Assign mir181 binding sites to a specific transcript

Melina Klostermann

11 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?

- Each binding site is assigned to a specific transcript isoform
- The following criteria are used to decide on the isoform:

1) Previously assigned region (see binding site definitions), if more then one transcript is possible:
2) if possible protein coding transcripts, if still more then one transcript is possible:
3) best transcript support level, if still more then one transcript is possible:
4) longest transcript

- Then the mir181 binding sites are mapped to their respective transcript coordinates.
- 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
anno$gene_id <- sub("\\..*", "", anno$gene_id)
anno$transcript_id <- sub("\\..*", "", anno$transcript_id)

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

3 Assign each binding site to a specific transcript

```
# get appris transcripts (when there are multiple take the longest)
transcripts <- anno[anno$type=="transcript"] %>% as.data.frame(.)
#transcripts$appris <- grepl(transcripts$tag, pattern= "appris_principal_1")


# get regional annotations
three_utr <- anno[anno$type=="three_prime_UTR"] %>% as.data.frame(.) %>%
  dplyr::select(seqnames, start, end, width, strand, transcript_id, gene_id) %>%
  left_join(transcripts %>% dplyr::select(., width, transcript_id, transcript_type, transcript_support_
  makeGRangesFromDataFrame(., keep.extra.columns = T)

cds <- anno[anno$type=="CDS"] %>% as.data.frame(.) %>%
  dplyr::select(seqnames, start, end, width, strand, transcript_id, gene_id) %>%
  left_join(transcripts %>% dplyr::select(., width, transcript_id, transcript_type, transcript_support_
  makeGRangesFromDataFrame(., keep.extra.columns = T)

five_utr <- anno[anno$type=="five_prime_UTR"] %>% as.data.frame(.) %>%
  dplyr::select(seqnames, start, end, width, strand, transcript_id, gene_id) %>%
  left_join(transcripts %>% dplyr::select(., width, transcript_id, transcript_type, transcript_support_
  makeGRangesFromDataFrame(., keep.extra.columns = T)


# select best transcript per bs
# 3'utr
bs_utr3 <- mir181_bs %>% subset(., region == "utr3") %>%
  makeGRangesFromDataFrame(., keep.extra.columns = T)
NROW(bs_utr3)
```

```
## [1] 5607
```

```
i_utr3 <- findOverlaps(bs_utr3, three_utr, type = "any")
t <- as.data.frame(three_utr[subjectHits(i_utr3)]) %>% dplyr::select(width_tx, transcript_type, transcri
t$transcript_type <- factor(t$transcript_type, levels = c("protein_coding"))

bs_utr3 <-  bs_utr3[queryHits(i_utr3)] %>%
  as.tibble()

bs_utr3 <- cbind(bs_utr3, t) %>%
  group_by(mir181BS_ID) %>%
  arrange(transcript_type, transcript_support_level,  width_tx, .by_group = T) %>%
  dplyr::slice(1)
nrow(bs_utr3)
```

```
## [1] 5607
```

```
# cds
bs_cds <- mir181_bs %>% subset(., region == "cds") %>%
  makeGRangesFromDataFrame(., keep.extra.columns = T)
NROW(bs_cds)
```

```
## [1] 4284
```

```
i_cds <- findOverlaps(bs_cds, cds, type = "any")
t <- as.data.frame(cds[subjectHits(i_cds)]) %>% dplyr::select(width_tx, transcript_type, transcript_sup

bs_cds <-  bs_cds[queryHits(i_cds)] %>%
  as.tibble()

bs_cds <- cbind(bs_cds, t) %>%
  group_by(mir181BS_ID) %>%
  arrange(transcript_type, transcript_support_level,  width_tx, .by_group = T) %>%
  dplyr::slice(1)
nrow(bs_cds)
```

```
## [1] 4284
```

```
# 5'utr
bs_utr5 <- mir181_bs %>% subset(., region == "utr5") %>%
  makeGRangesFromDataFrame(., keep.extra.columns = T)
NROW(bs_utr5)
```

```
## [1] 582
```

```
i_utr5 <- findOverlaps( bs_utr5, five_utr, type = "any")
t <- as.data.frame(five_utr[subjectHits(i_utr5)]) %>% dplyr::select(width_tx, transcript_type, transcri

bs_utr5 <-  bs_utr5[queryHits(i_utr5)] %>%
  as.tibble()

bs_utr5 <- cbind(bs_utr5, t) %>%
  group_by(mir181BS_ID) %>%
  arrange(transcript_type, transcript_support_level,  width_tx, .by_group = T) %>%
  dplyr::slice(1)
nrow(bs_utr5)
```

```
## [1] 582
```

```
mir181_bs <- rbind(bs_utr3, bs_cds, bs_utr5)
```

4 Get transcript coordinates

```
############################
# 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 <- makeGRangesFromDataFrame(mir181_bs, keep.extra.columns = T)
mir181_bs_tx <- mapToTranscripts(mir181_bs, txdb, extractor.fun = GenomicFeatures::exonsBy, ignore.stra
# Mapped position is computed by counting from the transcription start site (TSS) and is not affected b

# readd metadata
elementMetadata(mir181_bs_tx) <- c(elementMetadata(mir181_bs_tx), elementMetadata(mir181_bs[mir181_bs_t

# change the seqnames to the transcript names
names(mir181_bs_tx) <- 1: NROW(mir181_bs_tx)
mir181_bs_tx <- as.data.frame(mir181_bs_tx)
mir181_bs_tx$seqnames <- transcripts_txdb_mapper$tx_name[mir181_bs_tx$transcriptsHits]


mir181_bs_tx <- mir181_bs_tx %>% subset(seqnames == transcript_id)
```

- Number of mirBS: 10473
- Number of mirBS on transcripts (without intron): 10244
- Number of enriched mirBS: 4658
- Number of enriched mirBS on transcripts: 4519

Comment: we use findOverlaps with the center position of the binding site to assign the regions in the binding site definition scripts. The mapToTranscripts will only keep binding sites, that are completely inside the transcript. This is why we loose a few binding sites, when mapping to transcripts here.

```
saveRDS(mir181_bs_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: 2023-07-11 10:58:01 +0200 (Tue, 11 Jul 2023)
## # GenomicFeatures version at creation time: 1.50.4
## # RSQLite version at creation time: 2.3.1
## # DBSCHEMAVERSION: 1.2
```