MotifsLymphoblastoid

Dr. Nishat Mohammad

2025-01-10

Load Libraries

```r
library(compGenomRData)
library(GenomeInfoDb)
library(GenomicRanges)
library(GenomicAlignments)

library(pheatmap)
library(rtracklayer)
library(Gviz)
library(ggplot2)

library(BSgenome.Hsapiens.UCSC.hg38)
library(AnnotationHub)
library(tidyr)
library(dplyr)

library(GenomicFeatures)
library(normr)
library(txdbmaker)
library(MotifDb)
library(TFBSTools)

library(rGADEM)
library(BSgenome.Hsapiens.UCSC.hg19)
library(TFBSTools)
library(JASPAR2018)
```

1 EDA

```r
# Get data
data_path = system.file('extdata/chip-seq',package='compGenomRData')

# read  CTCF peaks created in peak calling part of tutorial
ctcf_peaks = read.table(file.path(data_path, 'CTCF_peaks.txt'), header=TRUE)

# convert peaks into a GRanges object
ctcf_peaks = makeGRangesFromDataFrame(ctcf_peaks, keep.extra.columns = TRUE)
```

```r
# order peaks by qvalue, and take top 250 peaks
ctcf_peaks = ctcf_peaks[order(ctcf_peaks$qvalue)]
ctcf_peaks = head(ctcf_peaks, n = 500)

# merge nearby CTCF peaks
ctcf_peaks = reduce(ctcf_peaks)

# expand CTCF peaks
ctcf_peaks_resized = resize(ctcf_peaks, width = 50, fix='center')
head(ctcf_peaks_resized)
```

```
## GRanges object with 6 ranges and 0 metadata columns:
##       seqnames            ranges strand
##          <Rle>         <IRanges>  <Rle>
##   [1]    chr21 14119351-14119400      *
##   [2]    chr21 14120226-14120275      *
##   [3]    chr21 14253851-14253900      *
##   [4]    chr21 14259851-14259900      *
##   [5]    chr21 14306726-14306775      *
##   [6]    chr21 14403851-14403900      *
##   -------
##   seqinfo: 1 sequence from an unspecified genome; no seqlengths
```

This code processes the CTCF peak data by filtering, merging, and resizing peaks, ensuring they are ready for subsequent analyses like motif discovery.

Motif Visualization

```r
# Get sequences
genome <- BSgenome.Hsapiens.UCSC.hg38
ctcf_sequences <- getSeq(genome, ctcf_peaks_resized)
ctcf_sequences_set <- DNAStringSet(ctcf_sequences)
class(ctcf_sequences_set)
```

```
## [1] "DNAStringSet"
## attr(,"package")
## [1] "Biostrings"
```

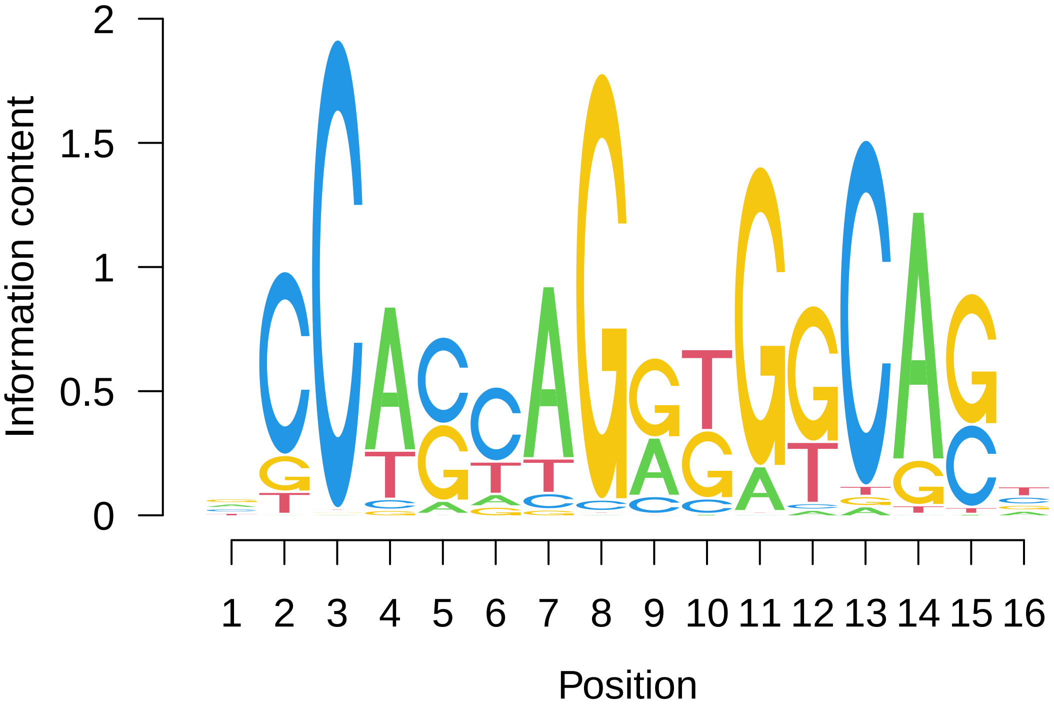
```r
length(ctcf_sequences_set)
```

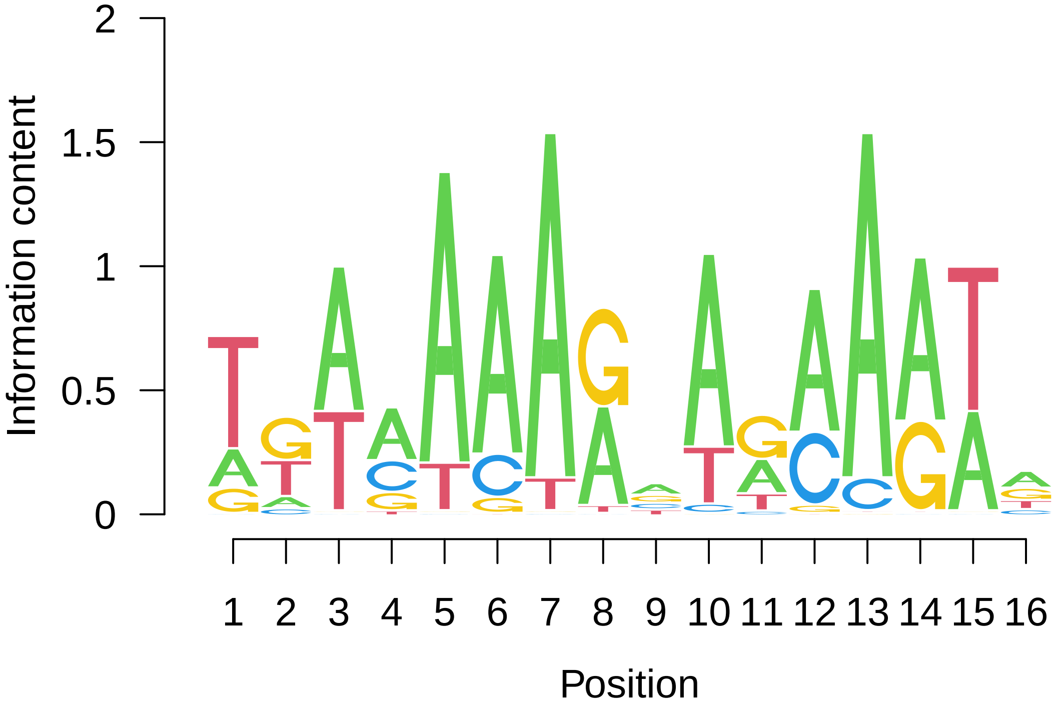
```
## [1] 334
```

```r
# Perform motif discovery
set.seed(1234)
novel_motifs <- GADEM(Sequences = ctcf_sequences_set,
                      genome = NULL,
                      seed = 1234,
                      nmotifs = 5)
```

```
## top 3  4, 5-mers: 12 40 52
## top 3  4, 5-mers: 12 36 34
## top 3  4, 5-mers: 12 36 28
```

```
# Visulaize motifs
plot(novel_motifs)
```





This code extracts sequences from the human genome corresponding to the CTCF peaks, performs motif discovery to identify the top 5 motifs, and visualizes the most enriched motifs.

- The R code chunk is made reproducible with `set.seed(1234)` for consistency.
- Visualization (`plot(novel_motifs)`) shows the discovered motifs for further interpretation.

3 Compare Motifs

```
# extract motif of interest from GADEM object
unknown_motif = getPWM(novel_motifs)[[1]]

# convert motif to a PWM matrix
unknown_pwm= PWMatrix(ID = 'unknown',
    profileMatrix = unknown_motif)

pwm_library = getMatrixSet(JASPAR2018,opts=list(collection = 'CORE',
                                                species    = 'Homo sapiens',
                                                matrixtype = 'PWM'))


# find most similar motif to our motif
pwm_sim = PWMSimilarity(pwm_library, unknown_pwm, method = 'Pearson')
```

```
# extract motif names from pwm library
pwm_library_list = lapply(pwm_library, function(x){
  data.frame(ID = ID(x), name = name(x))})

# combine list into one data frame
pwm_library_dt = dplyr::bind_rows(pwm_library_list)

# fetch similarity of each motif to our unknown motif
pwm_library_dt$similarity = pwm_sim[pwm_library_dt$ID]

# find most similar motif in the library
pwm_library_dt = pwm_library_dt[order(-pwm_library_dt$similarity),]
head(pwm_library_dt)
```

```
##           ID  name similarity
## 24  MA0139.1  CTCF  0.6715159
## 373 MA1102.1 CTCFL  0.6587983
## 281 MA0807.1  TBX5  0.4756392
## 370 MA1100.1 ASCL1  0.4662054
## 277 MA0803.1 TBX15  0.4541783
## 447 MA0693.2   VDR  0.4329790
```

This code compares the discovered motif with known motifs from the JASPAR database by calculating the Pearson similarity between them. The most similar motifs are then ordered and displayed.

- The unknown motif is extracted from the GADEM object and compared to known motifs from the JASPAR database.
- The `PWMSimilarity` function calculates the Pearson similarity between the unknown motif and the JASPAR motifs.
- The most similar motifs are ordered by similarity and displayed as the top results.

4 Conclusion

The motif analysis workflow involved several steps in processing and analyzing CTCF peak data.
- CTCF peaks were filtered, merged, and resized for motif discovery, ensuring high-quality input for subsequent analysis.
- Motif discovery was performed on the resulting sequences, identifying the top 5 enriched motifs.

- The motifs were then compared to known motifs from the JASPAR database using Pearson similarity, highlighting the closest matches. This shows the potential regulatory roles of CTCF motifs and their biological significance in gene regulation.