Supp Fig 3 C

RAC

14/12/2020

```
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)
library(viridis)
```

```
## Loading required package: viridisLite
```

```
COUNTS = "../../data/xiCLIP_all_5primepos.rRNAScaled.hg38_HeLa_trimmed_loci_major_primary_isoform_annota
EXPRESSION_VECTOR_FILEPATH = "../../data/log2_mean_cov_RNAseq_TTseq.RData"
ANNOTATION_BED_FILEPATH = "../../data/hg38_HeLa_trimmed_loci_major_primary_isoform_annotated.exonNumber
```

Load data frames containing expression of genes in HeLa cells, annotation bed file and count file

```
# Load 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))


# load 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", "ExonicDistance", "ExonDistFromTSS",
        "ExonStature", "GeneStructure"), sep = ":::") %>%
```

```
    mutate_at(vars(ExonDistFromTSS, ExonicDistance, ExonSize,
        TotalNumberOfExons, ExonNumber), .funs = as.numeric)


# load 3end count file ONLY ALYREF -----------------

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", "ExonicDistance", "ExonDistFromTSS",
        "ExonStature", "GeneStructure"), sep = ":::") %>%
    mutate_at(vars(ExonDistFromTSS, ExonicDistance, ExonSize,
        TotalNumberOfExons, ExonNumber), .funs = as.numeric) %>%
    filter(grepl("ALYREF", 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
```

```
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     .      -
```

```
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
```

```
# normalise counts to expression of genes, if gene is not
# present in expression vector then it is given the lowest
# expression present in the expression vector
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 value
    # in expression_vector
mutate_at(vars(ctrl_RNAseq_expr), ~replace(., is.na(.), min(expression_vector$ctrl_RNAseq_expr))) %>%
    mutate(norm_count = count/ctrl_RNAseq_expr)

head(norm_counts_to_gene_expression)
```

```
##                         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
```

```
##   ctrl_RNAseq_expr ctrl_TTseq_expr norm_count
## 1         2.456720        4.683736   3.571107
## 2         2.456720        4.683736   1.785556
## 3         2.456720        4.683736   1.785556
## 4         2.456720        4.683736   1.785556
## 5         2.456720        4.683736   1.785556
## 6         2.401283        4.697747   1.826778
```

```
# select exons over 50 and under 300nt in length, not
# single exon genes, and have at least an expression value
# of 1 at RNAseq.
geneIDs_from_over1log2Exp <- annoBed %>%
    left_join(expression_vector) %>%
    filter(TotalNumberOfExons > 1 & !grepl("snRNA|rRNA|TR_C_gene|IG_C_pseudogene|miRNA|misc_RNA",
        Biotype) & ExonSize %in% c(50:300)) %>%
    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")

number_of_exon_annotations <- annoBed %>%
    filter(geneID %in% geneIDs_from_over1log2Exp$geneID & ExonSize %in%
        c(50:300)) %>%
    group_by(GeneStructure, ExonSize) %>%
    summarise(exon_count = n())


for_graph <- norm_counts_to_gene_expression %>%
    filter(geneID %in% geneIDs_from_over1log2Exp$geneID & ExonSize %in%
        c(50:300) & GeneStructure == "multiExonicGene-internalExon") %>%
    group_by(Sample, GeneStructure, ExonSize, DistToLandmark) %>%
    summarise(sum_norm_count = sum(norm_count)) %>%
    ungroup() %>%
    group_by(Sample, GeneStructure, ExonSize) %>%
    mutate(max = max(sum_norm_count), norm_to_max = sum_norm_count/max(sum_norm_count)) %>%
    separate(Sample, c("Protein", "Rep", "Timepoint", "readType",
        "region"), sep = "_") %>%
    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", "t60", "DMSO")),
        Timepoint_f = factor(Timepoint_f, levels = c("negative",
            "t00", "t05", "t10", "t15", "t20", "t40", "t60",
            "DMSO"))) %>%
    group_by(Protein, Timepoint_f, region, GeneStructure, DistToLandmark,
        ExonSize) %>%
    summarise(mean = mean(norm_to_max)) %>%
    ungroup()
```

#Supp Fig 3 C #CLIP density over Exon 5' and 3' ends (facets) centred on 5' and 3' ends of exons (distance from landmark on x axis), stratified by exon lengths (yaxis). Only looking at internal exons between 50 and 300 nt in length

```r
# graph for ALYREF
for_graph %>%
    filter(Protein == "ALYREF" & Timepoint_f == "DMSO") %>%
    spread(DistToLandmark, mean, fill = 0) %>%
    gather("DistToLandmark", "norm_to_max", c(`-100`:`100`)) %>%
    mutate(DistToLandmark = as.numeric(DistToLandmark)) %>%
    ggplot() + geom_raster(aes(x = DistToLandmark, y = ExonSize,
    fill = norm_to_max)) + facet_grid(Timepoint_f ~ GeneStructure +
    region, scale = "free") + scale_fill_viridis_c(option = "inferno",
    direction = -1, na.value = "black") + theme_bw() + labs(subtitle = "ALYREF") +
    theme(axis.title = element_text(size = 6))
```