ABCB1_proximal_seq_tfs

charvi khanna

2025-11-03

```r
# Load necessary libraries
library(TFBSTools)
library(Biostrings)
```

```
## Loading required package: BiocGenerics
```

```
## Loading required package: generics
```

```
##
## Attaching package: 'generics'
```

```
## The following objects are masked from 'package:base':
##
##     as.difftime, as.factor, as.ordered, intersect, is.element, setdiff,
##     setequal, union
```

```
##
## Attaching package: 'BiocGenerics'
```

```
## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs
```

```
## The following objects are masked from 'package:base':
##
##     anyDuplicated, aperm, append, as.data.frame, basename, cbind,
##     colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,
##     get, grep, grepl, is.unsorted, lapply, Map, mapply, match, mget,
##     order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,
##     rbind, Reduce, rownames, sapply, saveRDS, table, tapply, unique,
##     unsplit, which.max, which.min
```

```
## Loading required package: S4Vectors
```

```
## Loading required package: stats4
```

```
##
## Attaching package: 'S4Vectors'
```

```
## The following object is masked from 'package:utils':
##
##     findMatches

## The following objects are masked from 'package:base':
##
##     expand.grid, I, unname

## Loading required package: IRanges

##
## Attaching package: 'IRanges'

## The following object is masked from 'package:grDevices':
##
##     windows

## Loading required package: XVector

## Loading required package: GenomeInfoDb

##
## Attaching package: 'Biostrings'

## The following object is masked from 'package:base':
##
##     strsplit
```

```r
library(JASPAR2022)
```

```
## Loading required package: BiocFileCache

## Loading required package: dbplyr
```

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:dbplyr':
##
##     ident, sql

## The following objects are masked from 'package:Biostrings':
##
##     collapse, intersect, setdiff, setequal, union

## The following object is masked from 'package:GenomeInfoDb':
##
##     intersect
```

```
## The following object is masked from 'package:XVector':
##
##     slice


## The following objects are masked from 'package:IRanges':
##
##     collapse, desc, intersect, setdiff, slice, union


## The following objects are masked from 'package:S4Vectors':
##
##     first, intersect, rename, setdiff, setequal, union


## The following objects are masked from 'package:BiocGenerics':
##
##     combine, intersect, setdiff, setequal, union


## The following object is masked from 'package:generics':
##
##     explain


## The following objects are masked from 'package:stats':
##
##     filter, lag


## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(writexl)
library(knitr)
library(kableExtra)
```

```
## Warning: package 'kableExtra' was built under R version 4.5.2


##
## Attaching package: 'kableExtra'


## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```r
# Global chunk options
knitr::opts_chunk$set(echo = TRUE, message = FALSE, warning = FALSE)

ABCB1_proximal <- DNAString("TCACAGGTAAAGAGTAATGATAGCCTTTAAAATGGTAATACAAGTGTTTATCCCAGTACCAGAGGAGGAGCTACA

# Get human motifs

opts <- list(species = 9606)
motifs <- getMatrixSet(JASPAR2022, opts)
```

```r
# Convert motifs to PWMs using log2 probability ratios

pwm <- toPWM(motifs, type = "log2probratio")

# Scan sequence for all motifs

hits <- searchSeq(
pwm,
ABCB1_proximal,
seqname = "ABCB1_proximal",
min.score = "95%",   # high threshold
strand = "*"         # both strands
)

# Keep only motifs with at least one hit

hits_filtered <- hits[sapply(hits, length) > 0]

# Convert to data.frame and add TF names

hits_df <- do.call(rbind, lapply(hits_filtered, as.data.frame))
hits_df$TF <- rep(names(hits_filtered), sapply(hits_filtered, length))

# Keep only desired columns

hits_df_small <- hits_df %>%
select(class, start, end, siteSeqs, absScore, relScore)

# Display top hits

kable(head(hits_df_small, 10), booktabs = TRUE, longtable = TRUE) %>%
kable_styling(latex_options = c("striped", "scale_down"))
```

```
## Warning in styling_latex_scale(out, table_info, "down"): Longtable cannot be
## resized.
```

	class	start	end	siteSeqs	absScore	relScore
MA0130.1.1	C2H2 zinc finger factors	486	491	ATCCAC	9.890853	1.0000000
MA0130.1.2	C2H2 zinc finger factors	135	140	CTCCAC	9.675124	0.9940480
MA0144.2	STAT domain factors	413	423	TTACTGGGAAG	10.125801	0.9534647
MA0599.1.1	C2H2 zinc finger factors	132	141	CCTCCACCCA	10.182292	0.9625844
MA0599.1.2	C2H2 zinc finger factors	143	152	CTCCCACCCT	10.356340	0.9637991
MA0599.1.3	C2H2 zinc finger factors	147	156	TCTACTCCCA	8.496201	0.9508171
MA0610.1	DM-type intertwined zinc finger factors	556	566	AATGTAACAAG	13.616380	0.9700553
MA0613.1	Fork head/winged helix factors	44	51	ATAAACAC	11.969470	0.9692009
MA0618.1.1	Homeo domain factors	241	248	TTAATGAG	10.576243	0.9666860
MA0618.1.2	Homeo domain factors	502	509	ATAATTAG	12.068281	0.9855810

```r
# Filter hits with absolute score greater than 10

hits_df_filtered_small <- hits_df_small %>% filter(absScore > 10)
```

```r
# Show filtered hits

kable(head(hits_df_filtered_small, 10), booktabs = TRUE, longtable = TRUE) %>%
kable_styling(latex_options = c("striped", "scale_down"))
```

```
## Warning in styling_latex_scale(out, table_info, "down"): Longtable cannot be
## resized.
```

	class	start	end	siteSeqs	absScore	relScore
MA0144.2	STAT domain factors	413	423	TTACTGGGAAG	10.12580	0.9534647
MA0599.1.1	C2H2 zinc finger factors	132	141	CCTCCACCCA	10.18229	0.9625844
MA0599.1.2	C2H2 zinc finger factors	143	152	CTCCCACCCT	10.35634	0.9637991
MA0610.1	DM-type intertwined zinc finger factors	556	566	AATGTAACAAG	13.61638	0.9700553
MA0613.1	Fork head/winged helix factors	44	51	ATAAACAC	11.96947	0.9692009
MA0618.1.1	Homeo domain factors	241	248	TTAATGAG	10.57624	0.9666860
MA0618.1.2	Homeo domain factors	502	509	ATAATTAG	12.06828	0.9855810
MA0628.1.1	Homeo domain factors	501	510	AATAATTAGT	13.17289	0.9737588
MA0628.1.2	Homeo domain factors	501	510	ACTAATTATT	13.17289	0.9737588
MA0634.1	Homeo domain factors	303	312	TTCAATTATA	10.59824	0.9542863

```r
write_xlsx(hits_df_filtered_small, path = "ABCB1_proximal_hits_filtered.xlsx")
```

```r
# Number of motifs scanned

length(pwm)
```

```
## [1] 692
```

```r
# Number of motifs with hits

length(hits_filtered)
```

```
## [1] 98
```

```r
# Number of filtered hits

nrow(hits_df_filtered_small)
```

```
## [1] 118
```