Working with motifs

Martin Morgan

2012-07-06 Fri

Contents

1 Background

ENCODE project

- Regulatory elements across many cell lines
- Our focus: ChIP-Seq for CTCF

CTCF

- Zinc finger transcription factor; sequence-specific DNA binding protein
- Insulator, blocking enhancer activity
- Original analysis: Illumina ChIP-seq; matching 'input' lanes of 1 or replicates from many cell lines
- GEO accession includes BAM and derived files

Questions: Are ChIP peaks enriched for CTCF motifs? Where are CTCF binding motifs located relative to transcription start sites?

2 Data

Called peaks (output of HotSpots peak caller) retrieved from ENCODE project web sites; common peaks across cell lines identified using ad hoc overlap criteria.

```
library(GenomicRanges)
stamFile <-
    system.file(package="SequenceAnalysisData", "data", "stam.Rda")
load(stamFile)
stam
ridx <- rowSums(assays(stam)[["Tags"]] > 0) == ncol(stam)
all <- stam[ridx,]
plot(density(width(rowData(all))),
     main="Peak widths; peaks shared by all samples")
```

3 Position weight matrices

Currated catalogs, e.g., JASPAR Forthcoming Bioc package: MotifDb Biostrings::PWM – estimate from DNAStringSet

```
load("2012-07-06-motifs-pwm.rda")
pwm[1:4, 1:5]
```

4 Questions

4.1 1. Do peaks contain CTCF motifs?

```
library(BSgenome.Hsapiens.UCSC.hg19)
peakSeqs <- getSeq(Hsapiens, rowData(all), as.character=FALSE)

## fudge: represent as a single sequence; should pad with N
allSeqs <- unlist(peakSeqs)

## score sequence for peaks
score <- function(pwm, seq) {
    hits <- matchPWM(pwm, seq)
    ## value <- PWMscoreStartingAt(pwm, subject(hits), start(hits))
    start(hits)
}
allScores <-
    c(score(pwm, allSeqs),                    # plus
      score(pwm, reverseComplement(allSeqs))) # minus

## break by peak widths
breaks <- c(0, cumsum(width(peakSeqs)))
scoredSeq <- cut(allScores, breaks, labels=FALSE)
motifsPerSequence <- tabulate(scoredSeq, length(peakSeqs))

plot(table(motifsPerSequence))
```

4.2 2. What is the pattern of ChIP peaks near transcription start sites?

```
## center of peak
peak <- resize(rowData(all), width=1, fix="center")

## start of transcript
library(TxDb.Hsapiens.UCSC.hg19.knownGene)
tx <- transcripts(TxDb.Hsapiens.UCSC.hg19.knownGene)
tss <- resize(tx, width=1)

## distance from center of peak to nearest tss
idx <- nearest(peak, tss)
sgn <- as.integer(ifelse(strand(tss)[idx] == "+", 1, -1))
dist <- (start(peak) - start(tss)[idx]) * sgn

## summarize peaks within 5k of TSS
bound <- 5000
ok <- abs(dist) < bound
table(ok)
dist5k <- dist[ok]
table(sign(dist5k))

plot(density(dist5k), xlim=range(dist5k))
abline(v=0, col="red", lty=2, lwd=2)
```

5 Further directions

- motIV for comparison of motifs, rGADEM for de novo discovery; MEME (non-Bioc) a common tool
- MotifDb for managing motifs

Downstream questions involving ranges and strings are very easily addressed.