metagene of genes

Manfred Schmid

Contents

03 July, 2020; 15:43

Setup

```
knitr::opts_chunk$set(fig.width=12, fig.height=8,
                      fig.path=paste0('../Figures/CLIP_genes/'),
                      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_tr_annotations.RData', verbose=T)
```

```
## Loading objects:
##   tr_anno_tbl
```

```
tr_anno_tbl
```

```
## # A tibble: 31,756 x 7
##    GeneID    exon_cnt tr_exons_width ENSEMBL  SYMBOL  gene_name gene_type
##    <chr>        <int>          <int> <chr>    <chr>   <chr>     <chr>
##  1 100287102        3           1649 ENSG000~ DDX11L1 DDX11L1   transcrib~
##  2 653635          11           1758 ENSG000~ WASH7P  WASH7P    unprocess~
##  3 102466751        1             67 ENSG000~ MIR685~ MIR6859-1 miRNA
##  4 100302278        1            137 ENSG000~ MIR130~ MIR1302-2 miRNA
```

```
##  5 645520           3           1127 ENSG000~ FAM138A FAM138A   lincRNA
##  6 79501            1            917 ENSG000~ OR4F5   OR4F5     protein_c~
##  7 729737           3           5471 <NA>     LOC729~ <NA>      <NA>
##  8 102725121        4           1169 <NA>     LOC102~ <NA>      <NA>
##  9 102723897       11           2109 ENSG000~ WASH9P  RP11-34P~ unprocess~
## 10 102465909        1             67 ENSG000~ MIR685~ MIR6859-2 miRNA
## # ... with 31,746 more rows
```

```
tr_anno_tbl %<>%
  filter(gene_type == 'protein_coding') %>%
  dplyr::select(GeneID, exon_cnt, tr_exons_width) %>%
  mutate(class = ifelse(exon_cnt == 1, 'monoexonic', 'multiexonic'))
```

```
pc_ids <- unique(tr_anno_tbl$GeneID)

length(pc_ids)
```

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

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_metagene.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"

#sed 1d $anno | sort -k2,2 -k1,1 -k3,3n | \
#awk '{
#   if(gene_id == $1){
#     starts=starts","($3-start-1)
#     sizes=sizes","($4-$3+1)
#     end=$4
#     n+=1
#   }else{
#     if(gene_id != ""){
#       print chr"\t"start"\t"end"\t"gene_id"\t0\t"strand"\t"start"\t"end"\t255,0,0\t"n"\t"sizes"\t"sta
#     }
#     gene_id=$1; chr=$2; start=$3-1; end=$4; strand=$5; n=1;
#     starts="0"; sizes=($4-$3+1)
#   }
# }END{print chr"\t"start"\t"end"\t"gene_id"\t0\t"strand"\t"start"\t"end"\t255,0,0\t"n"\t"sizes"\t"star
```

```
bed12="/home/schmidm/ms_tools/subread-2.0.0-Linux-x86_64/annotation/hg38_RefSeq_exon.bed"

#awk '{if($6 =="+"){print $0}}' $bed12 > ${bed12/.bed/_plus.bed}
#awk '{if($6 =="-"){print $0}}' $bed12 > ${bed12/.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 ${bed12/.bed/_plus.bed}
```

load to R

```
fname <- '/Volumes/GenomeDK/faststorage/CLIP/CLIPdb/scripts/deeptools_subReadanno_metagene_scaled.gz'

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

```
(df %<>%
   filter(id %in% pc_ids) %>%
    dplyr::mutate(sample_name = sub('.*\\/', '', sample_name) %>%
                          sub('_plus_hg38.bw', '', .)) %>%
  dplyr::select(id, sample_name, rel_pos, value) %>%
   dplyr::mutate(GeneID = as.character(id)) %>%
   left_join(., tr_anno_tbl))
```

```
## # A tibble: 293,794 x 8
##    id     sample_name   rel_pos value GeneID exon_cnt tr_exons_width class
##    <fct>  <chr>           <dbl> <dbl> <chr>     <int>          <int> <chr>
##  1 2784   C17orf85_PAR~   -1000 0.839 2784         10           1904 mult~
##  2 10847  C17orf85_PAR~   -1000 0.932 10847        34          10440 mult~
##  3 23644  C17orf85_PAR~   -1000 0.869 23644        29           4782 mult~
##  4 55009  C17orf85_PAR~   -1000 0.999 55009         3            892 mult~
##  5 153918 C17orf85_PAR~   -1000 0.756 153918        8            984 mult~
##  6 8349   C17orf85_PAR~   -1000 1.000 8349          1           2222 mono~
##  7 124923 C17orf85_PAR~   -1000 0.865 124923       12           3767 mult~
##  8 79132  C17orf85_PAR~   -1000 0.736 79132        14           2665 mult~
##  9 4358   C17orf85_PAR~   -1000 0.953 4358         10           3792 mult~
## 10 80115  C17orf85_PAR~   -1000 0.944 80115        20           3643 mult~
## # ... with 293,784 more rows
```

save

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

saveRDS(ncbp3, file='../data/NCBP3_CLIP_gene_metagene.rds')
```

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

saveRDS(eif4a3, file='../data/EIF4A3_CLIP_gene_metagene.rds')
```

NCBP2 from Giacometti et al

deeptools run

```
#!/bin/sh
##cd /home/schmidm/faststorage/CLIP/Giacometti_GSE94427/scripts
##sbatch --account=thj_common --mem=4g deeptools_subReadanno_metagene.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"

#sed 1d $anno | sort -k2,2 -k1,1 -k3,3n | \
#awk '{
#   if(gene_id == $1){
#     starts=starts","($3-start-1)
#     sizes=sizes","($4-$3+1)
#     end=$4
#     n+=1
#   }else{
#     if(gene_id != ""){
#       print chr"\t"start"\t"end"\t"gene_id"\t0\t"strand"\t"start"\t"end"\t255,0,0\t"n"\t"sizes"\t"sta
#     }
#     gene_id=$1; chr=$2; start=$3-1; end=$4; strand=$5; n=1;
#     starts="0"; sizes=($4-$3+1)
#   }
# }END{print chr"\t"start"\t"end"\t"gene_id"\t0\t"strand"\t"start"\t"end"\t255,0,0\t"n"\t"sizes"\t"star

bed12="/home/schmidm/ms_tools/subread-2.0.0-Linux-x86_64/annotation/hg38_RefSeq_exon.bed"

#awk '{if($6 =="+"){print $0}}' $bed12 > ${bed12/.bed/_plus.bed}
#awk '{if($6 =="-"){print $0}}' $bed12 > ${bed12/.bed/_minus.bed}

plus_bw=$(ls /home/schmidm/faststorage/CLIP/Giacometti_GSE94427/hg38/*_plus_hg38.bw | tr "\n" " ")
minus_bw=${plus_bw//_plus_hg38.bw/_minus_hg38.bw}

python ~/ms_tools/MS_Metagene_Tools/computeMatrixStranded.pyc scale-regions -Rp ${bed12/.bed/_plus.bed}
```

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

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

```
(cbp20 <- df %>%
   filter(grepl('CBP20', sample_name), id %in% pc_ids) %>%
    dplyr::mutate(sample_name = sub('.*GSM......._', '', sample_name) %>%
                          sub('_norm_plus_hg38.bw', '', .)) %>%
  dplyr::select(id, sample_name, rel_pos, value) %>%
   dplyr::mutate(GeneID = as.character(id)) %>%
   left_join(., tr_anno_tbl))
```

```
## # A tibble: 43,429 x 8
##    id     sample_name rel_pos value GeneID exon_cnt tr_exons_width class
##    <fct>  <chr>         <dbl> <dbl> <chr>     <int>          <int> <chr>
##  1 149478 CBP20_1       -1000  10.4 149478        8           2978 multie~
##  2 7389   CBP20_1       -1000  17.2 7389         10           1467 multie~
##  3 374973 CBP20_1       -1000  10.4 374973        3            991 multie~
```

```
##  4 148362 CBP20_1       -1000  52.0 148362       16           4994 multie~
##  5 116841 CBP20_1       -1000  10.4 116841        7           2786 multie~
##  6 11155  CBP20_1       -1000  10.4 11155        19           6746 multie~
##  7 54838  CBP20_1       -1000  10.4 54838         8           4723 multie~
##  8 118426 CBP20_1       -1000  20.8 118426        5           5906 multie~
##  9 51005  CBP20_1       -1000  20.8 51005        11           2272 multie~
## 10 51073  CBP20_1       -1000  10.4 51073         8           2396 multie~
## # ... with 43,419 more rows
```

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(id, sample_name, rel_pos, GeneID, exon_cnt, tr_exons_width, class) %>%
  summarize(value = sum(value)/2)
```

save cbp20 data

```
saveRDS(cbp20, file='../data/CBP20_CLIP_gene_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_metagene.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"

#sed 1d $anno | sort -k2,2 -k1,1 -k3,3n | \
#awk '{
#   if(gene_id == $1){
#     starts=starts","($3-start-1)
#     sizes=sizes","($4-$3+1)
#     end=$4
#     n+=1
#   }else{
#     if(gene_id != ""){
#       print chr"\t"start"\t"end"\t"gene_id"\t0\t"strand"\t"start"\t"end"\t255,0,0\t"n"\t"sizes"\t"sta
#     }
#     gene_id=$1; chr=$2; start=$3-1; end=$4; strand=$5; n=1;
#     starts="0"; sizes=($4-$3+1)
#   }
```

```
# }END{print chr"\t"start"\t"end"\t"gene_id"\t0\t"strand"\t"start"\t"end"\t255,0,0\t"n"\t"sizes"\t"star

bed12="/home/schmidm/ms_tools/subread-2.0.0-Linux-x86_64/annotation/hg38_RefSeq_exon.bed"

#awk '{if($6 =="+"){print $0}}' $bed12 > ${bed12/.bed/_plus.bed}
#awk '{if($6 =="-"){print $0}}' $bed12 > ${bed12/.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 ${bed12/.bed/_plus.bed}
```

```
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), id %in% pc_ids) %>%
    dplyr::mutate(sample_name = sub('.*GSE113896_', '', sample_name) %>%
                        sub('-union_hg38', '', .)) %>%
  dplyr::select(id, sample_name, rel_pos, value) %>%
   dplyr::mutate(GeneID = as.character(id)) %>%
   left_join(., tr_anno_tbl))
```

```
## # A tibble: 156,560 x 8
##    id     sample_name rel_pos value GeneID exon_cnt tr_exons_width class
##    <fct>  <chr>         <dbl> <dbl> <chr>     <int>          <int> <chr>
##  1 3151   Alyref-FLAG   -1000  1    3151          6           1963 multie~
##  2 149069 Alyref-FLAG   -1000  1    149069        9           1340 multie~
##  3 79729  Alyref-FLAG   -1000  1.5  79729        18           2905 multie~
##  4 6487   Alyref-FLAG   -1000  1    6487         26           4567 multie~
##  5 149473 Alyref-FLAG   -1000  1    149473        9           1454 multie~
##  6 5876   Alyref-FLAG   -1000  2    5876         11           1698 multie~
##  7 1945   Alyref-FLAG   -1000  1    1945          4           1253 multie~
##  8 23623  Alyref-FLAG   -1000  1    23623        10           4931 multie~
##  9 51093  Alyref-FLAG   -1000  1    51093         8           2274 multie~
## 10 2207   Alyref-FLAG   -1000  1.25 2207          5            586 multie~
## # ... with 156,550 more rows
```

save aly data

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

```
rm(df)
```

alternative starting point

```
ncbp3 <- readRDS('../data/NCBP3_CLIP_gene_metagene.rds')

ncbp3$sample_name <- 'NCBP3'
```

```
eif4a3 <- readRDS('../data/EIF4A3_CLIP_gene_metagene.rds')

eif4a3$sample_name <- 'EIF4A3'
```

```
cbp20 <- readRDS('../data/CBP20_CLIP_gene_metagene.rds')
```

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

aly$sample_name <- 'ALYREF'
```

combine

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

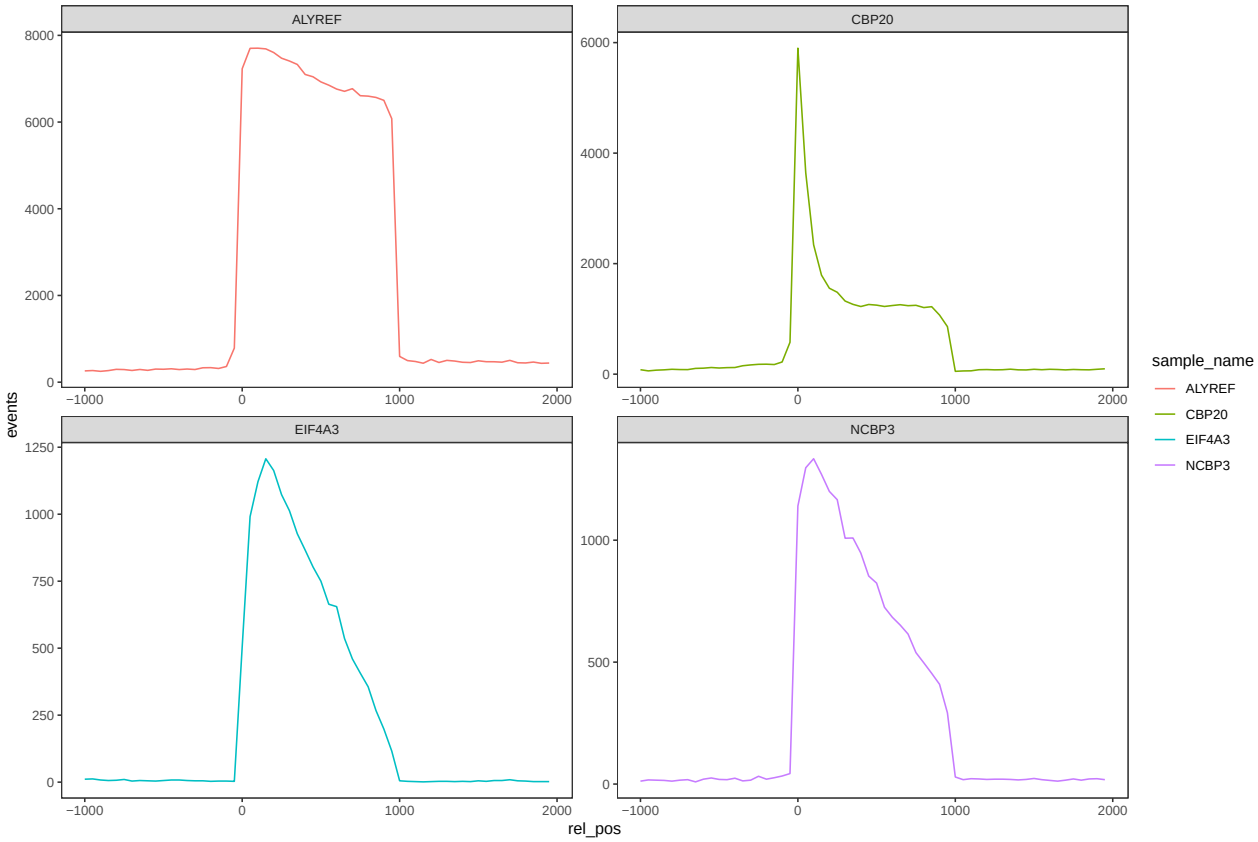
Plots

plot fun

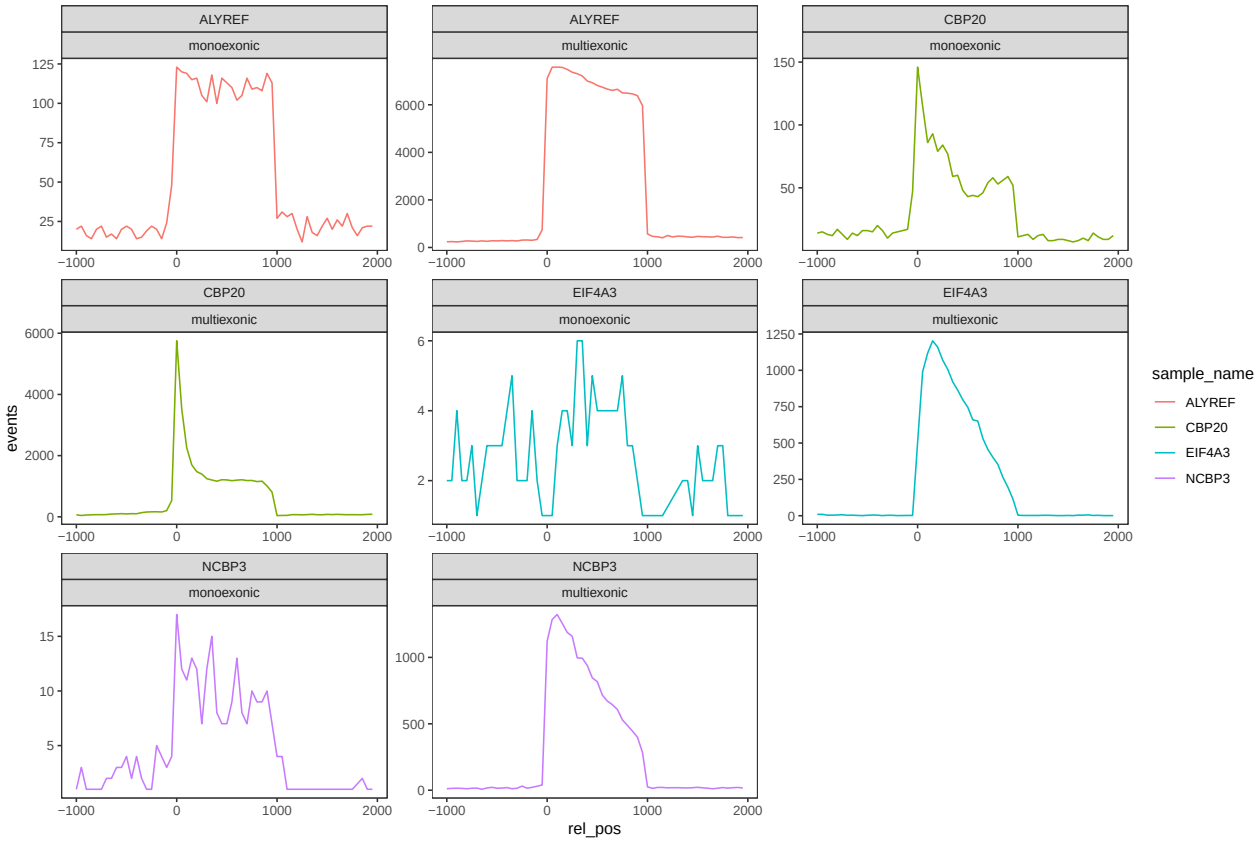
```
metaplot_all <- function(df) {
  df %>%
    group_by(sample_name, rel_pos) %>%
    summarize(events=n()) %>%
    ggplot(., aes(x=rel_pos, y=events, color=sample_name)) +
    geom_line() +
    facet_wrap(~sample_name, scales='free') +
    theme_bw() +
    theme(panel.grid=element_blank())
}
```

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

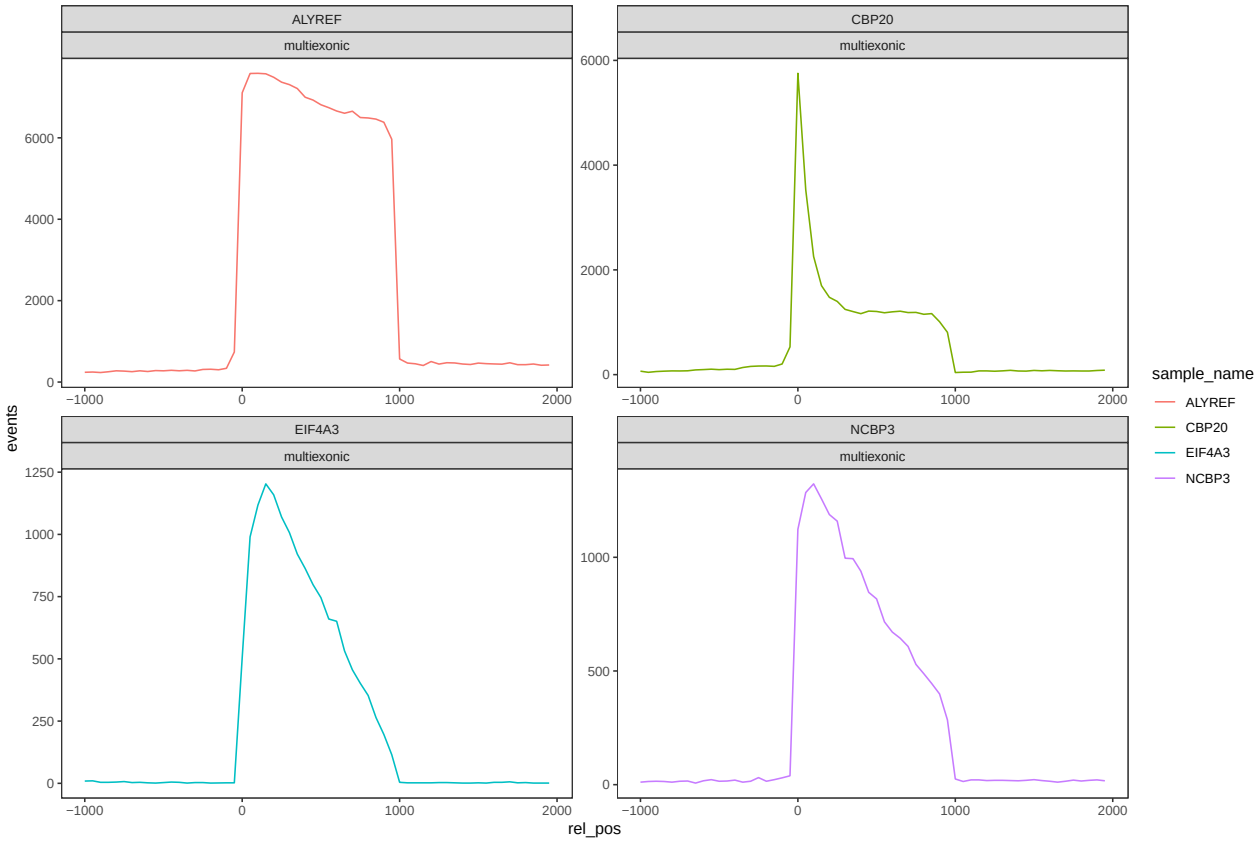
```
metaplot_all(df)
```



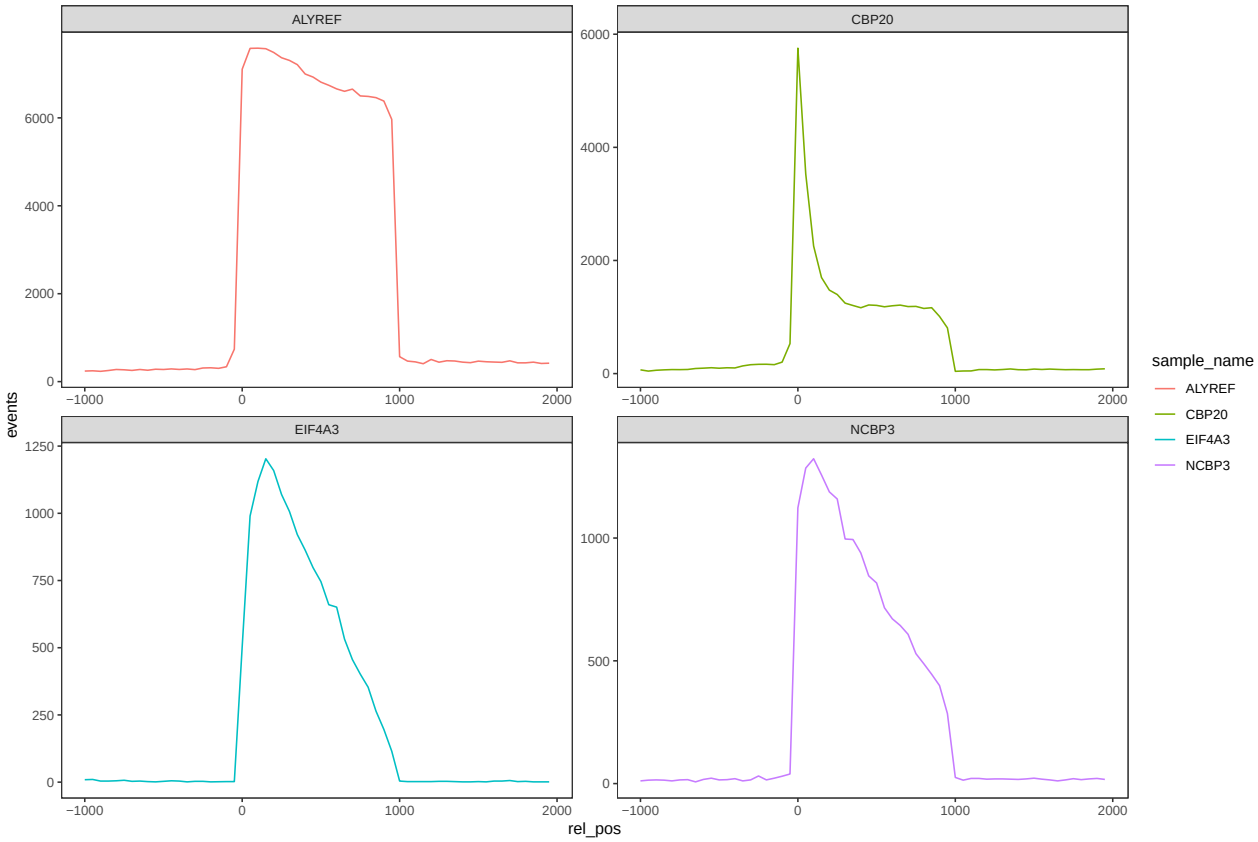

```
metaplot_perclass(df)
```




```
df %>%
  filter(class == 'multiexonic', tr_exons_width > 200) %>%
  metaplot_perclass
```




```
df %>%
  filter(class == 'multiexonic', tr_exons_width > 200) %>%
  metaplot_all
```



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
```