-01
## 4 -0.0005606123 2.972521e-01
## 5 -0.0005606123 2.972521e-01
## 6 -0.0204661503 5.951011e-01
##                                                                transcript_ids
## 1                                                              ENST00000370152.8
## 2   ENST00000370143.5,ENST00000370139.1,ENST00000370141.7,ENST00000482437.5
## 3   ENST00000370143.5,ENST00000370139.1,ENST00000370141.7,ENST00000482437.5
## 4   ENST00000370143.5,ENST00000370139.1,ENST00000370141.7,ENST00000482437.5
## 5   ENST00000370143.5,ENST00000370139.1,ENST00000370141.7,ENST00000482437.5
## 6 ENST00000672724.1,ENST00000253251.12,ENST00000377153.5,ENST00000343090.11
##   min_intron_number mode_intron_number                                biotype
## 1                 1                  1                          protein_coding
## 2                 1                  1 protein_coding,nonsense_mediated_decay
## 3                 1                  1 protein_coding,nonsense_mediated_decay
## 4                 1                  1 protein_coding,nonsense_mediated_decay
## 5                 1                  1 protein_coding,nonsense_mediated_decay
## 6                 1                  1                          protein_coding
##             genes_in_cluster is_first_intron pot_first_intron
## 1 SLC35A3,AC118553.2,MFSD14A            TRUE             TRUE
## 2                     TRMT13            TRUE             TRUE
## 3                     TRMT13            TRUE             TRUE
## 4                     TRMT13            TRUE             TRUE
## 5                     TRMT13            TRUE             TRUE
## 6                      UBE4B            TRUE             TRUE
##                             intron_id
## 1 chr1:100038316:100049909:clu_13961_+
## 2 chr1:100133315:100136882:clu_13963_+
## 3 chr1:100133315:100136882:clu_13963_+
## 4 chr1:100133315:100136882:clu_13963_+
## 5 chr1:100133315:100136882:clu_13963_+
## 6   chr1:10033694:10072028:clu_13109_+
```

```
table(all_first$annotation)
```

```
##
##    gencode     refseq fantom_cat
##      98661       5120       5215
```

```
table(filter(all_first, p.adjust < 0.05 & abs(deltapsi)>0.1) %>% pull("annotation"))
```

```
##
##    gencode     refseq fantom_cat
##       1027         24         44
```

```
table(filter(all_first, p.adjust < 0.05 & abs(deltapsi)>0.1) %>% select("annotation","intron_id") %>% di
```

```
##
##    gencode     refseq fantom_cat
##        411         23         37
```

```
rm(gencode_TSSes, refseq_TSSes, fantom_TSSes)
```

select background and save first exon files

```
all_first = select(all_first, chr, exon_start, exon_end, intron_id, deltapsi, strand, p.adjust, gene) %>
    mutate(exon_start = exon_start -1) %>% #adjustment from gtf to bed coordinates
    arrange(chr, exon_start, exon_end)
all_first = distinct(all_first)
nrow(all_first)
```

```
## [1] 108996
```

```
nrow(distinct(all_first[c("chr", "exon_start", "exon_end","strand")]))
```

```
## [1] 68805
```

```
nrow(filter(all_first, p.adjust < 0.05 & abs(deltapsi)>0.1)) #1031
```

```
## [1] 1095
```

```
nrow(distinct(filter(all_first, p.adjust < 0.05 & abs(deltapsi)>0.1) %>% select("intron_id"))) #363
```

```
## [1] 471
```

```
head(all_first)
```

```
##    chr exon_start exon_end                      intron_id      deltapsi strand
## 1 chr1      29320    29370   chr1:24891:29321:clu_27299_-  0.0005808663      -
## 2 chr1     168609   168767 chr1:168165:168610:clu_27301_-  0.0008945081      -
## 3 chr1     169048   169210 chr1:168165:169049:clu_27301_-  0.0139586116      -
## 4 chr1     169048   169240 chr1:168165:169049:clu_27301_-  0.0139586116      -
## 5 chr1     173752   173862 chr1:172688:173753:clu_27301_- -0.0233442066      -
## 6 chr1     195262   195411 chr1:188902:195263:clu_27299_-  0.0209026624      -
##      p.adjust       gene
## 1 0.175128408     WASH7P
## 2 0.007529064 AL627309.5
## 3 0.007529064 AL627309.5
## 4 0.007529064 AL627309.5
## 5 0.007529064 AL627309.5
## 6 0.175128408     WASH9P
```

```
head(all_first)
```

```
##    chr exon_start exon_end                       intron_id      deltapsi strand
## 1 chr1      29320    29370   chr1:24891:29321:clu_27299_-  0.0005808663      -
## 2 chr1     168609   168767 chr1:168165:168610:clu_27301_-  0.0008945081      -
## 3 chr1     169048   169210 chr1:168165:169049:clu_27301_-  0.0139586116      -
## 4 chr1     169048   169240 chr1:168165:169049:clu_27301_-  0.0139586116      -
## 5 chr1     173752   173862 chr1:172688:173753:clu_27301_- -0.0233442066      -
## 6 chr1     195262   195411 chr1:188902:195263:clu_27299_-  0.0209026624      -
##      p.adjust       gene
## 1 0.175128408     WASH7P
## 2 0.007529064 AL627309.5
## 3 0.007529064 AL627309.5
## 4 0.007529064 AL627309.5
## 5 0.007529064 AL627309.5
## 6 0.175128408     WASH9P
```

```
write.table(all_first, here("31_leafcutter", "histone_profile/lc_3db_all_first_exons.bed"),
            quote=F, sep="\t", row.names = F, col.names = F)
```

TSS only

```
all_TSSes = mutate(all_first, TSS_start = if_else(strand == "+", exon_start, exon_end-1),
                   TSS_end = if_else(strand=="+", exon_start+1, exon_end)) %>%
    arrange(chr, TSS_start, TSS_end)
head(all_TSSes)
```

```
##    chr exon_start exon_end                       intron_id      deltapsi strand
## 1 chr1      29320    29370   chr1:24891:29321:clu_27299_-  0.0005808663      -
## 2 chr1     168609   168767 chr1:168165:168610:clu_27301_-  0.0008945081      -
## 3 chr1     169048   169210 chr1:168165:169049:clu_27301_-  0.0139586116      -
## 4 chr1     169048   169240 chr1:168165:169049:clu_27301_-  0.0139586116      -
## 5 chr1     173752   173862 chr1:172688:173753:clu_27301_- -0.0233442066      -
## 6 chr1     195262   195411 chr1:188902:195263:clu_27299_-  0.0209026624      -
##      p.adjust       gene TSS_start TSS_end
## 1 0.175128408     WASH7P     29369   29370
## 2 0.007529064 AL627309.5    168766  168767
## 3 0.007529064 AL627309.5    169209  169210
## 4 0.007529064 AL627309.5    169239  169240
## 5 0.007529064 AL627309.5    173861  173862
## 6 0.175128408     WASH9P    195410  195411
```

```
length(unique(all_TSSes$intron_id[all_TSSes$p.adjust < 0.05 & abs(all_TSSes$deltapsi)>0.1]))
```

```
## [1] 471
```

Filter to 1 junction = 1 TSS

Pick the upstream TSS so we're consistent

```
filter_TSS = function(strand, coords){
    if (unique(strand) == "+"){return (min(coords))
        }else if (unique(strand) == "-"){return(max(coords))}
}

tss_filt = group_by(all_TSSes, intron_id, deltapsi, p.adjust, strand, chr) %>%
    summarise(TSS_start= filter_TSS(strand, TSS_start),  TSS_end = filter_TSS(strand, TSS_end), gene=pas
    select(chr, TSS_start, TSS_end, intron_id, deltapsi, strand, p.adjust, gene) %>%
    arrange(chr, TSS_start, TSS_end)
```

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

```
nrow(tss_filt)
```

```
## [1] 42397
```

```
head(tss_filt)
```

```
## # A tibble: 6 x 8
## # Groups:   intron_id, deltapsi, p.adjust, strand [6]
##   chr   TSS_start TSS_end intron_id               deltapsi strand p.adjust gene
##   <chr>     <dbl>   <dbl> <chr>                      <dbl> <chr>     <dbl> <chr>
## 1 chr1      29369   29370 chr1:24891:29321:clu_2~  5.81e-4 -       0.175   WASH~
## 2 chr1     168766  168767 chr1:168165:168610:clu~  8.95e-4 -       0.00753 AL62~
## 3 chr1     169239  169240 chr1:168165:169049:clu~  1.40e-2 -       0.00753 AL62~
## 4 chr1     173861  173862 chr1:172688:173753:clu~ -2.33e-2 -       0.00753 AL62~
## 5 chr1     195410  195411 chr1:188902:195263:clu~  2.09e-2 -       0.175   WASH~
## 6 chr1     199874  199875 chr1:195416:199837:clu~  8.39e-3 -       0.175   WASH~
```

```
nrow(filter(tss_filt, abs(deltapsi) > 0.1 & p.adjust < 0.05))
```

```
## [1] 471
```

```
white = filter(tss_filt, deltapsi < -0.1 & p.adjust < 0.05); nrow(white)
```

```
## [1] 239
```

```
beige = filter(tss_filt, deltapsi > 0.1 & p.adjust < 0.05); nrow(beige)
```

```
## [1] 232
```

```
head(beige)
```

```
## # A tibble: 6 x 8
## # Groups:   intron_id, deltapsi, p.adjust, strand [6]
##   chr   TSS_start  TSS_end intron_id               deltapsi strand p.adjust gene
##   <chr>     <dbl>    <dbl> <chr>                      <dbl> <chr>     <dbl> <chr>
```

```
## 1 chr1    6614856  6614857 chr1:6615591:6616757:~    0.102 +      1.03e- 2 ENSG~
## 2 chr1   14924129 14924130 chr1:14924572:1496068~    0.104 +      5.39e-13 KAZN
## 3 chr1   23800401 23800402 chr1:23799436:2380038~    0.128 -      1.24e- 7 GALE
## 4 chr1   45339954 45339955 chr1:45334511:4533989~    0.112 -      3.11e- 3 MUTYH
## 5 chr1   55215361 55215362 chr1:55178132:5521460~    0.155 -      1.26e- 8 ENSG~
## 6 chr1   87129764 87129765 chr1:87129994:8713361~    0.158 +      1.01e-39 LINC~
```

```
grep("PEMT|PPARG", white$gene, value=T)
```

```
## [1] "PEMT"  "PPARG"
```

```
grep("PEMT|PPARG", beige$gene, value=T)
```

```
## [1] "PEMT"  "PPARG"
```

```
length(unique(beige$intron_id))
```

```
## [1] 232
```

```
write.table(white, here("31_leafcutter", "histone_profile/lc_3db_white_TSSes.bed"),
            quote=F, sep="\t", row.names = F, col.names = F)
write.table(beige, here("31_leafcutter", "histone_profile/lc_3db_beige_TSSes.bed"),
             quote=F, sep="\t", row.names = F, col.names = F)
```

Not-significant TSSes

Minus DEGs

```
deg = read.delim(here("03limma/any_and_all_donor_DGE.tsv"))
s6_deg = filter(deg,adj.P.Val.s6 < 0.01)
is_deg = function(intron_id){
    gene = gsub("_.*","",intron_id)
    res = gene %in% s6_deg$gene_name
    return (res)
    }

nrow(tss_filt)
```

```
## [1] 42397
```

```
not_sig = filter(tss_filt, p.adjust > 0.05 & abs(deltapsi) < 0.1 & !is_deg(gene))
nrow(not_sig)#expressed, but not DSG or DEG in donor13
```

```
## [1] 23529
```

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

```
## # A tibble: 0 x 8
## # Groups:   intron_id, deltapsi, p.adjust, strand [0]
## # i 8 variables: chr <chr>, TSS_start <dbl>, TSS_end <dbl>, intron_id <chr>,
## #   deltapsi <dbl>, strand <chr>, p.adjust <dbl>, gene <chr>
```

```
write.table(not_sig, here("31_leafcutter", "histone_profile/lc_3db_not_sig_TSSes.bed"),
            quote=F, sep="\t", row.names = F, col.names = F)
```

TSS -2kb/+500b

```
promoters = mutate(.keep="none",
                   all_TSSes, chr,
                   promoter_start=if_else(strand == "+", TSS_start-2000, TSS_end-500),
                   promoter_end =if_else(strand=="+",TSS_start+500, TSS_end+2000),
                   intron_id, deltapsi,strand, p.adjust) %>%
    select(chr, promoter_start, promoter_end, intron_id, deltapsi,strand, p.adjust) %>%
    arrange(chr, promoter_start, promoter_end, strand)
head(promoters)
```

```
##    chr promoter_start promoter_end                      intron_id      deltapsi
## 1 chr1          28870        31370   chr1:24891:29321:clu_27299_-  0.0005808663
## 2 chr1         168267       170767 chr1:168165:168610:clu_27301_-  0.0008945081
## 3 chr1         168710       171210 chr1:168165:169049:clu_27301_-  0.0139586116
## 4 chr1         168740       171240 chr1:168165:169049:clu_27301_-  0.0139586116
## 5 chr1         173362       175862 chr1:172688:173753:clu_27301_- -0.0233442066
## 6 chr1         194911       197411 chr1:188902:195263:clu_27299_-  0.0209026624
##   strand    p.adjust
## 1      - 0.175128408
## 2      - 0.007529064
## 3      - 0.007529064
## 4      - 0.007529064
## 5      - 0.007529064
## 6      - 0.175128408
```

```
tail(promoters)
```

```
##         chr promoter_start promoter_end                         intron_id
## 108991 chrY       57205480     57207980 chrY:57207668:57208026:clu_9219_+
## 108992 chrY       57205480     57207980 chrY:57207668:57208201:clu_9219_+
## 108993 chrY       57205847     57208347 chrY:57209729:57209822:clu_9221_+
## 108994 chrY       57206518     57209018 chrY:57208665:57208843:clu_9220_+
## 108995 chrY       57206518     57209018 chrY:57208979:57209532:clu_9220_+
## 108996 chrY       57207305     57209805 chrY:57209733:57209822:clu_9221_+
##            deltapsi strand   p.adjust
## 108991 -0.058590556      + 0.08223072
## 108992 -0.018132674      + 0.08223072
## 108993  0.002203951      + 0.97892676
## 108994  0.014675882      + 0.01899764
## 108995 -0.016566101      + 0.01899764
## 108996 -0.002203951      + 0.97892676
```

```
#Convert coords to strings to avoid sci notation (bedtools won't have it)
promoters = mutate(promoters, promoter_start = format(promoter_start, scientific=F, trim=T),
                   promoter_end = format(promoter_end, scientific=F, trim=T))
```

```
white_promoters = filter(promoters, p.adjust < 0.05 & deltapsi <= -0.1)
nrow(white_promoters)
```

```
## [1] 567
```

```
beige_promoters = filter(promoters, p.adjust < 0.05 & deltapsi >= 0.1)
nrow(beige_promoters)
```

```
## [1] 528
```

```
write_tsv(white_promoters, here("31_leafcutter", "promoter_binding/lc_3db_white_promoters.bed"),
          col_names=F)
write_tsv(beige_promoters, here("31_leafcutter", "promoter_binding/lc_3db_beige_promoters.bed"),
              col_names = F)
```