

Supplementary Figure 1 J

RAC

08/05/2021

```
exonIntronReads <- read.csv("../data/xiCLIP_intronExon.200820.counts",
  sep = "\t", header = F)
totalCounts <- read.csv("../data/xiCLIP.read2.totalcounts.200402.tab",
  sep = "\t", header = F)
libraryScalings <- read.csv("../data/rRNAAFactor.tab", sep = " ",
  header = F)

exonAnno <- read.table("../data/hg38_HeLa_trimmed_loci_major_primary_isofrom_annotation_exon_numbered",
  header = F)
intronAnno <- read.table("../data/hg38_HeLa_trimmed_loci_major_primary_isofrom_annotation_intron_numbered",
  header = F)

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

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

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

make dataframes with meta data from annotation file

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"))

tcDf <- totalCounts %>%
  separate(Sample, into = c("Protein", "Rep", "Timepoint",
  "readType"))

# wrangle count file
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")

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
## 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

```

Supplementary figure 1 j

plotting log10 CLIP density vs total exons of TU, stratified by protein

```

library(dplyr)
library(tibble)
library(tidyr)
library(ggplot2)
library(Hmisc)
library(ggpointdensity)
library(viridis)

#create dataframe with gene size
gene_size<-
rbind(exonAnno,intronAnno) %>%
  separate(ID, c("GeneID","biotype"), sep = ":::") %>%
  group_by(GeneID) %>%
  summarise(gene_size = sum(end-start))

exonvstranscriptdensity<-
eIRDf %>%
  left_join(totalExons) %>%
  left_join(gene_size) %>%
  filter(Timepoint == "DMSO"
    & TotalExons > 1 ) %>% #select multiedxonic TUs

```

```

group_by(Protein, Rep, GeneID, Timepoint, Segment, TotalExons, gene_size) %>%
#calculate density of CLIP reads per gene
summarise(count = sum(count)) %>%
mutate(Densitykb = count/(as.numeric(gene_size)/1000)) %>%
ungroup() %>%
left_join(totalCountsDF) %>%
#normalise to rRNA factor
mutate(normTorRNAAAdj_Densitykb = as.numeric(Densitykb)*as.numeric(scaling),
       countNorm = as.numeric(count) * as.numeric(scaling)) %>%
mutate(Timepoint_r = case_when(
  Timepoint == "PBSDRB" ~ "t00",
  TRUE ~ Timepoint
)) %>%
mutate(Timepoint_r = factor(Timepoint_r, levels = c("negative", "t00", "t05", "t10", "t15", "t20", "t25", "t30", "t35", "t40", "t45", "t50", "t55", "t60", "t65", "t70", "t75", "t80", "t85", "t90", "t95", "t100", "t105", "t110", "t115", "t120")))
p<-
exonvstranscriptdensity %>%
filter(Segment == "exon") %>% #select only exonic regions
ggplot(aes(x=as.numeric(TotalExons), y=log2(normTorRNAAAdj_Densitykb))) +
  geom_pointdensity(adjust = 5, alpha = 0.25, size =0.1) +
  scale_color_viridis(option = "A",
                      breaks=c(1,40000),
                      labels=c("low","High")) +
  geom_smooth(method="lm") +
  facet_grid(Timepoint_r~Protein) +
  coord_equal(ratio = 5)

p +
theme_bw() +
theme(text = element_text(size = 8),
      axis.text.x = element_text(angle = 90, hjust = 1),
      legend.position = "none") +
xlab("Total Exons") +
ylab("Density (Log2; read/kb)")

```

