Data for mCSEA package

Jordi Martorell-Marugán [1] ***and Pedro Carmona-Sáez*** *[1]

[1]Bioinformatics Unit. GENYO, Centre for Genomics and Oncological Research

*pedro.carmona@genyo.es

5 November 2019

Abstract

mCSEAdata package contains the necessary files to run the core analysis in *mCSEA* package. It also contains example data used by *mCSEA* to show it's functionality.

Package

mCSEAdata 1.6.0

Contents

1 Package contents

```
library(mCSEAdata)
data(mcseadata)
data(bandTable)
```

Firstly, **betaTest**, **phenoTest** and **exprTest** are the objects necessary to run the examples in *mCSEA* package. **betaTest** is a matrix with the beta-values of 10000 EPIC probes for 20 samples. **exprTest** is a subset of 100 genes' expression from bone marrows of 10 healthy and 10 leukemia patients. **phenoTest** is a dataframe with the explanatory variable and covariates associated to the samples.

```
class(betaTest)
## [1] "matrix"
dim(betaTest)
## [1] 10000    20
head(betaTest, 3)
##                    1         2         3         4         5         6
## cg18478105 0.6845279 0.6917252 0.8622046 0.6966168 0.1204777 0.7670960
## cg10605442 0.1370685 0.8450987 0.5480076 0.8671236 0.8300113 0.1667405
## cg27657131 0.1333706 0.6745949 0.8702664 0.9338893 0.8788454 0.1853554
##                     7          8          9        10         11         12
## cg18478105 0.93804510 0.88166619 0.90385504 0.9287976 0.04052779 0.10765614
## cg10605442 0.08727434 0.10568040 0.11896201 0.1764874 0.73534148 0.05741730
## cg27657131 0.10463463 0.05660229 0.06469281 0.2235293 0.92030432 0.04618165
##                   13        14        15         16        17        18
## cg18478105 0.1459481 0.8334884 0.1209040 0.07747453 0.7001099 0.7528026
## cg10605442 0.8213965 0.8208602 0.1671381 0.10157830 0.8874912 0.1723724
## cg27657131 0.1374107 0.8432675 0.9642680 0.14536637 0.9372422 0.9315385
##                    19         20
## cg18478105 0.86687272 0.85999403
## cg10605442 0.88836050 0.06521765
## cg27657131 0.06357636 0.50609450
class(phenoTest)
## [1] "data.frame"
dim(phenoTest)
## [1] 20  2
head(phenoTest, 3)
##   expla cov1
## 1  Case    1
## 2  Case    2
## 3  Case    1
class(exprTest)
## [1] "matrix"
dim(exprTest)
## [1] 100  20
head(exprTest, 3)
##                         1        2        3        4        5        6
## ENSG00000179023 4.145748 4.388779 4.265583 4.374576 4.463465 4.078678
## ENSG00000179029 4.485414 5.044662 5.411474 5.590093 5.365381 4.951236
```

```
## ENSG00000179041 6.618769 6.443408 7.642324 7.989362 7.133374 7.224613
##                       7        8        9       10       11       12
## ENSG00000179023 4.335878 4.121601 4.163271 4.219654 4.340421 3.917131
## ENSG00000179029 6.626413 5.070305 5.582466 5.688895 5.675448 5.053258
## ENSG00000179041 5.853054 8.198245 6.847891 6.598557 6.546835 7.211352
##                      13       14       15       16       17       18
## ENSG00000179023 4.284802 4.161627 4.308718 4.074333 4.171878 4.083548
## ENSG00000179029 5.708689 5.170988 5.480265 5.118550 5.657001 5.257061
## ENSG00000179041 7.190893 6.825418 7.342032 7.309422 6.831020 7.728485
##                      19       20
## ENSG00000179023 4.549825 4.199466
## ENSG00000179029 5.677323 5.171198
## ENSG00000179041 7.214401 6.781880
```

On the other hand, there are 6 association objects. Each one is a list of features with their associated 450k or EPIC CpG probes. The features included are promoters (**assocPromoters450k** and **assocPromotersEPIC**), gene bodies (**assocGenes450k** and **assocGenesEPIC**) and CpG islands (**assocCGI450k** and **assocCGIEPIC**). These objects are internally used by *mCSEA.test* function in *mCSEA* package.

```
class(assocPromoters450k)
## [1] "list"
length(assocPromoters450k)
## [1] 20960
head(assocPromoters450k, 3)
## $FAM197Y2
## [1] "cg00050873" "cg03052502" "cg03443143" "cg17834650" "cg02802508"
## [6] "cg03535417" "cg08635406" "cg17769199"
##
## $TTTY14
## [1] "cg00212031" "cg15345074" "cg06628792" "cg11684211" "cg11816202"
##
## $TMSB4Y
## [1] "cg00214611" "cg02004872" "cg02730008" "cg26198148"
class(assocGenes450k)
## [1] "list"
length(assocGenes450k)
## [1] 19071
head(assocGenes450k, 3)
## $TSPY4
##  [1] "cg00050873" "cg03443143" "cg04016144" "cg05544622" "cg09350919"
##  [6] "cg15810474" "cg15935877" "cg17834650" "cg17837162" "cg25705492"
## [11] "cg00543493" "cg00903245" "cg01523029" "cg02606988" "cg02802508"
## [16] "cg03535417" "cg04958669" "cg08258654" "cg08635406" "cg10239257"
## [21] "cg13861458" "cg14005657" "cg25538674" "cg26475999"
##
## $TTTY14
## [1] "cg03244189" "cg05230942" "cg10811597" "cg13765957" "cg13845521"
## [6] "cg15281205" "cg26251715"
##
## $NLGN4Y
```

Data for mCSEA package

```
##  [1] "cg03706273" "cg25518695" "cg01073572" "cg01498999" "cg02340092"
##  [6] "cg03278611" "cg04419680" "cg05939513" "cg07795413" "cg08816194"
## [11] "cg09300505" "cg09748856" "cg09804407" "cg10990737" "cg18113731"
## [16] "cg19244032" "cg27214488" "cg27265812" "cg27443332"
class(assocCGI450k)
## [1] "list"
length(assocCGI450k)
## [1] 27176
head(assocCGI450k, 3)
## $`chrY:9363680-9363943`
## [1] "cg00050873" "cg03443143"
##
## $`chrY:21238448-21240005`
## [1] "cg00212031" "cg03244189" "cg15345074" "cg06628792" "cg10811597"
## [6] "cg11684211" "cg11816202" "cg13845521" "cg26251715"
##
## $`chrY:8147877-8148210`
## [1] "cg00213748" "cg02272584" "cg06237805" "cg08160949" "cg08702825"
## [6] "cg08739478"
class(assocPromotersEPIC)
## [1] "list"
length(assocPromotersEPIC)
## [1] 26208
head(assocPromotersEPIC, 3)
## $YTHDF1
##  [1] "cg18478105" "cg10605442" "cg27657131" "cg08514185" "cg13587582"
##  [6] "cg25802399" "cg22485414" "cg03501095" "cg24092253" "cg12589387"
##
## $EIF2S3
## [1] "cg09835024" "cg06127902" "cg12275687" "cg00914804" "cg27345735"
## [6] "cg12590845" "cg25034591" "cg16712639" "cg07622257"
##
## $PKN3
## [1] "cg14361672" "cg06550760" "cg14204415" "cg11056832" "cg14036226"
## [6] "cg22365023" "cg20593100"
class(assocGenesEPIC)
## [1] "list"
length(assocGenesEPIC)
## [1] 23772
head(assocGenesEPIC, 3)
## $CCDC57
##  [1] "cg01763666" "cg26701563" "cg16920238" "cg17286790" "cg11989942"
##  [6] "cg03388043" "cg05483915" "cg05915375" "cg04098763" "cg14090409"
## [11] "cg21295367" "cg20780302" "cg01465684" "cg18209359" "cg16578864"
## [16] "cg15754222" "cg21880101" "cg05522083" "cg12952529" "cg14673194"
## [21] "cg10477817" "cg17751591" "cg11719141" "cg26928858" "cg21698718"
## [26] "cg07310278" "cg13339291" "cg13367490" "cg12336460" "cg02208313"
## [31] "cg26507988" "cg15857073" "cg22476252" "cg11935831" "cg08864681"
## [36] "cg22167267" "cg14832684" "cg09804706" "cg24973483" "cg12486944"
## [41] "cg00412514" "cg13796123" "cg13000015" "cg04824810" "cg25639749"
## [46] "cg03789597" "cg14136083" "cg13855717" "cg25612997" "cg20880890"
```

Data for mCSEA package

```
## [51] "cg04955630" "cg19976037" "cg16849440" "cg25735697" "cg22312907"
## [56] "cg12223090" "cg02967812" "cg04210266" "cg26837952" "cg06493125"
## [61] "cg08047030" "cg20798760" "cg00755572" "cg25388952" "cg13198984"
## [66] "cg01216201" "cg19567758" "cg22882093" "cg24480260" "cg23985595"
## [71] "cg06073302" "cg16477682" "cg25532751" "cg20299209" "cg11716677"
## [76] "cg02094669" "cg11859384" "cg10505658" "cg21577598" "cg24963024"
## [81] "cg17251650" "cg24378699" "cg02262688" "cg06132853" "cg22491947"
## [86] "cg02200666" "cg07959490" "cg09163921" "cg18996153" "cg20197093"
## [91] "cg18151291" "cg22142205" "cg16124601" "cg26105045" "cg23522485"
## [96] "cg16279483" "cg26093898" "cg21565972"
##
## $INF2
##  [1] "cg12950382" "cg18425377" "cg09184385" "cg10533694" "cg20980960"
##  [6] "cg07039149" "cg18519050" "cg23206460" "cg05210373" "cg03576530"
## [11] "cg25592858" "cg18996808" "cg10345522" "cg08043200" "cg17331554"
## [16] "cg03719908" "cg18465331" "cg23956771" "cg21827986" "cg14377342"
## [21] "cg24404909" "cg00816970" "cg23601271" "cg04966159" "cg18924331"
## [26] "cg22090592" "cg04278105" "cg12031670" "cg26212352" "cg02878289"
## [31] "cg05018513" "cg06971503" "cg11290775" "cg23343291" "cg18447460"
##
## $PIP5K1C
##  [1] "cg26724186" "cg05809578" "cg05233128" "cg17845617" "cg25247177"
##  [6] "cg02322048" "cg19423978" "cg16583193" "cg20969388" "cg08145067"
## [11] "cg15564488" "cg07577499" "cg03228408" "cg24732692" "cg09288755"
## [16] "cg02952625" "cg13995193" "cg11243391" "cg00793543" "cg27490930"
## [21] "cg19841005" "cg10591771" "cg10490670" "cg06358131" "cg16019751"
## [26] "cg22848927" "cg21865657" "cg19736470" "cg11955890" "cg07479621"
## [31] "cg17791316" "cg05312862" "cg07907254" "cg15483758" "cg02818004"
## [36] "cg06724384" "cg07059636" "cg17097293" "cg22623033" "cg01168835"
## [41] "cg13588224" "cg15389497" "cg23249839" "cg23480820" "cg13561409"
## [46] "cg17820448" "cg08301518" "cg17698261" "cg22677650" "cg03540494"
## [51] "cg15200445" "cg16248034" "cg14093663" "cg16564917" "cg07963254"
## [56] "cg02859655" "cg02106453" "cg00376288" "cg17290669" "cg20750693"
## [61] "cg12557799" "cg01070272" "cg06497674" "cg04385058" "cg00438838"
## [66] "cg10996109" "cg18249653" "cg06587767" "cg13670756" "cg04801430"
## [71] "cg12298375" "cg23654206" "cg15080709" "cg15540507" "cg08267629"
## [76] "cg11976007" "cg11452653" "cg09547756" "cg04742624"
class(assocCGIEPIC)
## [1] "list"
length(assocCGIEPIC)
## [1] 27187
head(assocCGIEPIC, 3)
## $`chr20:61846843-61848103`
##  [1] "cg18478105" "cg10605442" "cg27657131" "cg08514185" "cg17364922"
##  [6] "cg13587582" "cg25802399" "cg22485414" "cg15407723" "cg03501095"
## [11] "cg02177162" "cg10201192" "cg13388572" "cg00624976" "cg24092253"
## [16] "cg12589387"
##
## $`chrX:24072558-24073135`
## [1] "cg09835024" "cg06127902" "cg12275687" "cg00914804" "cg27345735"
## [6] "cg12590845" "cg25034591" "cg16712639" "cg07622257"
```

```
##
## $`chr9:131464843-131465830`
## [1] "cg14361672" "cg06550760" "cg14204415" "cg07950002" "cg11056832"
## [6] "cg14036226" "cg22365023" "cg20593100" "cg13548833"
```

There are also 2 GRanges objects with the locations of 450K and EPIC probes, used by *mCSEAPlot()* and *mCSEAIntegrate()* functions:

```
class(annot450K)
## [1] "GRanges"
## attr(,"package")
## [1] "GenomicRanges"
head(annot450K, 3)
## GRanges object with 3 ranges and 0 metadata columns:
##              seqnames    ranges strand
##                 <Rle> <IRanges>  <Rle>
##   cg00050873     chrY   9363356      *
##   cg00212031     chrY  21239348      *
##   cg00213748     chrY   8148233      *
##   -------
##   seqinfo: 24 sequences from hg19 genome; no seqlengths
class(annotEPIC)
## [1] "GRanges"
## attr(,"package")
## [1] "GenomicRanges"
head(annotEPIC, 3)
## GRanges object with 3 ranges and 0 metadata columns:
##              seqnames    ranges strand
##                 <Rle> <IRanges>  <Rle>
##   cg18478105    chr20  61847650      *
##   cg09835024     chrX  24072640      *
##   cg14361672     chr9 131463936      *
##   -------
##   seqinfo: 24 sequences from hg19 genome; no seqlengths
```

Finally, **bandTable** object contains chromosomes band information and centromer location. It is used by mCSEAPlot() function to plot the chromosome track.

```
head(bandTable)
##   chrom chromStart chromEnd   name gieStain
## 1  chr1          0  2300000 p36.33     gneg
## 2  chr1    2300000  5400000 p36.32   gpos25
## 3  chr1    5400000  7200000 p36.31     gneg
## 4  chr1    7200000  9200000 p36.23   gpos25
## 5  chr1    9200000 12700000 p36.22     gneg
## 6  chr1   12700000 16200000 p36.21   gpos50
```

2 Sources

- Example objects:
  - **betaTest** contains simulated beta-values for EPIC platform probes.
  - **exprTest** contains expression data from Leukemia and healthy patients extracted from *leukemiaEset* package.
  - **phenoTest** contains arbitrary phenotypes for each sample.
- Association objects: They were all constructed from *IlluminaHumanMethylation450kanno.ilmn12.hg19* (Hansen 2016a) and *IlluminaHumanMethylationEPICanno.ilm10b2.hg19* (Hansen 2016b) packages annotation data. For that purpose, a *RGChannelSet* object was obtained with *minfi* (Aryee et al. 2014) package and *getAnnotation()* function was applied to such object in order to get the annotation DataFrame. That was done for both 450k and EPIC platforms. The annotation DataFrame contains several information about each CpG probe, and we used that information to associate each probe to one or more promoter, gene body or CpG Island following this scheme:

Region type	mCSEAdata association objects	Column from association DataFrame used	Column values	Feature name column
Promoters	assocPromoters450k and assocPromotersEPIC	UCSC_RefGene_Group	TSS1500, TSS200, 5'UTR or 1stExon	UCSC_RefGene_Name
Gene bodies	assocGenes450k and assocGenesEPIC	UCSC_RefGene_Group	Body	UCSC_RefGene_Name
CpG Islands	assocCGI450k and assocCGIEPIC	Relation_to_Island	Island, N_Shore, S_Shore, N_Shelf or S_Shelf	Islands_Name

For instance, *cg00212031* probe from 450k platform has the following annotation data in the association DataFrame:

UCSC_RefGene_Group	UCSC_RefGene_Name	Relation_to_Island	Islands_Name
TSS200	TTTY14	Island	chrY:21238448-21240005

So this probe is associated to TTTY14 promoter in assocPromoters450k object and to chrY:21238448-21240005 CpG Island in assocCGI450k object.

- Annotation objects (**annot450K** and **annotEPIC**): They were both constructed with *minfi* package. A *RGChannelSet* object was obtained for each platform and *getLocations()* function was applied to such objects.
- bandTable: It was constructed with *Gviz* package, concretely with *IdeogramTrack()* function.

3 Session info

```
## R version 3.6.1 (2019-07-05)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 18.04.3 LTS
##
## Matrix products: default
## BLAS:   /home/biocbuild/bbs-3.10-bioc/R/lib/libRblas.so
## LAPACK: /home/biocbuild/bbs-3.10-bioc/R/lib/libRlapack.so
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=C
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] parallel  stats4    stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
## [1] GenomicRanges_1.38.0 GenomeInfoDb_1.22.0  IRanges_2.20.0
## [4] S4Vectors_0.24.0     BiocGenerics_0.32.0  mCSEAdata_1.6.0
## [7] BiocStyle_2.14.0
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.2             knitr_1.25             XVector_0.26.0
##  [4] magrittr_1.5           zlibbioc_1.32.0        rlang_0.4.1
##  [7] stringr_1.4.0          tools_3.6.1            xfun_0.10
## [10] htmltools_0.4.0        yaml_2.2.0             digest_0.6.22
## [13] bookdown_0.14          GenomeInfoDbData_1.2.2 BiocManager_1.30.9
## [16] bitops_1.0-6           RCurl_1.95-4.12        evaluate_0.14
## [19] rmarkdown_1.16         stringi_1.4.3          compiler_3.6.1
```

References

Aryee, MJ, AE Jaffe, H Corrada-Bravo, C Ladd-Acosta, AP Feinberg, KD Hansen, and RA Irizarry. 2014. "Minfi: A Flexible and Comprehensive Bioconductor Package for the Analysis of Infinium DNA Methylation Microarrays." *Bioinformatics*.

Hansen, KD. 2016a. "IlluminaHumanMethylation450kanno.ilmn12.hg19: Annotation for Illumina's 450k Methylation Arrays." *R Package Version 0.6.0*.

———. 2016b. "IlluminaHumanMethylationEPICanno.ilm10b2.hg19: Annotation for Illumina's EPIC Methylation Arrays." *R Package Version 0.6.0*.