Annotation file preprocessing for binding sites

Melina Klostermann

31 March, 2023

Contents

1 Libraries and Settings

```
# ----------------------------------------
# libraries
# ----------------------------------------

library(GenomicRanges)
library(GenomicFeatures)
library(rtracklayer)
library(knitr)
library(dplyr)
library(kableExtra)

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

out <- "/Users/melinaklostermann/Documents/projects/AgoCLIP_miR181/R_github/miR181_paper/Methods/01_Anno
```

2 What was done?

- Annotations are downloaded as gff3 from GENCODE (Release m23 GRCm38.p6). This release is relatively old (05.2019), but it is consitent with current miRbase Release 22, which uses the GRCm38 (mm10) genome assembly.
- Annotations are kept as is with no further filtering.

- A txDb file of the annotation is made.
- Also the gene annotations are extracted as GRanges object.

- Both files are saved and later reloaded in the binding site definition scripts.

seqnames	start	end	width	strand	source	type	score	phase	ID
chr1	3073253	3074322	1070	+	HAVANA	gene	NA	NA	ENSMUSG00000102693.1
chr1	3073253	3074322	1070	+	HAVANA	transcript	NA	NA	ENSMUST00000193812.1
chr1	3073253	3074322	1070	+	HAVANA	exon	NA	NA	exon:ENSMUST0000019381
chr1	3102016	3102125	110	+	ENSEMBL	gene	NA	NA	ENSMUSG00000064842.1
chr1	3102016	3102125	110	+	ENSEMBL	transcript	NA	NA	ENSMUST00000082908.1
chr1	3102016	3102125	110	+	ENSEMBL	exon	NA	NA	exon:ENSMUST0000008290

3 Annotation txdb

```
# ----------------------------------------
# Process whole annotation
# ----------------------------------------

# load annotation
anno <- readGFFAsGRanges("/Users/melinaklostermann/Documents/projects/anno/gencodevM23/gencode.vM23.anno

# shorten gene ids
anno$geneID <- sub("\\..*", "", anno$gene_id)

# make txdb and save
annoDb = makeTxDbFromGRanges(anno)
saveDb(annoDb, paste0(out, "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: 2023-03-31 09:41:10 +0200 (Fri, 31 Mar 2023)
## # GenomicFeatures version at creation time: 1.50.4
## # RSQLite version at creation time: 2.2.20
## # DBSCHEMAVERSION: 1.2
```

```
kable(head(anno)) %>%
  kable_material(c("striped", "hover")) %>%
  scroll_box(width = "100%", height = "500px")
```

4 Get genes

```
# ----------------------------------------
# Retrieve genes
# ----------------------------------------

# get genes
gns = genes(annoDb)

# get back meta data
idx = match(gns$gene_id, anno$gene_id)
mcols(gns) = cbind(mcols(gns), mcols(anno)[idx,])
```

seqnames	start	end	width	strand	gene_id	gene_name	gene_type	ge
chr3	108107280	108146146	38867	-	ENSMUSG00000000001.4	Gnai3	protein_coding	EN
chrX	77837901	77853623	15723	-	ENSMUSG00000000003.15	Pbsn	protein_coding	EN
chr16	18780447	18811987	31541	-	ENSMUSG00000000028.15	Cdc45	protein_coding	EN
chr7	142575529	142578143	2615	-	ENSMUSG00000000031.16	H19	lncRNA	EN
chrX	161082525	161258213	175689	+	ENSMUSG00000000037.17	Scml2	protein_coding	EN
chr11	108343354	108414396	71043	+	ENSMUSG00000000049.11	Apoh	protein_coding	EN

```
# clean gene names and meta data
names(gns) = sub("\\..*", "", names(gns))
meta = data.frame(gene_id = gns$gene_id, gene_name = gns$gene_name, gene_type = gns$gene_type)
mcols(gns) = meta
gns$geneID = names(gns)

# save genes
saveRDS(gns, paste0(out, "gene_annotation.rds"))


kable(head(gns))  %>%
  kable_material(c("striped", "hover")) %>%
  scroll_box(width = "100%", height = "500px")
```