Binf ChIP-Seq Project

Stephen Pollo

March 12, 2018

Code for running csaw. Loosely based on the example in the csaw documentation

This is the final version that was used to generate the csaw analysis of the ChIP-Seq files

```r
library("csaw", lib.loc="~/R/win-library/3.4")
```

```
## Loading required package: GenomicRanges

## Loading required package: stats4

## Loading required package: BiocGenerics

## Loading required package: parallel

##
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:parallel':
##
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, parApply, parCapply, parLapply,
##     parLapplyLB, parRapply, parSapply, parSapplyLB

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

## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, cbind, colMeans,
##     colnames, colSums, do.call, duplicated, eval, evalq, Filter,
##     Find, get, grep, grepl, intersect, is.unsorted, lapply,
##     lengths, Map, mapply, match, mget, order, paste, pmax,
##     pmax.int, pmin, pmin.int, Position, rank, rbind, Reduce,
##     rowMeans, rownames, rowSums, sapply, setdiff, sort, table,
##     tapply, union, unique, unsplit, which, which.max, which.min

## Loading required package: S4Vectors

##
## Attaching package: 'S4Vectors'

## The following object is masked from 'package:base':
##
##     expand.grid

## Loading required package: IRanges
## Loading required package: GenomeInfoDb
```

```
## Loading required package: SummarizedExperiment
```

```
## Loading required package: Biobase
```

```
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.
```

```
## Loading required package: DelayedArray
```

```
## Loading required package: matrixStats
```

```
##
## Attaching package: 'matrixStats'
```

```
## The following objects are masked from 'package:Biobase':
##
##     anyMissing, rowMedians
```

```
##
## Attaching package: 'DelayedArray'
```

```
## The following objects are masked from 'package:matrixStats':
##
##     colMaxs, colMins, colRanges, rowMaxs, rowMins, rowRanges
```

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

```
## Loading required package: BiocParallel
```

```r
library("edgeR", lib.loc="~/R/win-library/3.4")
```

```
## Loading required package: limma
```

```
##
## Attaching package: 'limma'
```

```
## The following object is masked from 'package:BiocGenerics':
##
##     plotMA
```

```r
bam.files <- c("C:/Users/Stephen/Desktop/Binf Assignment/Sorted and indexed files/ENCFF828ZWQ_sorted.bam
design <- model.matrix(~factor(c('GM12878', 'GM12878', 'MCF-7', 'MCF-7')))
colnames(design) <- c("intercept", "cell.type")

param <- readParam(minq=50)
data <- windowCounts(bam.files, ext=110, width=10, param=param)

keep <- aveLogCPM(asDGEList(data)) >= -1
data <- data[keep,]

binned <- windowCounts(bam.files, bin=TRUE, width=10000, param=param)
data <- normOffsets(binned, se.out=data)

y <- asDGEList(data)
y <- estimateDisp(y, design)
```

```
fit <- glmQLFit(y, design, robust=TRUE)
results <- glmQLFTest(fit)

merged <- mergeWindows(rowRanges(data), tol=1000L, max.width=10000L)
tabcom <- combineTests(merged$id, results$table)

summary(width(merged$region))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    10.0    60.0   210.0   571.4   710.0  9860.0
```

```
is.sig <- tabcom$FDR <= 0.05
library("rtracklayer", lib.loc="~/R/win-library/3.4")
test <- merged$region[is.sig]
test$score <- -10*log10(tabcom$FDR[is.sig])
names(test) <- paste0("region", 1:sum(is.sig))
export(test, "CSAW GM12878 vs MCF-7 clusters.bed")

write.csv(merged$region, file = "GM12878 vs MCF-7 csaw.csv")
```

Code for running DIME from the example in the documentation on simulated data

All attempts at geneating input for DIME were far too large to even load into R

```
library("DIME", lib.loc="~/R/win-library/3.4")


#The following code is from the example in the DIME documentation to verify the installation

# generate simulated datasets with underlying exponential-normal components
N1 <- 1500; N2 <- 500; K <- 4; rmu <- c(-2.25,1.50); rsigma <- c(1,1);
rpi <- c(.05,.45,.45,.05); rbeta <- c(12,10);
set.seed(1234)
chr1 <- c(-rgamma(ceiling(rpi[1]*N1),shape = 1,scale = rbeta[1]),
rnorm(ceiling(rpi[2]*N1),rmu[1],rsigma[1]),
rnorm(ceiling(rpi[3]*N1),rmu[2],rsigma[2]),
rgamma(ceiling(rpi[4]*N1),shape = 1,scale = rbeta[2]));
chr2 <- c(-rgamma(ceiling(rpi[1]*N2),shape = 1,scale = rbeta[1]),
rnorm(ceiling(rpi[2]*N2),rmu[1],rsigma[1]),
rnorm(ceiling(rpi[3]*N2),rmu[2],rsigma[2]),
rgamma(ceiling(rpi[4]*N2),shape = 1,scale = rbeta[2]));
chr3 <- c(-rgamma(ceiling(rpi[1]*N2),shape = 1,scale = rbeta[1]),
rnorm(ceiling(rpi[2]*N2),rmu[1],rsigma[1]),
rnorm(ceiling(rpi[3]*N2),rmu[2],rsigma[2]),
rgamma(ceiling(rpi[4]*N2),shape = 1,scale = rbeta[2]));
# analyzing only chromosome 1 and chromosome 3
data <- list(chr1,chr3);
# run DIME with small maximum iteration and repetitions
```

```r
set.seed(1234);
test <- DIME(data,gng.max.iter=10,gng.rep=1,inudge.max.iter=10,inudge.rep=1,
nudge.max.iter=10,nudge.rep=1)
# get the name of the best fitted model
test$best$name
```

```
## [1] "GNG"
```

```r
# get classification based on inudge
test$inudge <- DIME.classify(data,test$inudge,obj.cutoff=0.1);
# vector of classification. 1 represents differential, 0 denotes non-differential
inudgeClass <- test$inudge$class
```

Code for running normR loosely following the example in the documentation

This is the final version that was used to generate the normR analysis of the ChIP-Seq files

```r
library("normr", lib.loc="~/R/win-library/3.4")
```

```
##
## Attaching package: 'normr'
```

```
## The following object is masked from 'package:edgeR':
##
##     getCounts
```

```
## The following object is masked from 'package:methods':
##
##     getClasses
```

```r
GMpooled <- "C:/Users/Stephen/Desktop/Binf Assignment/Sorted and indexed files/combined_ENCFF247RDS_ENC
MCFpooled <- "C:/Users/Stephen/Desktop/Binf Assignment/Sorted and indexed files/combined_ENCFF729OTK_EN

diffPooled <- diffR(treatment = GMpooled, control = MCFpooled, genome="hg19", countConfig = countConfig
```

```
## Getting genome coordinates for hg19 ...
```

```
## Warning in FUN(genome = names(SUPPORTED_UCSC_GENOMES)[idx], circ_seqs = supported_genome$circ_seqs,
##   NCBI assembly: chrM
```

```
## Counting on C:/Users/Stephen/Desktop/Binf Assignment/Sorted and indexed files/combined_ENCFF729OTK_E
```

```
## Warning in .local(bampath, gr, ...): some ranges' widths are not a multiple of the selected
##              binsize, some bins will correspond to less than binsize basepairs
```

```
## Warning in .local(bampath, gr, ...): some ranges' widths are not a multiple of the selected
##              binsize, some bins will correspond to less than binsize basepairs
```

```
## ... computing Q-values.
```

```
##
##
## +++ OVERALL RESULT ++++
## NormRFit-class object
##
## Type:                  'diffR'
## Number of Regions:     12382723
## Number of Components:  3
## Theta* (naive bg):     0.484
## Background component B: 2
##
## +++ Results of fit +++
## Mixture Proportions:
##    Class 1    Background       Class 2
##      12.6%         73.4%         14.0%
## Theta:
##    Class 1    Background       Class 2
##      0.169         0.469         0.748
##
## Bayesian Information Criterion:  36190025
##
## +++ Results of binomial test +++
## T-Filter threshold: 6
## Number of Regions filtered out: 10282729
## Significantly different from background B based on q-values:
## TOTAL:
##              ***          **           *           .
## Bins           2       18294       14710       24136       18998
## %       9.06e-05    8.28e-01    1.49e+00    2.59e+00    3.45e+00
##             n.s.
## Bins     2023854
## %       9.16e+01
## Class 1:
##              ***          **           *           .
## Bins           2        5233        4192        7954        5696
## %       9.52e-05    2.49e-01    2.00e-01    3.79e-01    2.71e-01
##             n.s.
## Bins     2076917
## %       9.89e+01
## Class 2:
##             ***         **          *          .                  n.s.
## Bins          0      13061      10518      16182      13302    2046931
## %         0.000      0.622      0.501      0.771      0.633     97.473
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '  ' 1 'n.s.'
```

```
exportR(x = diffPooled, filename = "normR GM12878 vs MCF-7 pooled regions.bed",fdr=0.01, type=c("bed"))
```