Figure 3D-F

RAC

24/08/2020

packages and filepaths

```
suppressMessages(library(dplyr))
suppressMessages(library(tidyr))
suppressMessages(library(ggplot2))
COUNTS="../../data/xiCLIP_all_5primepos.rRNAScaled.hg38_HeLa_trimmed_loci_major_primary_isoform_annotate
EXPRESSION_VECTOR_FILEPATH="../../data/log2_mean_cov_RNAseq_TTseq.RData"
ANNOTATION_BED_FILEPATH="../../data/hg38_HeLa_trimmed_loci_major_primary_isoform_annotated.exonNumber.s
```

load and wrangle data

```
#Load and wrangle expression vector -----------------------------------

load(EXPRESSION_VECTOR_FILEPATH)

expression_vector<-left_join(
  (as.data.frame(ctrl_RNAseq_expr) %>%
     add_rownames(var = "geneID")),
  (as.data.frame(ctrl_TTseq_expr) %>%
     add_rownames(var = "geneID"))
) %>%
  mutate(ctrl_RNAseq_expr = case_when(
    ctrl_RNAseq_expr ==0 ~ min(ctrl_RNAseq_expr[ctrl_RNAseq_expr > 0]),
    TRUE ~  ctrl_RNAseq_expr
  ))
```

```
## Warning: `add_rownames()` was deprecated in dplyr 1.0.0.
## i Please use `tibble::rownames_to_column()` instead.
```

```
## Joining, by = "geneID"
```

```
#load and wrangle annobed -----------------------

annoBed<-read.table(ANNOTATION_BED_FILEPATH, sep = "\t", header = F) %>%
setNames(c("chr", "start", "end", "geneID", "score", "strand")) %>%
    separate(geneID, into = c("geneID", "Biotype", "ExonNumber", "TotalNumberOfExons", "ExonSize", "Exon
  mutate_at(vars(ExonDistFromTSS,ExonicDistance,ExonSize,TotalNumberOfExons,ExonNumber), .funs = as.nume

#load and wrangle count file -----------------

counts<-
read.table(COUNTS, sep = "\t", header = F) %>%
  setNames(c("Sample","chr", "start", "end", "geneID", "DistToLandmark", "strand", "count")) %>%
    separate(geneID, into = c("geneID", "Biotype", "ExonNumber", "TotalNumberOfExons", "ExonSize", "Exon
    mutate_at(vars(ExonDistFromTSS,ExonicDistance,ExonSize,TotalNumberOfExons,ExonNumber), .funs = as.nu
```

```r
  filter(!grepl("CBP20_3", Sample))

head(expression_vector)
```

```
## # A tibble: 6 x 3
##   geneID             ctrl_RNAseq_expr ctrl_TTseq_expr
##   <chr>                         <dbl>           <dbl>
## 1 RN7SK                          12.9            4.20
## 2 RMRP,RNase_MRP                 12.5            3.12
## 3 Metazoa_SRP,RN7SL1             12.4            4.55
## 4 GAPDH                          10.7           10.1 
## 5 EEF1A1                         10.6           10.3 
## 6 TMSB10                         10.8            9.42
```

```r
head(annoBed)
```

```
##   chr   start     end geneID        Biotype ExonNumber TotalNumberOfExons
## 1   X 3608624 3608945   PRKX protein_coding          9                  9
## 2   X 3612176 3612325   PRKX protein_coding          8                  9
## 3   X 3615814 3615892   PRKX protein_coding          7                  9
## 4   X 3621258 3621316   PRKX protein_coding          6                  9
## 5   X 3626418 3626514   PRKX protein_coding          5                  9
## 6   X 3641851 3641971   PRKX protein_coding          4                  9
##   ExonSize ExonicDistance ExonDistFromTSS ExonStature
## 1      321           1817          104704   majorExon
## 2      149           1496          101324   majorExon
## 3       78           1347           97757   majorExon
## 4       58           1269           92333   majorExon
## 5       96           1211           87135   majorExon
## 6      120           1115           71678   majorExon
##                   GeneStructure score strand
## 1     multiExonicGene-lastExon     .      -
## 2 multiExonicGene-internalExon     .      -
## 3 multiExonicGene-internalExon     .      -
## 4 multiExonicGene-internalExon     .      -
## 5 multiExonicGene-internalExon     .      -
## 6 multiExonicGene-internalExon     .      -
```

```r
head(counts)
```

```
##                         Sample chr  start    end   geneID        Biotype
## 1 ALYREF_1_DMSO_5primepos_3end   1 184977 184978 NA.v1000 protein_coding
## 2 ALYREF_1_DMSO_5primepos_3end   1 184997 184998 NA.v1000 protein_coding
## 3 ALYREF_1_DMSO_5primepos_3end   1 185002 185003 NA.v1000 protein_coding
## 4 ALYREF_1_DMSO_5primepos_3end   1 185020 185021 NA.v1000 protein_coding
## 5 ALYREF_1_DMSO_5primepos_3end   1 185024 185025 NA.v1000 protein_coding
## 6 ALYREF_1_DMSO_5primepos_3end   1 189096 189097  NA.v999     intergenic
##   ExonNumber TotalNumberOfExons ExonSize ExonicDistance ExonDistFromTSS
## 1          3                  3      635           1990            2289
## 2          3                  3      635           1990            2289
## 3          3                  3      635           1990            2289
## 4          3                  3      635           1990            2289
## 5          3                  3      635           1990            2289
## 6          1                  1     2655           2655               0
##   ExonStature            GeneStructure DistToLandmark strand   count
## 1   majorExon multiExonicGene-lastExon           -53      - 8.77321
## 2   majorExon multiExonicGene-lastExon           -73      - 4.38661
## 3   majorExon multiExonicGene-lastExon           -78      - 4.38661
## 4   majorExon multiExonicGene-lastExon           -96      - 4.38661
## 5   majorExon multiExonicGene-lastExon          -100      - 4.38661
## 6   majorExon         singleExonicGene            97      - 4.38661
```

process data

```
#normalise counts to expression of the gene
norm_counts_to_gene_expression<-
  counts %>%
    left_join(expression_vector) %>%
    #this replaces NAs introduced by no value present in expression_vector, and replaces them with min
    mutate_at(vars(ctrl_RNAseq_expr), ~replace(., is.na(.), min(expression_vector$ctrl_RNAseq_expr))) %
    mutate(norm_count = count/ctrl_RNAseq_expr)
```

```
## Joining, by = "geneID"
```

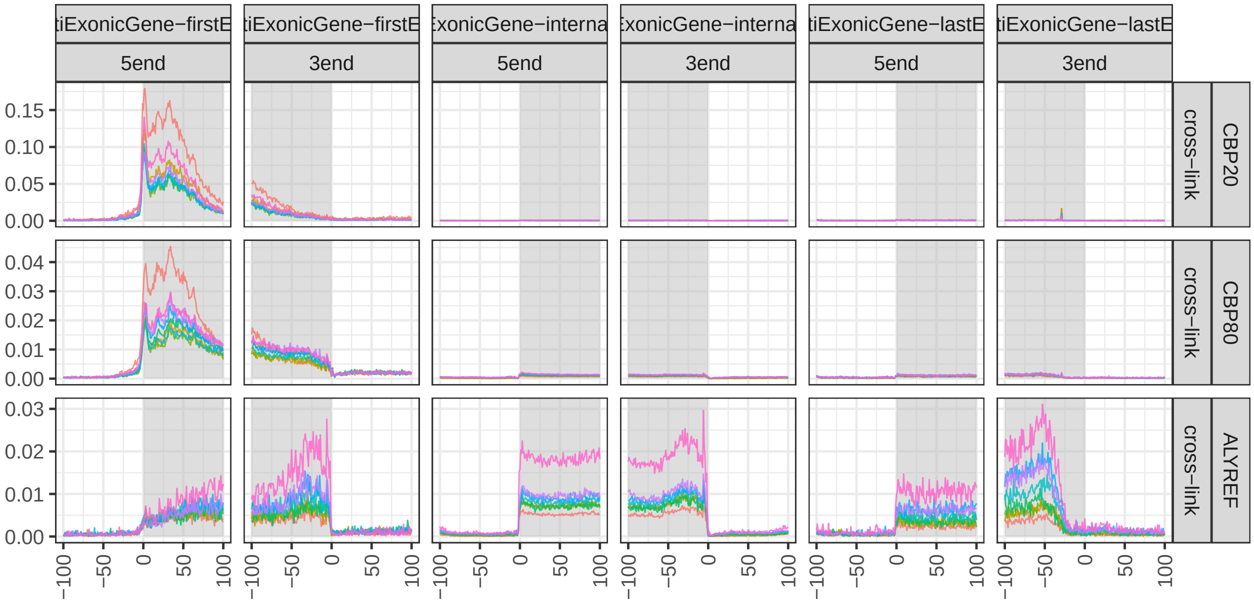
```
#only select multiexonic genes
gene_annotations_to_use<-
annoBed %>%
  left_join(expression_vector) %>%
  filter(TotalNumberOfExons > 1 & !grepl("snRNA|rRNA|TR_C_gene|IG_C_pseudogene|miRNA|misc_RNA", Biotype
  select(geneID, ctrl_RNAseq_expr, ctrl_TTseq_expr) %>%
  unique() %>%
  mutate(ctrl_RNAseq_expr = as.numeric(ctrl_RNAseq_expr)) %>%
  filter(ctrl_RNAseq_expr > 1 & geneID != "LINC00324")
```

```
## Joining, by = "geneID"
```

```
#number of annotations used for analysis
exon_annotations_pos_n<-
  annoBed %>%
  filter(geneID %in% gene_annotations_to_use$geneID & as.numeric(ExonSize) >99) %>%
  group_by(GeneStructure) %>%
  summarise(anno_count =n())

#number of TUs
GENECOUNT<-
   annoBed %>%
  filter(geneID %in% gene_annotations_to_use$geneID) %>%
  select(geneID) %>%
  unique() %>%
  summarise(geneCount =n())

#select genes
data_for_fig_3d_f<-
norm_counts_to_gene_expression %>%
  #filter for genes selected above
  filter(geneID %in% gene_annotations_to_use$geneID & as.numeric(ExonSize) >99) %>%
  group_by(Sample,GeneStructure, DistToLandmark) %>%
  summarise(sum_RNAseq_norm_count_norm_annotation_number = sum(norm_count)) %>%
  left_join(exon_annotations_pos_n) %>%
  #normalise to number of annotations
  mutate(sum_RNAseq_norm_count_norm_annotation_number = sum_RNAseq_norm_count_norm_annotation_number/ann
  #format data frame for plotting
  separate(Sample, c("Protein","Rep","Timepoint","readType","region"), sep = "_") %>%
  filter(Timepoint != "negative") %>%
  mutate(Timepoint_f = case_when(
    Timepoint == "PBSDRB" ~ "t00",
    TRUE ~ Timepoint )) %>%
  mutate(region = factor(region, levels = c("5end","3end")),
         Timepoint = factor(Timepoint, levels = c("PBSDRB", "t00", "t05", "t10", "t15", "t20", "t40", "
         Timepoint_f = factor(Timepoint_f, levels = c("t00", "t05", "t10", "t15", "t20", "t40", "t60",
         Protein = factor(Protein, levels = c("CBP20", "CBP80", "ALYREF")),
         readType = gsub("5primepos","cross-link", readType))
```

```
## `summarise()` has grouped output by 'Sample', 'GeneStructure'. You can override
## using the `.groups` argument.
## Joining, by = "GeneStructure"
```

```
fig3_d_f<-
data_for_fig_3d_f %>%
  filter(Protein %in% c("ALYREF","CBP20","CBP80")) %>%
  ggplot() +
  geom_rect(data = data.frame(region = "3end"), aes(xmin = -100, xmax = 0, ymin = 0, ymax = Inf), alpha
  geom_rect(data = data.frame(region = "5end"), aes(xmin = 0, xmax = 100, ymin = 0, ymax = Inf), alpha =
  geom_line(aes(x=DistToLandmark, y = sum_RNAseq_norm_count_norm_annotation_number, col = Timepoint_f),
  facet_grid(Protein+readType ~ GeneStructure + region, scale = "free") +
  xlab("distance (nt)") +
  ylab("mean coverage")+
  theme_bw()

fig3_d_f +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1),
        legend.position = "none") +
  xlab("") +
  ylab("")
```



```
#ggsave("figure_3_d-f.pdf", height = 3, width =7)
```

zz