m5C_sites_YBX1

Riccardo Mosca

2025-05-13

In this markdown the YBX1 T7 metafile is lifted over to GRCh19 genome build, to merge with the m5C dataset with the GRCh19 assembly (Chen et al., Nat. Cell. Bio., 2019)

```
library(GenomicRanges)
library(dplyr)
library(tidyr)
library(data.table)
library(GenomicFeatures)

YBX1 <-
read.table("/Users/riccardomosca/Desktop/RAPseq_PAPER/PEAKs/ANNOTATED/T7_Fig5
/Ybx1_T7_scored_annotated.txt",
    sep = "\t", header = T)
YBX1 <- YBX1 %>%
    dplyr::select(-chr, -start, -end, -strand)
YBX1 <- YBX1 %>%
    separate(peak_ID, into = c("chr", "start", "end", "strand"), sep = "_")
YBX1$start <- as.numeric(YBX1$start)
# The peak window is considered, not the summit
YBX1$start <- as.numeric(YBX1$start - 50)
YBX1$end <- as.numeric(YBX1$end)
YBX1$end <- YBX1$end + 50
YBX1$peak_ID <- paste(YBX1$chr, YBX1$start, YBX1$end, sep = "_")

YBX1_bed <- YBX1 %>%
    dplyr::select("chr", "start", "end")

# write.table(YBX1_bed, file = '/Users/riccardomosca/Desktop/Ybx1.bed',
# row.names = FALSE, col.names = T, sep = '\t', quote = FALSE)

# from the original positions chr17\tchr17\t37339830\t37339900 and chr17
# \t43241105\t43241176 chr19\t34399199\t34399348 chr19\t34399674\t34399814
# chr4\t184624377\t184624502 have been deleted because UCSC not able to lift.

original <-
read.table("/Users/riccardomosca/Desktop/RAPseq_PAPER/FIGUREs/FIGURE5/Ybx1.be
d",
    header = FALSE)
liftover <-
read.table("/Users/riccardomosca/Desktop/RAPseq_PAPER/FIGUREs/FIGURE5/Ybx1_19
.bed",
    header = FALSE)
```

```
# the two bed files are combined to keep the coordinates from both assemblies
combined <- cbind(original, liftover)
colnames(combined) <- c("chr", "start", "end", "chr19", "start19", "end19")
combined$peak_ID <- paste(combined$chr, combined$start, combined$end, sep =
"_")
combined <- combined %>%
    dplyr::select("chr19", "start19", "end19", "peak_ID")

YBX1$peak_ID <- gsub("_[+-]$", "", YBX1$peak_ID)

YBX1_19 <- merge(YBX1, combined, "peak_ID")
YBX1_19$start <- as.numeric(YBX1_19$start)
YBX1_19$start <- as.numeric(YBX1_19$start + 50)
YBX1_19$end <- as.numeric(YBX1_19$end)
YBX1_19$end <- YBX1_19$end - 50
YBX1_19$peak_ID <- paste(YBX1_19$chr, YBX1_19$start, YBX1_19$end,
YBX1_19$strand,
    sep = "_")

YBX1_19 <- YBX1_19 %>%
    dplyr::select(-chr, -start, -end)
write.table(YBX1_19,
"/Users/riccardomosca/Desktop/RAPseq_PAPER/FIGUREs/FIGURE5/Ybx1_metafile_19.t
xt",
    sep = "\t", row.names = FALSE, quote = FALSE)
```

Loading m5C dataset and annotating it

```
m5C <-
read.table("/Users/riccardomosca/Desktop/PhD/Literature/m5C_dataset/YBX1_NatC
ellBio_bladder/m5C_T24_NatCellBio.txt",
    sep = "\t", header = T)
m5C <- m5C %>%
    dplyr::select(1, 2, 3, 6)
colnames(m5C) <- c("chr19", "pos19", "strand", "m5clevel")
m5C <- na.omit(m5C)

txdb <- makeTxDbFromGFF(file =
"/Users/riccardomosca/Desktop/RAPseq_PAPER/FIGUREs/FIGURE5/NEW_FIGUREs/YBX1/g
encode.v19.annotation.gtf",
    format = "gtf")
Gencode_v33_IDs <- read.table(file =
"/Users/riccardomosca/Desktop/RAPseq_PAPER/ANNOTATIONs/hg19_gencode_annotatio
n.txt")
colnames(Gencode_v33_IDs) <- c("gene_ID", "transcript_ID", "gene_strand",
"gene_name",
    "gene_type")
Gencode_v33_IDs <- Gencode_v33_IDs[!duplicated(Gencode_v33_IDs), ]
Gencode_v33_IDs <- Gencode_v33_IDs[!is.na(Gencode_v33_IDs$transcript_ID), ]
```

```r
Intron_GR <- intronsByTranscript(txdb, use.names = TRUE)
Exon_GR <- exonsBy(txdb, by = "tx", use.names = TRUE)
ThreeUTR_GR <- threeUTRsByTranscript(txdb, use.names = TRUE)
FiveUTR_GR <- fiveUTRsByTranscript(txdb, use.names = TRUE)
CDS_GR <- cdsBy(txdb, by = "tx", use.names = TRUE)
pass_1 <- subsetByOverlaps(Exon_GR, CDS_GR, invert = T)
pass_2 <- subsetByOverlaps(pass_1, ThreeUTR_GR, invert = T)
Exon_GR <- subsetByOverlaps(pass_2, FiveUTR_GR, invert = T)
rm(pass_1)
rm(pass_2)


Introns <- as.data.frame(Intron_GR)[, c(3, 4, 5, 2, 7)]
Introns$feature <- rep("intron", nrow(Introns))
Exons <- as.data.frame(Exon_GR)[, c(3, 4, 5, 2, 7)]
Exons$feature <- rep("exon", nrow(Exons))
CDSs <- as.data.frame(CDS_GR)[, c(3, 4, 5, 2, 7)]
CDSs$feature <- rep("CDS", nrow(CDSs))
FiveUTRs <- as.data.frame(FiveUTR_GR)[, c(3, 4, 5, 2, 7)]
FiveUTRs$feature <- rep("5UTR", nrow(FiveUTRs))
ThreeUTRs <- as.data.frame(ThreeUTR_GR)[, c(3, 4, 5, 2, 7)]
ThreeUTRs$feature <- rep("3UTR", nrow(ThreeUTRs))
Features <- rbind(Introns, Exons, CDSs, FiveUTRs, ThreeUTRs)
colnames(Features) <- c("chr", "start", "end", "transcript_ID",
"feature_strand",
    "feature")

# rm(Introns, Intron_GR, Exons, Exon_GR, FiveUTRs, FiveUTR_GR, ThreeUTRs,
# ThreeUTR_GR, CDSs, CDS_GR)

Features <- merge(Features, Gencode_v33_IDs, by = "transcript_ID")
Features$gene_ID <- as.character(Features$gene_ID)
Features <- Features[, c(2, 3, 4, 7, 6, 5, 9, 10)]
colnames(Features) <- c("chr", "start", "end", "gene_ID", "feature",
"strand", "gene_name",
    "gene_type")
Features$IDs <- paste(Features[, 1], Features[, 2], Features[, 3], Features[,
4],
    Features[, 5], Features[, 6], Features[, 7], sep = "_")
Features <- Features[duplicated(Features$IDs) == "FALSE", ]
Features <- Features[, 1:8]
Features$chr <- as.character(Features$chr)
Features$strand <- as.character(Features$strand)
Features_GR <- makeGRangesFromDataFrame(Features[, 1:6])
colnames(Features) <- c("chr", "start", "end", "gene_ID", "feature",
"gene_strand",
    "gene_name", "gene_type")
Features$gene_type <- as.character(Features$gene_type)
```

```r
GR_m5C <- makeGRangesFromDataFrame(m5C[, c("chr19", "pos19", "strand",
"m5clevel")],
    seqnames.field = "chr19", start.field = "pos19", end.field = "pos19",
strand.field = "strand",
    keep.extra.columns = TRUE  # Keep extra columns (like RBPs, summit_ID)
)

m5C <- m5C[as.data.frame(findOverlaps(GR_m5C, Features_GR, type =
"within"))[, 1],
    ]
Annots <- Features[as.data.frame(findOverlaps(GR_m5C, Features_GR, type =
"within"))[,
    2], ][, 4:8]
m5C_annotated <- cbind(m5C, Annots)

# write.table(m5C_annotated,
#
'/Users/riccardomosca/Desktop/RAPseq_PAPER/FIGUREs/FIGURE5/m5C_annotated.txt'
,
# sep = '\t', row.names = FALSE, quote = FALSE)
```