

The `DMRcatedata` package user's guide

Peters TJ, Buckley MJ, Statham AL, Clark SJ, Molloy PL

October 24, 2015

Package Contents

`DMRcatedata` accompanies the `DMRcate` package, providing data for examples, probe filtering and transcript annotation.

```
library(DMRcatedata)
data(dmratedata)
```

Three objects are contained in `DMRcatedata`. `crosshyb` is a factor listing probe IDs potentially confounded by cross-hybridisation to other parts of the genome[1]. It is used internally by `rmSNPandCH()`.

```
str(crosshyb)
```

```
## Factor w/ 30969 levels "cg00001510","cg00003969",...: 1 2 3 4 5 6 7 8 9 10 ...
```

`illuminaSNPs` is a `data.frame` containing probes that are potentially confounded by a SNP. It lists the SNPID, distance (in nucleotides) to the CpG in question, and minor allele frequency for each associated SNP

```
class(illuminaSNPs)
```

```
## [1] "data.frame"
```

```
dim(illuminaSNPs)
```

```
## [1] 153113      3
```

```
head(illuminaSNPs)
```

```
##                               SNP Distance MinorAlleleFrequency
## cg00000108                rs9857774          16          0.056801
## cg00000109                rs9864492          17          0.010135
## cg00000165                rs76771611         25          0.027778
## cg00000292                rs62037371         31          0.333333
## cg00000769                rs74996347         51           0.5
## cg00000807 rs113612339;rs114523815         1, 38 0.500000, 0.025424
```

`myBetas` is a matrix of 450K probe beta values, matching to Homo Sapiens chromosome 20, sourced from the colon and rectal adenocarcinoma section of The Cancer Genome Atlas (TCGA) Repository. It contains 38 matched tumour/normal pairs.

```
class(myBetas)
## [1] "matrix"
dim(myBetas)
## [1] 10042 76
```

`CpGs` is a `GRanges` object containing simulated WGBS data, generated by `WGBSSuite[2]` v0.3 with the command `Rscript simulate_WGBS.R 100000 0.87605280264016 0.125787302952703 0.2 0.2 18.5989386821267 18.5989386821267 3 2 0.2 0.5 0.112588288740425,0.00330228672976263 . truncated.`

```
CpGs
## GRanges object with 100000 ranges and 12 metadata columns:
##           seqnames           ranges strand | Treatment1.C
##           <Rle>             <IRanges> <Rle> | <integer>
## [1] chr1 [ 1, 1] * | 11
## [2] chr1 [ 54, 54] * | 9
## [3] chr1 [ 58, 58] * | 14
## [4] chr1 [320, 320] * | 12
## [5] chr1 [325, 325] * | 10
## ...
## [99996] chr1 [19705499, 19705499] * | 13
## [99997] chr1 [19705511, 19705511] * | 11
## [99998] chr1 [19705521, 19705521] * | 15
## [99999] chr1 [19705567, 19705567] * | 19
## [100000] chr1 [19705760, 19705760] * | 11
##           Treatment1.cov Treatment2.C Treatment2.cov Treatment3.C
##           <integer> <integer> <integer> <integer>
## [1] 13 9 14 16
## [2] 15 16 26 18
## [3] 20 19 20 19
## [4] 15 14 14 17
## [5] 19 13 18 14
## ...
## [99996] 15 13 13 12
## [99997] 13 16 19 16
## [99998] 15 13 13 15
## [99999] 20 11 17 18
## [100000] 21 14 14 21
```

```

##          Treatment3.cov Control1.C Control1.cov Control2.C Control2.cov
##          <integer> <integer> <integer> <integer> <integer>
##      [1]             19         11          15          16          23
##      [2]             20         17          18          10          17
##      [3]             27         16          16          12          14
##      [4]             20         13          25          15          21
##      [5]             22          5          14          16          23
##      ...             ...         ...         ...         ...         ...
## [99996]             20         13          32          12          20
## [99997]             19         12          27          14          22
## [99998]             17         16          17           8          16
## [99999]             20         18          24          18          20
## [100000]            28         17          21          12          17
##          Control3.C Control3.cov
##          <integer> <integer>
##      [1]             11          14
##      [2]             19          21
##      [3]             15          19
##      [4]             18          22
##      [5]             20          26
##      ...             ...         ...
## [99996]             15          15
## [99997]             11          19
## [99998]             22          22
## [99999]             16          17
## [100000]            12          25
## -----
## seqinfo: 1 sequence from an unspecified genome; no seqlengths

```

tx.hg19, tx.hg38 and tx.mm10 are GRanges objects containing complete transcript annotation generated from ftp://ftp.ensembl.org/pub/release-75/gtf/homo_sapiens/Homo_sapiens.GRCh37.75.gtf.gz, ftp://ftp.ensembl.org/pub/release-81/gtf/homo_sapiens/Homo_sapiens.GRCh38.81.gtf.gz and ftp://ftp.ensembl.org/pub/release-81/gtf/mus_musculus/Mus_musculus.GRCm38.81.gtf.gz respectively.

```

tx.hg19
## GRanges object with 215170 ranges and 4 metadata columns:
##          seqnames          ranges strand
##          <Rle>             <IRanges> <Rle>
## ENST00000000233          chr7 [127228399, 127231759]      +
## ENST00000000412          chr12 [ 9092961, 9102551]      -
## ENST00000000442          chr11 [ 64073050, 64084210]      +
## ENST00000001008          chr12 [ 2904119, 2913124]      +
## ENST00000001146          chr2 [ 72356367, 72375167]      -

```

```

##      ...      ...      ...      ...
## ENST00000610276      chr21      [33108045, 33108720]      +
## ENST00000610277 chrHSCHR19LRC_LRC_I_CTG1      [54677109, 54693666]      -
## ENST00000610278      chr22      [21335650, 21336044]      -
## ENST00000610279      chr10      [69609283, 69610504]      +
## ENST00000610280      chr11      [58059298, 58060237]      -
##      |      gene_name      gene_type      gene_id
##      |      <character>      <character>      <character>
## ENST00000000233      |      ARF5      protein_coding      ENSG00000004059
## ENST00000000412      |      M6PR      protein_coding      ENSG00000003056
## ENST00000000442      |      ESRRA      protein_coding      ENSG00000173153
## ENST00000001008      |      FKBP4      protein_coding      ENSG00000004478
## ENST00000001146      |      CYP26B1      protein_coding      ENSG00000003137
##      ...      ...      ...      ...
## ENST00000610276      |      AP000255.6      lincRNA      ENSG00000273091
## ENST00000610277      |      MBOAT7      protein_coding      ENSG00000273130
## ENST00000610278      |      XXbac-B135H6.18      lincRNA      ENSG00000272829
## ENST00000610279      |      RP11-57G10.8      lincRNA      ENSG00000272892
## ENST00000610280      |      OR10Q2P      pseudogene      ENSG00000272900
##      |      tx_name
##      |      <character>
## ENST00000000233      |      ARF5-001
## ENST00000000412      |      M6PR-001
## ENST00000000442      |      ESRRA-002
## ENST00000001008      |      FKBP4-001
## ENST00000001146      |      CYP26B1-001
##      ...      ...
## ENST00000610276      |      AP000255.6-001
## ENST00000610277      |      MBOAT7-001
## ENST00000610278      |      XXbac-B135H6.18-001
## ENST00000610279      |      RP11-57G10.8-001
## ENST00000610280      |      OR10Q2P-001
## -----
## seqinfo: 265 sequences from an unspecified genome; no seqlengths

```

Sources

- myBetas sourced from The Cancer Genome Atlas (TCGA) data repository, colon and rectal adenocarcinoma
- illuminaSNPs sourced from http://supportres.illumina.com/documents/myillumina/88bab663-307c-444a-848e-0ed6c338ee4d/humanmethylation450_15017482_v.1.2.snpupdate.table.v3.txt, accessed February 2014

- `crosshyb` sourced from <http://www.sickkids.ca/MS-Office-Files/Research/WeksbergLab/48639-non-specific-probes-Illumina450k.xlsx>, accessed February 2014
- `tx.hg19`, `tx.hg38` and `tx.mm10` sourced from <ftp://ftp.ensembl.org>, accessed July 2015.

References

- [1] Chen YA, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, Gallinger S, Hudson TJ, Weksberg R. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics*. 2013 Jan 11;8(2).
- [2] Rackham, OJL, Dellaportas P, Petretto E, Bottolo, L. (2015). WGBSSuite: Simulating Whole Genome Bisulphite Sequencing data and benchmarking differential DNA methylation analysis tools. *Bioinformatics* (Oxford, England), (March). <http://doi.org/10.1093/bioinformatics/btv114>