Figure 3G

RAC

16/05/2021

```
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)

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

EXONCOUNTS="xiCLIP_all_Exon_splice.count"
SPLICEDEXONCOUNTS="xiCLIP_spliceSites10updown.count"
EXPRESSION_VECTOR_FILEPATH="../../data/log2_mean_cov_RNAseq_TTseq.RData"
ANNOTATION_BED_FILEPATH="../../data/hg38_HeLa_trimmed_loci_major_primary_isoform_annotated.exonNumber.s
SCALINGVECTOR="../../data/rRNAFactor.tab"
```

```
# scaling vector

scaling_rRNA <- read.table(SCALINGVECTOR) %>%
    setNames(c("Sample", "scalingFactorrRNA")) %>%
    separate(Sample, c("Protein", "Rep", "Timepoint"))


# 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))
```

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

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

```
# 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",
        "TotalNumberOfExon", "ExonSize", "ExonicDistance", "GenomicDistFromTSS",
        "ExonStature", "GeneStructure"), sep = ":::") %>%
    mutate_at(vars(ExonNumber, TotalNumberOfExon, ExonSize, ExonicDistance,
        GenomicDistFromTSS), .funs = as.numeric) %>%
    mutate(sizeRange = case_when(GenomicDistFromTSS < 10000 ~
        "<10k", GenomicDistFromTSS %in% c(10000:20000) ~ "10k-20k",
        GenomicDistFromTSS %in% c(20001:30000) ~ "20k-30k", GenomicDistFromTSS %in%
            c(30001:40000) ~ "30k-40k", GenomicDistFromTSS %in%
            c(40001:50000) ~ "40k-50k", GenomicDistFromTSS %in%
            c(50001:60000) ~ "50k-60k", GenomicDistFromTSS %in%
            c(60001:70000) ~ "60k-70k", GenomicDistFromTSS %in%
            c(70001:80000) ~ "70k-80k", GenomicDistFromTSS %in%
            c(80001:90000) ~ "80k-90k", GenomicDistFromTSS %in%
            c(90001:1e+05) ~ "90k-100k", GenomicDistFromTSS %in%
            c(100001:110000) ~ "100k-110k", GenomicDistFromTSS %in%
            c(110001:120000) ~ "110k-120k", GenomicDistFromTSS %in%
            c(120001:130000) ~ "120k-130k", GenomicDistFromTSS %in%
            c(130001:140000) ~ "130k-140k", GenomicDistFromTSS %in%
            c(140001:150000) ~ "140k-150k", GenomicDistFromTSS %in%
            c(150001:160000) ~ "150k-160k", GenomicDistFromTSS %in%
            c(160001:170000) ~ "160k-170k", GenomicDistFromTSS %in%
            c(170001:180000) ~ "170k-180k", GenomicDistFromTSS %in%
            c(180001:190000) ~ "180k-190k", GenomicDistFromTSS %in%
            c(190001:2e+05) ~ "190k-200k", GenomicDistFromTSS >
            2e+05 ~ ">200k"))



head(annoBed)
```

```
##   chr   start     end geneID        Biotype ExonNumber TotalNumberOfExon
## 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 GenomicDistFromTSS 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 sizeRange
## 1     multiExonicGene-lastExon     .      - 100k-110k
## 2 multiExonicGene-internalExon     .      - 100k-110k
## 3 multiExonicGene-internalExon     .      -  90k-100k
```

```
## 4 multiExonicGene-internalExon     .      -  90k-100k
## 5 multiExonicGene-internalExon     .      -   80k-90k
## 6 multiExonicGene-internalExon     .      -   70k-80k
```

```
head(EXPRESSION_VECTOR_FILEPATH)
```

```
## [1] "../../data/log2_mean_cov_RNAseq_TTseq.RData"
```

```
head(scaling_rRNA)
```

```
##   Protein Rep Timepoint scalingFactorrRNA
## 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 figure for 3G plot number of spliced reads over first and internal exons

```
# load data amd wrangle
SS_Exon_DF_3 <- read.table("../../data/xiCLIP_all_spliceSites.3endofExonand1ntdown.read2.count") %>%
    setNames(c("sampleInfo", "chr", "start", "end", "geneID",
        "score", "strand", "count")) %>%
    separate(sampleInfo, into = c("Protein", "Rep", "Timepoint",
        "spliceStatus", "readNumber")) %>%
    separate(geneID, into = c("geneID", "Biotype", "ExonNumber",
        "TotalNumberOfExon", "ExonSize", "ExonicDistance", "GenomicDistFromTSS",
        "ExonStature", "GeneStructure", "region"), sep = ":::") %>%
    mutate_at(vars(ExonNumber, TotalNumberOfExon, ExonSize, ExonicDistance,
        GenomicDistFromTSS), .funs = as.numeric) %>%
    select(-chr, -start, -end) %>%
    left_join(annoBed) %>%
    left_join(scaling_rRNA) %>%
    right_join(expression_vector) %>%
    drop_na() %>%
    unique() %>%
    # normalise counts to rRNA factor and gene expression
mutate(rRNAScaledCounts = (count * scalingFactorrRNA)/ctrl_RNAseq_expr) %>%
    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")),
        region = factor(region, c("5end", "3end")))
```

```
## Warning: Expected 5 pieces. Additional pieces discarded in 779008 rows [1, 2, 3,
## 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].
```

```
## Joining, by = c("geneID", "Biotype", "ExonNumber", "TotalNumberOfExon",
## "ExonSize", "ExonicDistance", "GenomicDistFromTSS", "ExonStature",
## "GeneStructure", "score", "strand")
## Joining, by = c("Protein", "Rep", "Timepoint")
## Joining, by = "geneID"
```

```
head(SS_Exon_DF_3)
```

```
##   Protein Rep Timepoint spliceStatus readNumber             geneID
## 1  ALYREF   1      DMSO   notspliced      read2 AC114498.1,MIR6723
```

```
## 2  ALYREF   1      DMSO   notspliced      read2              SDF4
## 3  ALYREF   1      DMSO   notspliced      read2              DVL1
## 4  ALYREF   1      DMSO   notspliced      read2             NA.v6
## 5  ALYREF   1      DMSO   notspliced      read2            NADK.v1
## 6  ALYREF   1      DMSO   notspliced      read2               SKI
##                 Biotype ExonNumber TotalNumberOfExon ExonSize ExonicDistance
## 1 unprocessed_pseudogene          1                1      475            475
## 2         protein_coding          7                7      757           1995
## 3         protein_coding         10               15       68           1406
## 4             intergenic          1                1     1193           1193
## 5         protein_coding          4               12      130            614
## 6         protein_coding          7                7     3539           6271
##   GenomicDistFromTSS ExonStature               GeneStructure region score
## 1                  0   majorExon            singleExonicGene   3end     .
## 2              14376   majorExon     multiExonicGene-lastExon   3end     .
## 3               9768   majorExon multiExonicGene-internalExon   3end     .
## 4                  0   majorExon            singleExonicGene   3end     .
## 5              21159   majorExon multiExonicGene-internalExon   3end     .
## 6              78544   majorExon     multiExonicGene-lastExon   3end     .
##   strand count chr   start     end sizeRange scalingFactorrRNA ctrl_RNAseq_expr
## 1      +     2   1  631073  631548      <10k          4.386606         7.626640
## 2      -     1   1 1216931 1217688   10k-20k          4.386606         5.667403
## 3      -     1   1 1339581 1339649      <10k          4.386606         4.857764
## 4      +     1   1 1356681 1357874      <10k          4.386606         2.416374
## 5      -     1   1 1757180 1757310   20k-30k          4.386606         4.433562
## 6      +     1   1 2306576 2310115   70k-80k          4.386606         2.703030
##   ctrl_TTseq_expr rRNAScaledCounts Timepoint_f
## 1        5.738012        1.1503379        DMSO
## 2        5.691092        0.7740064        DMSO
## 3        4.730549        0.9030093        DMSO
## 4        3.573424        1.8153672        DMSO
## 5        5.665138        0.9894090        DMSO
## 6        4.631011        1.6228477        DMSO
```

```
gene_with_first_exons_under_150nt <- SS_Exon_DF_3 %>%
    filter(as.numeric(ExonSize) < 150 & ExonNumber == "1" & GeneStructure ==
        "multiExonicGene-firstExon") %>%
    select(geneID) %>%
    unique()

head(gene_with_first_exons_under_150nt)
```

```
##          geneID
## 1        ZBTB48
## 2         TAF12
## 3        ZNF691
## 4         MUTYH
## 5 RABGGTB,ACADM
## 6          WDR3
```

```
number_of_annotations <- SS_Exon_DF_3 %>%
    filter(geneID %in% gene_with_first_exons_under_150nt$geneID) %>%
    select(region, GeneStructure, geneID, sizeRange) %>%
    unique() %>%
    right_join(expression_vector) %>%
```

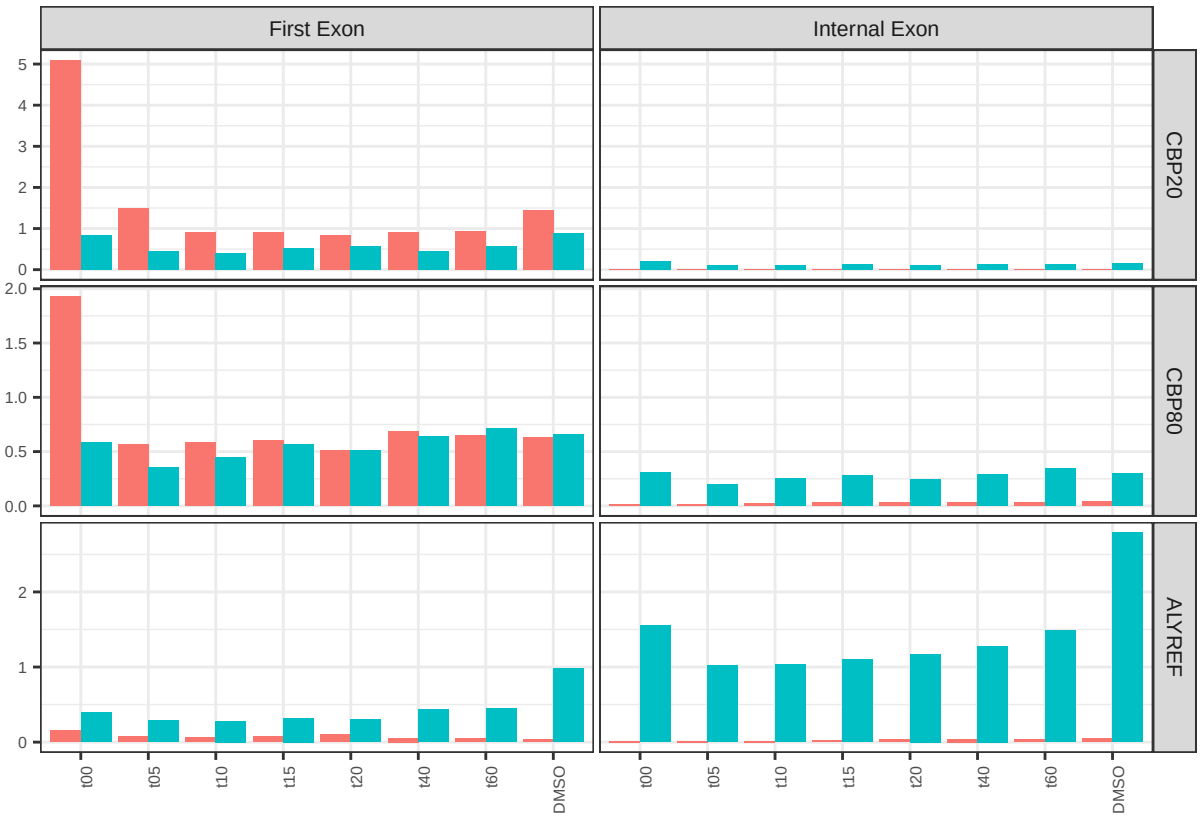
```
    group_by(region, GeneStructure) %>%
    summarise(n = n()) %>%
    drop_na()
```

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

```
fig_3g <- SS_Exon_DF_3 %>%
    # select for TUs with first exons under 150nt.
filter(geneID %in% gene_with_first_exons_under_150nt$geneID) %>%
    ungroup() %>%
    # mutate(density = as.numeric(rRNAScaledCounts)) %>%
group_by(Protein, Rep, Timepoint_f, spliceStatus, readNumber,
    region, GeneStructure) %>%
    summarise(count = sum(rRNAScaledCounts)) %>%
    # normalise to number of annotations used in analysis
left_join(number_of_annotations) %>%
    ungroup() %>%
    mutate(count = count/n) %>%
    # format plot
mutate_at("GeneStructure", ~replace(., GeneStructure == "multiExonicGene-internalExon",
    "Internal Exon")) %>%
    mutate_at("GeneStructure", ~replace(., GeneStructure == "multiExonicGene-firstExon",
        "First Exon")) %>%
    filter(!(GeneStructure %in% c("singleExonicGene", "multiExonicGene-lastExon"))) %>%
    filter(Protein %in% c("ALYREF", "CBP20", "CBP80")) %>%
    mutate(Protein = factor(Protein, levels = c("CBP20", "CBP80",
        "ALYREF"))) %>%
    filter(Timepoint_f != "negative") %>%
    # plot
ggplot(aes(x = Timepoint_f, y = count)) + geom_bar(aes(fill = spliceStatus),
    stat = "summary", fun = mean, position = "dodge") + facet_grid(Protein ~
    GeneStructure, scale = "free") + theme_bw() + theme(axis.text.x = element_text(angle = 90,
    hjust = 1), text = element_text(size = 8), legend.position = "none",
    panel.spacing = unit(0.15, "lines"), strip.text.x = element_text(size = 8),
    strip.text.y = element_text(size = 8)) + ylab("") + xlab("")
```

```
## `summarise()` has grouped output by 'Protein', 'Rep', 'Timepoint_f',
## 'spliceStatus', 'readNumber', 'region'. You can override using the `.groups`
## argument.
## Joining, by = c("region", "GeneStructure")
```

```
fig_3g
```




```
# fig_3g + ggsave('figure3g2.pdf', width = 2.75, height =
# 3.5)
```