NCBP3 TC conversions metagene

Manfred Schmid

Contents

03 July, 2020; 15:57

Setup

```
knitr::opts_chunk$set(fig.width=12, fig.height=8,
                      fig.path=paste0('../Figures/NCBP3_STARmap_TCconversions_metagene/'),
                      dev='pdf',
                      echo=TRUE, warning=FALSE, message=FALSE,
                      error=TRUE)
```

```
suppressWarnings(library('tidyverse'))
suppressWarnings(library('magrittr'))
suppressWarnings(library('knitr'))
```

```
#!/bin/sh
#srun --pty bash
cd /home/schmidm/faststorage/NCBP3/C17orf85_PARCLIP_Landthaler/scripts/
conda activate slamdunk

bam="/home/schmidm/faststorage/NCBP3/C17orf85_PARCLIP_Landthaler/data/STAR_map/SRR500480Aligned.sortedBy
bam="/home/schmidm/faststorage/NCBP3/C17orf85_PARCLIP_Landthaler/data/STAR_map/SRR500481Aligned.sortedBy

fasta="/home/schmidm/annotations/hg38/Genome/GRCh38.fa"


##do below for both bam files !!
samtools mpileup -B -A -f $fasta $bam > ${bam/.bam/_mpileup.out}

awk '$3 == "T" && !($5=="." || $5 ==",")' ${bam/.bam/_mpileup.out} | head
#-->check a few out using IGV; col5 capital 'C' are T>C conversions on plus strand
awk '$3 == "A" && !($5=="." || $5 ==",")' ${bam/.bam/_mpileup.out} | head
#-->check a few out using IGV; col5 lowercase 'g' are T>C conversions on minus strand

awk '{
  if($3 == "T" && ($5 ~ /C/)){
    split($5, chars, "");
    for(i=1;i<=length(chars);i++){
      if(chars[i] == "C"){
```

```
        TCcnt += 1;
      };
    }
    print $1"\t"$2-1"\t"$2"\t"TCcnt;
    TCcnt = 0;
  }
}' ${bam/.bam/_mpileup.out} > ${bam/.bam/_TCs_plus.bedgraph}

awk '{
  if($3 == "A" && ($5 ~ /g/)){
    split($5, chars, "");
    for(i=1;i<=length(chars);i++){
      if(chars[i] == "g"){
        TCcnt += 1;
      };
    }
    print $1"\t"$2-1"\t"$2"\t"TCcnt;
    TCcnt = 0;
  }
}' ${bam/.bam/_mpileup.out} > ${bam/.bam/_TCs_minus.bedgraph}

bam="/home/schmidm/faststorage/NCBP3/C17orf85_PARCLIP_Landthaler/data/STAR_map/SRR500480Aligned.sortedBy
wc -l ${bam/.bam/_TCs*.bedgraph}
# 1831 /home/schmidm/faststorage/NCBP3/C17orf85_PARCLIP_Landthaler/data/STAR_map/SRR500480Aligned.sorte
# 1867 /home/schmidm/faststorage/NCBP3/C17orf85_PARCLIP_Landthaler/data/STAR_map/SRR500480Aligned.sorte

awk '{sum+=$4}END{print sum}' ${bam/.bam/_TCs*.bedgraph}
#4027

bam="/home/schmidm/faststorage/NCBP3/C17orf85_PARCLIP_Landthaler/data/STAR_map/SRR500481Aligned.sortedBy

wc -l ${bam/.bam/_TCs*.bedgraph}
#22308 /home/schmidm/faststorage/NCBP3/C17orf85_PARCLIP_Landthaler/data/STAR_map/SRR500481Aligned.sorte
#22824 /home/schmidm/faststorage/NCBP3/C17orf85_PARCLIP_Landthaler/data/STAR_map/SRR500481Aligned.sorte
awk '{sum+=$4}END{print sum}' ${bam/.bam/_TCs*.bedgraph}
#47527




## intersect with exons


#anno="/home/schmidm/ms_tools/subread-2.0.0-Linux-x86_64/annotation/hg38_RefSeq_exon.txt"
#sed 1d $anno | awk '{if($1==gene_id){i+=1}else{i=1};gene_id=$1;print $2"\t"$3-1"\t"$4"\t"$1":"i"\t0\t"

exons="/home/schmidm/ms_tools/subread-2.0.0-Linux-x86_64/annotation/hg38_RefSeq_individualexons.bed"

awk '$6=="+"' $exons > ${exons/.bed/_plus.bed}
awk '$6=="-"' $exons > ${exons/.bed/_minus.bed}

bg="/home/schmidm/faststorage/NCBP3/C17orf85_PARCLIP_Landthaler/data/STAR_map/SRR500481Aligned.sortedBy
```

```
grep ^chr $bg | sort -k1,1 -k2,2n > tmp.bg
grep ^chr ${exons/.bed/_plus.bed} | sort -k1,1 -k2,2n -o ${exons/.bed/_plus.bed}
bedtools intersect -loj -a ${exons/.bed/_plus.bed} -b tmp.bg | \
awk '$8 != "-1"' > ${bg/.bedgraph/_intersectexons.txt}

wc -l ${bg/.bedgraph/_intersectexons.txt}
#7573 /home/schmidm/faststorage/NCBP3/C17orf85_PARCLIP_Landthaler/data/STAR_map/SRR500481Aligned.sorted

bg="/home/schmidm/faststorage/NCBP3/C17orf85_PARCLIP_Landthaler/data/STAR_map/SRR500481Aligned.sortedBy

grep ^chr $bg | sort -k1,1 -k2,2n > tmp.bg
grep ^chr ${exons/.bed/_minus.bed} | sort -k1,1 -k2,2n -o ${exons/.bed/_minus.bed}
bedtools intersect -loj -a ${exons/.bed/_minus.bed} -b tmp.bg | \
awk '$8 != "-1"' > ${bg/.bedgraph/_intersectexons.txt}

wc -l ${bg/.bedgraph/_intersectexons.txt}
#7389 /home/schmidm/faststorage/NCBP3/C17orf85_PARCLIP_Landthaler/data/STAR_map/SRR500481Aligned.sorted
```

load to R

```
TCplus <- read_tsv('/Volumes/GenomeDK/faststorage/NCBP3/C17orf85_PARCLIP_Landthaler/data/STAR_map/SRR50
         col_names = c('chr', 'start', 'end', 'name', 'score', 'strand', 'TCchrom', 'TCstart', 'TCend',

TCminus <- read_tsv('/Volumes/GenomeDK/faststorage/NCBP3/C17orf85_PARCLIP_Landthaler/data/STAR_map/SRR5
         col_names = c('chr', 'start', 'end', 'name', 'score', 'strand', 'TCchrom', 'TCstart', 'TCend',

(TC <- bind_rows(TCplus, TCminus))
```

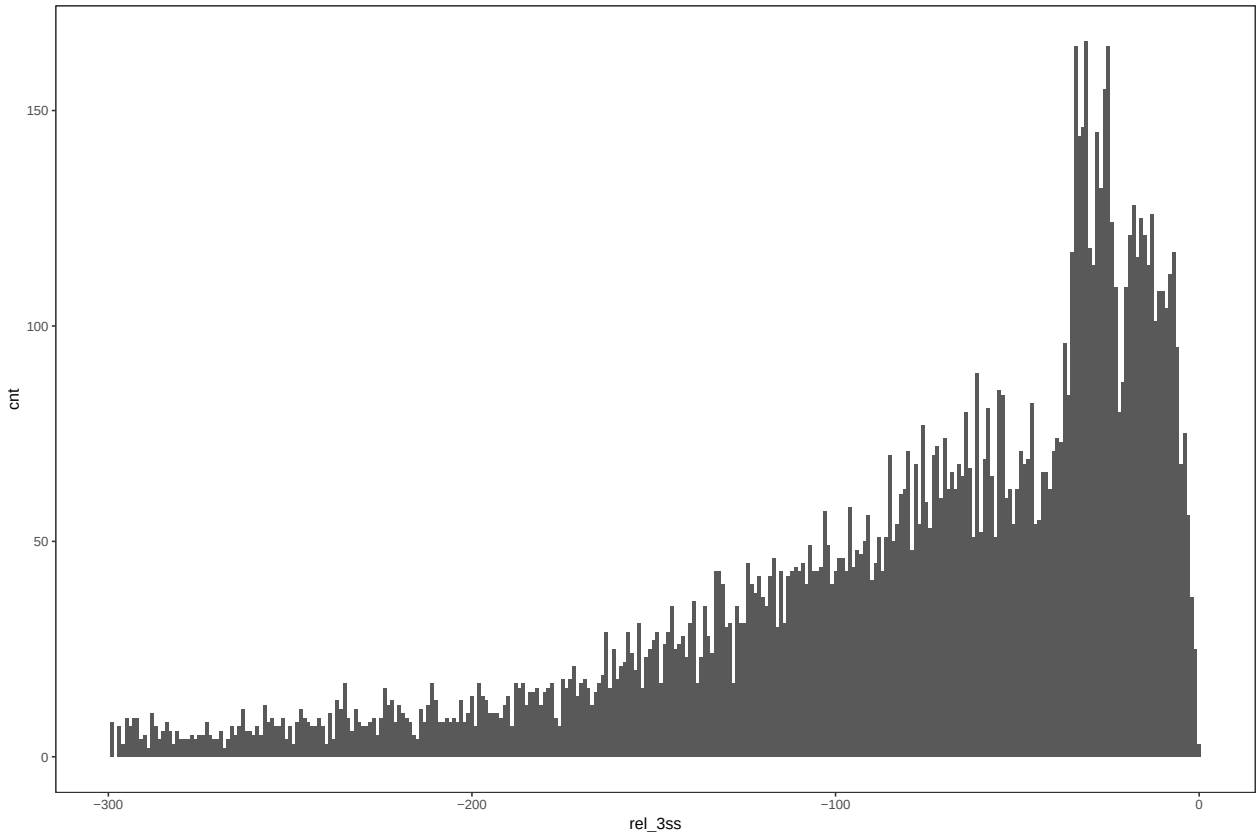
```
## # A tibble: 14,962 x 10
##    chr     start     end name    score strand TCchrom TCstart  TCend TCcnt
##    <chr>   <int>   <int> <chr>   <int> <chr>  <chr>     <int>  <int> <int>
##  1 chr1   941143  941306 148398~     0 +      chr1     941251 9.41e5     1
##  2 chr1   942558  943058 148398~     0 +      chr1     943006 9.43e5     1
##  3 chr1   961628  961750 339451~     0 +      chr1     961673 9.62e5     1
##  4 chr1   962703  962917 339451~     0 +      chr1     962743 9.63e5     1
##  5 chr1  1044333 1044439 375790~     0 +      chr1    1044411 1.04e6     1
##  6 chr1  1044333 1044439 375790~     0 +      chr1    1044415 1.04e6     1
##  7 chr1  1049902 1050037 375790~     0 +      chr1    1050000 1.05e6     1
##  8 chr1  1051452 1051645 375790~     0 +      chr1    1051620 1.05e6     1
##  9 chr1  1291253 1292029 6339:17     0 +      chr1    1291992 1.29e6     1
## 10 chr1  1327240 1328896 80772:5     0 +      chr1    1327343 1.33e6     1
## # ... with 14,952 more rows
```

metagene plots all

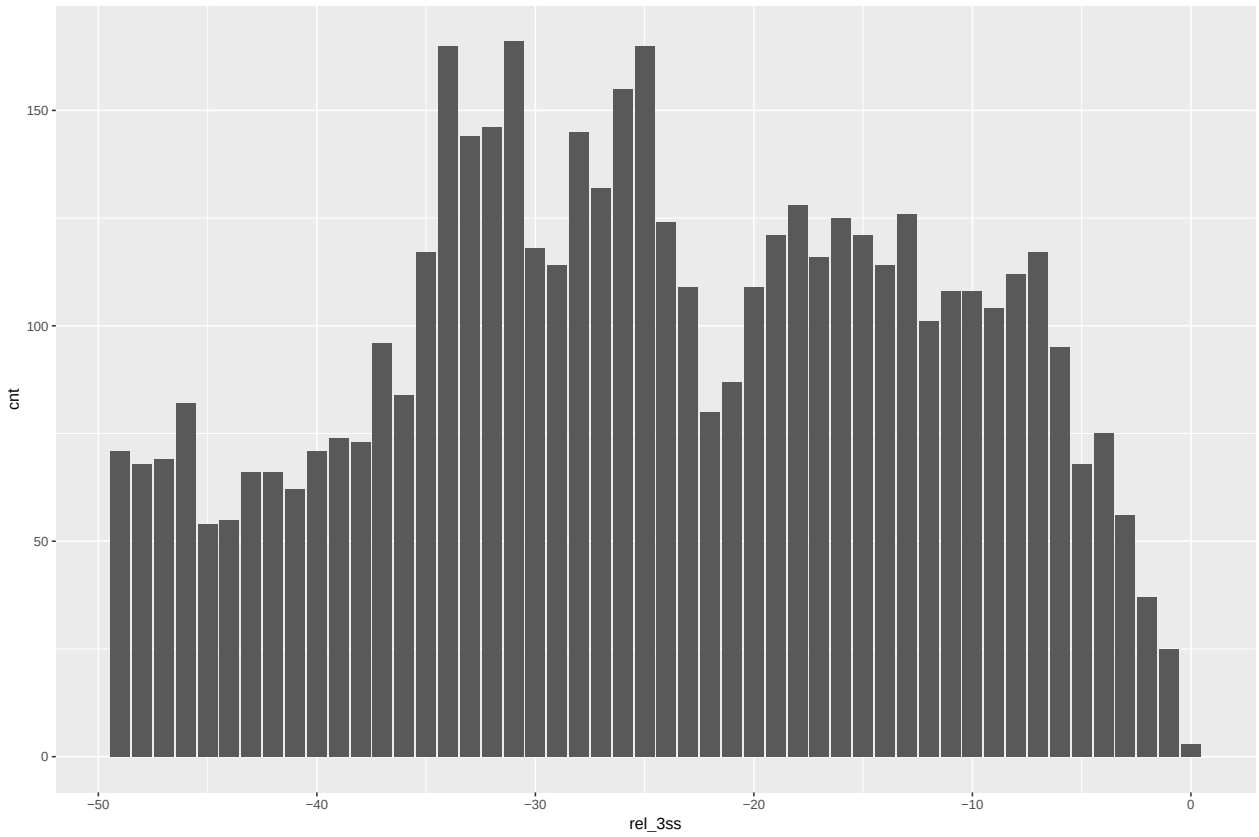
```
TCrel3ss <- TC %>%
  mutate(rel_3ss = ifelse(strand == '+', TCend-end, start-TCstart)) %>%
  group_by(rel_3ss) %>%
  summarize(cnt=n(),
            sum = sum(TCcnt))
```

```
TCrel3ss %>%
  filter(rel_3ss > -300) %>%
```

```
  ggplot(., aes(x=rel_3ss, y=cnt)) +
  geom_bar(stat='identity') +
  theme_bw() +
  theme(panel.grid=element_blank())
```




```
TCrel3ss %>%
  filter(rel_3ss > -50) %>%
  ggplot(., aes(x=rel_3ss, y=cnt)) +
  geom_bar(stat='identity')
```



metagene plots protein-coding

```
load('../../data/subRead_exon_annotations.RData', verbose=T)
```

```
## Loading objects:
##   exon_anno_tbl
```

```
exon_anno_tbl
```

```
## # A tibble: 4,801,919 x 12
##    GeneID  ExonID exon_nr width exon_cnt tr_exons_width class     ENSEMBL
##    <chr>   <chr>    <int> <int>    <int>          <int> <chr>     <chr>
##  1 100287~ 1            1   354        3           1649 multiexo~ ENSG000~
##  2 100287~ 2            2   109        3           1649 multiexo~ ENSG000~
##  3 100287~ 3            3  1189        3           1649 multiexo~ ENSG000~
##  4 653635  1           11   468       11           1758 multiexo~ ENSG000~
##  5 653635  2           10    69       11           1758 multiexo~ ENSG000~
##  6 653635  3            9   152       11           1758 multiexo~ ENSG000~
##  7 653635  4            8   159       11           1758 multiexo~ ENSG000~
##  8 653635  5            7   198       11           1758 multiexo~ ENSG000~
##  9 653635  6            6   136       11           1758 multiexo~ ENSG000~
## 10 653635  7            5   137       11           1758 multiexo~ ENSG000~
## # ... with 4,801,909 more rows, and 4 more variables: SYMBOL <chr>,
## #   REFSEQ <chr>, gene_name <chr>, gene_type <chr>
```

```
(exon_anno_tbl %<>%
  filter(gene_type == 'protein_coding') %>%
```

```
  dplyr::distinct(GeneID, ExonID, exon_nr, width, class))
```

```
## # A tibble: 233,610 x 5
##    GeneID ExonID exon_nr width class
##    <chr>  <chr>    <int> <int> <chr>
##  1 79501  1            1   918 monoexonic
##  2 729759 1            1   939 monoexonic
##  3 81399  1            1   939 monoexonic
##  4 148398 1            1    60 multiexonic first exon
##  5 148398 2            2    92 multiexonic internal
##  6 148398 3            3   182 multiexonic internal
##  7 148398 4            4    51 multiexonic internal
##  8 148398 5            5   125 multiexonic internal
##  9 148398 6            6    90 multiexonic internal
## 10 148398 7            7   186 multiexonic internal
## # ... with 233,600 more rows
```

```
pc_ids <- unique(exon_anno_tbl$GeneID)
length(pc_ids)
```

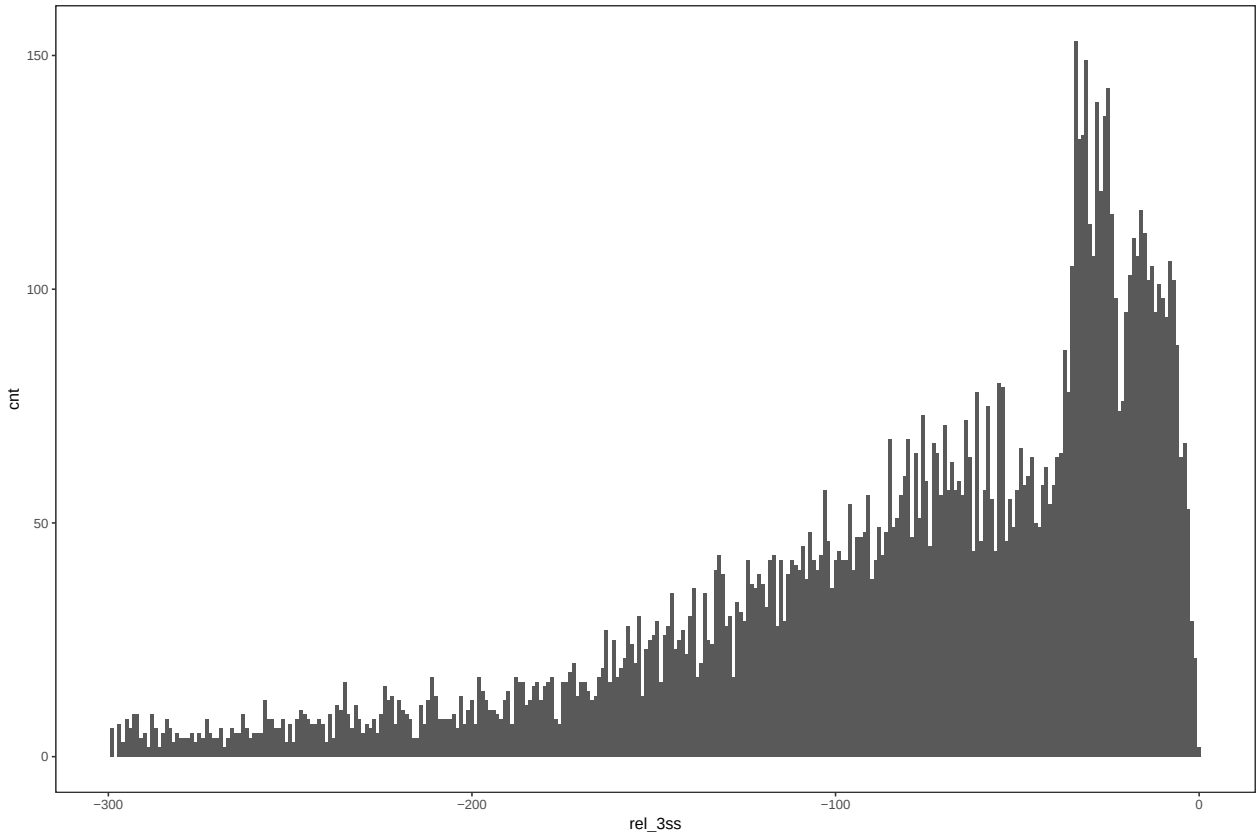
```
## [1] 19297
```

```
(pcTCrel3ss <- TC %>%
  tidyr::separate(name, c('GeneID', 'ExonID'), sep=':') %>%
  filter(GeneID %in% pc_ids) %>%
  left_join(., exon_anno_tbl) %>%
  mutate(rel_3ss = ifelse(strand == '+', TCend-end, start-TCstart)))
```

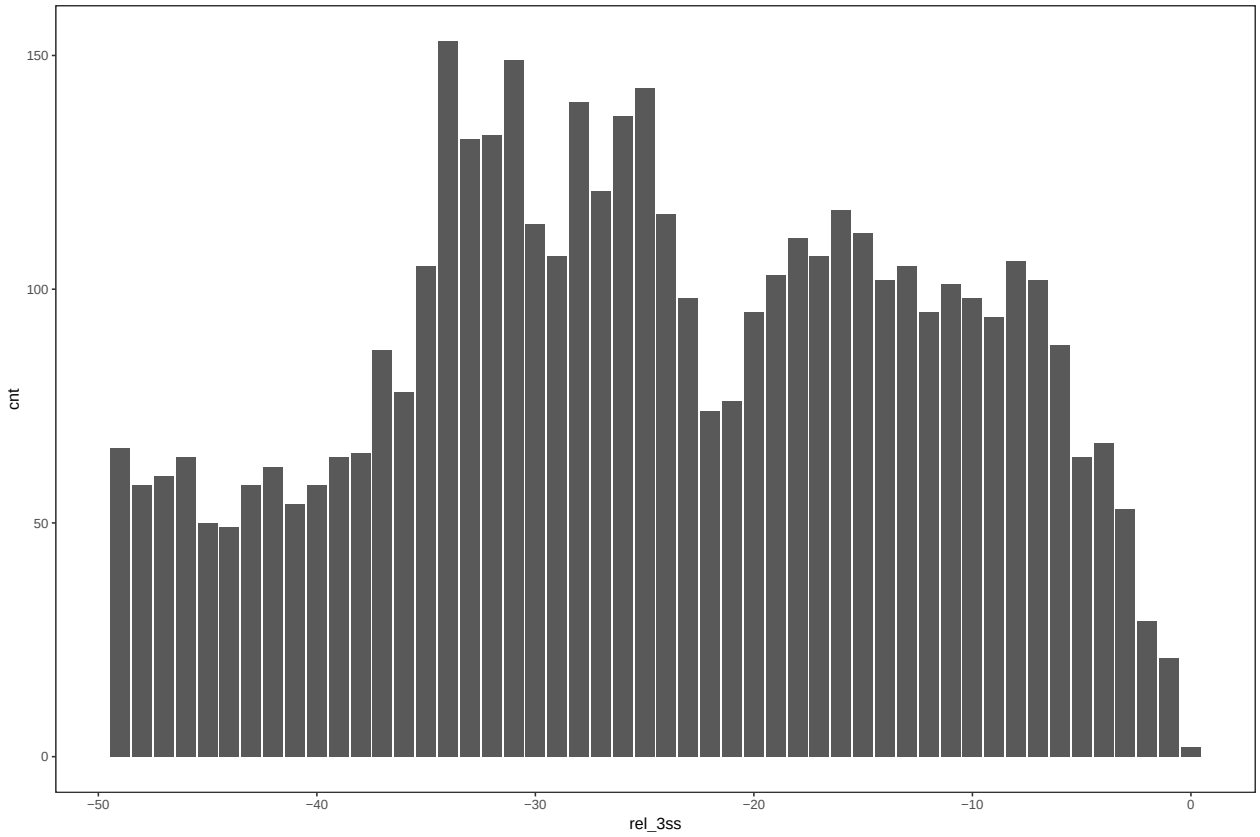
```
## # A tibble: 13,874 x 15
##    chr     start     end GeneID ExonID score strand TCchrom TCstart  TCend
##    <chr>   <int>   <int> <chr>  <chr>  <int> <chr>  <chr>     <int>  <int>
##  1 chr1   941143  941306 148398 8          0 +      chr1     941251 9.41e5
##  2 chr1   942558  943058 148398 11         0 +      chr1     943006 9.43e5
##  3 chr1   961628  961750 339451 4          0 +      chr1     961673 9.62e5
##  4 chr1   962703  962917 339451 7          0 +      chr1     962743 9.63e5
##  5 chr1  1044333 1044439 375790 13         0 +      chr1    1044411 1.04e6
##  6 chr1  1044333 1044439 375790 13         0 +      chr1    1044415 1.04e6
##  7 chr1  1049902 1050037 375790 28         0 +      chr1    1050000 1.05e6
##  8 chr1  1051452 1051645 375790 34         0 +      chr1    1051620 1.05e6
##  9 chr1  1291253 1292029 6339   17         0 +      chr1    1291992 1.29e6
## 10 chr1  1327240 1328896 80772  5          0 +      chr1    1327343 1.33e6
## # ... with 13,864 more rows, and 5 more variables: TCcnt <int>,
## #   exon_nr <int>, width <int>, class <chr>, rel_3ss <int>
```

```
pcTCrel3ss_meta <- pcTCrel3ss%>%
  group_by(rel_3ss) %>%
  summarize(cnt=n(),
            sum = sum(TCcnt))
```

```
pcTCrel3ss_meta %>%
  filter(rel_3ss > -300) %>%
  ggplot(., aes(x=rel_3ss, y=cnt)) +
  geom_bar(stat='identity') +
  theme_bw() +
  theme(panel.grid=element_blank())
```



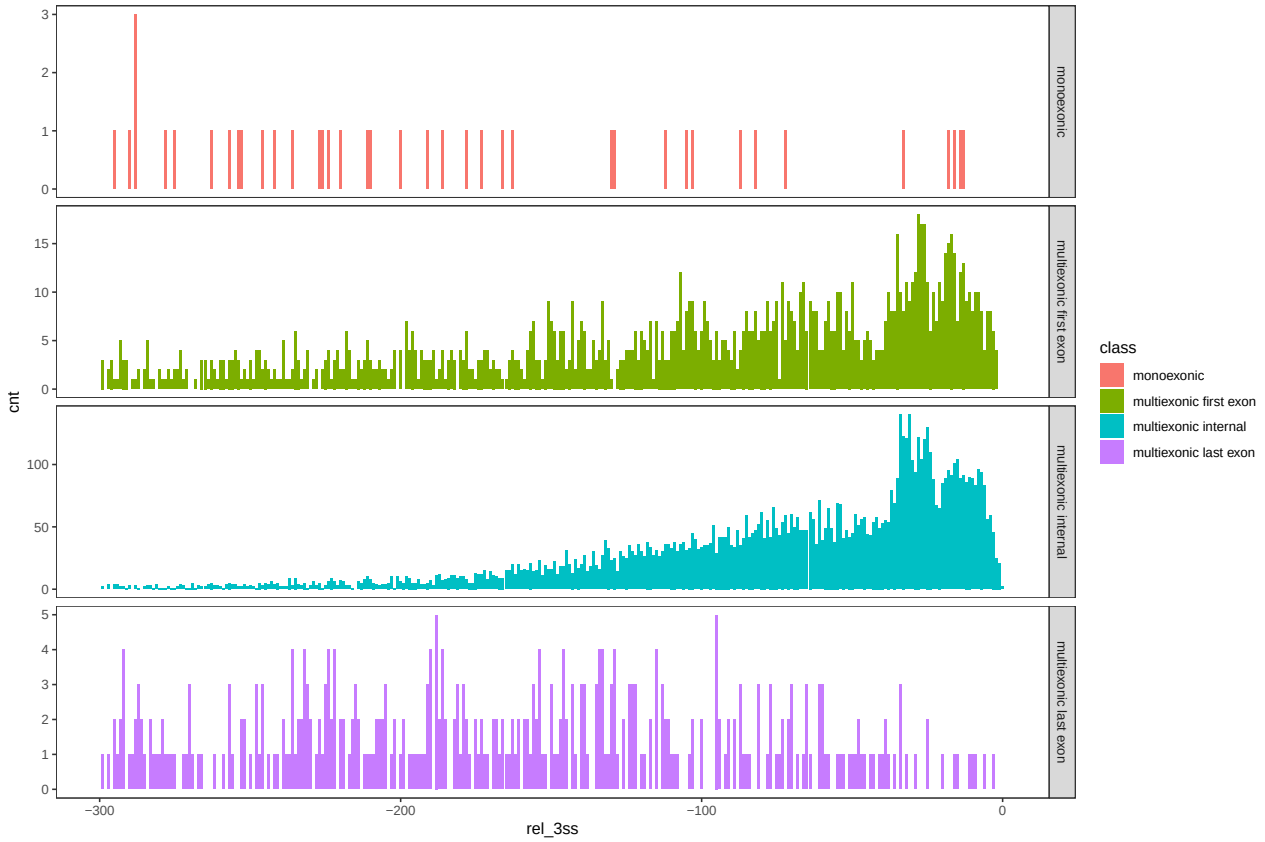

```
pcTCrel3ss_meta %>%
  filter(rel_3ss > -50) %>%
  ggplot(., aes(x=rel_3ss, y=cnt)) +
  geom_bar(stat='identity') +
  theme_bw() +
  theme(panel.grid=element_blank())
```



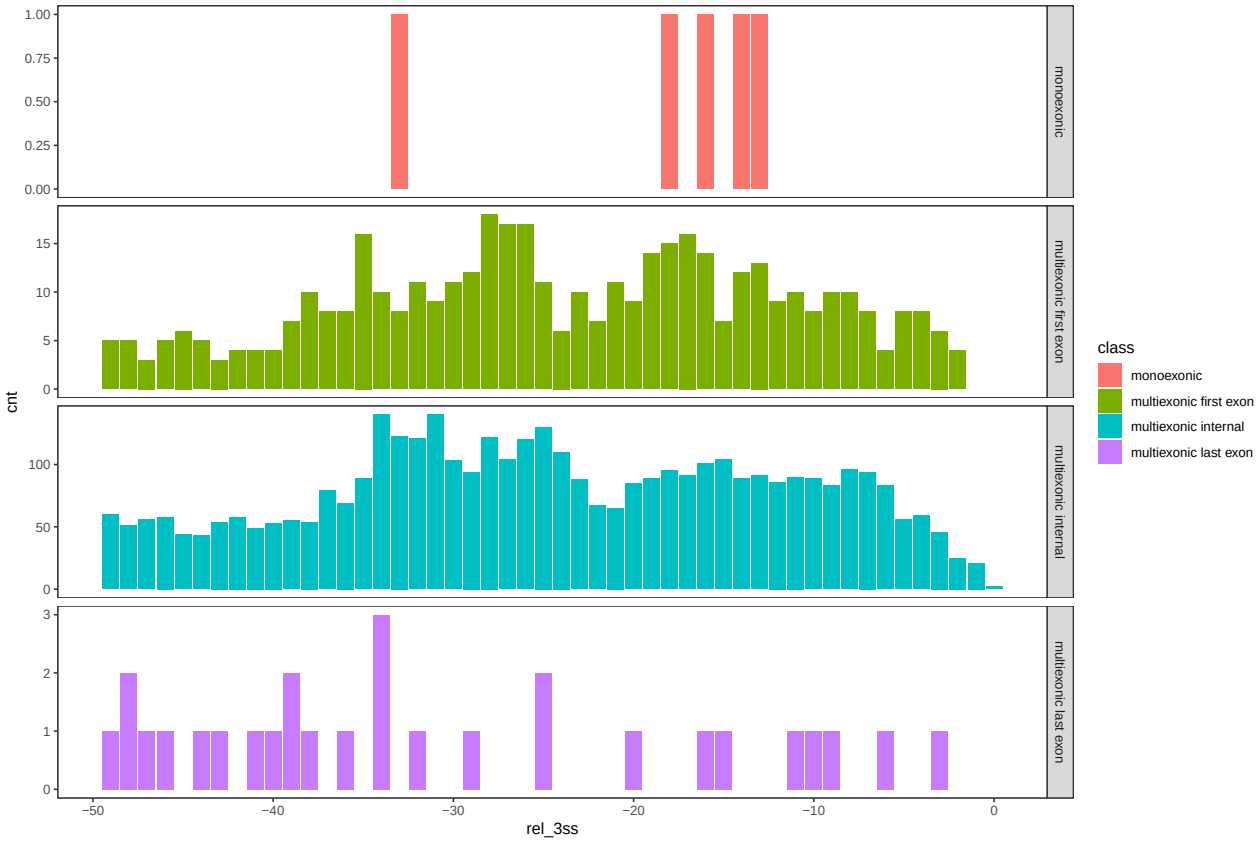
protein-coding per exon class

```
pcTCrel3ss_meta_per_class <- pcTCrel3ss%>%
  group_by(rel_3ss, class) %>%
  summarize(cnt=n(),
            sum = sum(TCcnt))
```

```
pcTCrel3ss_meta_per_class %>%
  filter(rel_3ss > -300) %>%
  ggplot(., aes(x=rel_3ss, y=cnt, fill=class)) +
  geom_bar(stat='identity') +
  facet_grid(class~., scales = 'free') +
  theme_bw() +
  theme(panel.grid=element_blank())
```



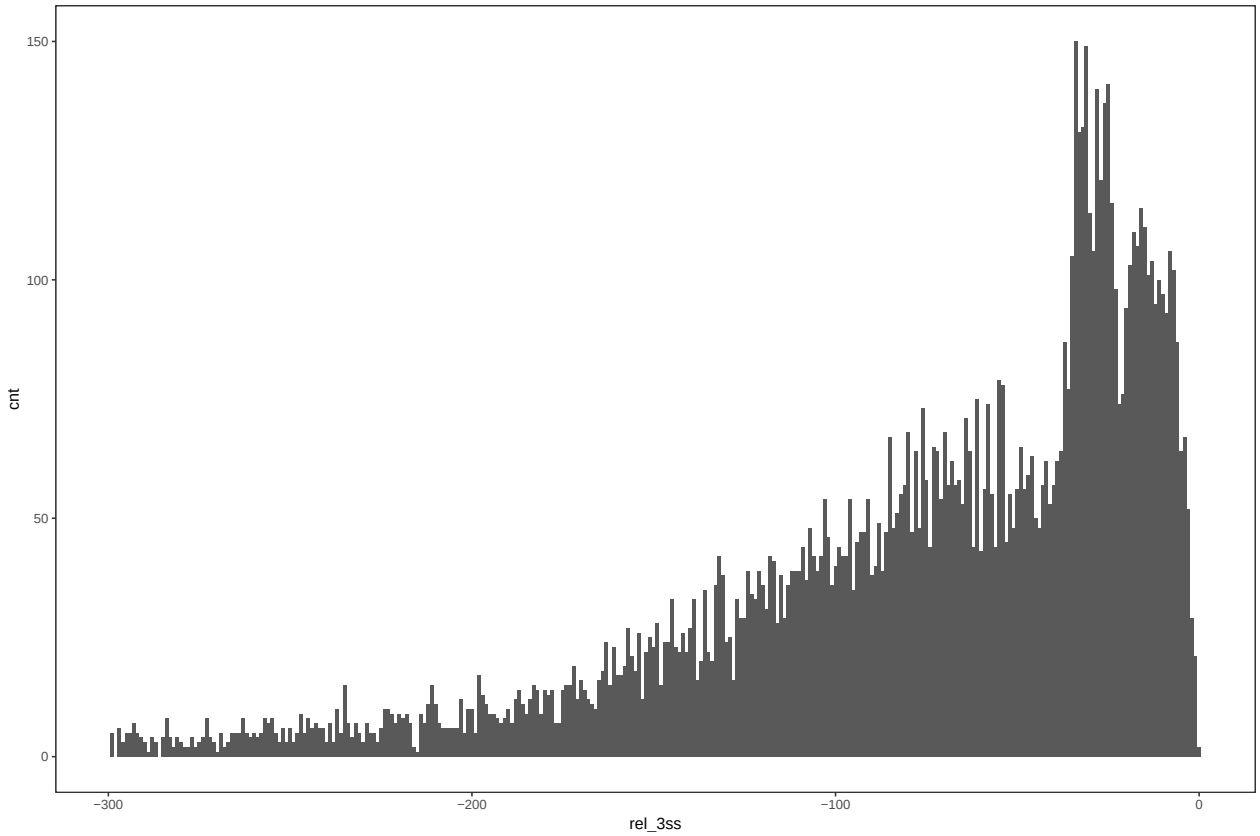

```
pcTCrel3ss_meta_per_class %>%
  filter(rel_3ss > -50) %>%
  ggplot(., aes(x=rel_3ss, y=cnt, fill=class)) +
  geom_bar(stat='identity') +
  facet_grid(class~., scales = 'free') +
  theme_bw() +
  theme(panel.grid=element_blank())
```



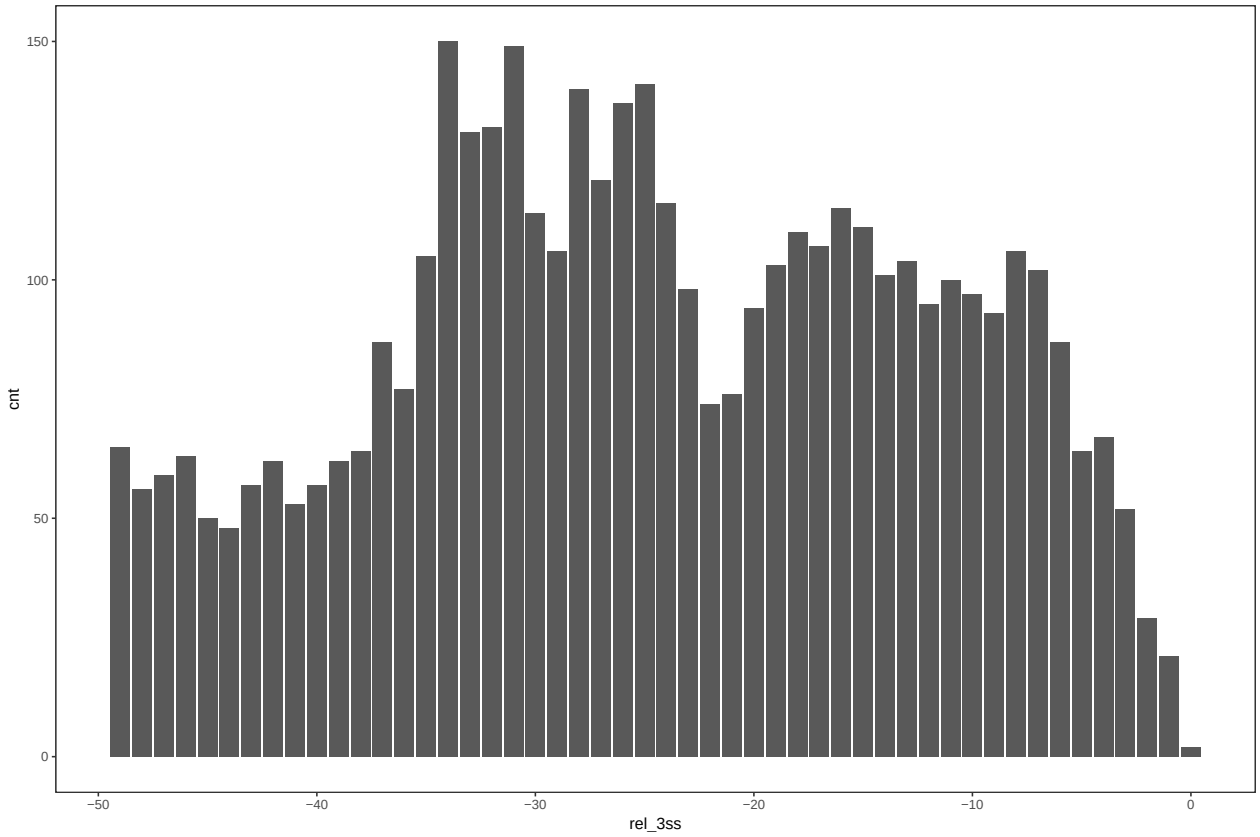
pc all first and internal exons

```
internal_pcTCrel3ss <- pcTCrel3ss %>%
  dplyr::filter(class == 'multiexonic first exon' | class == 'multiexonic internal') %>%
  group_by(rel_3ss) %>%
  summarize(cnt=n(),
            sum = sum(TCcnt))
```

```
internal_pcTCrel3ss %>%
  filter(rel_3ss > -300) %>%
  ggplot(., aes(x=rel_3ss, y=cnt)) +
  geom_bar(stat='identity') +
  theme_bw() +
  theme(panel.grid=element_blank())
```




```
internal_pcTCrel3ss %>%
  filter(rel_3ss > -50) %>%
  ggplot(., aes(x=rel_3ss, y=cnt)) +
  geom_bar(stat='identity') +
  theme_bw() +
  theme(panel.grid=element_blank())
```

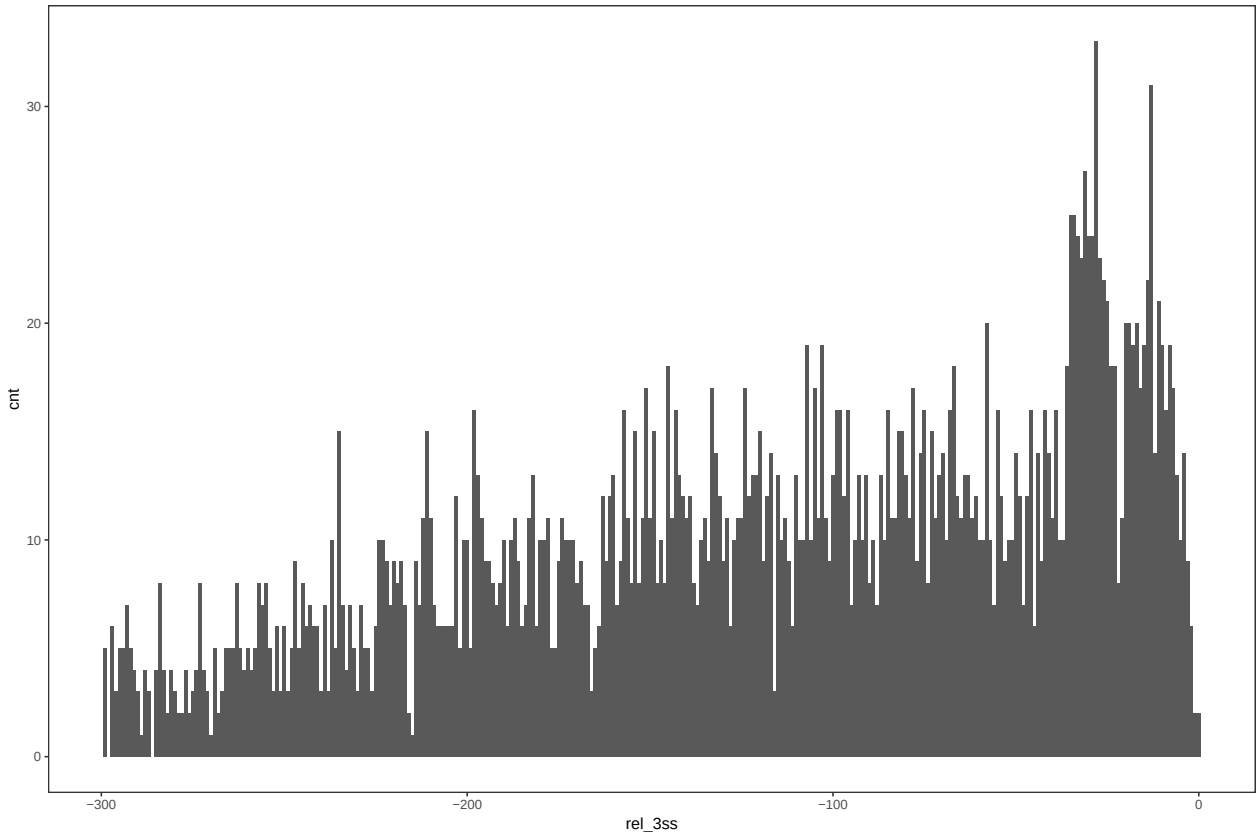


pc all internal exons greater 200bp

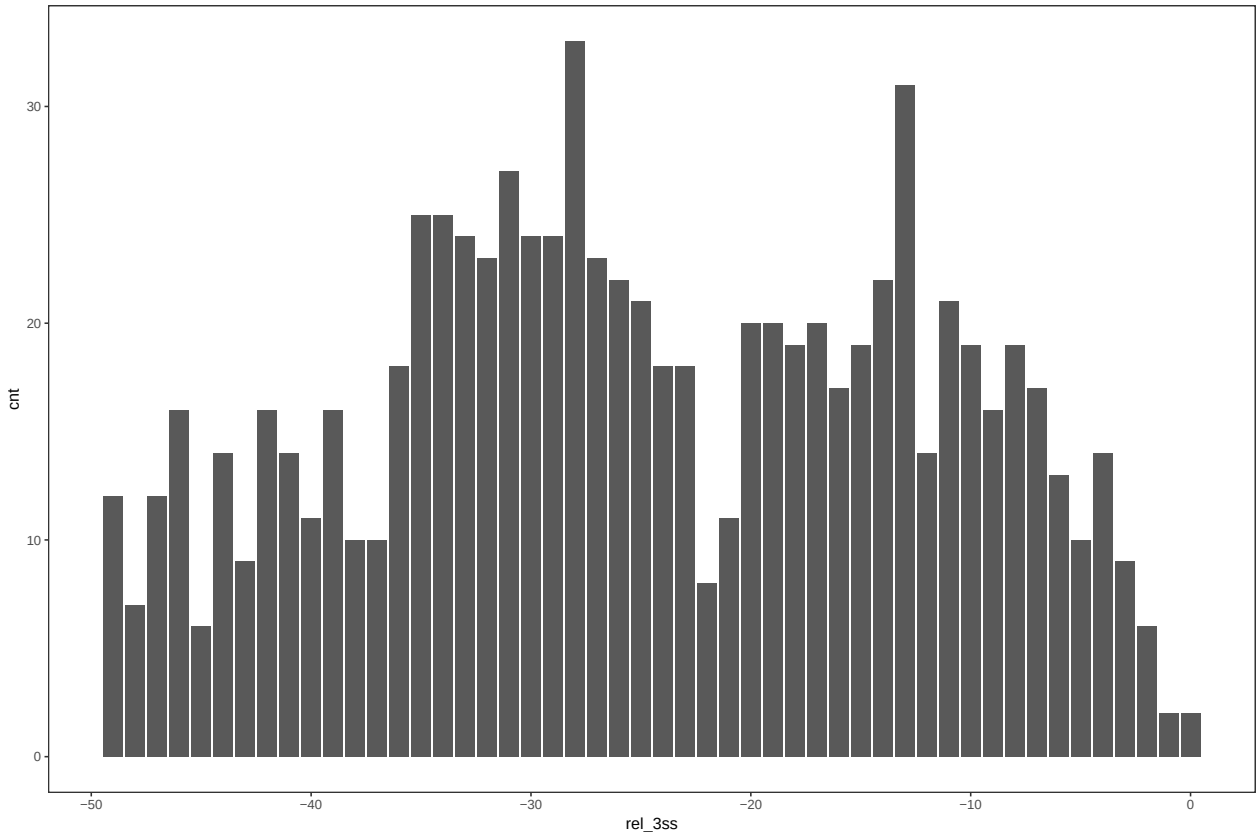
–> these are used in the paper.

```
g200_internal_pcTCrel3ss <- pcTCrel3ss %>%
  dplyr::filter(width > 200, class == 'multiexonic first exon' | class == 'multiexonic internal') %>%
  group_by(rel_3ss) %>%
  summarize(cnt=n(),
            sum = sum(TCcnt))
```

```
g200_internal_pcTCrel3ss %>%
  filter(rel_3ss > -300) %>%
  ggplot(., aes(x=rel_3ss, y=cnt)) +
  geom_bar(stat='identity') +
  theme_bw() +
  theme(panel.grid=element_blank())
```




```
g200_internal_pcTCrel3ss %>%
  filter(rel_3ss > -50) %>%
  ggplot(., aes(x=rel_3ss, y=cnt)) +
  geom_bar(stat='identity')+
  theme_bw() +
  theme(panel.grid=element_blank())
```

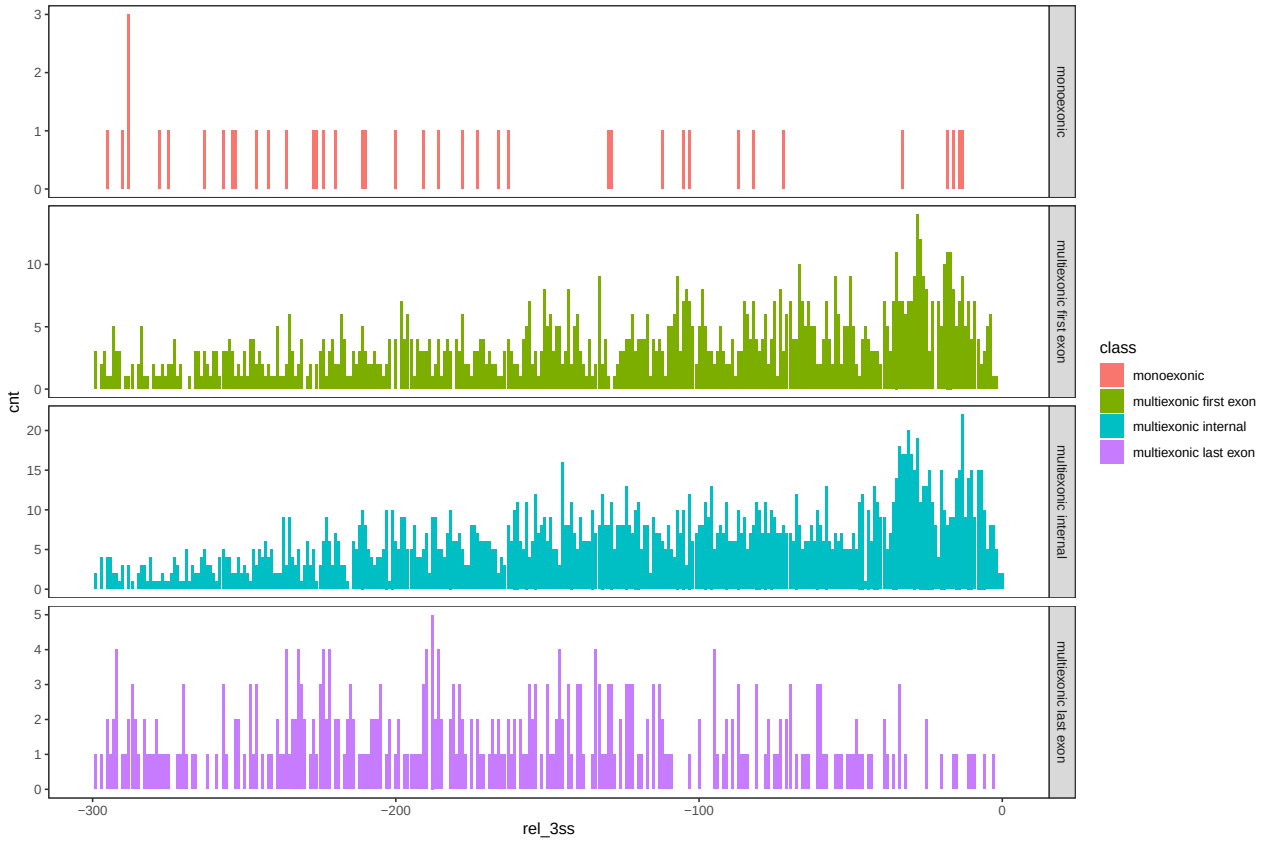


pc all internal exons greater 200bp per class

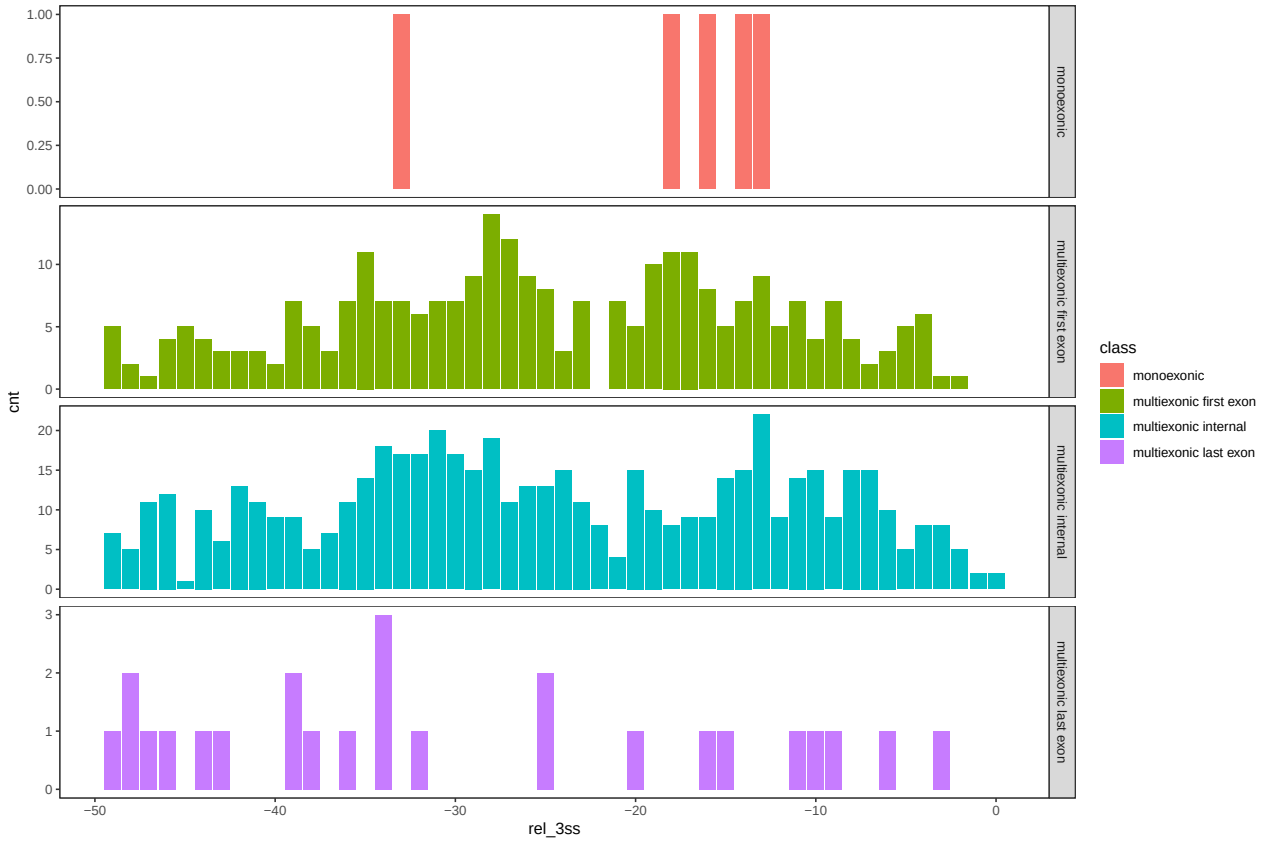
–> these are used in the paper.

```
g200_internal_pcTCrel3ss_perclass <- pcTCrel3ss %>%
  dplyr::filter(width > 200) %>%
  group_by(rel_3ss, class) %>%
  summarize(cnt=n(),
            sum = sum(TCcnt))
```

```
g200_internal_pcTCrel3ss_perclass %>%
  filter(rel_3ss > -300) %>%
  ggplot(., aes(x=rel_3ss, y=cnt, fill=class)) +
  geom_bar(stat='identity') +
  facet_grid(class~., scales = 'free')+
  theme_bw() +
  theme(panel.grid=element_blank())
```




```
g200_internal_pcTCrel3ss_perclass %>%
  filter(rel_3ss > -50) %>%
  ggplot(., aes(x=rel_3ss, y=cnt, fill=class)) +
  geom_bar(stat='identity') +
  facet_grid(class~., scales = 'free')+
  theme_bw() +
  theme(panel.grid=element_blank())
```




```
sessionInfo()
```

```
## R version 3.5.0 (2018-04-23)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS  10.14.6
##
## Matrix products: default
## BLAS: /Library/Frameworks/R.framework/Versions/3.5/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.5/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] bindrcpp_0.2.2  knitr_1.20      magrittr_1.5    forcats_0.3.0
##  [5] stringr_1.3.1   dplyr_0.7.5     purrr_0.2.5     readr_1.1.1
##  [9] tidyr_0.8.1     tibble_1.4.2    ggplot2_3.1.0   tidyverse_1.2.1
##
## loaded via a namespace (and not attached):
##  [1] tidyselect_0.2.4 reshape2_1.4.3   haven_1.1.1      lattice_0.20-35
##  [5] colorspace_1.3-2 htmltools_0.3.6  yaml_2.1.19      utf8_1.1.4
##  [9] rlang_0.2.1      pillar_1.2.3     foreign_0.8-70   glue_1.2.0
## [13] withr_2.1.2      modelr_0.1.2     readxl_1.1.0     bindr_0.1.1
## [17] plyr_1.8.4       munsell_0.5.0    gtable_0.2.0     cellranger_1.1.0
```

```
## [21] rvest_0.3.2      psych_1.8.4      evaluate_0.10.1  labeling_0.3
## [25] parallel_3.5.0   broom_0.4.4      Rcpp_0.12.17     scales_0.5.0
## [29] backports_1.1.2  jsonlite_1.5     mnormt_1.5-5     hms_0.4.2
## [33] digest_0.6.15    stringi_1.2.3    grid_3.5.0       rprojroot_1.3-2
## [37] cli_1.0.0        tools_3.5.0      lazyeval_0.2.1   crayon_1.3.4
## [41] pkgconfig_2.0.1  xml2_1.2.0       lubridate_1.7.4  assertthat_0.2.0
## [45] rmarkdown_1.10   httr_1.3.1       rstudioapi_0.7   R6_2.2.2
## [49] nlme_3.1-137     compiler_3.5.0
```