Package 'EpiDISH'

April 11, 2018

Title Epigenetic Dissection of Intra-Sample-Heterogeneity

Version 1.0.0

Description EpiDISH is a R package to infer the proportions of a priori known cell subtypes present in a sample representing a mixture of such cell-types. Inference proceeds via one of 3 methods (Robust Partial Correlations-RPC, Cibersort (CBS), Constrained Projection (CP)), as determined by user.

Date 2017-06-26

Author Andrew E. Teschendorff <a.teschendorff@ucl.ac.uk>, Shijie C. Zheng<shijieczheng@gmail.com>

Maintainer Shijie Charles Zheng<shijieczheng@gmail.com>

Depends R (>= 3.4)

Imports MASS, e1071, quadprog

Suggests roxygen2, GEOquery, BiocStyle, knitr, rmarkdown, Biobase, testthat

VignetteBuilder knitr

License GPL-2

LazyData true

NeedsCompilation no

RoxygenNote 6.0.1

URL https://github.com/sjczheng/EpiDISH

BugReports https://github.com/sjczheng/EpiDISH/issues

biocViews DNAMethylation, MethylationArray, Epigenetics, DifferentialMethylation

R topics documented:

centDHSbloodDMC.m 2
DummyBeta.m 3
epidish 3

Index **5**

`centDHSbloodDMC.m` *Whole blood reference of 333 tsDHS-DMCs and 8 blood cell subtypes*

Description

Reference-based cell proportions estimation algorithms rely on a prior defined reference matrix. We leveraged cell-type specific DNAse Hypersensitive Site (DHS) information from the NIH Epigenomics Roadmap, and used 450k purified blood cell types dataset from Reinius et al (2012) to construct this improved reference DNA methylation dataset, as described in Teschendorff et al (2017). It contains 333 tsDHS-DMCs of 8 blood cell subtypes(You can select the corresponding cell types existing in the dataset, which you would like to estimate cell proportions, to use. Notice that Granulocytes consist of Neutrophils and Eosinophils. So you should include either of Granulocytes column or Neutrophils and Eosinophils columns.):

Usage

```
data(centDHSbloodDMC.m)
```

Format

A matrix with 333 rows and 8 columns

Details

- B-cells
- CD4+ T-cells
- CD8+ T-cells
- NK-cells
- Monocytes
- Neutrophils
- Eosinophils
- Granulocytes

References

Teschendorff AE, Breeze CE, Zheng SC, Beck S. *A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies.* BMC Bioinformatics (2017) 18: 105. doi:10.1186/s12859-017-1511-5.

Reinius LE, Acevedo N, Joerink M, Pershagen G, Dahlen S-E, Greco D, Soderhall C, Scheynius A, Kere J. *Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility.* PLoS ONE (2012) 7: e41361. doi:10.1371/journal.pone.0041361.

DummyBeta.m *Dummy beta value matrix*

Description

A dataset containing dummy whole blood beta value matrix The dataset is part of the 450k GEO dataset GSE80559 The tissue type is whole blood To reduce the data size, only 1000 probes are included 330 probes are overlapped with centDHSblood.m You can get the whole beta value matrix by `exprs(GEOquery::getGEO('GSE80559')[[1]])`

Usage

```
data(DummyBeta.m)
```

Format

A matrix with 1000 rows and 2 columns

Details

- beta value matrix of 1000 probes and 2 samples

epidish *Epigenetic Dissection of Intra-Sample-Heterogeneity*

Description

A reference-based function to infer the proportions of a priori known cell subtypes present in a sample representing a mixture of such cell-types. Inference proceeds via one of 3 methods (Robust Partial Correlations-RPC, Cibersort (CBS), Constrained Projection (CP)), as determined by user.

Usage

```
epidish(avdata.m, ref.m, method = c("RPC", "CBS", "CP"), maxit = 50,
  nu.v = c(0.25, 0.5, 0.75), constraint = c("inequality", "equality"))
```

Arguments

avdata.m	A data matrix with rows labeling the molecular features (should use same ID as in cent.m) and columns labeling samples (e.g. primary tumour specimens). No missing values are allowed and all values should be positive or zero. In the case of DNA methylation, these are beta-values.
ref.m	A matrix of reference 'centroids', i.e. representative molecular profiles, for a number of cell subtypes. rows label molecular features (e.g. CpGs,...) and columns label the cell-type. IDs need to be provided as rownames and colnames, respectively. No missing values are allowed, and all values in this matrix should be positive or zero. For DNAm data, values should be beta-values.
method	Chioce of a reference-based method ('RPC','CBS','CP')
maxit	Used in RPC mode, the limit on the number of IWLS iterations

nu.v	This is only used for CBS mode. It is a vector of several nv values. nu is parameter needed for nu-classification, nu-regression, and one-classification in svm
constraint	For CP mode, you can choose either of 'inequality' or 'equality' normalization constraint. The default is 'inequality' (i.e sum of weights adds to a number less or equal than 1), which was implemented in Houseman et al (2012).

Value

CP-mode A list with the following entries: estF: the estimated cell fraction matrix; ref: the reference centroid matrix used; dataREF: the input data matrix over the probes defined in the reference matrix

CBS-mode A list with the following entries: estF: the estimated cell fraction matrix; nu: a vector of 'best' nu-parameter for each sample; ref: the reference centroid matrix used; dataREF: the input data matrix over the probes defined in the reference matrix

RPC-mode A list with the following entries: estF: the estimated cell fraction matrix; ref: the reference centroid matrix used; dataREF: the input data matrix over the probes defined in the reference matrix

References

Teschendorff AE, Breeze CE, Zheng SC, Beck S. *A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies.* BMC Bioinformatics (2017) 18: 105. doi:10.1186/s12859-017-1511-5.

Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, Wiencke JK, Kelsey KT. *DNA methylation arrays as surrogate measures of cell mixture distribution.* BMC Bioinformatics (2012) 13: 86. doi:10.1186/1471-2105-13-86.

Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, Hoang CD, Diehn M, Alizadeh AA. *Robust enumeration of cell subsets from tissue expression profiles.* Nat Methods (2015) 12: 453-457. doi:10.1038/nmeth.3337.

Examples

```
data(centDHSbloodDMC.m)
data(DummyBeta.m)
out.l <- epidish(DummyBeta.m, centDHSbloodDMC.m[,1:6], method = 'RPC')
estF.m <- out.l$estF
## avdata.m is from samples you would like to infer weights of cell subtypes
## estF.m is the inferred proportions
```

Index

∗Topic **datasets**
centDHSbloodDMC.m, 2
DummyBeta.m, 3

centDHSbloodDMC.m, 2

DummyBeta.m, 3

epidish, 3