RNAduplex on all binding sites, (1007nt window)

Melina Klostermann

30 October, 2023

Contents

1 Libraries and settings

```
# ----------------------------------------
# libraries
# ----------------------------------------
library(tidyverse)
library(GenomicRanges)
library(colorspace)
library(gghalves)
library(BSgenome.Mmusculus.UCSC.mm10)
library(Biostrings)
library(ComplexHeatmap)
library(ggpubr)
library(circlize)


here <- here::here()

# ----------------------------------------
# settings
# ----------------------------------------

out <- paste0(here,"/Figure6/02_other_window_size_for_revision/")
source(paste0(here,"/Supporting_scripts/themes/theme_paper.R"))
source(paste0(here,"/Supporting_scripts/themes/CustomThemes.R"))

set.seed(2)
```

2 What was done?

- for revision
- run RNAduplex on a region of 500nt before until 500nt after all 6mer seeds in the expressed transcriptome
- show how many duplexes start at the seed position

3 Files

```
# ----------------------------------------
# MREs
# ----------------------------------------

mir181_bs <- readRDS(paste0(here, "/Figure5/01_Seed_motifes/mir181_bs_with_seeds_transcripts.rds"))

mir181_enriched_set <- mir181_bs %>%
  as.data.frame(.) %>%
  subset(set %in% c("ago_bs_mir181_chi&mir181_enriched", "mir181_enriched"))

# transcript seqeuces
transcript_fasta <- readDNAStringSet("/Users/melinaklostermann/Documents/projects/anno/gencodevM23/genc

# annotation
anno <- readRDS(paste0(here,"/Supporting_scripts/annotation_preprocessing/annotation.rds"))

# expressed_genes
expressed_genes <- readRDS(paste0(here, "/Supporting_scripts/TPMs-RNAseq/expressed_genes.rds"))

# seeds
seeds <- read.csv(paste0(here,"/Figure3/01_Ribosome_profiling_pipeline/RPF_masterframe.csv"))

# Ribofootprint
rfp <- read.csv(paste0(here,"/Figure3/01_Ribosome_profiling_pipeline/RPF_masterframe.csv") )
rfp <- rfp %>% mutate(Gene = sub("\\..*", "", Gene))
```

4 Run RNAplfold on all 6mers

4.1 Get transcript sequences

```
###################
# Get sequences of mature transcripts
##################

# expressed transcripts ( = transcripts with any crosslinks)
annotation_transcripts <- anno[anno$type == "transcript"]
expressed_transcripts <- annotation_transcripts[annotation_transcripts$geneID %in% expressed_genes]


# transcript seqeunces
transcript_anno_meta <- names(transcript_fasta)
transcript_anno_meta <- data.frame(all = transcript_anno_meta) %>%
  tidyr::separate(., col = all,
                  into = c("transcript_id", "gene_id", "a", "b", "isoform_name", "gene_name", "entrez_g
```

```r
names(transcript_fasta) <- sub("\\..*", "", transcript_anno_meta$transcript_id)

# get transcript_id and transcript lengths from fasta names
transcript_fasta_df <- data.frame(tx_name = names(transcript_fasta), width = width(transcript_fasta))
```

4.2 write fasta

```r
###################
# 1006 nt window
##################

# make window around mir181 binding sites
w <- 1007
mir181_enriched_set_1007nt <- mir181_enriched_set %>%
  left_join(transcript_fasta_df, by= c(seqnames = "tx_name"), suffix = c(".bs", ".tx")) %>%
  mutate(end = end + 500, start = start -500, strand = "*") %>%
  dplyr::filter((end <  width.tx)  & (start > 0)) %>%
  makeGRangesFromDataFrame(., keep.extra.columns = T) %>%
  unique(.)

mir181_enriched_set_1007nt <- mir181_enriched_set_1007nt[width(mir181_enriched_set_1007nt)==w]

mir181_enriched_set_1007nt_seqs <- Biostrings::getSeq(x = transcript_fasta, names = mir181_enriched_set_

NROW(mir181_enriched_set_1007nt)
```

```
## [1] 2865
```

```r
# oneline fasta
#writeXStringSet(mir181_enriched_set_1007nt_seqs, filepath = paste0(out,"mir181_enriched_set_1007nt.fas
```

```r
# specific column for 6mer seeds
seed_from_1007nt <- as.data.frame(mir181_enriched_set_1007nt) %>%
  mutate(end = end - 200, start = start +30) %>%
  unnest(all_seeds_200down)

seed_from_1007nt <- seed_from_1007nt %>%
    subset(., ((.$Seeds_200down.type %in% c("seed_6mer", "seed_6mer_wobble")) | is.na(.$Seeds_200down.ty
  group_by(mir181BS_ID) %>%
  arrange(Seeds_200down.start, .by_group = T) %>%
  dplyr::slice(1) %>%
  ungroup() %>%
  mutate(Seeds_200down.type = case_when(is.na(Seeds_200down.type) ~ "no_seed",
                                        T ~ Seeds_200down.type))
```

4.3 RNAduplex output

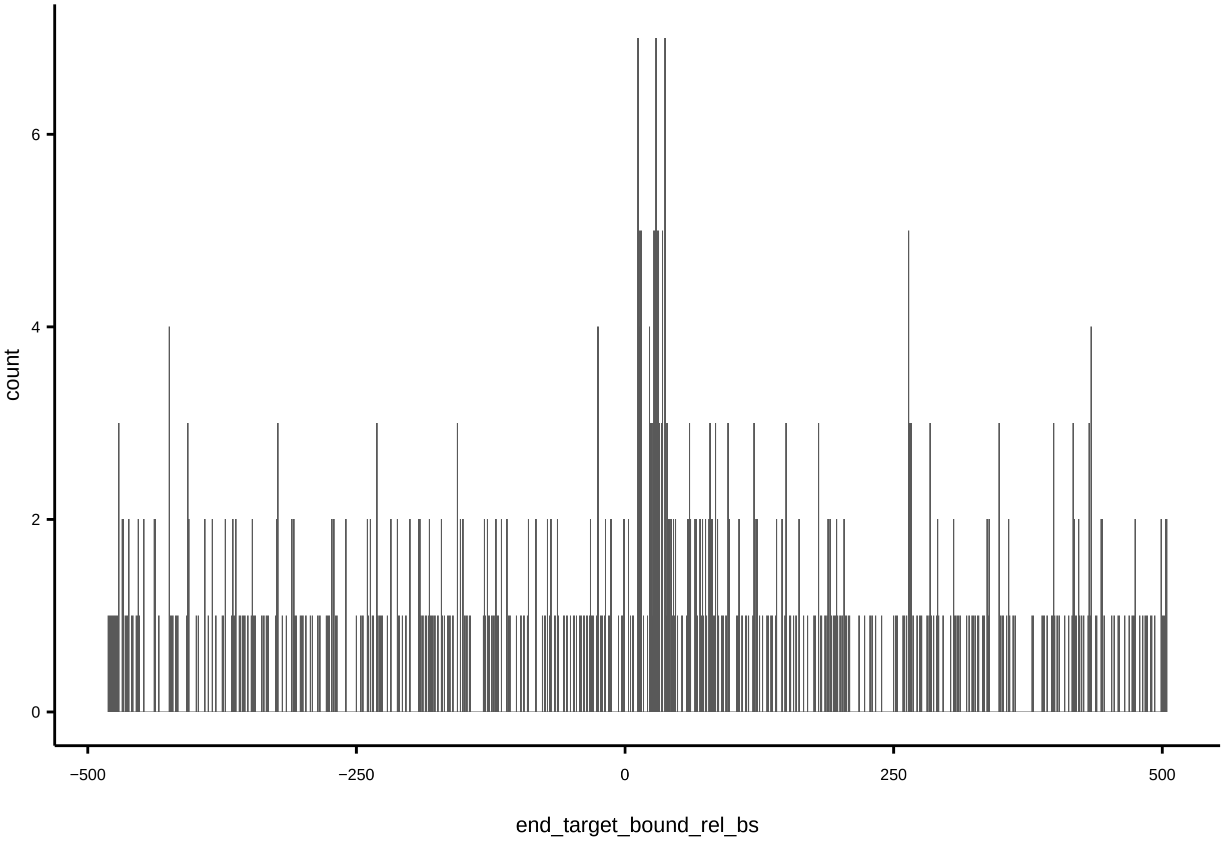
```r
# --------------------
# Read in RNAduplex output and clean
# --------------------

struct <- read_table(paste0(out,"/mir181_enriched_set_1007nt.struct"), col_names = c("seq", "mir", "stru
struct <- struct %>%
  rowwise(.) %>%
  mutate(struct_mir = str_split_1(structure, pattern = "&")[2],
         struct_target = str_split_1(structure, pattern = "&")[1],
         start_mir = str_split_1(position_mir, pattern = ",")[1] %>% as.numeric(.),
         end_mir = str_split_1(position_mir, pattern = ",")[2] %>% as.numeric(.),
         start_target = str_split_1(position_seq, pattern = ",")[1] %>% as.numeric(.),
         end_target = str_split_1(position_seq, pattern = ",")[2] %>% as.numeric(.),
         min_free_energy = gsub("[()]", "", min_free_energy) %>% as.numeric(.),
         norm_free_energy = min_free_energy / (nchar(structure)-1),
         # the last bound position in the target = the position that is bound by the beginning of the m
         end_target_bound = end_target - nchar(str_split_1(rev(struct_target), pattern = "[(]")[1]))

struct <- struct %>%
  mutate(struct_bound_mir_full = paste0(
    paste0( rep(".", start_mir), collapse = ""),
    struct_mir,
    paste0(rep(".", (23 - end_mir)), collapse = ""),
    collapse = ""))

struct$mir181BS_ID <- mir181_enriched_set_1007nt$mir181BS_ID

head(struct)
```

```
## # A tibble: 6 x 17
## # Rowwise:
##   seq   mir   structure          position_seq x     position_mir min_free_energy
##   <chr> <chr> <chr>              <chr>        <chr> <chr>                  <dbl>
## 1 >seq  >mir  .((((.((....(((((~ 750,778      :     2,23                   -18.2
## 2 >seq  >mir  .(((.(((((((.....~ 511,540      :     1,23                   -21.1
## 3 >seq  >mir  .(((((((..(((.(..~ 330,348      :     6,23                   -13.6
## 4 >seq  >mir  .(((.((((..(((((.~ 91,116       :     2,23                   -18.9
## 5 >seq  >mir  .(((((((((..((((.~ 416,432      :     1,20                   -18  
## 6 >seq  >mir  .(((((.((((((((((~ 504,526      :     1,23                   -16.8
## # i 10 more variables: struct_mir <chr>, struct_target <chr>, start_mir <dbl>,
## #   end_mir <dbl>, start_target <dbl>, end_target <dbl>,
## #   norm_free_energy <dbl>, end_target_bound <dbl>,
## #   struct_bound_mir_full <chr>, mir181BS_ID <int>
```

```
# --------------------
# make structur matrix of mir
# --------------------
struct_bound_mir_mat <- data.frame(s = struct$struct_bound_mir_full)
struct_bound_mir_mat <- struct_bound_mir_mat %>% separate(., s, as.character(1:25), sep = "")
struct_bound_mir_mat <- as.matrix(struct_bound_mir_mat)
struct_bound_mir_mat <- struct_bound_mir_mat[,-1]
n <- ncol(struct_bound_mir_mat)

struct_bound_mir_mat[struct_bound_mir_mat == ")"] = 1
struct_bound_mir_mat[struct_bound_mir_mat == "."] = 0
struct_bound_mir_mat[struct_bound_mir_mat == ""] = NA
struct_bound_mir_mat <- as.numeric(struct_bound_mir_mat) %>% matrix(., ncol = n)

# --------------------
# make data frame with extra info
# --------------------
struct_bound_mir_df <- as.data.frame(struct_bound_mir_mat)

st <- struct %>%
  as.data.frame(.) %>%
  dplyr::select( min_free_energy, start_target, end_target, end_target_bound, norm_free_energy, mir181BS

struct_bound_mir_df <- cbind(struct_bound_mir_df, struct)

# info from mir binding sites
s <- seed_from_1007nt %>%
  dplyr::select( seqnames, scoreMax, region, resBs.log2FoldChange, Seeds_200down.type, Seeds_200down.sta

struct_bound_mir_df <- left_join(struct_bound_mir_df, s, by = "mir181BS_ID")
```

4.4 Duplex start in relation to 6mer start

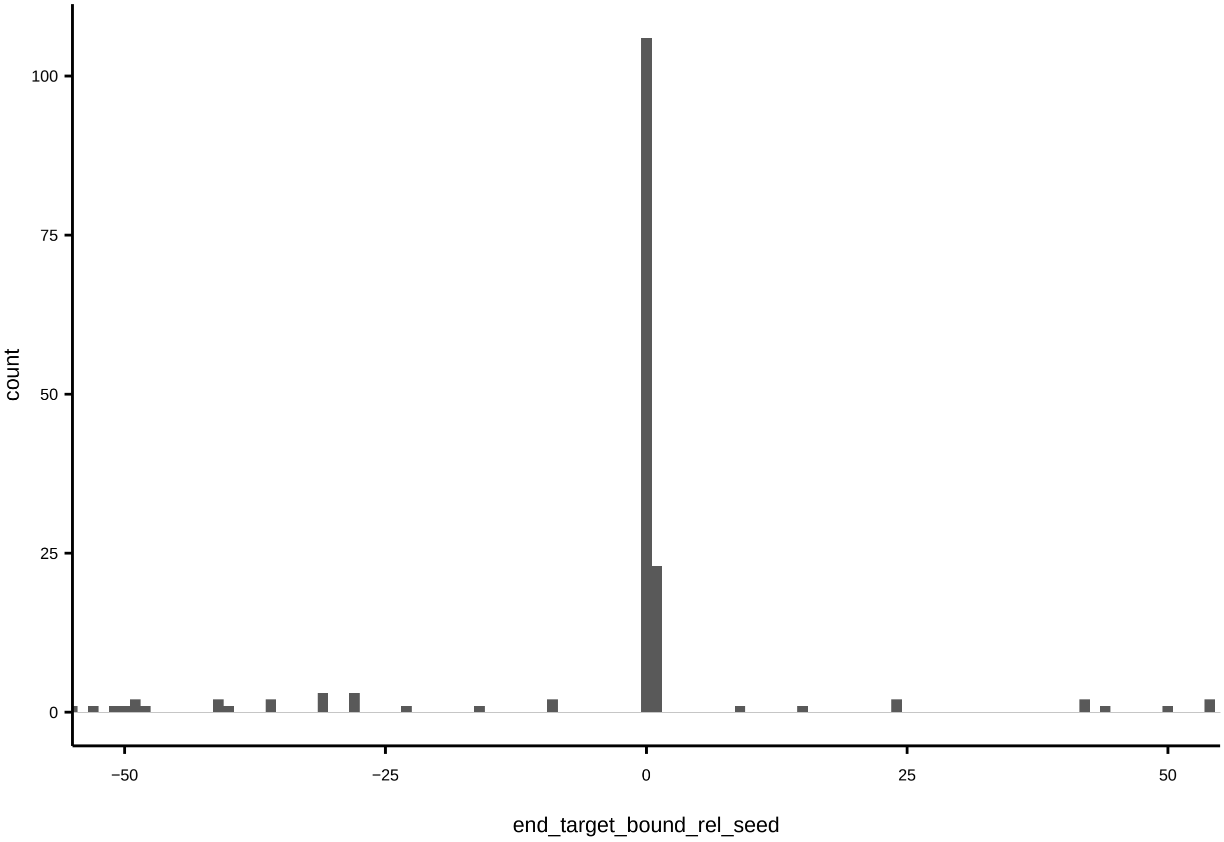
```
struct_bound_mir_df$end_target_bound_rel_bs <- struct_bound_mir_df$end_target_bound -500
struct_bound_mir_df$end_target_bound_rel_seed <- struct_bound_mir_df$end_target_bound_rel_bs - struct_bo

struct_bound_mir_df_6mer <- struct_bound_mir_df %>% subset(Seeds_200down.type == "seed_6mer")

ggplot(struct_bound_mir_df_6mer, aes(x = end_target_bound_rel_bs ))+
  geom_histogram( binwidth = 1)+
  theme_paper()
```



```
nrow(struct_bound_mir_df_6mer)
```

```
## [1] 623
```

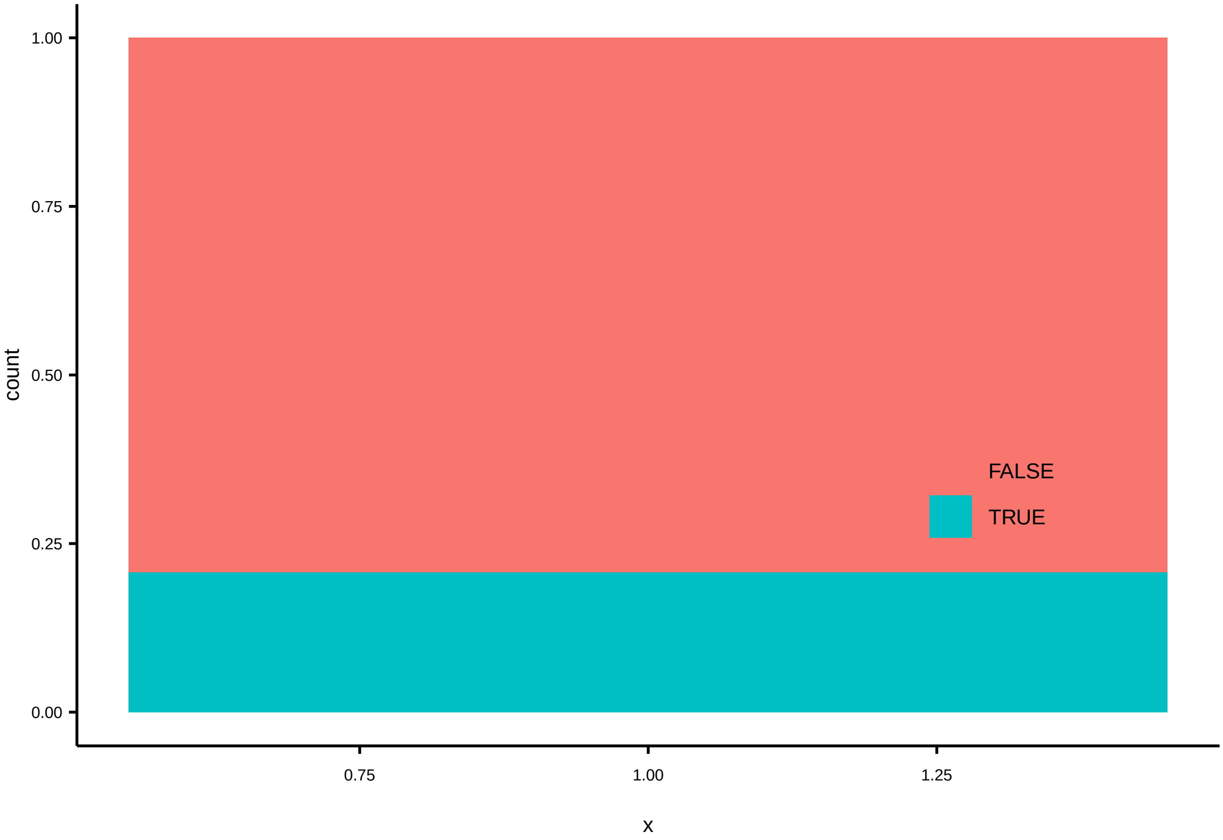
```
p1 <- ggplot(struct_bound_mir_df_6mer, aes(x = end_target_bound_rel_seed ))+
  geom_histogram( binwidth = 1)+
  theme_paper()+
  coord_cartesian(xlim=c(-50,50))

p1
```



```
ggsave(p1, filename = paste0(out, "Revision_SuppFigure6C_duplex_start_position_1007nt.pdf"), width = 6,
```

4.5 Duplexes that use correct 6mer

```
struct_bound_mir_df_6mer <- struct_bound_mir_df_6mer %>%
  rowwise(.) %>%
  mutate(correct_duplex_end = end_target_bound_rel_seed %in% c(0,1),
         paired_6mer = all(across(V3:V8) == 1),
         canonical_duplex = all(c(correct_duplex_end, paired_6mer)))


p2 <- ggplot(struct_bound_mir_df_6mer, aes(x = 1, fill = canonical_duplex)) +
  geom_bar( position = "fill")+
  theme_paper()

p2
```



```
t <- table(struct_bound_mir_df_6mer$canonical_duplex)
t
```

```
##
## FALSE  TRUE
##   494   129
```

```
t/sum(t)
```

```
##
##     FALSE      TRUE
## 0.7929374 0.2070626
```

```
ggsave(p2, filename = paste0(out, "Revision_SuppFigure6C_canonical_duplex_seeds_bar_1007nt.pdf"), width
```