Global tiCLIP coverage over TUs stratified by biotype, exonic type, and coding status

RAC

08/05/2021

#Figures 1 C-D & Supplementary Fig. 1 G-I

```
library(dplyr)
library(tidyr)
library(ggplot2)
```

#load annotations and data

```
# data load and add col names (remove snRNA and histone
# RNAs from count files)
exonIntronReads <- read.csv("../../data/xiCLIP_intronExon.200820.counts",
    sep = "\t", header = F) %>%
    filter(!grepl(":::snRNA|:::histone_coding", V5))

totalCounts <- read.csv("../../data/xiCLIP.read2.totalcounts.200402.tab",
    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("../../data/rRNAFactor.tab", sep = " ",
    header = F)
colnames(libraryScalings) <- c("Sample", "scaling")

# load annotation files and add col names
exonAnno <- read.table("../../data/hg38_HeLa_trimmed_loci_major_primary_isoform_annotated_exon_numbered
    header = F)
intronAnno <- read.table("../../data/hg38_HeLa_trimmed_loci_major_primary_isoform_annotated_intron_numb
    header = F)

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

# data example
head(exonIntronReads)
```

```
##                Sample chr  start    end
## 1 ALYREF_1_DMSO_read2   1 184924 185559
```

```
## 2 ALYREF_1_DMSO_read2   1 186316 187577
## 3 ALYREF_1_DMSO_read2   1 189193 191848
## 4 ALYREF_1_DMSO_read2   1 629639 630560
## 5 ALYREF_1_DMSO_read2   1 631073 631548
## 6 ALYREF_1_DMSO_read2   1 633695 634374
##                                                       ID segmentNumber strand
## 1                                NA.v1000:::protein_coding        exon_3      -
## 2                                NA.v1000:::protein_coding        exon_2      -
## 3                                     NA.v999:::intergenic        exon_1      -
## 4                           MTND2P28:::unprocessed_pseudogene        exon_1      +
## 5                AC114498.1,MIR6723:::unprocessed_pseudogene        exon_1      +
## 6 MTATP6P1,MTATP8P1,RP5-857K21.11:::unprocessed_pseudogene        exon_1      +
##   count
## 1     7
## 2     3
## 3     3
## 4    22
## 5    17
## 6    17
```

```
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
```

```
head(exonAnno)
```

```
##    chr  start    end                                         ID segmentNumber
## 1 chr1 184924 185559                NA.v1000:::protein_coding        exon_3
## 2 chr1 186316 187577                NA.v1000:::protein_coding        exon_2
## 3 chr1 187754 187848                NA.v1000:::protein_coding        exon_1
## 4 chr1 189193 191848                     NA.v999:::intergenic        exon_1
## 5 chr1 629639 630560           MTND2P28:::unprocessed_pseudogene        exon_1
## 6 chr1 631073 631548 AC114498.1,MIR6723:::unprocessed_pseudogene        exon_1
##   strand
## 1      -
## 2      -
## 3      -
## 4      -
## 5      +
## 6      +
```

```
head(intronAnno)
```

```
##    chr  start    end                                ID segmentNumber strand
## 1 chr1 185559 186316          NA.v1000:::protein_coding      intron_2      -
## 2 chr1 187577 187754          NA.v1000:::protein_coding      intron_1      -
## 3 chr1 773107 774170            RP11-206L10.2:::lincRNA      intron_1      -
## 4 chr1 827775 829002   LINC01128:::processed_transcript      intron_1      +
## 5 chr1 915471 915749 RP11-54O7.1,RP11-54O7.16:::lincRNA      intron_1      +
## 6 chr1 926013 930154            SAMD11:::protein_coding      intron_1      +
```

#wrangle data and annotation files

```
# catagorise gene annotation based on multiexonic or
# monoexonic
totalExons <- exonAnno %>%
    separate(ID, into = c("GeneID", "Biotype"), sep = ":::") %>%
    separate(segmentNumber, into = c("Segment", "SegmentNumber"),
        sep = "_") %>%
    group_by(GeneID, Biotype) %>%
    summarise(TotalExons = max(as.numeric(SegmentNumber))) %>%
    mutate(Exonic = case_when(TotalExons > 1 ~ "multiExonic",
        TRUE ~ "monoExonic"))

# catagorise genes by class: lncRNA, pcRNA, sncRNA
classes <- exonAnno %>%
    separate(ID, into = c("GeneID", "Biotype"), sep = ":::") %>%
    separate(segmentNumber, into = c("Segment", "SegmentNumber")) %>%
    mutate(size = as.numeric(end) - as.numeric(start)) %>%
    mutate(class = case_when(grepl("protein|histone", Biotype) ~
        "pcRNA", (!grepl("protein|histone", Biotype) & size >
        250) ~ "lncRNA", TRUE ~ "sncRNA")) %>%
    select(GeneID, class) %>%
    unique()


# wrangle total counts data
tcDf <- totalCounts %>%
    separate(Sample, into = c("Protein", "Rep", "Timepoint",
        "readType"))

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

# total nucleotides covered intron and exonic gene segments
segmentCovered_sizes <- eIRDf %>%
    select(start, end, GeneID, Segment) %>%
    unique() %>%
    group_by(Segment) %>%
    summarise(size = sum(end - start)) %>%
    mutate(sizeKb = size/1000)
```

```
# total nucleotides covered by lncRNA, pcRNA, sncRNA
segmentCovered_sizes_classes <- eIRDf %>%
    select(start, end, GeneID, Segment) %>%
    left_join(classes) %>%
    unique() %>%
    group_by(Segment, class) %>%
    summarise(size = sum(end - start)) %>%
    mutate(sizeKb = size/1000)

# total nucleotides covered by intronic, mono-, and
# multi-exonic gene segements
segmentCovered_sizes_Exonic <- eIRDf %>%
    select(start, end, GeneID, Segment) %>%
    left_join(totalExons) %>%
    unique() %>%
    group_by(Segment, Exonic) %>%
    summarise(size = sum(end - start)) %>%
    mutate(sizeKb = size/1000)

# prep rRNA library scaling
libScale <- libraryScalings %>%
    separate(Sample, into = c("Protein", "Rep", "Timepoint"))

# make df with total counts and the rRNA scaling
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)
```

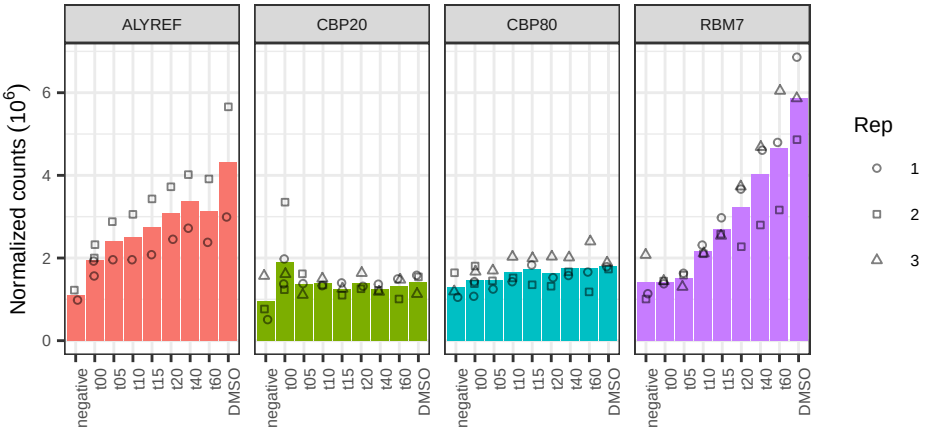
```
##   Protein Rep Timepoint readType chr  start    end
## 1  ALYREF   1      DMSO    read2   1 184924 185559
## 2  ALYREF   1      DMSO    read2   1 186316 187577
## 3  ALYREF   1      DMSO    read2   1 189193 191848
## 4  ALYREF   1      DMSO    read2   1 629639 630560
## 5  ALYREF   1      DMSO    read2   1 631073 631548
## 6  ALYREF   1      DMSO    read2   1 633695 634374
##                            GeneID               Biotype Segment SegmentNumber
## 1                       NA.v1000        protein_coding    exon             3
## 2                       NA.v1000        protein_coding    exon             2
## 3                        NA.v999            intergenic    exon             1
## 4                       MTND2P28 unprocessed_pseudogene    exon             1
## 5             AC114498.1,MIR6723 unprocessed_pseudogene    exon             1
## 6 MTATP6P1,MTATP8P1,RP5-857K21.11 unprocessed_pseudogene    exon             1
##   strand count
## 1      -     7
```

```
## 2      -     3
## 3      -     3
## 4      +    22
## 5      +    17
## 6      +    17
```

#Figure 1 C Normalise total counts to rRNA factors and display average of replicates as a bar graph

```
Fig1C <- totalCountsDF %>%
    mutate(normalised = scaling * TotalCount) %>%
    rename(raw = TotalCount) %>%
    gather(value = "counts", key = "method", c(normalised, raw)) %>%
    filter(method == "normalised") %>%
    mutate(Timepoint_f = case_when(Timepoint == "PBSDRB" ~ "t00",
        TRUE ~ Timepoint)) %>%
    mutate(Timepoint_f = factor(Timepoint_f, levels = c("negative",
        "t00", "t05", "t10", "t15", "t20", "t40", "t60", "DMSO"))) %>%
    ggplot(aes(x = Timepoint_f, y = counts/1e+06)) + geom_bar(aes(fill = Protein),
    stat = "summary", fun = "mean", show.legend = FALSE) + facet_grid(. ~
    Protein, scales = "free_y") + geom_point(aes(shape = Rep),
    position = position_jitterdodge(jitter.width = 0.2, dodge.width = 0),
    alpha = 0.5, size = 1) + scale_shape_manual(values = c(21,
    22, 24)) + theme_bw() + theme(text = element_text(size = 8),
    axis.text.x = element_text(angle = 90, hjust = 1)) + ylab(expression("Normalized counts" ~
    (10^6))) + xlab("")
```

```
Fig1C
```

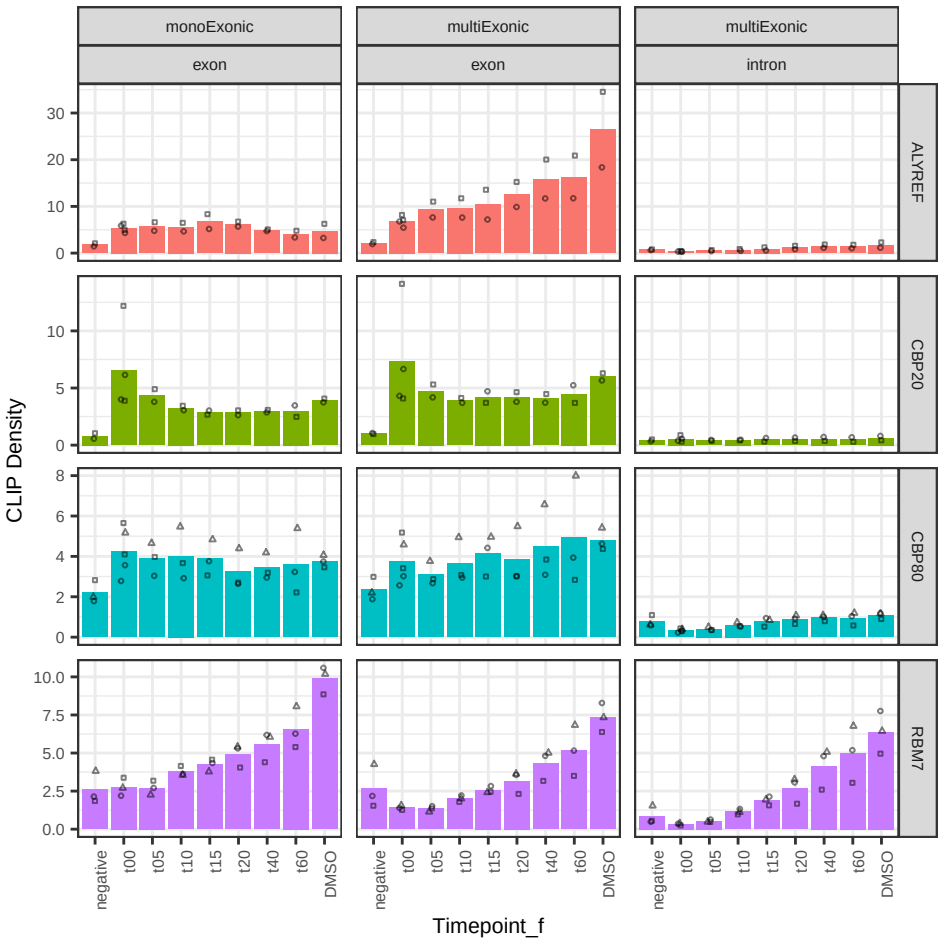


#Figure 1 D Normalise total counts to rRNA factors and display average of replicates as a bar graph. Results are normalised to rRNA factor and total nucleotides of each segment Segments are defined as multiExonic-Introns, multiExonic-exons or monoExonic-exons

```
Fig1D<-
eIRDf %>%
  left_join(totalExons) %>%
  group_by(Protein, Rep, Timepoint, Segment, Exonic) %>%
  summarise(segSum = sum(count)) %>% #sum up all counts associated with segment (multiexonic-intron, mu
  left_join(totalCountsDF) %>%  #total counts/rRNA factor
  mutate(normTorRNAAdj = as.numeric(segSum)*as.numeric(scaling)) %>% #normalise data to rRNA factor
  left_join(segmentCovered_sizes_Exonic) %>%  #join the total number of nucleotides present in each seg
  mutate(density_kb_all  = normTorRNAAdj/sizeKb) %>% # normalise reads to total number of nucleotides p
    mutate(Timepoint_f = case_when(Timepoint == "PBSDRB" ~ "t00", TRUE ~ Timepoint)) %>%
```

```
    mutate(Timepoint_f = factor(Timepoint_f, levels=c("negative", "t00", "t05", "t10", "t15", "t20", "t4
  ggplot(aes(x=Timepoint_f, y=density_kb_all)) +
  geom_bar(aes(fill = Protein), stat="summary", fun = "mean") +
  #stat_summary(fun.data="mean_cl_boot", geom="errorbar", width =0.1 )+
   geom_point(aes(shape = Rep),
              position=position_jitterdodge(jitter.width = 0.2,
                                            dodge.width=0), alpha = 0.5, size=0.5 ) +
  scale_shape_manual(values = c(21,22,24)) +
  theme_bw()+
  theme(legend.position = "none",
        text = element_text(size=8),
        axis.text.x = element_text(angle = 90, hjust = 1)) +
  ylab("CLIP Density")+
  facet_grid(Protein~Exonic+Segment, scales="free_y")
```

Fig1D

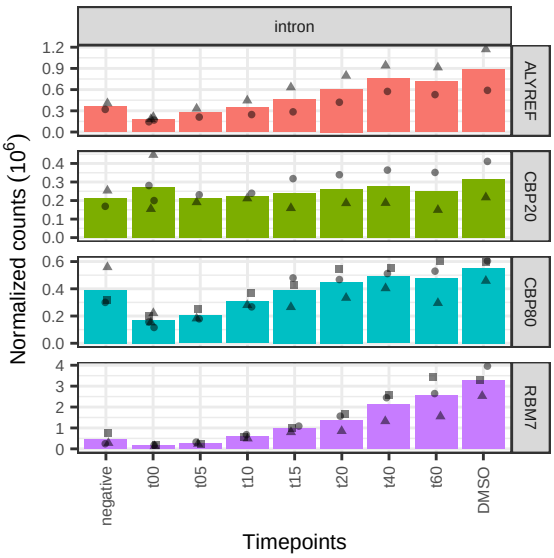


##Supplementry Figure 1 G Read density over introns.

```
suppFig1B <- eIRDf %>%
    left_join(totalExons) %>%
    group_by(Protein, Rep, Timepoint, Segment, Exonic) %>%
    summarise(SegmentSum = sum(count)) %>%
    ungroup() %>%
    left_join(totalCountsDF) %>%
```

```
    mutate(normTorRNAAdj = as.numeric(SegmentSum) * as.numeric(scaling)) %>%
    # filter(Protein == 'RBM7' & Timepoint == 'DMSO') %>%
mutate(Timepoint_r = case_when(Timepoint == "PBSDRB" ~ "t00",
    TRUE ~ Timepoint)) %>%
    mutate(Timepoint_r = factor(Timepoint_r, levels = c("negative",
        "t00", "t05", "t10", "t15", "t20", "t40", "t60", "DMSO"))) %>%
    filter(Segment == "intron") %>%
    # remove PHAX
filter(Protein != "PHAX") %>%
    ggplot(aes(x = Timepoint_r, y = normTorRNAAdj/10^6, fill = Protein)) +
    geom_bar(stat = "summary", fun = "mean") + geom_point(aes(shape = Rep),
    position = position_jitterdodge(jitter.width = 0.2, dodge.width = 0),
    alpha = 0.5, size = 1) + theme_bw() + scale_x_discrete(guide = guide_axis(angle = 90)) +
    facet_grid(Protein ~ Segment, scales = "free_y") + theme(legend.position = "none",
    text = element_text(size = 8)) + xlab("Timepoints") + ylab(expression("Normalized counts" ~
    (10^6)))
```

suppFig1B

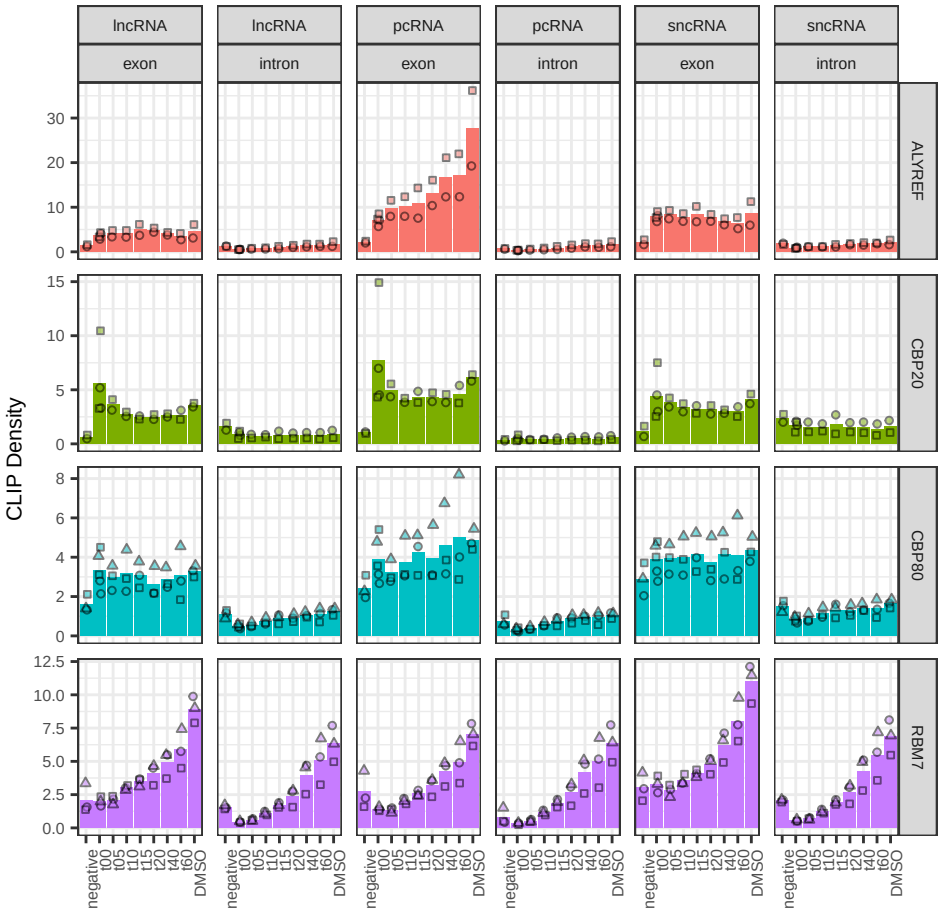


#Supplementry Figure 1 H same as figure 1 D except stratifying by coding potential and TU size.

```
suppFig1H <- eIRDf %>%
    left_join(classes) %>%
    group_by(Protein, Rep, Timepoint, Segment, class) %>%
    summarise(segSum = sum(count)) %>%
    left_join(totalCountsDF) %>%
    mutate(normTorRNAAdj = as.numeric(segSum) * as.numeric(scaling)) %>%
    left_join(segmentCovered_sizes_classes) %>%
    mutate(density_kb_all = normTorRNAAdj/sizeKb) %>%
    mutate(Timepoint_f = case_when(Timepoint == "PBSDRB" ~ "t00",
        TRUE ~ Timepoint)) %>%
    # remove PHAX
filter(Protein != "PHAX") %>%
    mutate(Timepoint_f = factor(Timepoint_f, levels = c("negative",
        "t00", "t05", "t10", "t15", "t20", "t40", "t60", "DMSO"))) %>%
    ggplot(aes(x = Timepoint_f, y = density_kb_all, fill = Protein)) +
```

```
    geom_bar(stat = "summary", fun = "mean") + geom_point(aes(shape = Rep),
    position = position_jitterdodge(jitter.width = 0.2, dodge.width = 0),
    alpha = 0.5, size = 1) + scale_shape_manual(values = c(21,
    22, 24)) + theme_bw() + theme(axis.text.x = element_text(angle = 90,
    hjust = 1), legend.position = "none", text = element_text(size = 8)) +
    ylab("CLIP Density") + xlab("") + facet_grid(Protein ~ class +
    Segment, scales = "free")
```

suppFig1H



#Supp Figure 1 I Calculating binding density for ALYREF-DMSO and RBM7-DMSO over exonic regions of the genome, but relative to CBP20-DMSO

```
# calculate binding densities over multiexonic
# introns/exons and monoexonic exons
df_densities <- eIRDf %>%
    filter(Timepoint == "DMSO") %>%
    left_join(totalExons) %>%
    group_by(Protein, Rep, Timepoint, Segment, Exonic) %>%
    summarise(segSum = sum(count)) %>%
    left_join(totalCountsDF) %>%
    mutate(normTorRNAAdj = as.numeric(segSum) * as.numeric(scaling)) %>%
    left_join(segmentCovered_sizes_Exonic) %>%
    mutate(density_kb_all = normTorRNAAdj/sizeKb) %>%
```

```
    mutate(Timepoint_f = case_when(Timepoint == "PBSDRB" ~ "t00",
        TRUE ~ Timepoint)) %>%
    mutate(Timepoint_f = factor(Timepoint_f, levels = c("negative",
        "t00", "t05", "t10", "t15", "t20", "t40", "t60", "DMSO"))) %>%
    group_by(Protein, Timepoint, Segment, Exonic) %>%
    summarise(mean = mean(density_kb_all)) %>%
    ungroup()

head(df_densities)
```

```
## # A tibble: 6 x 5
##   Protein Timepoint Segment Exonic        mean
##   <chr>   <chr>     <chr>   <chr>        <dbl>
## 1 ALYREF  DMSO      exon    monoExonic   4.75
## 2 ALYREF  DMSO      exon    multiExonic 26.4
## 3 ALYREF  DMSO      intron  multiExonic  1.73
## 4 CBP20   DMSO      exon    monoExonic   3.91
## 5 CBP20   DMSO      exon    multiExonic  5.98
## 6 CBP20   DMSO      intron  multiExonic  0.616
```

```
# pull out CBP20 data density data
CBP20 <- df_densities %>%
    filter(Protein == "CBP20") %>%
    mutate(CBP20 = mean) %>%
    select(-Protein, -mean)

head(CBP20)
```

```
## # A tibble: 3 x 4
##   Timepoint Segment Exonic      CBP20
##   <chr>     <chr>   <chr>       <dbl>
## 1 DMSO      exon    monoExonic  3.91
## 2 DMSO      exon    multiExonic 5.98
## 3 DMSO      intron  multiExonic 0.616
```

```
# normalise ALYREF and RBM7 binding density to CBP20
RBM7_ALYREF_densities <- df_densities %>%
    left_join(CBP20) %>%
    mutate(rel_to_CBP20 = mean/CBP20) %>%
    dplyr::filter(!(Protein %in% c("CBP20", "CBP80")) & Segment ==
        "exon")

head(RBM7_ALYREF_densities)
```

```
## # A tibble: 4 x 7
##   Protein Timepoint Segment Exonic       mean CBP20 rel_to_CBP20
##   <chr>   <chr>     <chr>   <chr>       <dbl> <dbl>        <dbl>
## 1 ALYREF  DMSO      exon    monoExonic   4.75  3.91         1.22
## 2 ALYREF  DMSO      exon    multiExonic 26.4   5.98         4.42
## 3 RBM7    DMSO      exon    monoExonic   9.88  3.91         2.52
## 4 RBM7    DMSO      exon    multiExonic  7.33  5.98         1.23
```

```
SupFig1I <- RBM7_ALYREF_densities %>%
    ggplot(aes(x = Protein, y = rel_to_CBP20)) + geom_hline(yintercept = 1,
    linetype = "dotted") + geom_bar(aes(fill = Exonic), stat = "summary",
    fun = "mean", position = position_dodge2()) + theme_bw() +
```

```
    theme(text = element_text(size = 7), axis.text.x = element_text(angle = 90,
        hjust = 1), legend.position = c(0.75, 0.8)) + ylab("CLIP density (relative to CBP20)")
```

SupFig1I