Untitled

RAC

24/08/2020

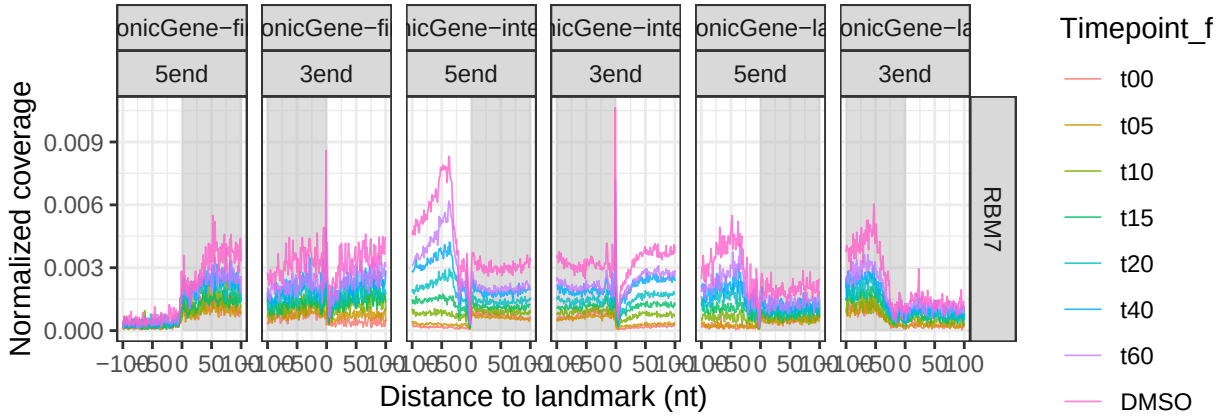
```
knitr::opts_chunk$set(warning=FALSE, message=FALSE, tidy.opts = list(width.cutoff = 60), tidy = TRUE)

COUNTS="../../data/xiCLIP_all_5primepos.rRNAScaled.hg38_HeLa_trimmed_loci_major_primary_isoform_annotat

EXPRESSION_VECTOR_FILEPATH="../../data/log2_mean_cov_RNAseq_TTseq.RData"

ANNOTATION_BED_FILEPATH="../../data/hg38_HeLa_trimmed_loci_major_primary_isoform_annotated.exonNumber.si
```

#Figure 5 B

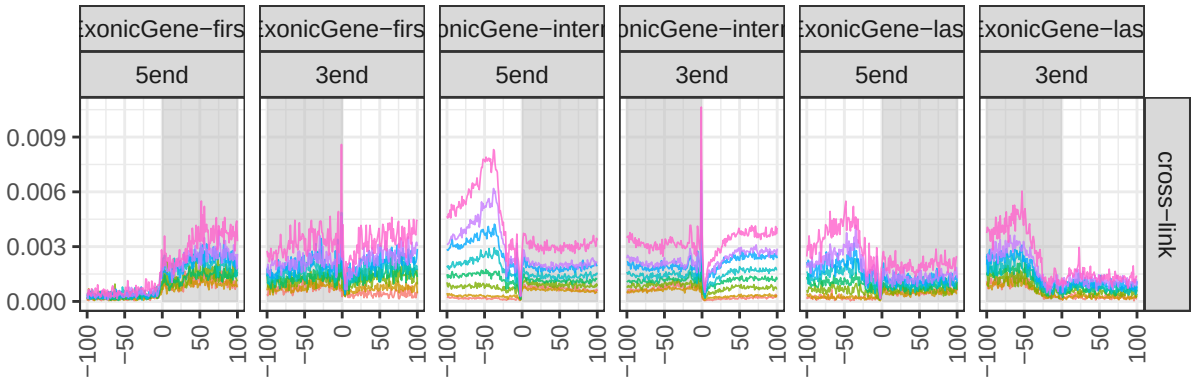
```
suppressMessages(library(ggplot2))
suppressMessages(library(dplyr))
suppressMessages(library(tidyr))

# 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 count file -----------------

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("CBP20_3", Sample))


# multiexonic genes

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)

expressed_genes <- annoBed %>%
    left_join(expression_vector) %>%
    filter(TotalNumberOfExons > 1 & !grepl("snRNA|rRNA|TR_C_gene|IG_C_pseudogene|miRNA|misc_RNA",
        Biotype) & as.numeric(ExonSize) > 99) %>%
    select(geneID, ctrl_RNAseq_expr, ctrl_TTseq_expr) %>%
    unique() %>%
    mutate(ctrl_RNAseq_expr = as.numeric(ctrl_RNAseq_expr)) %>%
    arrange(desc(ctrl_RNAseq_expr)) %>%
    filter(ctrl_RNAseq_expr > 1 & geneID != "LINC00324")

number_of_intron_annotations <- annoBed %>%
    filter(TotalNumberOfExons > 1 & !grepl("snRNA|rRNA|TR_C_gene|IG_C_pseudogene|miRNA|misc_RNA",
        Biotype) & as.numeric(ExonSize) > 99) %>%
    filter(geneID %in% expressed_genes$geneID) %>%
    group_by(GeneStructure) %>%
    summarise(intron_count = n())

GENECOUNT <- annoBed %>%
    filter(geneID %in% expressed_genes$geneID) %>%
    select(geneID) %>%
    unique() %>%
    summarise(geneCount = n())

plot <- norm_counts_to_gene_expression %>%
    filter(grepl("RBM7", Sample)) %>%
    filter(geneID %in% expressed_genes$geneID) %>%
    group_by(Sample, GeneStructure, DistToLandmark) %>%
    summarise(sum_RNAseq_norm_count_norm_annotation_number = sum(norm_count)) %>%
    left_join(number_of_intron_annotations) %>%
    mutate(sum_RNAseq_norm_count_norm_annotation_number = sum_RNAseq_norm_count_norm_annotation_number/
    separate(Sample, c("Protein", "Rep", "Timepoint", "readType",
        "region"), sep = "_") %>%
    filter(Timepoint != "negative") %>%
    mutate(Timepoint_f = case_when(Timepoint == "PBSDRB" ~ "t00",
        TRUE ~ Timepoint)) %>%
    mutate(region = factor(region, levels = c("5end", "3end")),
        readType = gsub("5primepos", "cross-link", readType),
```

```
        Timepoint = factor(Timepoint, levels = c("PBSDRB", "t00",
            "t05", "t10", "t15", "t20", "t40", "t60", "DMSO")),
        Timepoint_f = factor(Timepoint_f, levels = c("t00", "t05",
            "t10", "t15", "t20", "t40", "t60", "DMSO"))) %>%
    ggplot() + geom_rect(data = data.frame(region = "3end"),
    aes(xmin = -100, xmax = 0, ymin = 0, ymax = Inf), alpha = 0.5,
    fill = "grey") + geom_rect(data = data.frame(region = "5end"),
    aes(xmin = 0, xmax = 100, ymin = 0, ymax = Inf), alpha = 0.5,
    fill = "grey") + geom_line(aes(x = DistToLandmark, y = sum_RNAseq_norm_count_norm_annotation_number
    col = Timepoint_f), stat = "summary", fun = "mean", alpha = 0.8,
    size = 0.3) + facet_grid(Protein ~ GeneStructure + region,
    scale = "free") + xlab("Distance to landmark (nt)") + ylab("Normalized coverage") +
    theme_bw()

print(plot)
```




```
plot + theme(axis.text.x = element_text(angle = 90, vjust = 0.5,
    hjust = 1), legend.position = "none") + xlab("") + ylab("") +
    facet_grid(readType ~ GeneStructure + region, scale = "free")
```




```
# ggsave('fig5b.pdf', height = 2, width =6)
```

#Supplementary figure 5 C #plot cross-link sites around 5'SS

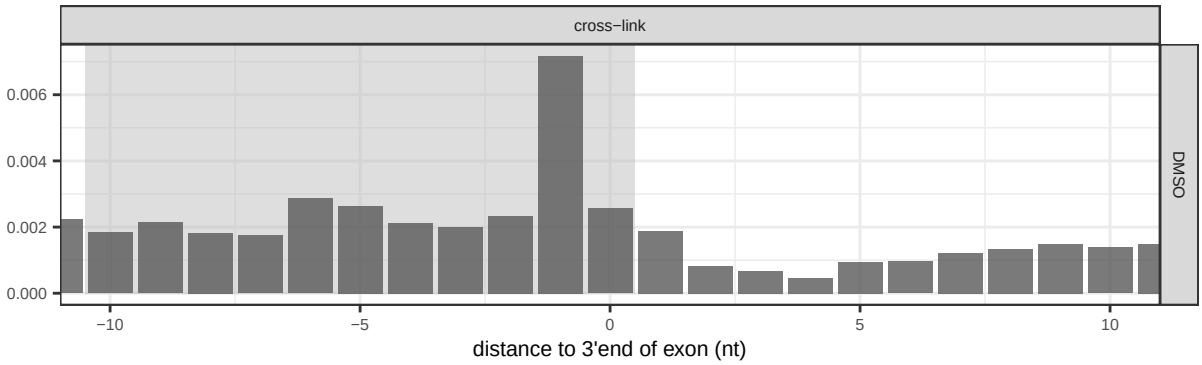
```
expressed_genes <- annoBed %>%
    left_join(expression_vector) %>%
    filter(TotalNumberOfExons > 1 & !grepl("snRNA|rRNA|TR_C_gene|IG_C_pseudogene|miRNA|misc_RNA",
        Biotype) & as.numeric(ExonSize) > 99) %>%
    select(geneID, ctrl_RNAseq_expr, ctrl_TTseq_expr) %>%
```

```
    unique() %>%
    mutate(ctrl_RNAseq_expr = as.numeric(ctrl_RNAseq_expr)) %>%
    arrange(desc(ctrl_RNAseq_expr)) %>%
    filter(ctrl_RNAseq_expr > 1 & geneID != "LINC00324")

number_of_intron_annotations <- annoBed %>%
    filter(geneID %in% expressed_genes$geneID & ExonNumber !=
        TotalNumberOfExons) %>%
    summarise(intron_count = n())

number_of_intron_annotations[, 1]
```

```
## [1] 104821
```

```
GENECOUNT <- annoBed %>%
    filter(geneID %in% expressed_genes$geneID) %>%
    select(geneID) %>%
    unique() %>%
    summarise(geneCount = n())


for_graph <- norm_counts_to_gene_expression %>%
    filter(geneID %in% expressed_genes$geneID & grepl("RBM7",
        Sample) & grepl("3end", Sample) & grepl("DMSO", Sample) &
        ExonNumber != TotalNumberOfExons) %>%
    group_by(Sample, DistToLandmark) %>%
    summarise(sum_RNAseq_norm_count_norm_annotation_number = sum(norm_count)) %>%
    # left_join(number_of_intron_annotations) %>%
mutate(sum_RNAseq_norm_count_norm_annotation_number = sum_RNAseq_norm_count_norm_annotation_number/numbe
    1]) %>%
    separate(Sample, c("Protein", "Rep", "Timepoint", "readType",
        "region"), sep = "_") %>%
    filter(Timepoint != "negative") %>%
    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("t00", "t05",
            "t10", "t15", "t20", "t40", "t60", "DMSO"))) %>%
    mutate(readType = gsub("5primepos", "cross-link", gsub("3endOfRead2",
        "3'CLIP", readType)))

p <- for_graph %>%
    ggplot() + geom_rect(data = data.frame(region = "3end"),
    aes(xmin = -10.5, xmax = 0.5, ymin = 0, ymax = Inf), alpha = 0.5,
    fill = "grey") + geom_bar(aes(x = DistToLandmark, y = sum_RNAseq_norm_count_norm_annotation_number)
    stat = "summary", fun = "mean", alpha = 0.8, size = 0.3) +
    coord_cartesian(xlim = c(-10, 10)) + facet_grid(Timepoint_f ~
    . + readType, scale = "free") + xlab("distance to 3'end of exon (nt)") +
    ylab("") + theme_bw() + theme(text = element_text(size = 8),
    legend.position = "right", panel.spacing.y = unit(0.4, "lines"),
    panel.spacing.x = unit(0.8, "lines"))
```

```
print(p)
```


cross−link
DMSO
0.006
0.004
0.002
0.000
−10
−5
0
5
10
distance to 3'end of exon (nt)