Creating Response Vector Y

Spiro Stilianoudakis

August 16, 2018

Loading Libraries

```
#library(MultiAssayExperiment)
library(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
```

```
#library(IRanges)
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.4.4
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
## Warning: package 'ggplot2' was built under R version 3.4.4
library(data.table)
## Warning: package 'data.table' was built under R version 3.4.4
##
## Attaching package: 'data.table'
## The following object is masked from 'package:GenomicRanges':
##
##     shift
## The following object is masked from 'package:IRanges':
##
##     shift
## The following objects are masked from 'package:S4Vectors':
##
##     first, second
library(gbm)
## Loading required package: survival
##
## Attaching package: 'survival'
## The following object is masked from 'package:caret':
##
##     cluster
## Loading required package: splines
## Loaded gbm 2.1.3
library(pROC)
## Warning: package 'pROC' was built under R version 3.4.4
## Type 'citation("pROC")' for a citation.
##
## Attaching package: 'pROC'
## The following objects are masked from 'package:IRanges':
##
##     cov, var
## The following objects are masked from 'package:S4Vectors':
##
##     cov, var
## The following object is masked from 'package:BiocGenerics':
##
##     var
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(plyr)
```

```
##
## Attaching package: 'plyr'
## The following object is masked from 'package:IRanges':
##
##     desc
## The following object is masked from 'package:S4Vectors':
##
##     rename
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.4
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
## The following objects are masked from 'package:data.table':
##
##     between, first, last
## The following objects are masked from 'package:GenomicRanges':
##
##     intersect, setdiff, union
## The following object is masked from 'package:GenomeInfoDb':
##
##     intersect
## 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, union
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Reading in binned genome in the form of contact matrix at 10kb resolution

```
binslist10 <- read.table("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/data/10kb_bins/chr22/chr22_1

dim(binslist10)
```

```
## [1] 4105209       4
```

```
#ordering the bins according to left endpoint
binslist10 <- binslist10[order(binslist10$V2, decreasing=FALSE),]

#removing duplicate left endpoints
binslist10 <- binslist10[!duplicated(binslist10$V2),]

#extracting and renaming first 2 columns
binslist10 <- binslist10[,1:2]
colnames(binslist10) <- c("Chromosome", "Coordinate")

dim(binslist10)
```

```
## [1] 3493    2
```

```
#creating a granges object from binned genome
binslist10 <- GRanges(seqnames = binslist10$Chromosome, ranges = IRanges(start = binslist10$Coordinate,
                                                                         width = 10000))


saveRDS(binslist10, "C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/10kb_bins/chr22/bins
```

Reading in TAD data

```
#setwd("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/data")
setwd("C:/Users/Spiro Stilianoudakis/Documents/TAD_data_analysis/data")

domains <- read.table("C:/Users/Spiro Stilianoudakis/Documents/TAD_data_analysis/data/arrowhead_data.tx
domains <- domains[,1:3]
head(domains)
```

```
##   Chromosome   Start     End
## 1       chr1  915000 1005000
## 2       chr1 1030000 1235000
## 3       chr1 1255000 1450000
## 4       chr1 1710000 1840000
## 5       chr1 1860000 2055000
## 6       chr1 1865000 1985000
```

```
dim(domains)
```

```
## [1] 9274    3
```

```
#9274   3
```

```
#keeping only chr22
domains <- domains[which(domains$Chromosome=="chr22"),]

#creating granges object out of tad boundary data
coords <- domains
colnames(coords)[2:3] <- c("coordinate", "coordinate")
coords <- rbind.data.frame(coords[,c(1,2)],coords[,c(1,3)])
#remove duplicates for coordinates that are conjoined
coords <- coords[!duplicated(coords),]
coords <- coords[order(as.numeric(substr(coords$Chromosome,4,5)), coords$coordinate, decreasing = FALSE
dim(coords)
```

```
## [1] 314   2
```

```
coords$Chromosome <- as.character(coords$Chromosome)
bounds <- GRanges(seqnames=coords$Chromosome, ranges=IRanges(start=coords$coordinate, width=1))

saveRDS(bounds, "C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/10kb_bins/chr22//bounds
```

Creating Response Vector Y (1 if tad boundary is in bin; 0 if not)

```
y <- countOverlaps(binslist10, bounds)
length(y) #3493
```

```
## [1] 3493
```

```
table(y)
```

```
## y
##    0    1    2
## 3195  282   16
```

```
#   0    1    2
#3195  282   16
y <- ifelse(y>=1,1,0)
prop.table(table(y))
```

```
## y
##          0          1
## 0.91468652 0.08531348
```

```
mcols(binslist10)$y <- y
```

Creating the data frame for modeling

```
gm12878_10kb <- data.frame(y = y)

#setwd("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/data/10kb_bins/chr22")

saveRDS(gm12878_10kb, "C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/10kb_bins/chr22/g
```