CLIP exon/intron density plots in figure 3 ab, and supp. fig. 3b

RAC

20/08/2020

```
knitr::opts_chunk$set(warning=FALSE, message=FALSE, tidy.opts = list(width.cutoff = 60), tidy = TRUE)

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)

#filepaths
exonIntronReadsFP="../../data/xiCLIP_intron_exclusiveExon.200820.counts"
totalCountsFP="../../data/xiCLIP.read2.totalcounts.200402.tab"
exonAnnotationFP="../../data/hg38_HeLa_trimmed_loci_major_primary_isoform_annotated_exon_numbered.bed"
intronAnnotationFP="../../data/hg38_HeLa_trimmed_loci_major_primary_isoform_annotated_intron_numbered.b
rRNAScalingFP="../../data/rRNAFactor.tab"
```

#load count tables, dataframes containing library scaling and annotation files

```
# data load
exonIntronReads <- read.csv(exonIntronReadsFP, sep = "\t", header = F)
totalCounts <- read.csv(totalCountsFP, sep = "\t", header = F)

colnames(exonIntronReads) <- c("Sample", "chr", "start", "end",
    "ID", "segmentNumber", "strand", "count")
colnames(totalCounts) <- c("Sample", "TotalCount")

# load rRNA scalings
libraryScalings <- read.csv(rRNAScalingFP, sep = " ", header = F)
colnames(libraryScalings) <- c("Sample", "scaling")

# load filepath
exonAnno <- read.table(exonAnnotationFP, header = F)
intronAnno <- read.table(intronAnnotationFP, header = F)

colnames(exonAnno) <- c("chr", "start", "end", "ID", "segmentNumber",
```

```
    "strand")
colnames(intronAnno) <- c("chr", "start", "end", "ID", "segmentNumber",
    "strand")

head(exonIntronReads)
```

```
##                Sample chr  start    end                     ID segmentNumber
## 1 ALYREF_1_DMSO_read2   1 945653 946172 NOC2L:::protein_coding     intron_16
## 2 ALYREF_1_DMSO_read2   1 946545 948130 NOC2L:::protein_coding     intron_14
## 3 ALYREF_1_DMSO_read2   1 948603 951126 NOC2L:::protein_coding     intron_12
## 4 ALYREF_1_DMSO_read2   1 951238 951999 NOC2L:::protein_coding     intron_11
## 5 ALYREF_1_DMSO_read2   1 953892 954003 NOC2L:::protein_coding      intron_7
## 6 ALYREF_1_DMSO_read2   1 954082 955922 NOC2L:::protein_coding      intron_6
##   strand count
## 1      -     1
## 2      -     1
## 3      -     1
## 4      -     1
## 5      -     1
## 6      -     4
```

```
head(totalCounts)
```

```
##                    Sample TotalCount
## 1     ALYREF_1_DMSO_read2     682096
## 2 ALYREF_1_negative_read2      11853
## 3   ALYREF_1_PBSDRB_read2     217702
## 4      ALYREF_1_t00_read2     383332
## 5      ALYREF_1_t05_read2     410081
## 6      ALYREF_1_t10_read2     582600
```

```
head(libraryScalings)
```

```
##           Sample  scaling
## 1    CBP80_3_t05 1.655995
## 2    CBP80_3_t00 1.699813
## 3 CBP20_1_PBSDRB 1.729306
## 4   CBP80_3_DMSO 1.750802
## 5    CBP20_1_t00 1.974724
## 6    CBP20_1_t20 2.008973
```

Load all of the annotation files and wrange the count files

```
# make df containing: total exons,
totalExons <- exonAnno %>%
    separate(ID, into = c("GeneID", "Biotype"), sep = ":::") %>%
    separate(segmentNumber, into = c("Segment", "SegmentNumber")) %>%
    mutate(SegmentNumber = as.numeric(SegmentNumber)) %>%
    group_by(GeneID, Biotype, Segment) %>%
    summarise(TotalExons = max(SegmentNumber)) %>%
    mutate(Exonic = case_when(TotalExons > 1 ~ "multiExonic",
        TRUE ~ "monoExonic"))

# exon/intron sizes
segmentSizes <- rbind(exonAnno, intronAnno) %>%
    separate(ID, into = c("GeneID", "Biotype"), sep = ":::") %>%
```

```
    separate(segmentNumber, into = c("Segment", "SegmentNumber")) %>%
    mutate(SegmentNumber = as.numeric(SegmentNumber), end = as.numeric(end),
        start = as.numeric(start)) %>%
    group_by(Segment) %>%
    summarise(size = sum(end - start))


# wrangle counts table
eIRDf <- exonIntronReads %>%
    separate(Sample, into = c("Protein", "Rep", "Timepoint",
        "readType")) %>%
    separate(ID, into = c("GeneID", "Biotype"), sep = ":::") %>%
    separate(segmentNumber, into = c("Segment", "SegmentNumber")) %>%
    mutate(SegmentNumber = as.numeric(SegmentNumber)) %>%
    filter(!(Protein == "CBP20" & Rep == "3"))

# wrangle total counts for library and library scaling,
# then join
tcDf <- totalCounts %>%
    separate(Sample, into = c("Protein", "Rep", "Timepoint",
        "readType"))
libScale <- libraryScalings %>%
    separate(Sample, into = c("Protein", "Rep", "Timepoint"))
totalCountsDF <- merge(tcDf, libScale)

head(totalCountsDF)
```

```
##   Protein Rep Timepoint readType TotalCount   scaling
## 1  ALYREF   1      DMSO    read2     682096  4.386606
## 2  ALYREF   1  negative    read2      11853 82.872928
## 3  ALYREF   1    PBSDRB    read2     217702  8.841733
## 4  ALYREF   1       t00    read2     383332  4.082188
## 5  ALYREF   1       t05    read2     410081  4.774029
## 6  ALYREF   1       t10    read2     582600  3.360968
```

```
head(eIRDf)
```

```
##   Protein Rep Timepoint readType chr  start    end GeneID        Biotype
## 1  ALYREF   1      DMSO    read2   1 945653 946172  NOC2L protein_coding
## 2  ALYREF   1      DMSO    read2   1 946545 948130  NOC2L protein_coding
## 3  ALYREF   1      DMSO    read2   1 948603 951126  NOC2L protein_coding
## 4  ALYREF   1      DMSO    read2   1 951238 951999  NOC2L protein_coding
## 5  ALYREF   1      DMSO    read2   1 953892 954003  NOC2L protein_coding
## 6  ALYREF   1      DMSO    read2   1 954082 955922  NOC2L protein_coding
##   Segment SegmentNumber strand count
## 1  intron            16      -     1
## 2  intron            14      -     1
## 3  intron            12      -     1
## 4  intron            11      -     1
## 5  intron             7      -     1
## 6  intron             6      -     4
```

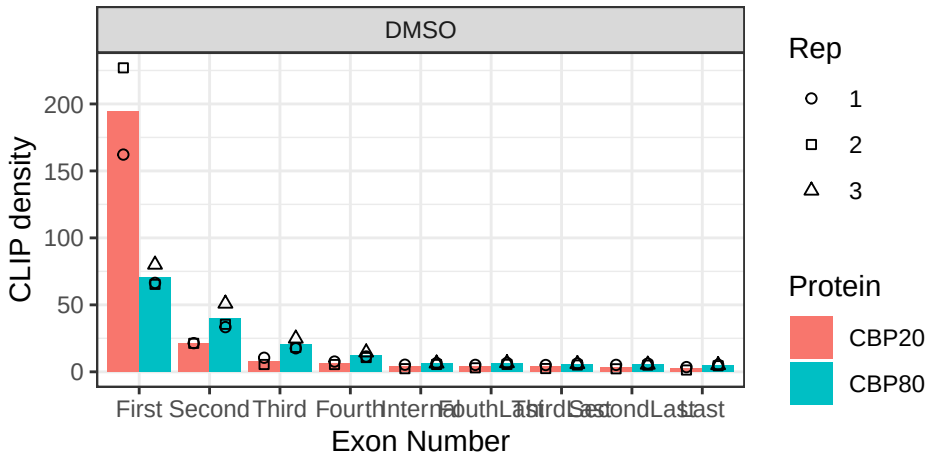
#wrangle datasets for CLIP density over first 4, last 4 and internal exons for Figure 3A-B, and Supplementary Figure 3B

```
# size of exons or introns, segment number = exon or intron
# number
segmentSizes <- rbind(exonAnno, intronAnno) %>%
    separate(ID, into = c("GeneID", "Biotype"), sep = ":::") %>%
    separate(segmentNumber, into = c("Segment", "SegmentNumber"),
        sep = "_") %>%
    mutate_at(vars(c("start", "end", "SegmentNumber")), funs(as.numeric)) %>%
    group_by(Segment, SegmentNumber, GeneID) %>%
    summarise(size = as.numeric(end - start))

# total number of exons or introns
totalSegments <- rbind(exonAnno, intronAnno) %>%
    separate(ID, into = c("GeneID", "Biotype"), sep = ":::") %>%
    separate(segmentNumber, into = c("Segment", "SegmentNumber"),
        sep = "_") %>%
    mutate_at(vars(c("start", "end", "SegmentNumber")), funs(as.numeric)) %>%
    group_by(Segment, GeneID, Biotype) %>%
    summarise(totalSegments = max(SegmentNumber))

# annotate position of exon or intron
metaDataForSegments <- totalSegments %>%
    left_join(segmentSizes) %>%
    mutate(Exonic = case_when(totalSegments > 1 ~ "multiExonic",
        TRUE ~ "monoExonic")) %>%
    mutate(ExonDescription = case_when(Exonic == "multiExonic" &
        totalSegments == SegmentNumber ~ "Last", Exonic == "multiExonic" &
        SegmentNumber == "1" ~ "First", Exonic == "multiExonic" &
        SegmentNumber == "2" ~ "Second", Exonic == "multiExonic" &
        SegmentNumber == "3" ~ "Third", Exonic == "multiExonic" &
        SegmentNumber == "4" ~ "Fourth", Exonic == "multiExonic" &
        SegmentNumber == (totalSegments - 4) ~ "FouthLast", Exonic ==
        "multiExonic" & SegmentNumber == (totalSegments - 2) ~
        "SecondLast", Exonic == "multiExonic" & SegmentNumber ==
        (totalSegments - 3) ~ "ThirdLast", totalSegments == 1 ~
        "monoExonic", TRUE ~ "Internal")) %>%
    mutate(ExonDescription = factor(ExonDescription, levels = c("First",
        "Second", "Third", "Fourth", "Internal", "FouthLast",
        "ThirdLast", "SecondLast", "Last")))

# number of annotations
annotation_number <- metaDataForSegments %>%
    filter(totalSegments > 8) %>%
    group_by(Exonic, ExonDescription, Segment) %>%
    summarise(number_of_annotations = n())

# remove excluded dataset, scale data to rRNA reads
eIRDf_filtered_scaled <- eIRDf %>%
    select(-c(chr, start, end, strand)) %>%
    left_join(metaDataForSegments) %>%
    filter(!(Protein == "CBP20" & Rep == "3")) %>%
    left_join(libScale) %>%
    mutate(scaled_counts = count * scaling, scaled_density = (count/(size/1000)) *
        scaling)
```

```
# generate dataframe for fig 3ab, selecting exons and genes
# > 8 exons long. Average density of reads per exon.
df_for_fig_3ab <- eIRDf_filtered_scaled %>%
    filter(Timepoint == "DMSO" & totalSegments > 8 & Segment ==
        "exon") %>%
    group_by(Protein, Rep, Timepoint, Segment, ExonDescription) %>%
    summarise(sum_scaled_density = sum(scaled_density)) %>%
    left_join(annotation_number) %>%
    mutate(sum_scaled_density_norm_to_anno_n = sum_scaled_density/number_of_annotations) %>%
    mutate(Timepoint_f = case_when(Timepoint == "PBSDRB" ~ "t00",
        TRUE ~ Timepoint)) %>%
    mutate(Timepoint_f = factor(Timepoint_f, levels = c("t00",
        "t05", "t10", "t15", "t20", "t40", "t60", "DMSO")))
```

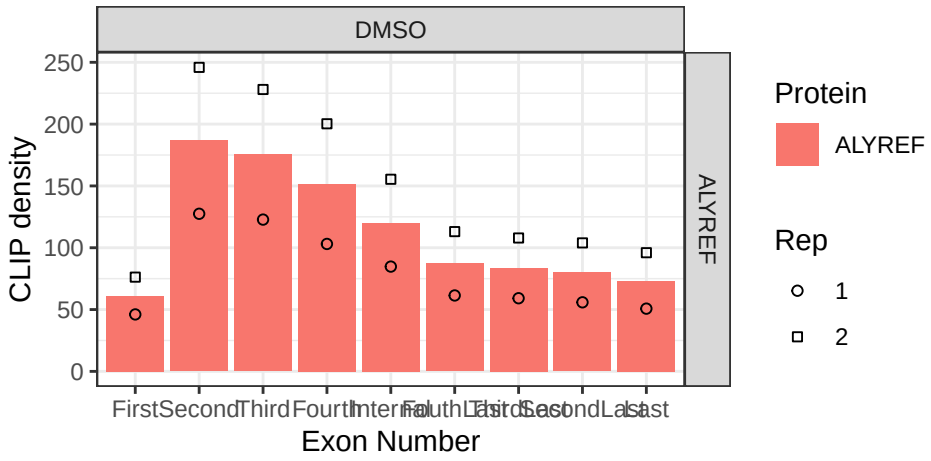
#Figure 3A

```
df_for_fig_3ab %>%
    filter(grepl("CBP", Protein)) %>%
    ggplot() + geom_bar(aes(x = ExonDescription, y = sum_scaled_density_norm_to_anno_n,
    fill = Protein), stat = "summary", fun = mean, position = position_dodge()) +
    geom_point(aes(x = ExonDescription, y = sum_scaled_density_norm_to_anno_n,
        shape = Rep, group = Protein), position = position_dodge(width = 0.9)) +
    scale_shape_manual(values = c(21, 22, 24)) + facet_grid(. ~
    Timepoint_f, scales = "free") + ylab("CLIP density") + xlab("Exon Number") +
    theme_bw()
```



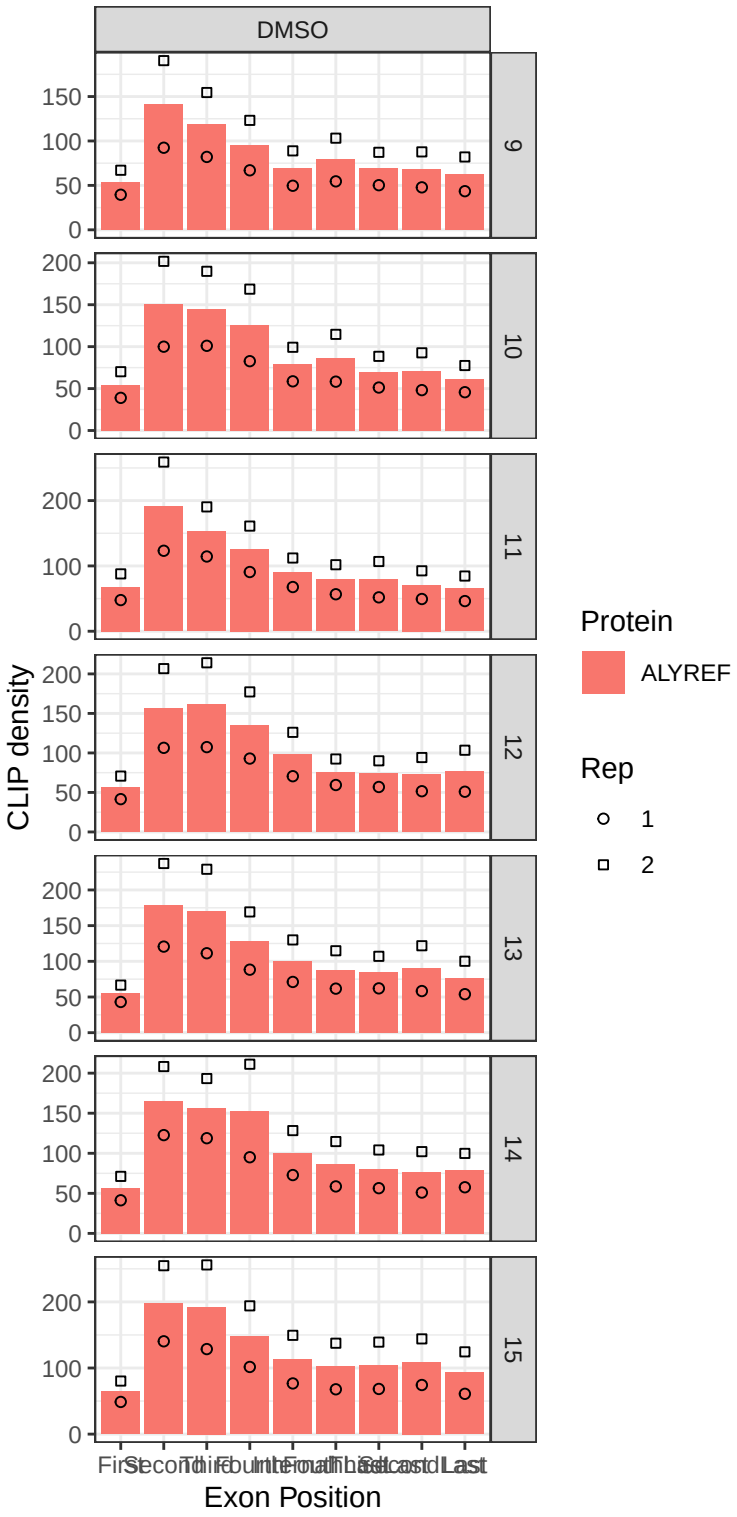
#Figure 3b

```
df_for_fig_3ab %>%
    filter(grepl("ALYREF", Protein)) %>%
    ggplot() + geom_bar(aes(x = ExonDescription, y = sum_scaled_density_norm_to_anno_n,
    fill = Protein), stat = "summary", fun = mean, position = position_dodge()) +
    geom_point(aes(x = ExonDescription, y = sum_scaled_density_norm_to_anno_n,
        shape = Rep, group = Protein), position = position_dodge(width = 0.9)) +
    scale_shape_manual(values = c(21, 22, 24)) + facet_grid(Protein ~
    Timepoint_f, scales = "free") + ylab("CLIP density") + xlab("Exon Number") +
    theme_bw()
```



#Supplementary Figure 3A

```
# additional helper dataframe to calculate number of
# annotations, stratified by total exon number, and
# position of exon.
annotation_number <- metaDataForSegments %>%
    group_by(Exonic, Segment, ExonDescription, totalSegments) %>%
    summarise(number_of_annotations = n())

# dataframe for supp fig 3b, select genes with 9-15 exons
# long
df_for_Supp_fig_3a <- eIRDf_filtered_scaled %>%
    filter(Timepoint == "DMSO" & Protein == "ALYREF" & totalSegments %in%
        c(9:15) & Segment == "exon") %>%
    group_by(Protein, Rep, Timepoint, Segment, ExonDescription,
        totalSegments) %>%
    summarise(sum_scaled_density = sum(scaled_density)) %>%
    left_join(annotation_number) %>%
    mutate(sum_scaled_density_norm_to_anno_n = sum_scaled_density/number_of_annotations) %>%
    mutate(Timepoint_f = case_when(Timepoint == "PBSDRB" ~ "t00",
        TRUE ~ Timepoint)) %>%
    mutate(Timepoint_f = factor(Timepoint_f, levels = c("t00",
        "t05", "t10", "t15", "t20", "t40", "t60", "DMSO")))

df_for_Supp_fig_3a %>%
    ggplot() + geom_bar(aes(x = ExonDescription, y = sum_scaled_density_norm_to_anno_n,
    fill = Protein), stat = "summary", fun = mean, position = position_dodge()) +
    geom_point(aes(x = ExonDescription, y = sum_scaled_density_norm_to_anno_n,
        shape = Rep, group = Protein), position = position_dodge(width = 0.9)) +
    scale_shape_manual(values = c(21, 22, 24)) + facet_grid(totalSegments ~
    Timepoint_f, scales = "free") + ylab("CLIP density") + xlab("Exon Position") +
    theme_bw()
```


DMSO
9
10
11
12
13
14
15
CLIP density
Exon Position
Protein
ALYREF
Rep
1
2