ELMER.data: Supporting data for the ELMER package

Lijing Yao [cre, aut], Ben Berman [aut], Peggy Farnham [aut]
Hui Shen [ctb], Peter Laird [ctb], Simon Coetzee [ctb]

May 28, 2016

Contents

1 Introduction

This document provides an introduction of the *ELMER.data*, which contains supporting data for *ELMER*. *ELMER* is package using DNA methylation to identify enhancers, and correlates enhancer state with expression of nearby genes to identify one or more transcriptional targets. Transcription factor (TF) binding site analysis of enhancers is coupled with expression analysis of all TFs to infer upstream regulators. *ELMER* provide 5 necessary data for *ELMER* analysis:

1. Probes.motif: motif occurences within -/+100bp of probe sites on HM450K array.
2. human.TF: All human transcription factors.
3. Combined.TSS: Human TSS regions consist of TSS from hg19 UCSC gene and GENCODE V15.
4. motif.relavent.TFs: TFs may recognize the same motif.
5. Union.enhancer: A comprehensive list of genomic strong enhancers.

1.1 Installing and loading ELMER.data

To install this package, start R and enter

```
source("http://bioconductor.org/biocLite.R")
biocLite("ELMER.data")
```

```
library("ELMER.data")
library("GenomicRanges")
```

```
## 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, xtabs

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

## Loading required package:  S4Vectors

## Loading required package:  stats4

##
## Attaching package:  'S4Vectors'

## The following objects are masked from 'package:base':
##
##    colMeans, colSums, expand.grid, rowMeans, rowSums

## Loading required package:  IRanges

## Loading required package:  GenomeInfoDb
```

2 Contents

2.1 Probes.motif

Probes.motif provide information for motif occurences within -/+100bp of probe sites on HM450K array. FIMO was used with a p-value ¡ 0.0001 to scan a +/- 100bp region around each probe on HM450K using Factorbook motif position weight matrices (PWMs) and Jasper core human motif PWMs generated from the *MotifDb*. This data set is used in get.enriched.motif function in *ELMER* to calculate Odds Ratio of motif enrichments for a given set of probes. This data is storaged in a matrix with 485512 row and 91 column. Each row is each probe regions and each column is motif from Factorbook and Jasper. The value 1 indicates the occurrence of a motif in a particular probe and 0 means no occurrence.

```
data("Probes.motif")
dim(Probes.motif)

## [1] 485512     91

str(Probes.motif)

##  int [1:485512, 1:91] 0 0 0 0 0 0 0 0 0 0 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:485512] "cg00035864" "cg00050873" "cg00061679" "cg00063477" ...
##   ..$ : chr [1:91] "AP1" "AP2" "BARHL2" "BHLHE40" ...
```

2.2 human.TF

human.TF list all human transcription factors information such as gene id, symbol. It is a data frame contains symbols and gene ids for 1982 human trancription factor. This data is used for get.TFs function in *ELMER* to identify the TFs whose expression highly associate with average DNA methylation motif sties.

```
data("human.TF")
dim(human.TF)
```

```
## [1] 1982    2
```

```
head(human.TF)
```

```
##   GeneID   Symbol
## 1   2023     ENO1
## 2   3006 HIST1H1C
## 3  10155   TRIM28
## 4   3151    HMGN2
## 5    468     ATF4
## 6  10856   RUVBL2
```

2.3 Combined.TSS

Combined.TSS provide a comprehensive list of human hg19 annotated TSSs from UCSC gene and GENCODE V15. This is used for identification of distal elements in get.feature.probe in *ELMER*. This data is stored as a GRanges object containing coordinates of combined TSS.

```
data("Combined.TSS")
Combined.TSS
```

```
## GRanges object with 255540 ranges and 0 metadata columns:
##            seqnames              ranges strand
##               <Rle>           <IRanges>  <Rle>
##        [1]     chr1       [11868, 11869]      *
##        [2]     chr1       [11871, 11872]      *
##        [3]     chr1       [11873, 11874]      *
##        [4]     chr1       [12009, 12010]      *
##        [5]     chr1       [29369, 29370]      *
##        ...      ...                  ...    ...
##   [255536]     chr6 [90539619, 90539619]      *
##   [255537]    chr22 [50964033, 50964033]      *
##   [255538]    chr22 [50964034, 50964034]      *
##   [255539]    chr22 [50964570, 50964570]      *
##   [255540]    chr22 [50964905, 50964905]      *
##   -------
##   seqinfo: 48 sequences from an unspecified genome; no seqlengths
```

2.4 motif.relavent.TFs

motif.relavent.TFs list each motif and TFs that binding the motis. Multiple TFs may recognize a same motif such as TF family. This data is stored as a list whose elements are motifs and contents for each element are TFs which recognize the same motif that is the name of the element. This data is used in function get.TFs in *ELMER* to identify the real regulator TF whose motif is enriched in a given set of probes and expression associate with average DNA methylation of these motif sites.

```
data("motif.relavent.TFs")
motif.relavent.TFs[1:2]
```

```
## $AP1
## [1] "FOS"   "FOSB"  "FOSL1" "FOSL2" "JUND"  "JUNB"  "JUN"
##
## $AP2
## [1] "TFAP2A" "TFAP2B" "TFAP2C" "TFAP2D" "TFAP2E"
```

2.5 Union.enhancer

Union.enhancer is a comprehensive list of genomic strong enhancers which comes from a combination of enhancers from the Roadmap Epigenomics Mapping Consortium (REMC) and the Encyclopedia of DNA Elements (ENCODE) Project, in which enhancers were identified using ChromHMM for 98 tissues or cell lines. We used the union of genomic elements labeled as EnhG1, EnhG2, EnhA1 or EnhA2 (representing intergenic and intragenic active enhancers) in any of the 98 cell types, resulting in a total of 389,967 non-overlapping enhancer regions. FANTOM5 enhancers (43,011) identified by eRNAs for 400 distinct cell types were added to this list. This data is stored as a GRanges object contains coordinates of this human comprehensive genomic enhancer set. It will be used in get.feature.probe in *ELMER* to identify the distal enhancer probes which are at least 2Kb awary from TSS regions and overlap with comprehensive enhancers.

```
data("Union.enhancer")
Union.enhancer
```

```
## GRanges object with 571084 ranges and 0 metadata columns:
##            seqnames               ranges strand
##               <Rle>            <IRanges>  <Rle>
##        [1]     chr1       [10002, 10400]      *
##        [2]     chr1       [14802, 15200]      *
##        [3]     chr1       [16002, 16600]      *
##        [4]     chr1       [20002, 20400]      *
##        [5]     chr1       [79002, 79800]      *
##        ...      ...                  ...    ...
##   [571080]     chrY [59002202, 59002400]      *
##   [571081]     chrY [59019347, 59019376]      *
##   [571082]     chrY [59019812, 59020026]      *
##   [571083]     chrM [       2,     1000]      *
##   [571084]     chrM [    1002,    16400]      *
##   -------
##   seqinfo: 25 sequences from an unspecified genome; no seqlengths
```

```
sessionInfo()
```

```
## R version 3.3.0 (2016-05-03)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 14.04.4 LTS
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C               LC_TIME=en_US.UTF-8
##  [4] LC_COLLATE=C               LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C                  LC_ADDRESS=C
## [10] LC_TELEPHONE=C             LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats4    parallel  stats     graphics  grDevices utils     datasets  methods
## [9] base
##
## other attached packages:
## [1] GenomicRanges_1.24.0 GenomeInfoDb_1.8.2   IRanges_2.6.0        S4Vectors_0.10.1
## [5] BiocGenerics_0.18.0  ELMER.data_1.2.2     knitr_1.13
##
## loaded via a namespace (and not attached):
##  [1] codetools_0.2-14 digest_0.6.9     formatR_1.4      magrittr_1.5     evaluate_0.9
##  [6] highr_0.6        zlibbioc_1.18.0  stringi_1.0-1    XVector_0.12.0   BiocStyle_2.0.2
## [11] tools_3.3.0      stringr_1.0.0
```