Myc Mel Peakset Analysis

m2407447

```
# Load libraries (install these first manually in the Console!)
library(ChIPseeker)
library(GenomicRanges)
library(BSgenome.Hsapiens.UCSC.hg19)
library(GenomicFeatures)
```

1. Read in Myc Mel Replicate Peak Files

```
# Load peak files
rep1_file <- "C:/Users/Asus/OneDrive/Desktop/RStudio assessment/Chip-seq/mycmelrep1_peaks.xls"
rep2_file <- "C:/Users/Asus/OneDrive/Desktop/RStudio assessment/Chip-seq/mycmelrep2_peaks.xls"

rep1_peaks <- readPeakFile(rep1_file)
rep2_peaks <- readPeakFile(rep2_file)
```

2. Find Common Peaks

```
# Find overlapping/common peaks
overlaps <- findOverlaps(rep1_peaks, rep2_peaks)
common_peaks <- rep1_peaks[queryHits(overlaps)]

# View the first few common peaks
head(common_peaks)
```

```
## GRanges object with 6 ranges and 7 metadata columns:
##       seqnames           ranges strand |    length abs_summit    pileup
##          <Rle>        <IRanges>  <Rle> | <integer>  <integer> <numeric>
##   [1]        1 4775338-4775959      * |       623    4775616        28
##   [2]        1 4847545-4847931      * |       388    4847795        39
##   [3]        1 5073029-5073344      * |       317    5073202        41
##   [4]        1 7078802-7079170      * |       370    7078892        13
##   [5]        1 7387588-7388483      * |       897    7387940        53
##   [6]        1 7606349-7606524      * |       177    7606476        18
##       X.log10.pvalue. fold_enrichment X.log10.qvalue.               name
##             <numeric>       <numeric>       <numeric>        <character>
##   [1]        22.79415         9.52390        19.64777  mycmelrep1_peak_4
##   [2]        24.16184         7.45675        20.97327  mycmelrep1_peak_5
##   [3]        31.73000        10.20078        28.34149  mycmelrep1_peak_7
##   [4]         6.38932         4.00702         3.99521 mycmelrep1_peak_12
##   [5]        57.84166        19.01092        53.95182 mycmelrep1_peak_13
##   [6]        12.40486         6.45661         9.64372 mycmelrep1_peak_14
##   -------
##   seqinfo: 22 sequences from an unspecified genome; no seqlengths
```

3. Rank Peaks by Fold Enrichment and Select Top 500

```
# Rank by fold enrichment (descending)
ranked_peaks <- common_peaks[order(common_peaks$fold_enrichment, decreasing = TRUE)]

# Select top 500 peaks
top_500_peaks <- head(ranked_peaks, 500)

# View top ranked peaks
head(top_500_peaks)
```

```
## GRanges object with 6 ranges and 7 metadata columns:
##       seqnames              ranges strand |    length abs_summit    pileup
##          <Rle>           <IRanges>  <Rle> | <integer>  <integer> <numeric>
##   [1]        4   45965698-45967126      * |      1430   45966486       248
##   [2]        9   21155249-21158095      * |      2848   21157016       228
##   [3]        9   21155249-21158095      * |      2848   21157016       228
##   [4]       12 114345161-114346639      * |      1480  114345880       183
##   [5]        3   87846836-87847851      * |      1017   87847065       205
##   [6]        5 136577955-136578699      * |       746  136578211       175
##       X.log10.pvalue. fold_enrichment X.log10.qvalue.                  name
##             <numeric>       <numeric>       <numeric>           <character>
##   [1]         488.149        123.1159         479.841 mycmelrep1_peak_33018
##   [2]         437.832        111.5354         430.369 mycmelrep1_peak_48303
##   [3]         437.832        111.5354         430.369 mycmelrep1_peak_48303
##   [4]         360.959        104.7455         354.200 mycmelrep1_peak_11995
##   [5]         378.718         96.9533         371.892 mycmelrep1_peak_30691
##   [6]         336.006         96.8103         329.475 mycmelrep1_peak_38571
##   -------
##   seqinfo: 22 sequences from an unspecified genome; no seqlengths
```

4. Resize Peaks to 200bp Around Center

```
# Resize each peak to 200bp centered on its midpoint
resized_peaks <- resize(top_500_peaks, width = 200, fix = "center")

# Get chromosome lengths from the hg19 genome
genome_lengths <- seqlengths(BSgenome.Hsapiens.UCSC.hg19)

# Ensure chromosome names match
seqlevelsStyle(resized_peaks) <- "UCSC"

# Assign seqlengths to resized_peaks so we can validate them
seqlengths(resized_peaks) <- genome_lengths[names(seqlengths(resized_peaks))]
```

```r
# Keep only peaks that are within the chromosome boundaries
valid_peaks <- resized_peaks[start(resized_peaks) > 0 & end(resized_peaks) <= seqlengths(resized_peaks)

# Check how many peaks are valid
length(valid_peaks)
```

```
## [1] 484
```

5. Extract DNA Sequences from hg19

```r
# Fix chromosome naming style
seqlevelsStyle(resized_peaks) <- "UCSC"

# Extract DNA sequences using hg19 reference genome
seqs <- getSeq(BSgenome.Hsapiens.UCSC.hg19, valid_peaks)

# Check first few sequences
head(seqs)
```

```
## DNAStringSet object of length 6:
##     width seq
## [1]   200 ACAGCTTTTGCTCATTCAGTATGATGATGGCTGT...TTTTGTCTTTAGTTCTGTTTATGTGATAAACCA
## [2]   200 GCATGGAATGAAATGAACCTCTGATACTTGGAGT...ATGAATATATATTTAAAACCACAACAAAACACA
## [3]   200 GCATGGAATGAAATGAACCTCTGATACTTGGAGT...ATGAATATATATTTAAAACCACAACAAAACACA
## [4]   200 GGAGGAGACGACCTGTGCAGAGGAGAGACACCTG...TCTTCAGGGAAAGCCTGGAGAATGGGAAGTCTT
## [5]   200 TCAGGTCTTAATGTCATTCGGATCATCACTCTCA...TCTCCCTCACCAGTTCCAACTCCTGTATGTCTA
## [6]   200 GCAGTATGAAAATGGACTAATACACTTGTCTCTA...TCCAGCCTGGGTGACAGGGTGAAACCCTGTACC
```

6. Write Sequences to FASTA File

```r
# Create a unique identifier for each peak (e.g., peak name or ID)
names(seqs) <- paste("peak", seq_along(seqs), sep="_")

# Write the extracted sequences to a FASTA file
fasta_file <- "extracted_sequences.fasta"
writeXStringSet(seqs, filepath = fasta_file)

# Check that the FASTA file is written in the working directory
list.files()
```

```
## [1] "Chip-seq.html"              "Chip-seq.Rmd"
## [3] "Chip-seq.Rproj"             "extracted_sequences.fasta"
## [5] "MEME results"               "mycmelrep1_peaks.xls"
## [7] "mycmelrep2_peaks.xls"
```

7. Final Check: Number of Sequences

```r
# Ensure 500 sequences are present
length(seqs)
```

```
## [1] 484
```