Assign mir181 binding sites to a specific transcript

Melina Klostermann

06 July, 2023

Contents

1 Libraries and settings

```
# ----------------------------------------
# libraries
# ----------------------------------------
library(tidyverse)
library(GenomicRanges)
library(GenomicFeatures)


# ----------------------------------------
# settings
# ----------------------------------------
out <- "/Users/melinaklostermann/Documents/projects/AgoCLIP_miR181/R_github/miR181_paper/Figure2/02_ass
```

2 What was done?

- The main expressed transcript isoform (as defined by APRIS) of each mir181 binding site is obtained
- Then the mir181 binding sites are mapped to their respective transcript
- The transcript annotations are later used for motiv discovery and structure predictions

```
#-------------------------------------
# Files
#-------------------------------------
anno <- readRDS("/Users/melinaklostermann/Documents/projects/AgoCLIP_miR181/R_github/miR181_paper/Metho

mir181_bs <- readRDS("/Users/melinaklostermann/Documents/projects/AgoCLIP_miR181/R_github/miR181_paper/
```

```
# get appris transcripts (when there are multiple take the longest)
transcripts <- anno[anno$type=="transcript"] %>% as.data.frame(.)

transcripts$transcript_id <- sub("\\..*", "", transcripts$transcript_id)

transcripts_appris <- transcripts[grepl(transcripts$tag, pattern= "appris_principal_1"),] %>%
  group_by(geneID) %>%
```

```
  arrange(desc(width), .by_group = T) %>%
  dplyr::slice(1)


# add transcript id to binding sites
transcripts_appris <- makeGRangesFromDataFrame(transcripts_appris, keep.extra.columns = T)
mir181_bs <- makeGRangesFromDataFrame(mir181_bs, keep.extra.columns = T)

idx <- findOverlaps(mir181_bs, transcripts_appris)

transcripts_appris <- as.data.frame(transcripts_appris) %>%
  dplyr::select(seqnames, start, end, width, strand, geneID, transcript_id)

colnames(transcripts_appris) <- paste0(colnames(transcripts_appris), "_tx")

mir181_bs_appris <- as.data.frame(mir181_bs)

mir181_bs_appris <- cbind(mir181_bs_appris[queryHits(idx),], transcripts_appris[subjectHits(idx),])

# get mir181 bs position relative to transcript
# (start of transcript is 1, strand is always +)

# rel_mir181_bs_appris_p <- mir181_bs_appris %>%
#   subset(strand == "+") %>%
#   rowwise(.) %>%
#   mutate(start = start - start_tx,
#          end = end - start_tx,
#          seqnames = transcript_id_tx) %>%
#   subset(start > 0)
#
# rel_mir181_bs_appris_m <- mir181_bs_appris %>%
#   subset(strand == "-") %>%
#   rowwise(.) %>%
#   mutate(start_genomic = start,
#          start = -(end - end_tx),
#          end = -(start_genomic - end_tx),
#          strand = "+",
#          seqnames = transcript_id_tx)  %>%
#   subset(end > 0)
#
# rel_mir181_bs_appris_m$start_genomic <- NULL
#
# rel_mir181_bs_appris <- rbind(rel_mir181_bs_appris_p, rel_mir181_bs_appris_m)
#
# rel_mir181_bs_appris <- rel_mir181_bs_appris %>%
#   subset(end <= width_tx)
```

```
############################
# BS sequence considering mature transcripts
############################
# prepare a txdb of expressed transcripts
anno_transcripts_exons <- anno[anno$type != "gene"]
anno_transcripts_exons$transcript_id <- sub("\\..*", "", anno_transcripts_exons$transcript_id)
anno_transcripts_GR_list <- anno_transcripts_exons %>%
```

```
  splitAsList(., f = .$transcript_id) %>%
  GRangesList(.)

txdb <- makeTxDbFromGRanges(unlist(anno_transcripts_GR_list))

# prepare a transcript mapper (contains transcript ids and names together with genomic positions of tra
transcripts_txdb_mapper <- transcripts(txdb)

# get transcript-relative coordinates of BS
mir181_bs_appris <- makeGRangesFromDataFrame(mir181_bs_appris, keep.extra.columns = T)
mir181_bs_appris_tx <- mapToTranscripts(mir181_bs_appris, txdb, extractor.fun = GenomicFeatures::exonsB

# readd metadata
elementMetadata(mir181_bs_appris_tx) <- c(elementMetadata(mir181_bs_appris_tx), elementMetadata(mir181_

# change the seqnames to the transcript names
names(mir181_bs_appris_tx) <- 1: NROW(mir181_bs_appris_tx)
mir181_bs_appris_tx <- as.data.frame(mir181_bs_appris_tx)
mir181_bs_appris_tx$seqnames <- transcripts_txdb_mapper$tx_name[mir181_bs_appris_tx$transcriptsHits]

mir181_bs_appris_tx <- mir181_bs_appris_tx %>% subset(seqnames == transcript_id_tx)
```

- Number of mirBS: 10989
- Number of mirBS on appris transcrips: 6937
- Number of enriched mirBS: 4960
- Number of enriched mirBS on appris transcripts: 3179

```
saveRDS(mir181_bs_appris_tx, paste0(out,"mir181_bs_on_transcripts.rds"))
saveDb(txdb, paste0(out,"transcript_annotation.db"))
```

```
## TxDb object:
## # Db type: TxDb
## # Supporting package: GenomicFeatures
## # Genome: NA
## # Nb of transcripts: 142351
## # Db created by: GenomicFeatures package from Bioconductor
## # Creation time: 2023-07-06 10:15:47 +0200 (Thu, 06 Jul 2023)
## # GenomicFeatures version at creation time: 1.50.4
## # RSQLite version at creation time: 2.3.1
## # DBSCHEMAVERSION: 1.2
```