snoRNA_intron

RAC

24/08/2020

```
knitr::opts_chunk$set(tidy.opts=list(width.cutoff=60), tidy=TRUE)
#!/usr/bin/env Rscript
#
#suppressMessages(library(dplyr))
#suppressMessages(library(tidyr))
#suppressMessages(library(fuzzyjoin))
#suppressMessages(library(stringr))
#args = commandArgs(trailingOnly=TRUE)
#
#ANNOTATION_BED_FILEPATH<-args[1]
#EXPRESSION_VECTOR_FILEPATH<-args[2]
#IN_FILE_PATH<-args[3]
#HOST_GENE_INDEX_FILEPATH<-args[4]
#OUTFILE_PATH<-args[5]
#
##Function to process count files --------------------------
#wrangle_bed_counts <- function(dataframe) {
#  dataframe %>%
#    #the gene ID is complicated and has different number of columns for some snoRNAs, best to label by
#    mutate(snoRNALocation = case_when(
#      grepl(":::intronic:::", geneID) ~ "intronic",
#      grepl(":::nonintronic:::", geneID) ~ "non_intronic"
#    )) %>%
#    mutate(region = case_when(
#      grepl("intron3ss", geneID) ~ "intron3ss",
#      grepl("intron5ss", geneID) ~ "intron5ss",
#      grepl("mature3end", geneID) ~ "mature3end",
#      grepl("mature5end", geneID) ~ "mature5end",
#    )) %>%
#      mutate(snoRNAType = case_when(
#      grepl("SNORD|snoU|U3|U8|snoMe28S-Am2634|snoMBII|snoZ|snosnR66", geneID) ~ "cdBox",
#      grepl("SNORA|ACA|RNU105C|RNU105B", geneID) ~ "HACA",
#      grepl("SCARNA", geneID) ~ "scaRNA",
#      TRUE ~ "other"
#    )) %>%
#    mutate(geneID = as.character(geneID)) %>%
#    #fuzzy_left join allows string matching rather than exact matching. Noticed more output afterwards
#    fuzzy_left_join(host_gene_index, by =c("geneID" = "geneID"), match_fun = str_detect) %>%
#    separate(hostGeneID, into=c("hostGeneID","transcriptBiotype","intronNumber","distFromTSS"),sep = "
#    unique() %>%
#    #left join the expression vector using specific predefined columns.
#    left_join(expression_vector, by = c("hostGeneID" = "geneID")) %>%
#    mutate(norm_count = count/ctrl_RNAseq_expr) %>%
#    select(-count, -ctrl_RNAseq_expr) %>%
#    group_by(Sample,region,snoRNAType, snoRNALocation, 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
#}
#
#
#
#
##Load expression vector ------------------------------------
#print("load expression vector")
#
#load(EXPRESSION_VECTOR_FILEPATH)
#
#expression_vector<-left_join(
#  (as.data.frame(ctrl_RNAseq_expr) %>%
#     tibble::rownames_to_column(var = "geneID")),
#  (as.data.frame(ctrl_TTseq_expr) %>%
#     tibble::rownames_to_column(var = "geneID"))
#) %>%
#  select(-ctrl_TTseq_expr) %>%
#  mutate(ctrl_RNAseq_expr = case_when(
#    ctrl_RNAseq_expr ==0 ~ min(ctrl_RNAseq_expr[ctrl_RNAseq_expr > 0]),
#    TRUE ~  ctrl_RNAseq_expr
#  ))
#
#
##load annobed ------------------------
#print("load annotation bed")
#
#annoBed<-read.table(ANNOTATION_BED_FILEPATH, sep = "\t", header = F)
#print(ncol(annoBed))
#colnames(annoBed)<-c("chr", "start", "end", "geneID", "score", "strand")
#
#number_of_intron_annotations<-
#  annoBed %>%
#  mutate(snoRNALocation = case_when(
#    grepl(":::intronic", geneID) ~ "intronic",
#    grepl(":::nonintronic", geneID) ~ "non_intronic"
#  )) %>%
#  mutate(snoRNAType = case_when(
#      grepl("SNORD|snoU|U3|U8|snoMe28S-Am2634|snoMBII|snoZ|snosnR66", geneID) ~ "cdBox",
#      grepl("SNORA|ACA|RNU105C|RNU105B", geneID) ~ "HACA",
#      grepl("SCARNA", geneID) ~ "scaRNA",
#      TRUE ~ "other"
#    )) %>%
#  group_by(snoRNALocation, snoRNAType) %>%
#  summarise(intron_count =n())
#
##load index file ------------------
#print("load index file")
```

```r
#
#host_gene_index<-read.table(HOST_GENE_INDEX_FILEPATH, sep = "\t", header = F)
#colnames(host_gene_index)<-c("geneID", "hostGeneID")
#host_gene_index$geneID<-as.character(host_gene_index$geneID)
#
#
##load count file -----------------
#print("load count file")
#
#int_exon_junction_coverage<-read.table(IN_FILE_PATH, sep = "\t", header = F)
#colnames(int_exon_junction_coverage)<-c("Sample","chr", "start", "end", "geneID", "DistToLandmark", "s
#
#int_exon_junction_coverage<-
#  wrangle_bed_counts(int_exon_junction_coverage)
#
##rbind the files together and write to output filename
#
#OUTFILE<-int_exon_junction_coverage
#
#write.table(OUTFILE, OUTFILE_PATH, quote = F, sep = "\t", row.names = F)
```

```r
# suppressMessages(library(dplyr))
# suppressMessages(library(tidyr))
# suppressMessages(library(stringr))
# suppressMessages(library(ggplot2)) dir.create('figs/')
# dir.create('figs/snoRNA_coverage_plots') load in the file
# annotation_file<-read.table('snoRNAs.GRCh38andrefGene.mature.bed')
# colnames(annotation_file)<-c('chr', 'start', 'end',
# 'geneID', 'score', 'strand') count number of snoRNAs to
# provide a number to normalise to. annotation_number<-
# annotation_file %>% mutate(snoRNALocation = case_when(
# grepl(':::intronic', geneID) ~ 'intronic',
# grepl(':::nonintronic', geneID) ~ 'non_intronic' )) %>%
# mutate(snoRNAType = case_when(
# grepl('SNORD|snoU|U3|U8|snoMe28S-Am2634|snoMBII|snoZ|snosnR66',
# geneID) ~ 'cdBox', grepl('SNORA|ACA|RNU105C|RNU105B',
# geneID) ~ 'HACA', grepl('SCARNA', geneID) ~ 'scaRNA',
# TRUE ~ 'other' )) %>% group_by(snoRNALocation) %>%
# summarise(n=n())

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

```
## Registering fonts with R
```

```r
# read in count file

snoRNA_and_host_introns <- read.table("../../data/xiCLIP_all_hg38_snoRNAs_and_host_introns_no_gene_norm_
    header = T)
```

```r
head(snoRNA_and_host_introns)
```

```
##                       Sample    region snoRNAType snoRNALocation DistToLandmark
## 1 ALYREF_1_DMSO_3endOfRead2 intron3ss      cdBox       intronic           -100
## 2 ALYREF_1_DMSO_3endOfRead2 intron3ss      cdBox       intronic            -99
## 3 ALYREF_1_DMSO_3endOfRead2 intron3ss      cdBox       intronic            -98
## 4 ALYREF_1_DMSO_3endOfRead2 intron3ss      cdBox       intronic            -97
## 5 ALYREF_1_DMSO_3endOfRead2 intron3ss      cdBox       intronic            -96
## 6 ALYREF_1_DMSO_3endOfRead2 intron3ss      cdBox       intronic            -95
##   sum_count intron_count sum_count_norm_annotation_number
## 1 1175.6097          393                        2.9913734
## 2 1315.9832          393                        3.3485577
## 3  407.9558          393                        1.0380555
## 4  140.3714          393                        0.3571792
## 5  109.6652          393                        0.2790464
## 6  741.3364          393                        1.8863521
```

```r
# group by snoRNA type, location and each bin and the sum
# the count. This is then joined to the appropriate
# annotation number and futher normalised

# wrangle in the rest of the factors
snoRNA_and_host_introns_wrangled_sum_for_graph <- snoRNA_and_host_introns %>%
    separate(Sample, c("Protein", "Rep", "Timepoint", "Read")) %>%
    mutate(region = factor(region, c("intron5ss", "mature5end",
        "mature3end", "intron3ss")), Timepoint = factor(Timepoint,
        c("negative", "PBSDRB", "t00", "t05", "t10", "t15", "t20",
            "t40", "t60", "DMSO"))) %>%
    mutate(Timepoint_f = case_when(Timepoint == "PBSDRB" ~ "t00",
        TRUE ~ as.character(Timepoint))) %>%
    mutate(Timepoint_f = factor(Timepoint_f, levels = c("negative",
        "t00", "t05", "t10", "t15", "t20", "t40", "t60", "DMSO")))

head(snoRNA_and_host_introns_wrangled_sum_for_graph)
```

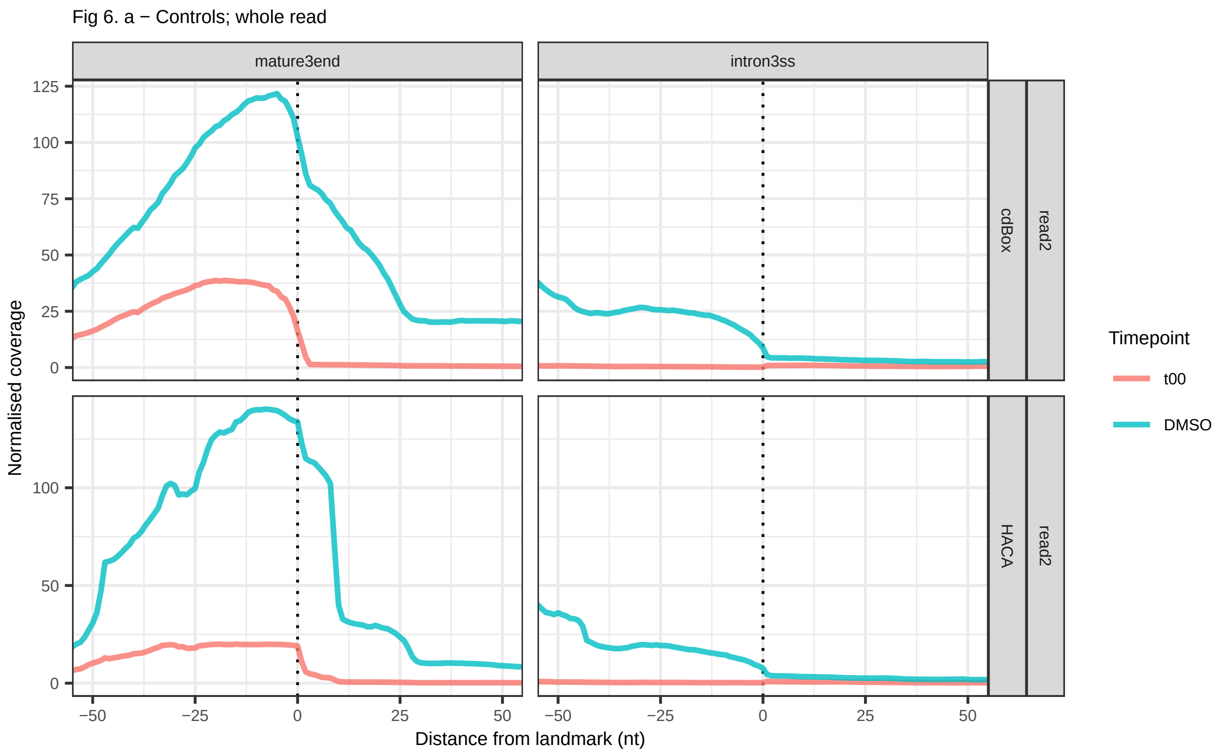
```
##   Protein Rep Timepoint        Read    region snoRNAType snoRNALocation
## 1  ALYREF   1      DMSO 3endOfRead2 intron3ss      cdBox       intronic
## 2  ALYREF   1      DMSO 3endOfRead2 intron3ss      cdBox       intronic
## 3  ALYREF   1      DMSO 3endOfRead2 intron3ss      cdBox       intronic
## 4  ALYREF   1      DMSO 3endOfRead2 intron3ss      cdBox       intronic
## 5  ALYREF   1      DMSO 3endOfRead2 intron3ss      cdBox       intronic
## 6  ALYREF   1      DMSO 3endOfRead2 intron3ss      cdBox       intronic
##   DistToLandmark sum_count intron_count sum_count_norm_annotation_number
## 1           -100 1175.6097          393                        2.9913734
## 2            -99 1315.9832          393                        3.3485577
## 3            -98  407.9558          393                        1.0380555
## 4            -97  140.3714          393                        0.3571792
## 5            -96  109.6652          393                        0.2790464
## 6            -95  741.3364          393                        1.8863521
##   Timepoint_f
## 1        DMSO
## 2        DMSO
## 3        DMSO
## 4        DMSO
## 5         DMSO
## 6         DMSO
```

```
figure_data <- snoRNA_and_host_introns_wrangled_sum_for_graph %>%
    filter(snoRNAType %in% c("cdBox", "HACA") & snoRNALocation ==
        "intronic" & Timepoint != "negative" & Protein == "RBM7" &
        region %in% c("mature3end", "intron3ss"))

head(figure_data)
```

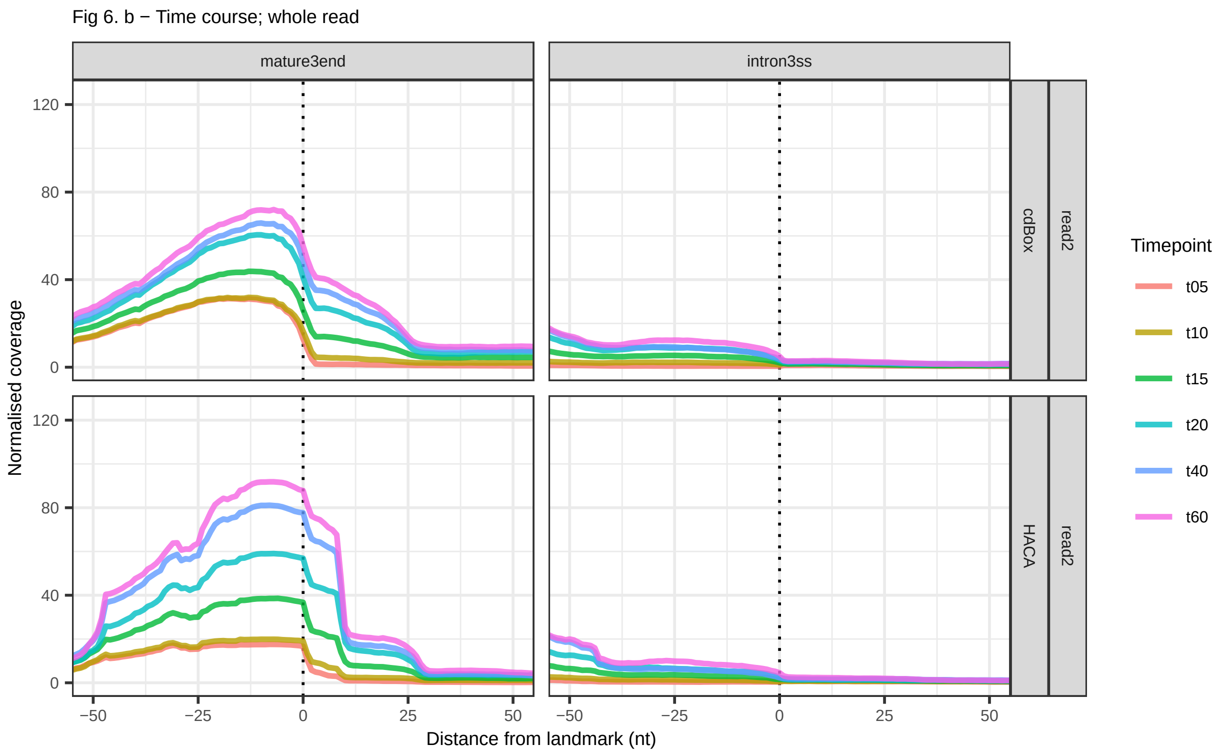
```
##   Protein Rep Timepoint        Read    region snoRNAType snoRNALocation
## 1    RBM7   1      DMSO 3endOfRead2 intron3ss      cdBox       intronic
## 2    RBM7   1      DMSO 3endOfRead2 intron3ss      cdBox       intronic
## 3    RBM7   1      DMSO 3endOfRead2 intron3ss      cdBox       intronic
## 4    RBM7   1      DMSO 3endOfRead2 intron3ss      cdBox       intronic
## 5    RBM7   1      DMSO 3endOfRead2 intron3ss      cdBox       intronic
## 6    RBM7   1      DMSO 3endOfRead2 intron3ss      cdBox       intronic
##   DistToLandmark sum_count intron_count sum_count_norm_annotation_number
## 1           -100  817.5904          393                        2.0803827
## 2            -99  524.9302          393                        1.3357003
## 3            -98  376.2774          393                        0.9574489
## 4            -97  283.3694          393                        0.7210417
## 5            -96  329.8234          393                        0.8392453
## 6            -95  548.1572          393                        1.3948020
##   Timepoint_f
## 1        DMSO
## 2        DMSO
## 3        DMSO
## 4        DMSO
## 5        DMSO
## 6        DMSO
```

#Figure 6 A-D. #Calculate coverage of read2 and 3'CLIP at nucleotide resolution over 101nt windows centred on the 3' end of snoRNAs or its corresponding downstream intron-exon junction. SnoRNAs are stratified by class (H/ACA or cdBox). Normalisation: read coverage over snoRNA window to host gene expression, aggregate reads, and divide by the total number of snoRNA annotations.

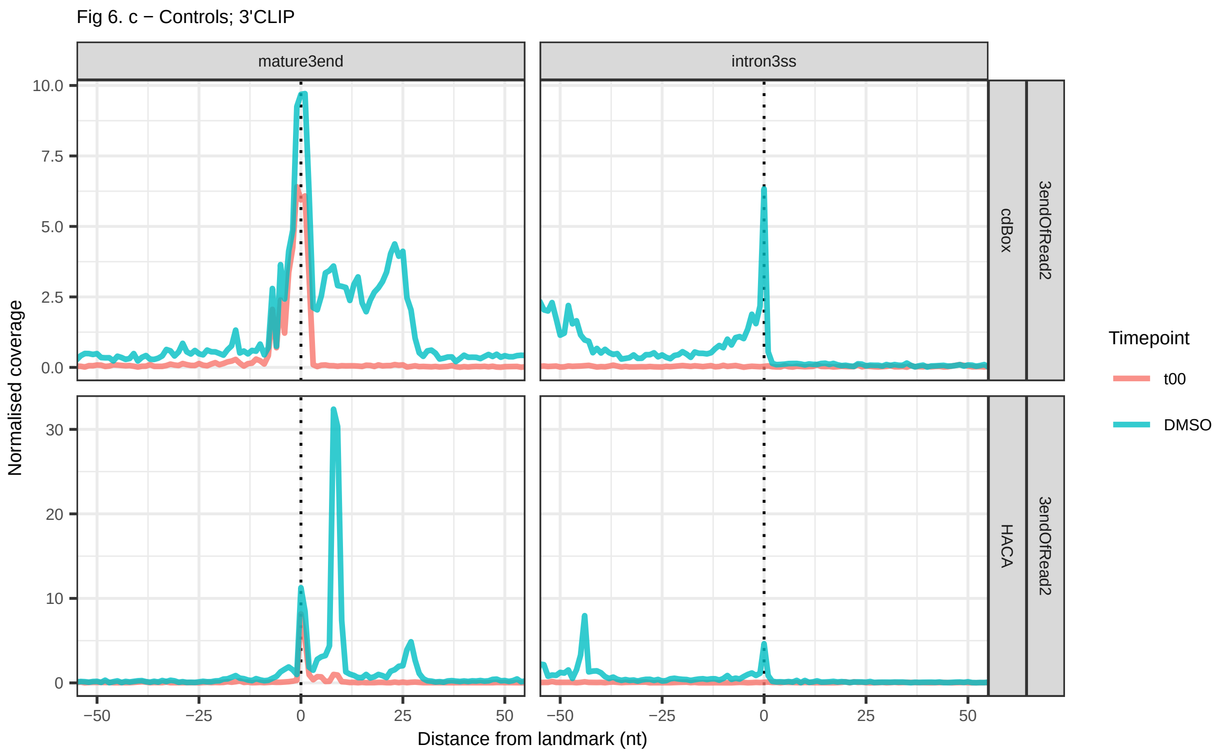
```
# Figure 6 A whole read plots
figure_data %>%
    filter(snoRNAType %in% c("HACA", "cdBox") & Read == "read2" &
        Timepoint %in% c("t00", "DMSO")) %>%
    ggplot() + geom_vline(xintercept = 0, linetype = "dotted",
    size = 0.5) + geom_line(aes(x = DistToLandmark, y = sum_count_norm_annotation_number,
    col = Timepoint), stat = "summary", fun = "mean", alpha = 0.8,
    size = 1) + facet_grid(Read + snoRNAType ~ region, scales = "free_y") +
    theme_bw() + coord_cartesian(xlim = c(-50, 50)) + xlab("Distance from landmark (nt)") +
    ylab("Normalised coverage") + labs(subtitle = "Fig 6. a - Controls; whole read") +
    theme(text = element_text(size = 7, family = "Arial"))
```



```r
# Figure 6 B
dummy_data <- figure_data %>%
    filter(snoRNAType %in% c("HACA", "cdBox") & Read == "read2" &
        Timepoint %in% c("t00", "DMSO"))

# whole read plots
figure_data %>%
    filter(snoRNAType %in% c("HACA", "cdBox") & Read == "read2" &
        !(Timepoint %in% c("t00", "DMSO"))) %>%
    ggplot() + geom_vline(xintercept = 0, linetype = "dotted",
    size = 0.5) + geom_line(aes(x = DistToLandmark, y = sum_count_norm_annotation_number,
    col = Timepoint), stat = "summary", fun = "mean", alpha = 0.8,
    size = 1) + facet_grid(Read + snoRNAType ~ region, scales = "free_y") +
    theme_bw() + coord_cartesian(xlim = c(-50, 50), ylim = c(0,
    125)) + xlab("Distance from landmark (nt)") + ylab("Normalised coverage") +
    labs(subtitle = "Fig 6. b - Time course; whole read") + theme(text = element_text(size = 7,
    family = "Arial"))
```

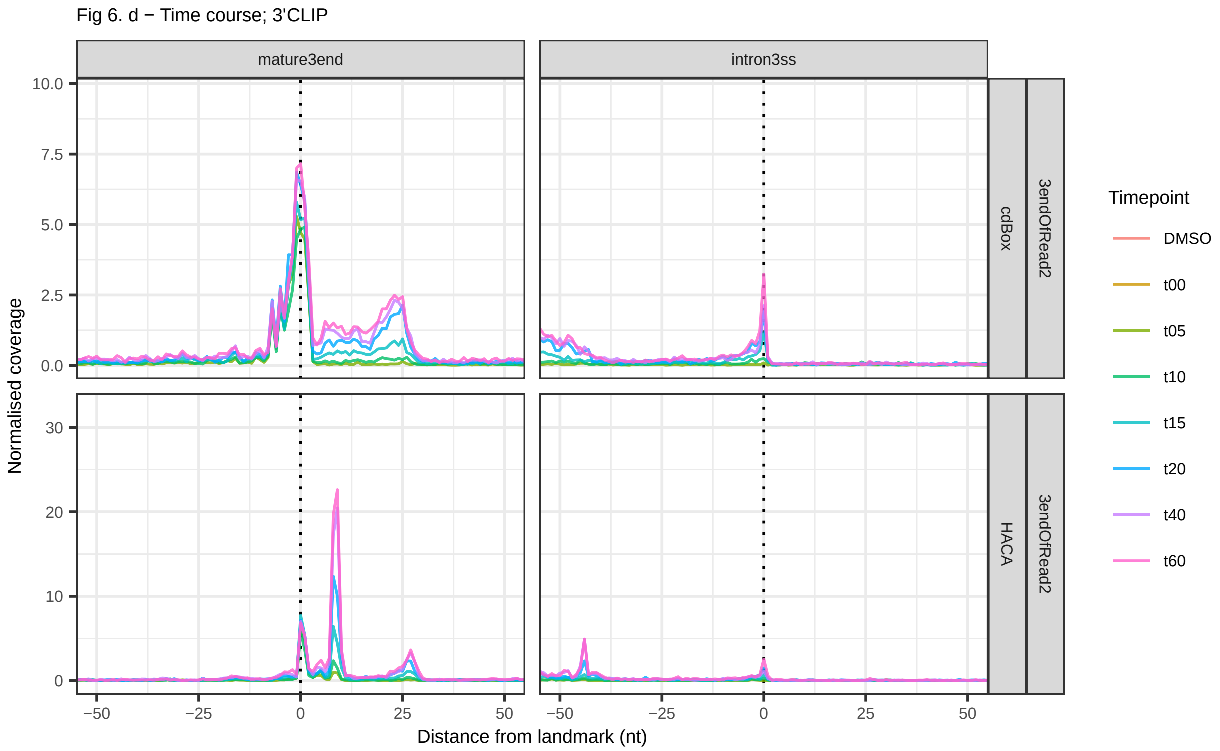


```r
# Figure 6 C whole read plots, 3'CLIP
figure_data %>%
    filter(snoRNAType %in% c("HACA", "cdBox") & Read == "3endOfRead2" &
        (Timepoint %in% c("t00", "DMSO"))) %>%
    ggplot() + geom_vline(xintercept = 0, linetype = "dotted",
    size = 0.5) + geom_line(aes(x = DistToLandmark, y = sum_count_norm_annotation_number,
    col = Timepoint), stat = "summary", fun = "mean", alpha = 0.8,
    size = 1) + facet_grid(Read + snoRNAType ~ region, scales = "free_y") +
    theme_bw() + coord_cartesian(xlim = c(-50, 50)) + xlab("Distance from landmark (nt)") +
    ylab("Normalised coverage") + labs(subtitle = "Fig 6. c - Controls; 3'CLIP") +
    theme(text = element_text(size = 7, family = "Arial"))
```



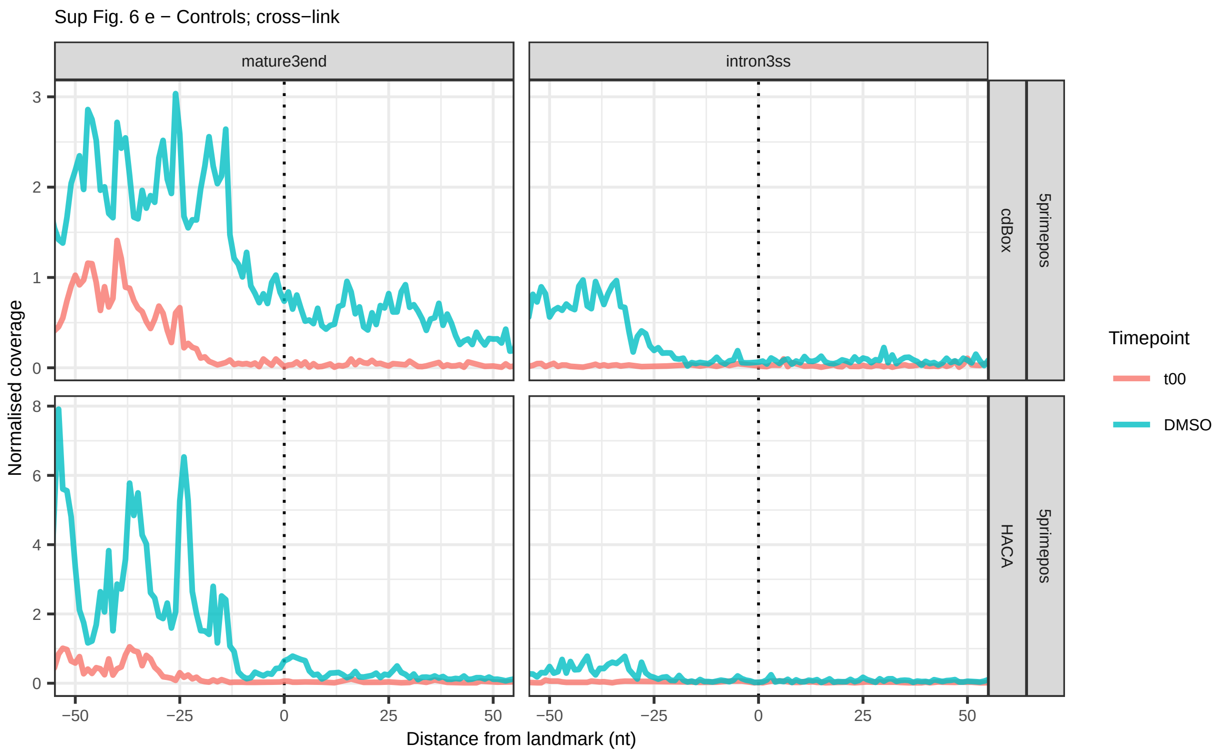
```r
# Figure 6 D invisible data to keep yaxis same
dummy_data <- figure_data %>%
    filter(snoRNAType %in% c("HACA", "cdBox") & Read == "3endOfRead2" &
        (Timepoint %in% c("t00", "DMSO")))

# whole read plots
figure_data %>%
    filter(snoRNAType %in% c("HACA", "cdBox") & Read == "3endOfRead2" &
        !(Timepoint %in% c("t00", "DMSO"))) %>%
    ggplot() + geom_vline(xintercept = 0, linetype = "dotted",
    size = 0.5) + geom_line(aes(x = DistToLandmark, y = sum_count_norm_annotation_number,
    col = Timepoint), stat = "summary", fun = "mean", alpha = 0.8,
    size = 0.5) + geom_blank(data = dummy_data, aes(x = DistToLandmark,
    y = sum_count_norm_annotation_number, col = Timepoint), stat = "summary",
    fun = "mean", alpha = 0.8, size = 1) + facet_grid(Read +
    snoRNAType ~ region, scales = "free_y") + theme_bw() + coord_cartesian(xlim = c(-50,
    50)) + xlab("Distance from landmark (nt)") + ylab("Normalised coverage") +
    labs(subtitle = "Fig 6. d - Time course; 3'CLIP") + theme(text = element_text(size = 7,
    family = "Arial"))
```

```
## Warning: Ignoring unknown parameters: alpha, size
```



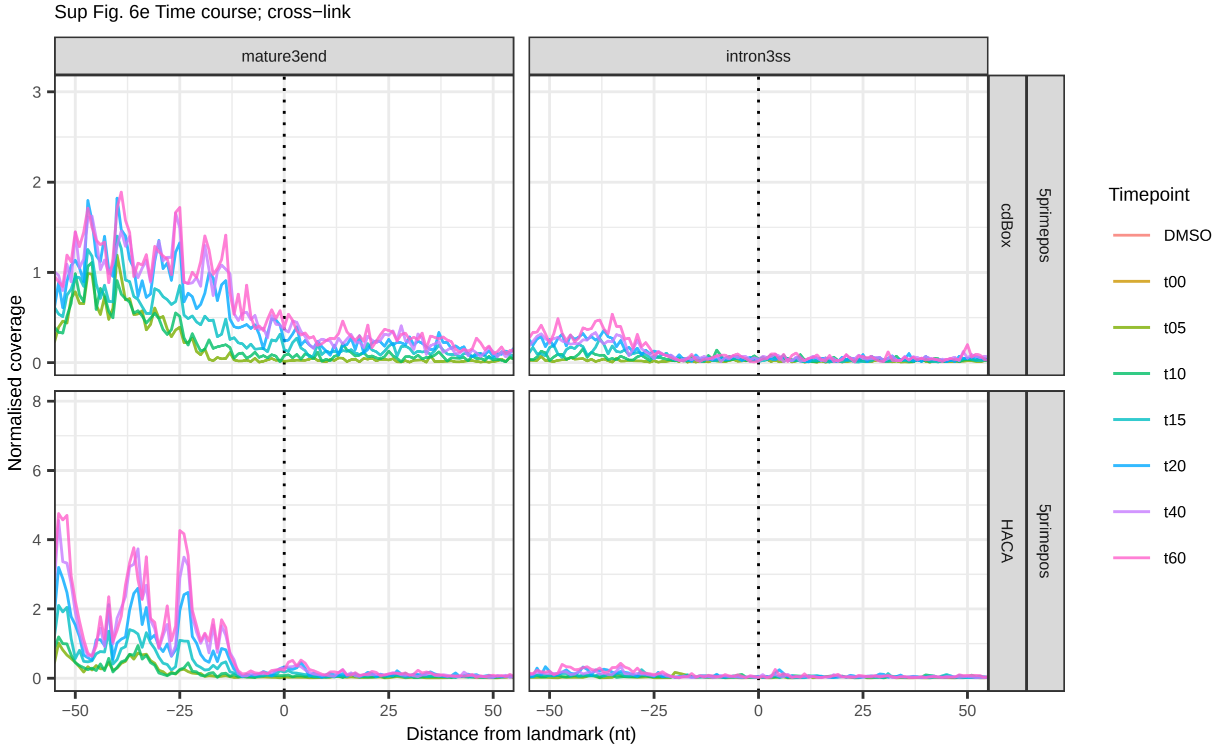
#Supplementary Figure 6 c - e coverage as above over snoRNAs, except showing cross-link sites also showing cross-link sites at 5' end, and upstream 5'SS for snoRNA containing introns

```r
# Supplementary Figure 6 e cross-link
figure_data %>%
    filter(snoRNAType %in% c("HACA", "cdBox") & Read == "5primepos" &
        (Timepoint %in% c("t00", "DMSO"))) %>%
    ggplot() + geom_vline(xintercept = 0, linetype = "dotted",
    size = 0.5) + geom_line(aes(x = DistToLandmark, y = sum_count_norm_annotation_number,
    col = Timepoint), stat = "summary", fun = "mean", alpha = 0.8,
    size = 1) + facet_grid(Read + snoRNAType ~ region, scales = "free_y") +
    theme_bw() + coord_cartesian(xlim = c(-50, 50)) + xlab("Distance from landmark (nt)") +
    ylab("Normalised coverage") + labs(subtitle = "Sup Fig. 6 e - Controls; cross-link") +
    theme(text = element_text(size = 7, family = "Arial"))
```

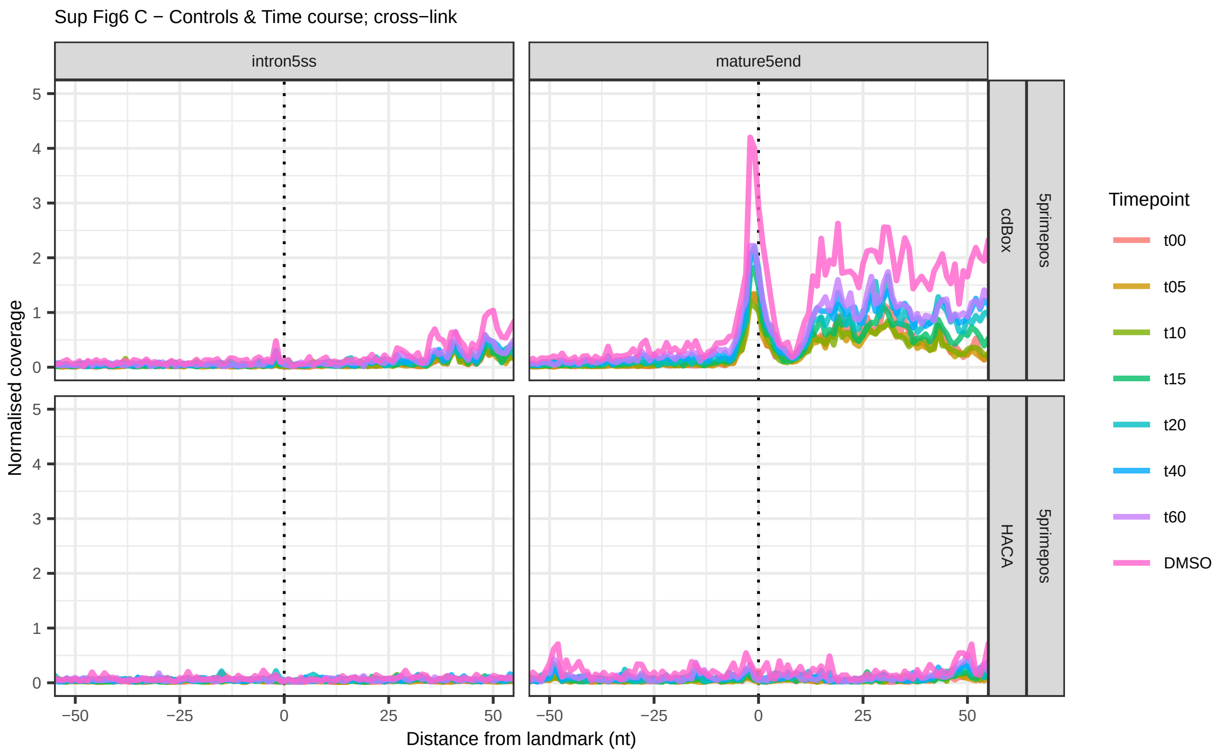


```
# Supplementary Figure 6 e invisible data to keep yaxis
# same
dummy_data <- figure_data %>%
    filter(snoRNAType %in% c("HACA", "cdBox") & Read == "5primepos" &
        (Timepoint %in% c("t00", "DMSO")))

# cross-link
figure_data %>%
    filter(snoRNAType %in% c("HACA", "cdBox") & Read == "5primepos" &
        !(Timepoint %in% c("t00", "DMSO"))) %>%
    ggplot() + geom_vline(xintercept = 0, linetype = "dotted",
    size = 0.5) + geom_line(aes(x = DistToLandmark, y = sum_count_norm_annotation_number,
    col = Timepoint), stat = "summary", fun = "mean", alpha = 0.8,
    size = 0.5) + geom_blank(data = dummy_data, aes(x = DistToLandmark,
    y = sum_count_norm_annotation_number, col = Timepoint), stat = "summary",
    fun = "mean", alpha = 0.8, size = 1) + facet_grid(Read +
    snoRNAType ~ region, scales = "free_y") + theme_bw() + coord_cartesian(xlim = c(-50,
    50)) + xlab("Distance from landmark (nt)") + ylab("Normalised coverage") +
    labs(subtitle = "Sup Fig. 6e Time course; cross-link") +
    theme(text = element_text(size = 7, family = "Arial"))
```

```
## Warning: Ignoring unknown parameters: alpha, size
```



```
supplementary_figure_data <- snoRNA_and_host_introns_wrangled_sum_for_graph %>%
    filter(snoRNAType %in% c("cdBox", "HACA") & snoRNALocation ==
        "intronic" & Timepoint != "negative" & Protein == "RBM7" &
        region %in% c("mature5end", "intron5ss"))

# supplementary figure 6 c

supplementary_figure_data %>%
    filter(snoRNAType %in% c("HACA", "cdBox") & Read == "5primepos") %>%
    ggplot() + geom_vline(xintercept = 0, linetype = "dotted",
    size = 0.5) + geom_line(aes(x = DistToLandmark, y = sum_count_norm_annotation_number,
    col = Timepoint), stat = "summary", fun = "mean", alpha = 0.8,
    size = 1) + facet_grid(Read + snoRNAType ~ region, scales = "free_y") +
    theme_bw() + coord_cartesian(xlim = c(-50, 50), ylim = c(0,
    5)) + xlab("Distance from landmark (nt)") + ylab("Normalised coverage") +
    labs(subtitle = "Sup Fig6 C - Controls & Time course; cross-link") +
    theme(text = element_text(size = 7, family = "Arial"))
```


Sup Fig6 C − Controls & Time course; cross−link
intron5ss
mature5end
cdBox
5primepos
HACA
5primepos
Normalised coverage
0
1
2
3
4
5
Distance from landmark (nt)
−50
−25
0
25
50
Timepoint
t00
t05
t10
t15
t20
t40
t60
DMSO