Package 'deconvR'

October 18, 2022

Title Simulation and Deconvolution of Omic Profiles

Version 1.2.0

Date 2022-18-03

Description This package provides a collection of functions designed for analyzing deconvolution of the bulk sample(s) using an atlas of reference omic signature profiles and a user-selected model. Users are given the option to create or extend a reference atlas and,also simulate the desired size of the bulk signature profile of the reference cell types.The package includes the cell-type-specific methylation atlas and, Illumina Epic B5 probe ids that can be used in deconvolution. Additionally,we included BSmeth2Probe, to make mapping WGBS data to their probe IDs easier.

License Artistic-2.0

URL https://github.com/BIMSBbioinfo/deconvR

BugReports https://support.bioconductor.org/t/deconvR

biocViews DNAMethylation, Regression, GeneExpression, RNASeq, SingleCell, StatisticalMethod, Transcriptomics

Encoding UTF-8

LazyData false

Roxygen list(markdown = TRUE)

RoxygenNote 7.1.2

Depends R (>= 4.1), data.table (>= 1.14.0)

Imports S4Vectors (>= 0.30.0), methylKit (>= 1.18.0), IRanges (>= 2.26.0), GenomicRanges (>= 1.44.0), BiocGenerics (>= 0.38.0), stats, methods, foreach (>= 1.5.1), magrittr (>= 2.0.1), matrixStats (>= 0.61.0), e1071 (>= 1.7.9), quadprog (>= 1.5.8), nnls (>= 1.4), rsq (>= 2.2), MASS, utils, dplyr (>= 1.0.7), tidyr (>= 1.1.3), assertthat

Suggests testthat (>= 3.0.0), roxygen2 (>= 7.1.2), doParallel (>= 1.0.16), parallel, knitr (>= 1.34), BiocStyle (>= 2.20.2), reshape2 (>= 1.4.4), ggplot2 (>= 3.3.5), rmarkdown, devtools (>= 2.4.2), sessioninfo (>= 1.1.1), covr, RefManageR

VignetteBuilder knitr

Config/testthat/edition 3

git_url https://git.bioconductor.org/packages/deconvR

git_branch RELEASE_3_15

git_last_commit a682bc4

git_last_commit_date 2022-05-11

Date/Publication 2022-10-18

Author İrem B. Gündüz [aut, cre] (<https://orcid.org/0000-0003-2641-0916>),
Veronika Ebenal [aut] (<https://orcid.org/0000-0001-7976-3964>),
Altuna Akalin [aut] (<https://orcid.org/0000-0002-0468-0117>)

Maintainer İrem B. Gündüz <irembgunduz@gmail.com>

R topics documented:

`BSmeth2Probe` *A function to map WGBS methylation data to Illumina Probe IDs*

Description

A function to map WGBS methylation data to Illumina Probe IDs

Usage

```
BSmeth2Probe(
  probe_id_locations,
  WGBS_data,
  cutoff = 10,
  multipleMapping = FALSE
)
```

Arguments

probe_id_locations
: Either a dataframe or GRanges object containing probe IDs and their locations. If dataframe: must contain columns named "ID", "seqnames", "Start", "End", and "Strand". If GRanges: should have locations ("seqnames", "ranges", "strand"), as well as metadata column "ID". Start and end locations should be 1-based coordinates. Note that any row with NA values will not be used.

WGBS_data
: Either a GRanges object or methylKit object (methylRaw, methylBase, methylRawDB, or methylBaseDB) of CpG locations and their methylation values. Contains locations ("seqnames", "ranges", "strand") and metadata column(s) of methylation values of sample(s) (i.e. one column per sample). These methylation values must be between 0 and 1.

cutoff
: The maximum number of basepairs distance to consider for probes which have not been directly covered in the WGBS data. Default value is 10.

multipleMapping
: When searching for matches for probes not directly covered in WGBS data, should WGBS CpGs which have already been mapped to another probe still be considered? If TRUE, then yes. If FALSE, then no. Default value is FALSE.

Value

A dataframe with first column "IDs" for CpG IDs, then 1 or more columns for methylation values of sample(s) (same number of samples as in WGBS_data) ID for each probe which was mapped, and then methylation value(s) of the WGBS CpG to which it was matched (where either it overlapped or the gap was < cutoff). If it matched to more than one CpG, the mean methylation value is taken.

Examples

```
data("IlluminaMethEpicB5ProbeIDs")
load(system.file("extdata", "WGBS_GRanges.rda", package = "deconvR"))
meth_probres <- BSmeth2Probe(
    probe_id_locations = IlluminaMethEpicB5ProbeIDs,
    WGBS_data = WGBS_GRanges
)
methp_cut <- BSmeth2Probe(
    probe_id_locations = IlluminaMethEpicB5ProbeIDs,
    WGBS_data = WGBS_GRanges[5:1000],
    cutoff = 2
)
```

---

deconvolute
: *A function to deconvolute of bulk samples to their origin proportions using data from reference atlas (e.g. methylation signatures) Results of model are returned in dataframe "results". Summary of partial R-squared values of model (min, median, mean, max) are printed upon completion.*

Description

A function to deconvolute of bulk samples to their origin proportions using data from reference atlas (e.g. methylation signatures) Results of model are returned in dataframe "results". Summary of partial R-squared values of model (min, median, mean, max) are printed upon completion.

Usage

```
deconvolute(reference, vec = NULL, bulk, model = "nnls")
```

Arguments

`reference`	A dataframe containing signatures of different cell types (e.g. methylation signature) used to train the model. The first column should contain a unique ID (e.g. target ID) to match rows of the reference to rows of the bulk. All subsequent columns are cell types. One row per unit of the signature (e.g. CpG). Each cell contains the value for the cell type of this unit (e.g. methylation value of the CpG). If not given, defaults to a reference atlas which is included in this package. This reference atlas comes from Moss et al. (2018)
`vec`	The user may provide a vector with which partial R-squared of the results will be calculated. The length must match the number of rows of the reference and bulk tables merged on the ID column (with NAs removed). Defaults to row means of reference.
`bulk`	A dataframe containing signatures of bulk samples used to test to model. Should be dataframe with first column with unique IDs (does not need to exactly match list of IDs in reference,but should have significant overlap), and rest of columns = samples. Should not have duplicate IDs. May use simulateCellMix function to create this dataframe.
`model`	A string indicating which model is used to deconvolute the samples. Can be either "nnls" (for non-negative least squares) or "svr" (support vector regression) or "qp" (quadratic programming) or "rlm" (robust linear regression). If not given, defaults to "nnls".

Details

deconvolute checks if deconvolution brings advantages on top of the basic bimodal profiles through partial R-squares. The reference matrix usually follows a bimodal distribution in the case of methylation,and taking the average of the rows of methylation matrix might give a pretty similar profile to the bulk methylation profile you are trying to deconvolute.If the deconvolution is advantageous, partial R-squared is expect to be high.

Value

A list, first is a dataframe called proportions which contains predicted cell-type proportions of bulk sample profiles in "bulk", second is called rsq,containing partial-rsq values of results, one value per sample.

References

Moss, J. et al. (2018). Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. Nature communications, 9(1), 1-12. https://doi.org/10.1038/s41467-018-07466-6

Examples

```
data("HumanCellTypeMethAtlas")
bulk_data <- simulateCellMix(10, reference = HumanCellTypeMethAtlas)[[1]]
# non-least negative square regression
results_nnls <- deconvolute(
    bulk = bulk_data,
    reference = HumanCellTypeMethAtlas
)
# Quadric programming
results_qp <- deconvolute(
    reference = HumanCellTypeMethAtlas,
    bulk = bulk_data, model = "qp"
)
```

---

findSignatures *A function to construct a signature matrix*

Description

A function to construct a signature matrix

Usage

```
findSignatures(samples, sampleMeta, atlas = NULL, variation_cutoff = NULL)
```

Arguments

samples	dataframe, has first column IDs, rest of columns are samples (must have column name as sample accession ID which should be found in sampleMeta), rows are units of signature (e.g. CpGs)
sampleMeta	dataframe, must have first column for accession ID of each sample, and second column for cell type of sample, rows are samples
atlas	dataframe, the reference atlas to which new signatures can be added, if not present then a new reference atlas will be created using sample(s). Should be dataframe with column for each cell type, rows units of signature (e.g. CpGs)
variation_cutoff	either a number between 0 to 1, or NULL.For multiple samples from the same cell type, ignore CpGs with variation > variation_cutoff with that cell type. defaults to NULL (i.e. no cutoff)

Value

A dataframe extendedAtlas which contains all cell types in atlas (if given), and those in samples added by cell type, has first column "IDs", rest of columns are cell types, rows are have first cell with the ID (e.g. CpG ID) and then values of signature (e.g. methylation values)

Examples

```
data("HumanCellTypeMethAtlas")
exampleSamples <- simulateCellMix(1,
    reference = HumanCellTypeMethAtlas
)$simulated
exampleMeta <- data.table(
    "Experiment_accession" = "example_sample",
    "Biosample_term_name" = "example_cell_type"
)
colnames(exampleSamples)[-1] <- c("example_sample")
signatures <- findSignatures(
    samples = exampleSamples,
    sampleMeta = exampleMeta
)
signatures <- findSignatures(
    samples = exampleSamples, sampleMeta = exampleMeta,
    atlas = HumanCellTypeMethAtlas
)
```

HumanCellTypeMethAtlas

The comprehensive human methylome reference atlas

Description

The comprehensive human methylome reference atlas

Usage

```
data("HumanCellTypeMethAtlas")
```

Format

data.frame object with 6000 CpG loci and 25 human cell types column:

cell type columns For each cell type and CpG locus, a methylation valuebetween 0 and 1 is provided. This value represents the fraction of methylated bases of the CpG locus.

IDs CpG loci IDs for each cell type. ...

Source

https://doi.org/10.1038/s41467-018-07466-6

References

Moss, J. et al. (2018). Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. Nature communications, 9(1), 1-12.

`IlluminaMethEpicB5ProbeIDs` *A dataset Illumina probe IDs of 400000 genomic loci (identified using the "seqnames", "ranges", and "strand" values).*

Description

A dataset Illumina probe IDs of 400000 genomic loci (identified using the "seqnames", "ranges", and "strand" values).

Usage

```
data(IlluminaMethEpicB5ProbeIDs)
```

Format

GRanges object with 400000 ranges and 1 metadata column:

metadata contains seqnames ranges strand and ID ...

Source

https://support.illumina.com/downloads/infinium-methylationepic-v1-0-product-files.html

---

`simulateCellMix` *A function to generate a dataframe of mixed cell-type origin simulated samples using given reference atlas.*

Description

A function to generate a dataframe of mixed cell-type origin simulated samples using given reference atlas.

Usage

```
simulateCellMix(numberOfSamples, mixingVector = NULL, reference)
```

Arguments

numberOfSamples	The number of simulated samples to be generated in the dataframe.
mixingVector	Specify the cell origin proportions.If numberOfSamples = 1, this can be a vector of length = number of cell types in reference.Otherwise, this is a dataframe with rows for cell types (must be equal to cell types in reference) and columns for samples. Cells contain the proportion of the sample from the cell type. Use zeros for any unused cell type. If this object is not given,will use random values for the simulation.
reference	A dataframe containing signatures of different cell types used to generate the simulation. The first column should contain a unique ID (e.g. CpG target ID) which can be used in deconvolution to match rows of the reference to rows of the bulk. All subsequent columns are cell types. Rows are units of the signature. Each cell contains the value for the cell type and signature unit (e.g. methylation value at this CpG).

Value

A list containing two data frames. simulated: A dataframe which contains mixed cell-type origin simulated samples. The first column contains a unique ID (used from reference) which can be used in deconvolution to match rows of the reference to rows of the bulk.All subsequent columns are cell types. Rows are units of signature (e.g. CpGs). Each cell contains the value for the cell type and unit (e.g. methylation value at this CpG) proportions: A dataframe with the cell proportions of the generated samples. Each row is a sample. Columns are cell types.

References

Moss, J. et al. (2018). Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. Nature communications, 9(1), 1-12. https://doi.org/10.1038/s41467-018-07466-6

Examples

```
data("HumanCellTypeMethAtlas")
bulk_mix50 <- simulateCellMix(50, reference = HumanCellTypeMethAtlas)

bulk_mixVec <- simulateCellMix(1, mixingVector = c(
    0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
    0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
), reference = HumanCellTypeMethAtlas)
```

Index