Untitled

RAC

24/08/2020

```
COUNTS="../../data/xiCLIP.3endOfRead2.rRNAScaled.hg38_HeLa_trimmed_loci_major_primary_isoform_annotated

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

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

#Figure 5 C

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



#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", "TotalNumberOfExons", "ExonSize", "Exon
  mutate_at(vars(ExonDistFromTSS,ExonicDistance,ExonSize,TotalNumberOfExons,ExonNumber), .funs = as.nume

#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", "Exon
    mutate_at(vars(ExonDistFromTSS,ExonicDistance,ExonSize,TotalNumberOfExons,ExonNumber), .funs = as.nu
  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
    mutate_at(vars(ctrl_RNAseq_expr), ~replace(., is.na(.), min(expression_vector$ctrl_RNAseq_expr))) %
    mutate(norm_count = count/ctrl_RNAseq_expr)
```

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

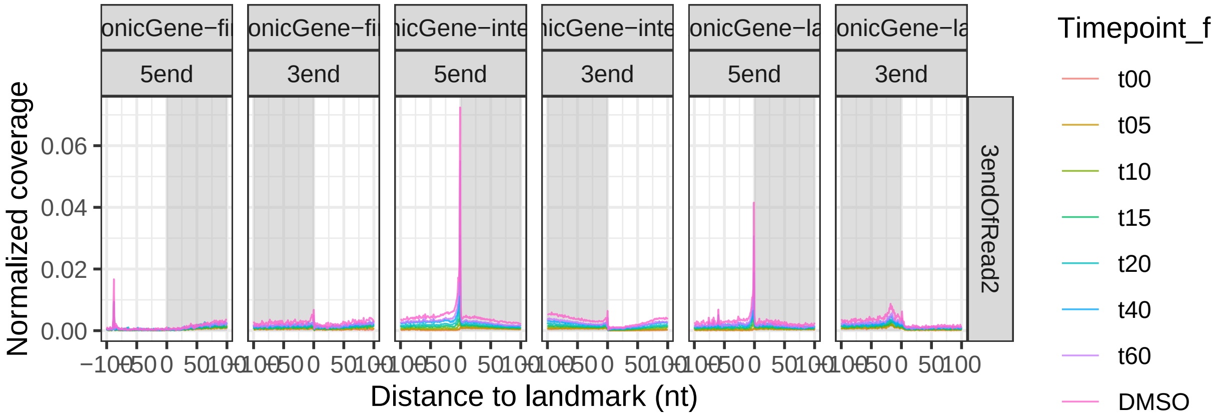
```
expressed_genes<-
annoBed %>%
  left_join(expression_vector) %>%
  filter(TotalNumberOfExons > 1 & !grepl("snRNA|rRNA|TR_C_gene|IG_C_pseudogene|miRNA|misc_RNA", Biotype
  select(geneID, ctrl_RNAseq_expr, ctrl_TTseq_expr) %>%
  unique() %>%
  mutate(ctrl_RNAseq_expr = as.numeric(ctrl_RNAseq_expr))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

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

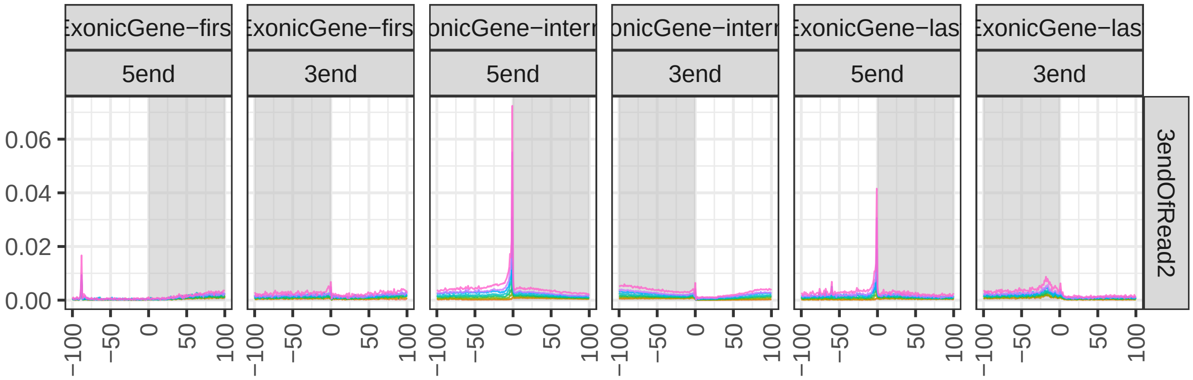
```
number_of_intron_annotations<-
  annoBed %>%
  filter(TotalNumberOfExons > 1 & !grepl("snRNA|rRNA|TR_C_gene|IG_C_pseudogene|miRNA|misc_RNA", Biotype
  filter(geneID %in% expressed_genes$geneID) %>%
  group_by(GeneStructure) %>%
  summarise(intron_count =n())

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$geneID) %>%
  group_by(Sample,GeneStructure, DistToLandmark) %>%
  summarise(sum_RNAseq_norm_count_norm_annotation_number = sum(norm_count)) %>%
  left_join(number_of_intron_annotations) %>%
  mutate(sum_RNAseq_norm_count_norm_annotation_number = sum_RNAseq_norm_count_norm_annotation_number/in
  separate(Sample, c("Protein","Rep","Timepoint","readType","region"), sep = "_") %>%
  filter(Timepoint != "negative") %>%
  mutate(Timepoint_f = case_when(
    Timepoint == "PBSDRB" ~ "t00",
    TRUE ~ Timepoint )) %>%
  mutate(readType = gsub("3CLIP", "3endOfRead2", readType),
         region = factor(region, levels = c("5end","3end")),
         Timepoint = factor(Timepoint, levels = c("PBSDRB", "t00", "t05", "t10", "t15", "t20", "t40", "
         Timepoint_f = factor(Timepoint_f, levels = c("t00", "t05", "t10", "t15", "t20", "t40", "t60",
  ggplot() +
  geom_rect(data = data.frame(region = "3end"), aes(xmin = -100, xmax = 0, ymin = 0, ymax = Inf), alpha
  geom_rect(data = data.frame(region = "5end"), aes(xmin = 0, xmax = 100, ymin = 0, ymax = Inf), alpha
  geom_line(aes(x=DistToLandmark, y = sum_RNAseq_norm_count_norm_annotation_number, col = Timepoint_f),
  facet_grid(readType ~ GeneStructure + region, scale = "free") +
 # labs(subtitle=paste0("exonsize>99, filtered out non-RNAPII genes, Excluded LINC00324, over 1 RNAseq"
#                       caption="normalised to gene expression, aggrigated reads, divided by number of
#                       n=",GENECOUNT$geneCount)) +
  xlab("Distance to landmark (nt)") +
  ylab("Normalized coverage")+
  theme_bw()
```

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

```
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("fig5c.pdf",  height = 2, width =6)
```