MMSAT4 & MurSatRep1

Melina Klostermann

22 September, 2023

Contents

1 Libraries and settings

```
# ----------------------------------------
# libraries
# ----------------------------------------
library(tidyverse)
library(GenomicRanges)
library(colorspace)
library(gghalves)
library(BSgenome.Mmusculus.UCSC.mm10)
library(Biostrings)
library(plyranges)
library(AnnotationHub)
library(GenomicFeatures)

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

here <- here::here()

source(paste0(here,"/Supporting_scripts/themes/theme_paper.R"))
source(paste0(here,"/Supporting_scripts/themes/CustomThemes.R"))


out <- paste0(here,"/Figure4/04_MMSAT4/")
```

2 What was done?

The repeats from Repeat masker are mapped onto mature transcripts.

3 Files

```
# ------------------------------------
# transcript sequences
# ------------------------------------
transcript_fasta <- readDNAStringSet("/Users/melinaklostermann/Documents/projects/anno/gencodevM23/genco

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

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

# add N in beginning in end to not run out of transcripts when search motif
n200 <- c(rep("N",200)) %>%
  paste(., collapse = "") %>%
  RNAStringSet()

transcript_fasta <- xscat(n200, transcript_fasta, n200)
names(transcript_fasta) <- names_transcript_fasta

transcript_fasta_df <- data.frame(tx_name = names(transcript_fasta), width = width(transcript_fasta))

# ----------------------------------------
# MREs
# ----------------------------------------

mir181_bs <- readRDS(paste0(here, "/Figure4/03_assign_transcripts/mir181_bs_on_transcripts.rds"))

mir181_enriched_set <- mir181_bs %>%
  subset(set %in% c("ago_bs_mir181_chi&mir181_enriched", "mir181_enriched")) %>%
  makeGRangesFromDataFrame(., keep.extra.columns = T)


# ----------------------------------------
# annotation
# ----------------------------------------

anno <- readRDS(paste0(here,"/Supporting_scripts/annotation_preprocessing/annotation.rds"))
anno$gene_id <- sub("\\..*", "", anno$gene_id)
anno$transcript_id <- sub("\\..*", "", anno$transcript_id)
```

4 Overlap of mir181 binding sites to Satelite sequences

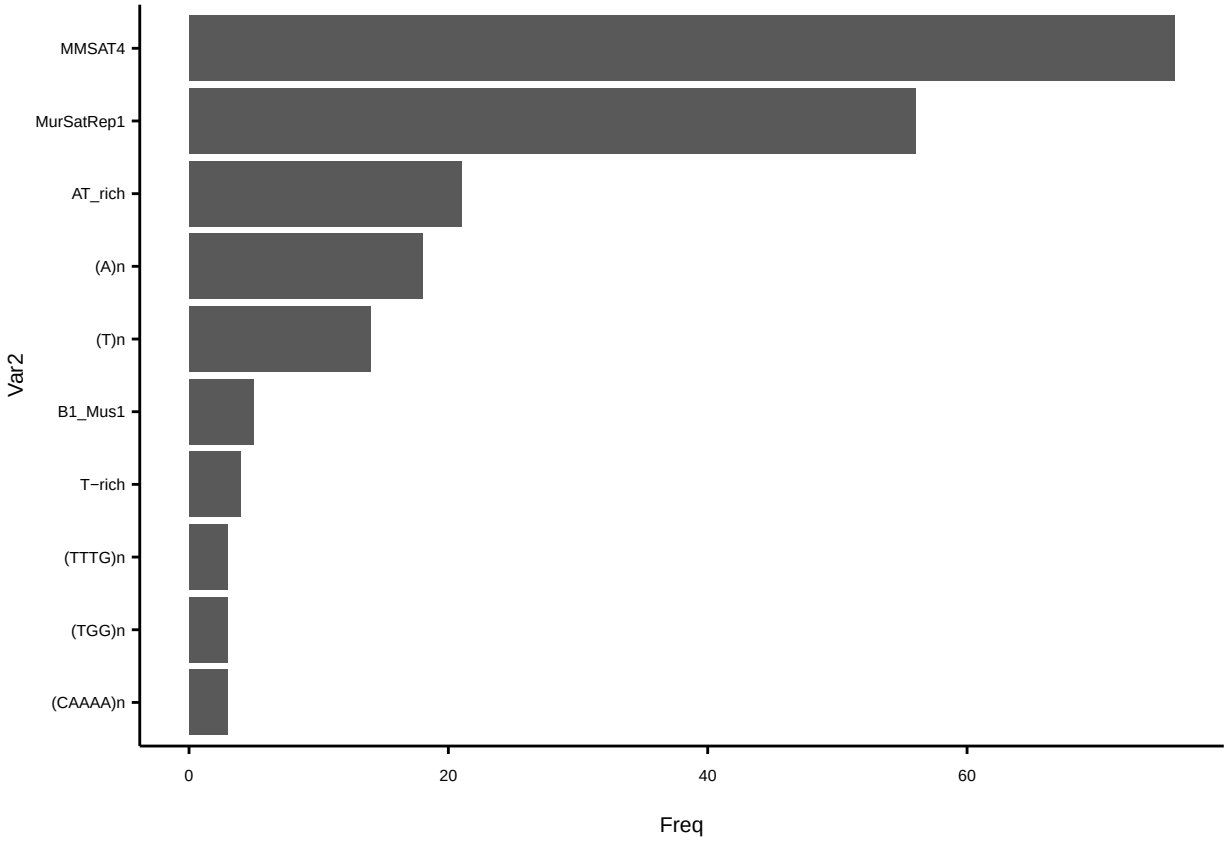
```
# ------------------------
# Get repeat Masker annotation for mouse
# ------------------------
```

```
ah = AnnotationHub()
repeat_masker <- ah[["AH99012"]]

# ------------------------
# map repeats to 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
repeat_masker_tx <- mapToTranscripts(repeat_masker, txdb, extractor.fun = GenomicFeatures::exonsBy, ign
# Mapped position is computed by counting from the transcription start site (TSS) and is not affected b

# readd metadata
elementMetadata(repeat_masker_tx) <- c(elementMetadata(repeat_masker_tx), elementMetadata(repeat_masker

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


repeat_masker_tx <- makeGRangesFromDataFrame(repeat_masker_tx, keep.extra.columns = T)
repeat_masker_tx$rep_id <-  paste0(repeat_masker_tx$repName, "-", 1:NROW(repeat_masker_tx))

gene_id_mapper <- anno[anno$type == "transcript"] %>%
  as.data.frame(.) %>%
  dplyr::select(gene_id, transcript_id)

repeat_masker_tx <- as.data.frame(repeat_masker_tx) %>%
  left_join(., gene_id_mapper, by = c(seqnames = "transcript_id")) %>%
  makeGRangesFromDataFrame(., keep.extra.columns = T)


saveRDS(repeat_masker_tx, paste0(out, "rep_on_transcripts.rds"))
```

5 Number of MREs overlapping with repeat elements

```
bound_repeats <- subsetByOverlaps(repeat_masker_tx, mir181_enriched_set)

df <- table(bound_repeats$repName) %>% t() %>%
  as.data.frame() %>%
```

```
  arrange(desc(Freq)) %>%
  .[1:10,]

df <- df %>%
  arrange(Freq)

df$Var2 <- factor(df$Var2, levels = df$Var2)


ggplot(df,  aes(x = Var2, y = Freq  ))+
  geom_col()+
  coord_flip()+
  theme_paper()
```




```
ggsave(paste0(out, "Figure4G_MREs_overlapping_with_repeat_elements.pdf"), width = 4, height = 6, units =
```

6 Session Info

```
sessionInfo()
```

```
## R version 4.2.2 (2022-10-31)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Big Sur ... 10.16
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRblas.0.dylib
```

```
## LAPACK: /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats4    stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
##  [1] GenomicFeatures_1.50.4            AnnotationDbi_1.60.2
##  [3] Biobase_2.58.0                    AnnotationHub_3.6.0
##  [5] BiocFileCache_2.6.1               dbplyr_2.3.3
##  [7] plyranges_1.18.0                  BSgenome.Mmusculus.UCSC.mm10_1.4.3
##  [9] BSgenome_1.66.3                   rtracklayer_1.58.0
## [11] Biostrings_2.66.0                 XVector_0.38.0
## [13] gghalves_0.1.4                    colorspace_2.1-0
## [15] GenomicRanges_1.50.2              GenomeInfoDb_1.34.9
## [17] IRanges_2.32.0                    S4Vectors_0.36.2
## [19] BiocGenerics_0.44.0               lubridate_1.9.2
## [21] forcats_1.0.0                     stringr_1.5.0
## [23] dplyr_1.1.2                       purrr_1.0.1
## [25] readr_2.1.4                       tidyr_1.3.0
## [27] tibble_3.2.1                      ggplot2_3.4.2
## [29] tidyverse_2.0.0                   knitr_1.43
##
## loaded via a namespace (and not attached):
##  [1] ggsignif_0.6.4               rjson_0.2.21
##  [3] ellipsis_0.3.2               rprojroot_2.0.3
##  [5] rstudioapi_0.15.0            farver_2.1.1
##  [7] ggpubr_0.6.0                 bit64_4.0.5
##  [9] interactiveDisplayBase_1.36.0 fansi_1.0.4
## [11] xml2_1.3.5                   codetools_0.2-19
## [13] cachem_1.0.8                 Rsamtools_2.14.0
## [15] broom_1.0.5                  png_0.1-8
## [17] shiny_1.7.4.1                BiocManager_1.30.21
## [19] compiler_4.2.2               httr_1.4.6
## [21] backports_1.4.1              Matrix_1.5-4.1
## [23] fastmap_1.1.1                cli_3.6.1
## [25] later_1.3.1                  htmltools_0.5.5
## [27] prettyunits_1.1.1            tools_4.2.2
## [29] gtable_0.3.3                 glue_1.6.2
## [31] GenomeInfoDbData_1.2.9       rappdirs_0.3.3
## [33] Rcpp_1.0.11                  carData_3.0-5
## [35] vctrs_0.6.3                  xfun_0.39
## [37] timechange_0.2.0             mime_0.12
## [39] lifecycle_1.0.3              restfulr_0.0.15
## [41] rstatix_0.7.2                XML_3.99-0.14
## [43] zlibbioc_1.44.0              scales_1.2.1
## [45] ragg_1.2.5                   hms_1.1.3
## [47] promises_1.2.0.1             MatrixGenerics_1.10.0
## [49] parallel_4.2.2               SummarizedExperiment_1.28.0
## [51] yaml_2.3.7                   curl_5.0.1
## [53] memoise_2.0.1                biomaRt_2.54.1
```

```
## [55] stringi_1.7.12              RSQLite_2.3.1
## [57] highr_0.10                  BiocVersion_3.16.0
## [59] BiocIO_1.8.0                filelock_1.0.2
## [61] BiocParallel_1.32.6         systemfonts_1.0.4
## [63] rlang_1.1.1                 pkgconfig_2.0.3
## [65] matrixStats_1.0.0           bitops_1.0-7
## [67] evaluate_0.21               lattice_0.21-8
## [69] labeling_0.4.2              GenomicAlignments_1.34.1
## [71] bit_4.0.5                   tidyselect_1.2.0
## [73] here_1.0.1                  magrittr_2.0.3
## [75] R6_2.5.1                    generics_0.1.3
## [77] DelayedArray_0.24.0         DBI_1.1.3
## [79] pillar_1.9.0                withr_2.5.0
## [81] KEGGREST_1.38.0             abind_1.4-5
## [83] RCurl_1.98-1.12             crayon_1.5.2
## [85] car_3.1-2                   utf8_1.2.3
## [87] tzdb_0.4.0                  rmarkdown_2.23
## [89] progress_1.2.2              grid_4.2.2
## [91] blob_1.2.4                  digest_0.6.33
## [93] xtable_1.8-4                httpuv_1.6.11
## [95] textshaping_0.3.6           munsell_0.5.0
```