metagene of exons Viphakone

Manfred Schmid

Contents

03 July, 2020; 15:12

Setup

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

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

load Annotation

```
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>
```

```
(exons_per_class <- table(distinct(exon_anno_tbl, GeneID, ExonID, class)$class))
```

```
##
##              monoexonic multiexonic first exon   multiexonic internal
##                    4900                  23494                 209862
##  multiexonic last exon
##                   23496
```

Load CLIP data

NCBP3 and EIF4A3 from CLIPdb

Data for NCBP3 (c17orf85 and EIF4A3) are from CLIPdb study, mapped to hg38

metagene values using deeptools

```
#!/bin/sh
##cd /home/schmidm/faststorage/CLIP/CLIPdb/scripts
##sbatch --account=thj_common --mem=4g deeptools_subReadanno_perexon.sh

. /home/schmidm/miniconda2/etc/profile.d/conda.sh
conda activate deeptools3

#these annotations are shipped with subRead
#anno="/home/schmidm/ms_tools/subread-2.0.0-Linux-x86_64/annotation/hg38_RefSeq_exon.txt"

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

#awk '{if($6 =="+"){print $0}}' $bed > ${bed/.bed/_plus.bed}
#awk '{if($6 =="-"){print $0}}' $bed > ${bed/.bed/_minus.bed}

plus_bw=$(ls /home/schmidm/faststorage/CLIP/CLIPdb/hg38_bw/*_plus_hg38.bw | awk '$1 ~ /C17orf85/ || $1

minus_bw=${plus_bw//_plus_hg38.bw/_minus_hg38.bw}

python ~/ms_tools/MS_Metagene_Tools/computeMatrixStranded.pyc scale-regions -Rp ${bed/.bed/_plus.bed} -

python ~/ms_tools/MS_Metagene_Tools/computeMatrixOperationsMS.py -m deeptools_subReadanno_individualexo
```

load to R

```
fname <- '/Volumes/GenomeDK/faststorage/CLIP/CLIPdb/scripts/deeptools_subReadanno_individualexons_scaled

df <- RMetaTools::load_deeptoolsmatrix3(fname)
```

```
(df %<>%
   tidyr::separate(id, c('GeneID', 'ExonID'), sep=':') %>%
    dplyr::mutate(sample_name = sub('.*\\/', '', sample_name) %>%
                        sub('_plus_hg38.bw', '', .)) %>%
  dplyr::select(GeneID, ExonID, sample_name, rel_pos, value))
```

```
## # A tibble: 21,875,040 x 5
##    GeneID ExonID sample_name                rel_pos value
##    <chr>  <chr>  <chr>                        <dbl> <dbl>
##  1 643837 4      C17orf85_PARCLIP_PARalyzer    -100     0
##  2 643837 8      C17orf85_PARCLIP_PARalyzer    -100     0
##  3 148398 11     C17orf85_PARCLIP_PARalyzer    -100     0
##  4 148398 14     C17orf85_PARCLIP_PARalyzer    -100     0
##  5 339451 6      C17orf85_PARCLIP_PARalyzer    -100     0
##  6 339451 12     C17orf85_PARCLIP_PARalyzer    -100     0
##  7 9636   1      C17orf85_PARCLIP_PARalyzer    -100     0
##  8 9636   2      C17orf85_PARCLIP_PARalyzer    -100     0
##  9 375790 2      C17orf85_PARCLIP_PARalyzer    -100     0
## 10 375790 8      C17orf85_PARCLIP_PARalyzer    -100     0
## # ... with 21,875,030 more rows
```

select only protein-coding genes and positions with CLIP signal

```
(df %<>% filter(value > 0) %>%
  left_join(., exon_anno_tbl) %>%
  filter(gene_type == 'protein_coding'))
```

```
## # A tibble: 13,399,566 x 15
##    GeneID ExonID sample_name          rel_pos value exon_nr width exon_cnt
##    <chr>  <chr>  <chr>                  <dbl> <dbl>   <int> <int>    <int>
##  1 375790 35     C17orf85_PARCLIP_PA~    -100 0.763      35    88       40
##  2 375790 35     C17orf85_PARCLIP_PA~    -100 0.763      35    88       40
##  3 375790 35     C17orf85_PARCLIP_PA~    -100 0.763      35    88       40
##  4 375790 35     C17orf85_PARCLIP_PA~    -100 0.763      35    88       40
##  5 375790 35     C17orf85_PARCLIP_PA~    -100 0.763      35    88       40
##  6 375790 35     C17orf85_PARCLIP_PA~    -100 0.763      35    88       40
##  7 375790 35     C17orf85_PARCLIP_PA~    -100 0.763      35    88       40
##  8 375790 35     C17orf85_PARCLIP_PA~    -100 0.763      35    88       40
##  9 375790 35     C17orf85_PARCLIP_PA~    -100 0.763      35    88       40
## 10 375790 35     C17orf85_PARCLIP_PA~    -100 0.763      35    88       40
## # ... with 13,399,556 more rows, and 7 more variables:
## #   tr_exons_width <int>, class <chr>, ENSEMBL <chr>, SYMBOL <chr>,
## #   REFSEQ <chr>, gene_name <chr>, gene_type <chr>
```

save

```
ncbp3 <- filter(df, sample_name == 'C17orf85_PARCLIP_PARalyzer')

saveRDS(ncbp3, file='../data/NCBP3_CLIP_exon_metagene.rds')
```

```
eif4a3 <- filter(df, sample_name == 'EIF4A3_HITSCLIP_Piranha_001')

saveRDS(eif4a3, file='../data/EIF4A3_CLIP_exon_metagene.rds')
```

NCBP2 from Giacometti et al

deeptools run

```
#!/bin/sh
##cd /home/schmidm/faststorage/CLIP/Giacometti/scripts
##sbatch --account=thj_common --mem=4g deeptools_subReadanno_perexon.sh

. /home/schmidm/miniconda2/etc/profile.d/conda.sh
conda activate deeptools3

#these annotations are shipped with subRead
#anno="/home/schmidm/ms_tools/subread-2.0.0-Linux-x86_64/annotation/hg38_RefSeq_exon.txt"

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

#awk '{if($6 =="+"){print $0}}' $bed > ${bed/.bed/_plus.bed}
#awk '{if($6 =="-"){print $0}}' $bed > ${bed/.bed/_minus.bed}

plus_bw=$(ls /home/schmidm/faststorage/CLIP/Giacometti/hg38_bw/*_plus_hg38.bw | awk '$1 ~ /C17orf85/ ||

minus_bw=${plus_bw//_plus_hg38.bw/_minus_hg38.bw}

python ~/ms_tools/MS_Metagene_Tools/computeMatrixStranded.pyc scale-regions -Rp ${bed/.bed/_plus.bed} -

python ~/ms_tools/MS_Metagene_Tools/computeMatrixOperationsMS.py -m deeptools_subReadanno_individualexon
```

```
fname <- '/Volumes/GenomeDK/faststorage/CLIP/Giacometti_GSE94427/scripts/deeptools_subReadanno_individua

df <- RMetaTools::load_deeptoolsmatrix3(fname)
```

```
(cbp20 <- df %>%
   filter(grepl('CBP20', sample_name)) %>%
   tidyr::separate(id, c('GeneID', 'ExonID'), sep=':') %>%
    dplyr::mutate(sample_name = sub('.*GSM......._', '', sample_name) %>%
                    sub('_norm_plus_hg38.bw', '', .)) %>%
  dplyr::select(GeneID, ExonID, sample_name, rel_pos, value))
```

```
## # A tibble: 9,234,720 x 5
##    GeneID    ExonID sample_name rel_pos value
##    <chr>     <chr>  <chr>         <dbl> <dbl>
##  1 100287102 3      CBP20_1        -100     0
##  2 100302278 1      CBP20_1        -100     0
##  3 79501     1      CBP20_1        -100     0
##  4 400728    2      CBP20_1        -100     0
##  5 643837    1      CBP20_1        -100     0
##  6 643837    2      CBP20_1        -100     0
##  7 643837    3      CBP20_1        -100     0
##  8 643837    4      CBP20_1        -100     0
##  9 643837    5      CBP20_1        -100     0
## 10 643837    6      CBP20_1        -100     0
## # ... with 9,234,710 more rows
```

select only protein-coding genes and positions with CLIP signal

```
(cbp20 %<>% filter(value > 0) %>%
  left_join(., exon_anno_tbl) %>%
  filter(gene_type == 'protein_coding'))
```

```
## # A tibble: 1,417,732 x 15
##    GeneID ExonID sample_name rel_pos value exon_nr width exon_cnt
##    <chr>  <chr>  <chr>         <dbl> <dbl>   <int> <int>    <int>
##  1 93611  2      CBP20_1        -100  10.4       2    96        9
##  2 93611  2      CBP20_1        -100  10.4       2    96        9
##  3 93611  2      CBP20_1        -100  10.4       2    96        9
##  4 93611  2      CBP20_1        -100  10.4       2    96        9
##  5 93611  2      CBP20_1        -100  10.4       2    96        9
##  6 93611  2      CBP20_1        -100  10.4       2    96        9
##  7 93611  2      CBP20_1        -100  10.4       2    96        9
##  8 93611  2      CBP20_1        -100  10.4       2    96        9
##  9 93611  2      CBP20_1        -100  10.4       2    96        9
## 10 93611  2      CBP20_1        -100  10.4       2    96        9
## # ... with 1,417,722 more rows, and 7 more variables:
## #   tr_exons_width <int>, class <chr>, ENSEMBL <chr>, SYMBOL <chr>,
## #   REFSEQ <chr>, gene_name <chr>, gene_type <chr>
```

average replicates

The datasets are replicates that behave nicely (not shown here), so we simply average over the 2 replicates.

```
cbp20 %<>%
  mutate(sample_name = sub('_.*', '', sample_name)) %>%
  group_by(GeneID, ExonID, sample_name, rel_pos, exon_nr, width, class) %>%
  summarize(value = sum(value)/2)
```

save cbp20 data

```
saveRDS(cbp20, file='../data/CBP20_CLIP_exon_metagene.rds')
```

```
rm(df)
```

ALY data from Viphakone et al

deeptools run

```
#!/bin/sh
##cd /project/THJ_common/faststorage/people/MS/Yuhui/Viphakone_etal
##sbatch --account=thj_common --mem=4g deeptools_subReadanno_perexon.sh

. /home/schmidm/miniconda2/etc/profile.d/conda.sh
conda activate deeptools3



#these annotations are Viphakonepped with subRead
```

```
#anno="/home/schmidm/ms_tools/subread-2.0.0-Linux-x86_64/annotation/hg38_RefSeq_exon.txt"

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

#awk '{if($6 =="+"){print $0}}' $bed > ${bed/.bed/_plus.bed}
#awk '{if($6 =="-"){print $0}}' $bed > ${bed/.bed/_minus.bed}

plus_bw=$(ls /home/schmidm/THJ_common/faststorage/data/Human/GEO/GSE113896/hg38/*plus*.bw | tr "\n" " "
minus_bw=${plus_bw//_hg38_plus.bw/_hg38_minus.bw}

python ~/ms_tools/MS_Metagene_Tools/computeMatrixStranded.pyc scale-regions -Rp ${bed/.bed/_plus.bed} -

python ~/ms_tools/MS_Metagene_Tools/computeMatrixOperationsMS.py -m deeptools_subReadanno_individualexon
```

```
fname <- '/Volumes/GenomeDK/THJ_common/faststorage/people/MS/Yuhui/Viphakone_etal/deeptools_subReadanno_

df <- RMetaTools::load_deeptoolsmatrix3(fname)
```

```
(aly <- df %>%
   filter(grepl('Alyref', sample_name), value > 0) %>%
   tidyr::separate(id, c('GeneID', 'ExonID'), sep=':') %>%
   dplyr::mutate(sample_name = sub('.*GSE113896_', '', sample_name) %>%
                              sub('-union_hg38', '', .)) %>%
  dplyr::select(GeneID, ExonID, sample_name, rel_pos, value))
```

```
## # A tibble: 438,052 x 5
##    GeneID ExonID sample_name rel_pos value
##    <chr>  <chr>  <chr>         <dbl> <dbl>
##  1 148398 2      Alyref-FLAG    -100     1
##  2 148398 6      Alyref-FLAG    -100     1
##  3 148398 7      Alyref-FLAG    -100     1
##  4 339451 8      Alyref-FLAG    -100     1
##  5 339451 9      Alyref-FLAG    -100     1
##  6 339451 10     Alyref-FLAG    -100     1
##  7 9636   1      Alyref-FLAG    -100     1
##  8 9636   2      Alyref-FLAG    -100     1
##  9 375790 2      Alyref-FLAG    -100     1
## 10 375790 8      Alyref-FLAG    -100     1
## # ... with 438,042 more rows
```

```
(aly %<>%
  left_join(., exon_anno_tbl) %>%
  filter(gene_type == 'protein_coding'))
```

```
## # A tibble: 7,811,563 x 15
##    GeneID ExonID sample_name rel_pos value exon_nr width exon_cnt
##    <chr>  <chr>  <chr>         <dbl> <dbl>   <int> <int>    <int>
##  1 148398 2      Alyref-FLAG    -100     1       2    92       14
##  2 148398 2      Alyref-FLAG    -100     1       2    92       14
##  3 148398 6      Alyref-FLAG    -100     1       6    90       14
##  4 148398 6      Alyref-FLAG    -100     1       6    90       14
##  5 148398 7      Alyref-FLAG    -100     1       7   186       14
##  6 148398 7      Alyref-FLAG    -100     1       7   186       14
##  7 339451 8      Alyref-FLAG    -100     1       8   473       12
##  8 339451 8      Alyref-FLAG    -100     1       8   473       12
##  9 339451 8      Alyref-FLAG    -100     1       8   473       12
## 10 339451 8      Alyref-FLAG    -100     1       8   473       12
## # ... with 7,811,553 more rows, and 7 more variables:
## #   tr_exons_width <int>, class <chr>, ENSEMBL <chr>, SYMBOL <chr>,
## #   REFSEQ <chr>, gene_name <chr>, gene_type <chr>
```

save aly data

```
saveRDS(aly, file='../data/ALY_CLIP_exon_metagene.rds')
```

```
rm(df)
```

alternative starting point

```
ncbp3 <- readRDS('../data/NCBP3_CLIP_exon_metagene.rds')

ncbp3$sample_name <- 'NCBP3'
```

```
eif4a3 <- readRDS('../data/EIF4A3_CLIP_exon_metagene.rds')

eif4a3$sample_name <- 'EIF4A3'
```

```
cbp20 <- readRDS('../data/CBP20_CLIP_exon_metagene.rds')
```

```
aly <- readRDS('../data/ALY_CLIP_exon_metagene.rds')

aly$sample_name <- 'ALYREF'
```

combine

```
df <- bind_rows(ncbp3, eif4a3) %>%
  bind_rows(., cbp20) %>%
  bind_rows(., aly)
```

Plots

plot fun

```
metaplot_all <- function(df) {
  df %>%
    group_by(sample_name, rel_pos) %>%
    summarize(events=n()) %>%
    ggplot(., aes(x=rel_pos, y=events, color=sample_name)) +
    geom_line() +
    facet_wrap(~sample_name, scales='free') +
    theme_bw() +
    theme(panel.grid=element_blank())
}
```

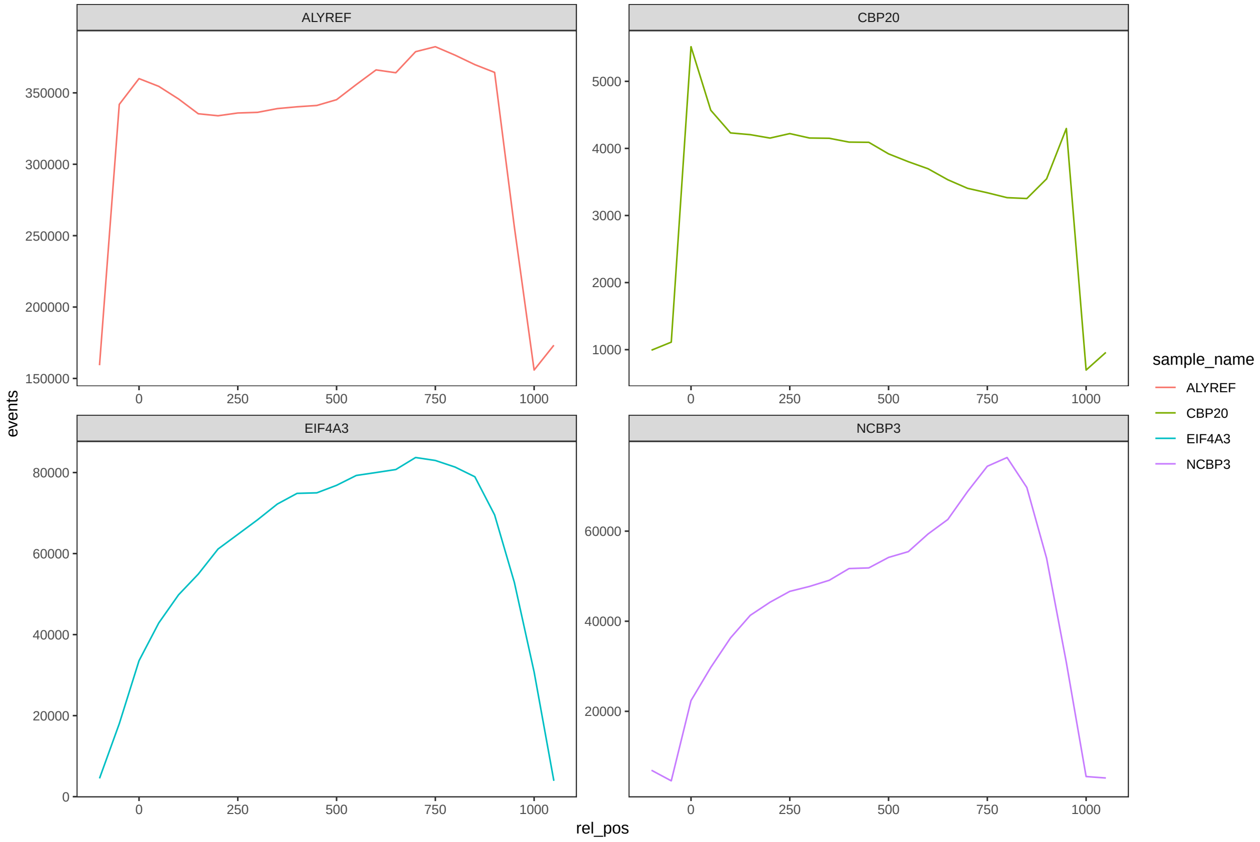
```
metaplot_perclass <- function(df, overlay=FALSE, exonsperclass = exons_per_class) {

  if(overlay){
    p <- df %>%
    group_by(class, sample_name, rel_pos) %>%
    summarize(events=n()) %>%
    mutate(exons_per_class = exonsperclass[class],
           events_per_exon = events/exons_per_class) %>%
    ggplot(., aes(x=rel_pos, y=events_per_exon, color=class)) +
    geom_line() +
    facet_wrap(~sample_name, scales='free') +
    theme_bw() +
    theme(panel.grid=element_blank())
  }else{
    p <- df %>%
    group_by(class, sample_name, rel_pos) %>%
    summarize(events=n()) %>%
    mutate(exons_per_class = exons_per_class[class],
           events_per_exon = events/exons_per_class) %>%
    ggplot(., aes(x=rel_pos, y=events_per_exon, color=sample_name)) +
    geom_line() +
    facet_grid(class~sample_name) +
    theme_bw() +
    theme(panel.grid=element_blank())
  }

  p
}
```

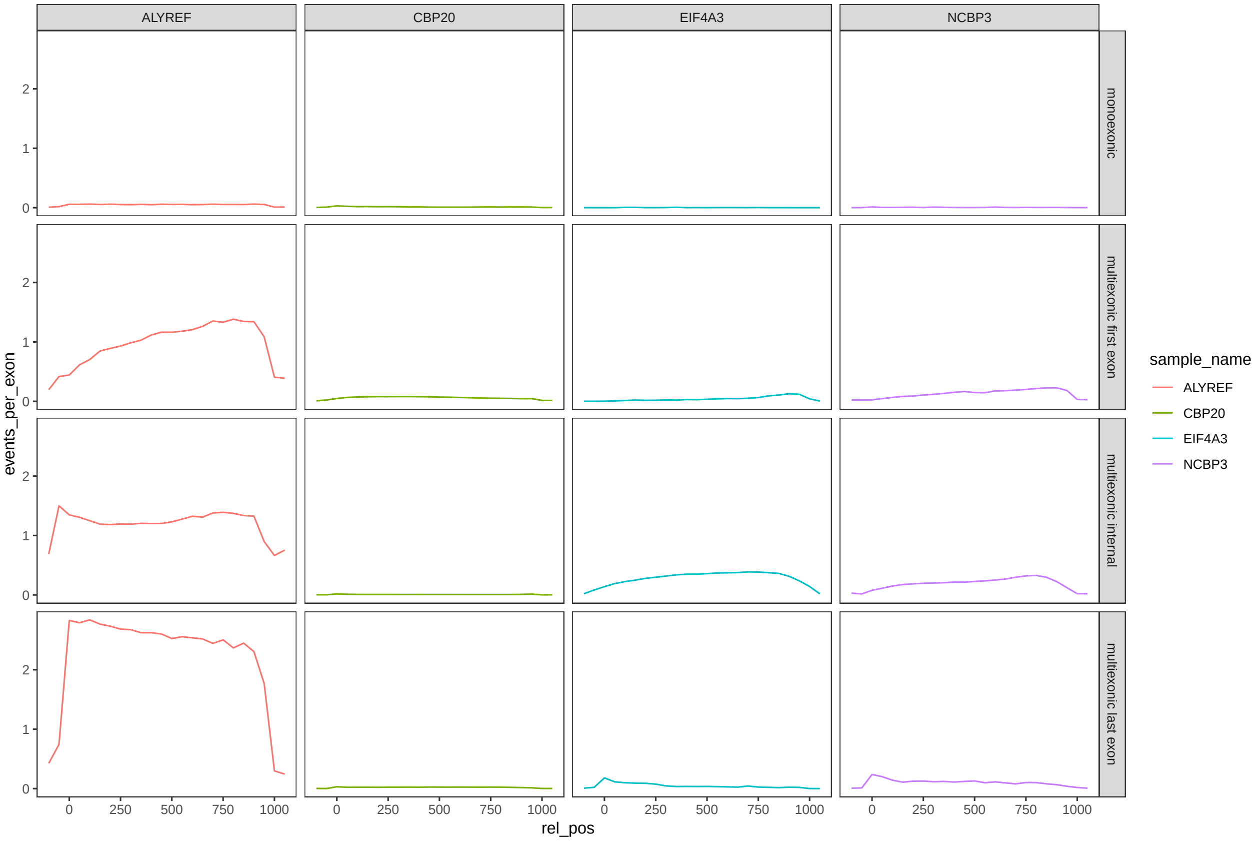
per exon metagene plot

```
metaplot_all(df)
```



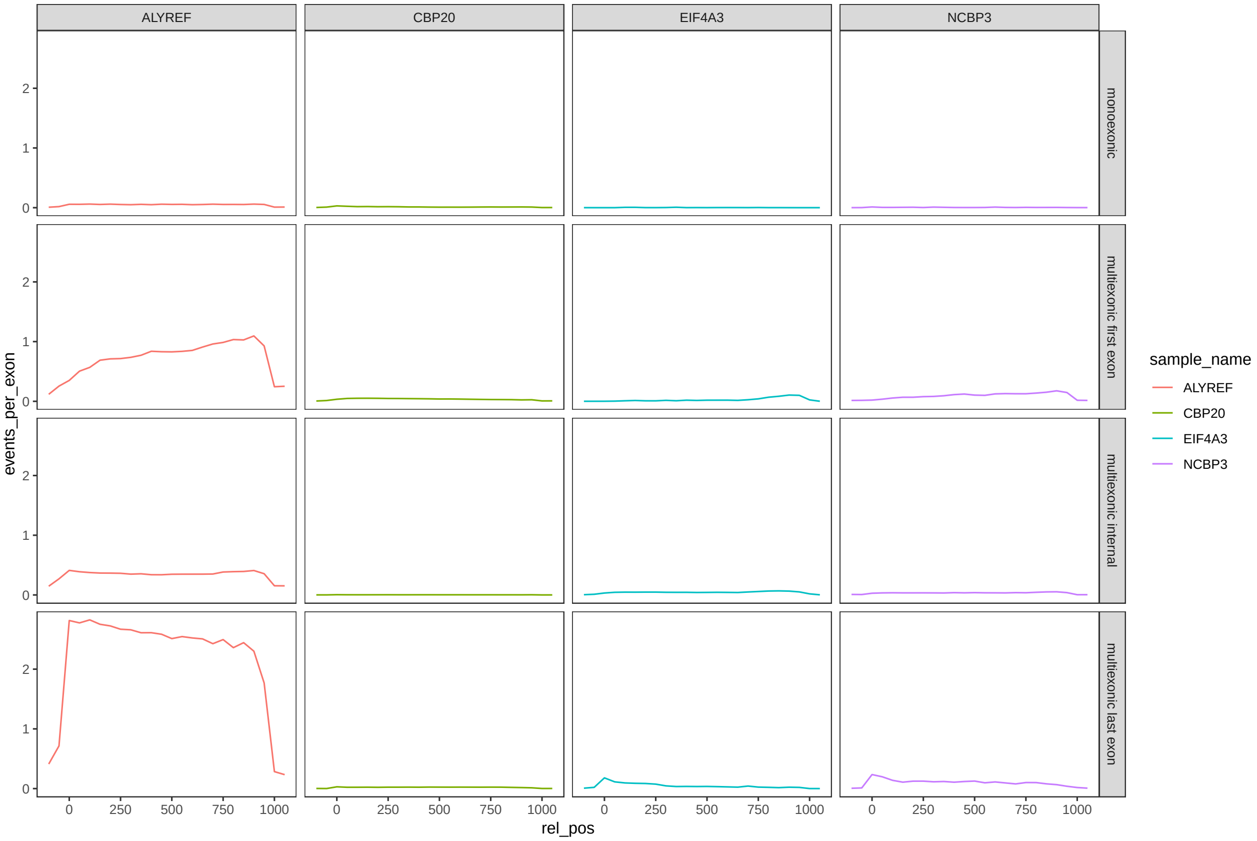
first vs internal vs last

```
df %>%
  metaplot_perclass
```



mono vs multiexonic only exons g200nt

```
df %>%
  filter(width > 200) %>%
  metaplot_perclass
```

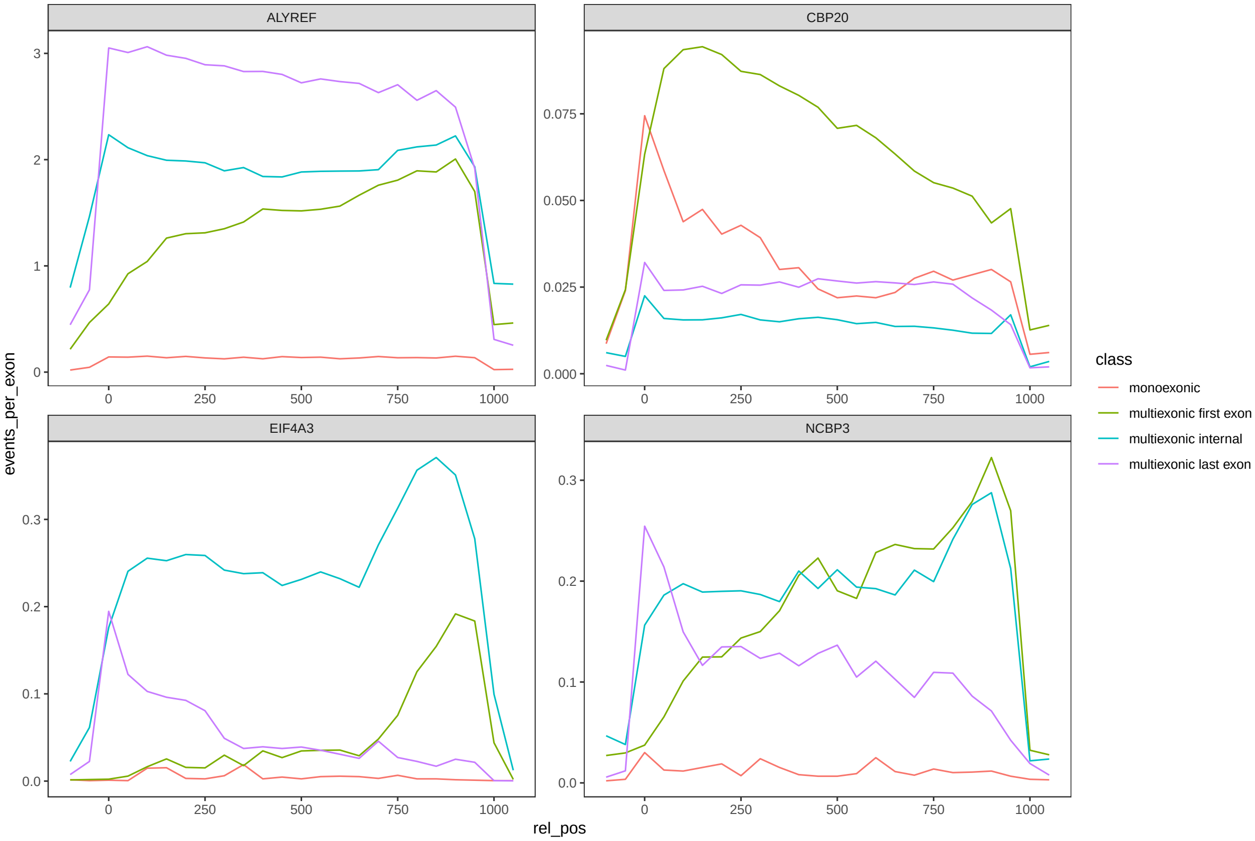


as overlay for paper:

```
(exons_per_classg200 <- table(filter(exon_anno_tbl, width > 200) %>%
                                distinct(GeneID, ExonID, class) %$%
                                class))
```

```
##
##             monoexonic multiexonic first exon   multiexonic internal
##                   1961                  12823                  38628
##  multiexonic last exon
##                  21673
```

```
df %>%
  filter(width > 200) %>%
  metaplot_perclass(., overlay = T, exons_per_classg200)
```



sessionInfo

```
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] parallel  stats4    stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] bindrcpp_0.2.2       RMetaTools_0.1       jsonlite_1.5
##  [4] rtracklayer_1.40.3   GenomicRanges_1.32.3 GenomeInfoDb_1.16.0
##  [7] IRanges_2.14.10      S4Vectors_0.18.3     BiocGenerics_0.26.0
## [10] broom_0.4.4          knitr_1.20           magrittr_1.5
## [13] forcats_0.3.0        stringr_1.3.1        dplyr_0.7.5
## [16] purrr_0.2.5          readr_1.1.1          tidyr_0.8.1
## [19] tibble_1.4.2         ggplot2_3.1.0        tidyverse_1.2.1
##
## loaded via a namespace (and not attached):
##  [1] Biobase_2.40.0              httr_1.3.1
##  [3] modelr_0.1.2                assertthat_0.2.0
##  [5] GenomeInfoDbData_1.1.0      cellranger_1.1.0
##  [7] Rsamtools_1.32.0            yaml_2.1.19
##  [9] pillar_1.2.3                backports_1.1.2
## [11] lattice_0.20-35             glue_1.2.0
## [13] digest_0.6.15               XVector_0.20.0
## [15] rvest_0.3.2                 colorspace_1.3-2
## [17] htmltools_0.3.6             Matrix_1.2-14
## [19] plyr_1.8.4                  psych_1.8.4
## [21] XML_3.98-1.11               pkgconfig_2.0.1
## [23] haven_1.1.1                 zlibbioc_1.26.0
## [25] scales_0.5.0                BiocParallel_1.14.1
## [27] withr_2.1.2                 SummarizedExperiment_1.10.1
## [29] lazyeval_0.2.1              cli_1.0.0
## [31] mnormt_1.5-5                crayon_1.3.4
## [33] readxl_1.1.0                evaluate_0.10.1
## [35] nlme_3.1-137                xml2_1.2.0
## [37] foreign_0.8-70              tools_3.5.0
## [39] hms_0.4.2                   matrixStats_0.53.1
## [41] munsell_0.5.0               DelayedArray_0.6.0
## [43] Biostrings_2.48.0           compiler_3.5.0
## [45] rlang_0.2.1                 grid_3.5.0
## [47] RCurl_1.95-4.10             rstudioapi_0.7
## [49] labeling_0.3                bitops_1.0-6
## [51] rmarkdown_1.10              gtable_0.2.0
## [53] reshape2_1.4.3              R6_2.2.2
## [55] GenomicAlignments_1.16.0    lubridate_1.7.4
## [57] utf8_1.1.4                  bindr_0.1.1
## [59] rprojroot_1.3-2             stringi_1.2.3
## [61] Rcpp_0.12.17                tidyselect_0.2.4
```