# Package 'EpiDISH'

October 17, 2020

**Title** Epigenetic Dissection of Intra-Sample-Heterogeneity

**Version** 2.4.0

**Description** EpiDISH is a R package to infer the proportions of a priori known cell-types present in a sample representing a mixture of such cell-types. Right now, the package can be used on DNAm data of whole blood, generic epithelial tissue and breast tissue. Besides, the package provides a function that allows the identification of differentially methylated cell-types and their directionality of change in Epigenome-Wide Association Studies.

**Date** 2020-03-01

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

**Maintainer** Shijie Charles Zheng<shijieczheng@gmail.com>

**Depends** R (>= 3.5)

**Imports** MASS, e1071, quadprog, parallel, stats, matrixStats, stringr, locfdr

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

**VignetteBuilder** knitr

**License** GPL-2

**LazyData** true

**RoxygenNote** 7.0.2

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

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

**biocViews** DNAMethylation, MethylationArray, Epigenetics, DifferentialMethylation, ImmunoOncology

**git_url** https://git.bioconductor.org/packages/EpiDISH

**git_branch** RELEASE_3_11

**git_last_commit** f7bf105

**git_last_commit_date** 2020-04-27

**Date/Publication** 2020-10-16

# R topics documented:



---

`CellDMC` *A function that allows the identification of differentially methylated cell-types in in Epigenome-Wide Association Studies(EWAS)*

---

## Description

An outstanding challenge of Epigenome-Wide Association Studies performed in complex tissues is the identification of the specific cell-type(s) responsible for the observed differential methylation. CellDMC is a novel statistical algorithm, which is able to identify not only differentially methylated positions, but also the specific cell-type(s) driving the methylation change.

## Usage

```
CellDMC(
  beta.m,
  pheno.v,
  frac.m,
  adjPMethod = "fdr",
  adjPThresh = 0.05,
  cov.mod = NULL,
  sort = FALSE,
  mc.cores = 1
)
```

## Arguments

| | |
|---|---|
| `beta.m` | A beta value matrix with rows labeling the CpGs and columns labeling samples. |
| `pheno.v` | A vector of phenotype. CellDMC can handle both of binary and continuous/oderinal phenotypes. `NA` is not allowed in `pheno.v`. |
| `frac.m` | A matrix contains fractions of each cell-type. Each row labels a sample, with the same order of the columns in beta.m. Each column labels a cell-type. Column names, which are the names of cell-types, are required. The rowSums of frac.m should be 1 or close to 1. |
| `adjPMethod` | The method used to adjust p values. The method can be any of method accepted by `p.adjust`. |

| | |
|---|---|
| adjPThresh | A numeric value, default as 0.05. This is used to call DMCTs. For each cell-type respectively, the CpG with the adjusted p values less than this threshold will be reported as DMCTs (-1 or 1) in the 'dmct' matrix in the returned list. |
| cov.mod | A design matrix from model.matrix, which contains other covariates to be adjusted. For example, input model.matrix(~ geneder,data = pheno.df) to adjust gender. Do not put cell-type fraction here! |
| sort | Default as FALSE. If TRUE, the data.frame in coe list will be sorted based on p value of each CpG. The order of rows in 'dmct' will not change since the orders of each cell-type are different. |
| mc.cores | The number of cores to use, i.e. at most how many threads will run simultaneously. The defatul is 1, which means no parallelization. |

## Value

A list with the following two items.

dmct A matrix gives wheter the input CpGs are DMCTs and DMCs. The first column tells whether a CpG is a DMC or not. If the CpG is called as DMC, the value will be 1, otherwise it is 0. The following columns give DMCTs for each cell-type. If a CpG is a DMCT, the value will be 1 (hypermethylated for case compared to control) or -1 (hypomethylated for case compared to control). Otherwise, the value is 0 (non-DMCT). The rows of this matrix are ordered as the same as that of the input beta.m.

coe This list contains several dataframes, which correspond to each cell-type in frac.m. Each dataframe contains all CpGs in input beta.m. All dataframes contain estimated DNAm changes (Estimate), standard error (SE), estimated t statistics (t), raw P values (p), and multiple hypothesis corrected P values (adjP).

## References

Zheng SC, Breeze CE, Beck S, Teschendorff AE. *Identification of differentially methylated cell-types in Epigenome-Wide Association Studies.* Nat Methods (2018) 15: 1059-1066 doi:10.1038/s41592-018-0213-x.

## Examples

```
data(centEpiFibIC.m)
data(DummyBeta.m)
out.l <- epidish(DummyBeta.m, centEpiFibIC.m, method = 'RPC')
frac.m <- out.l$estF
pheno.v <- rep(c(0, 1), each = 5)
celldmc.o <- CellDMC(DummyBeta.m, pheno.v, frac.m)
# Pls note this is a faked beta value matrix.
```

## Description

This reference is a subset of `centDHSbloodDMC.m`, and contains 188 DMCs which exhibit similar median DNAm values across epithelial cells, fibroblasts and ICs to ensure that the estimation of IC subtype fractions is not confounded by the epithelial and fibroblast cells in the sample. It should be used in the `hepidish` function to estimate fractions of immunce cell subtypes.

## Usage

```
data(centBloodSub.m)
```

## Format

A matrix with 188 rows and 7 columns

## Details

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

## References

Zheng SC, Webster AP, Dong D, Feber A, Graham DG, Sullivan R, Jevons S, Lovat LB, Beck S, Widschwendter M, Teschendorff AE *A novel cell-type deconvolution algorithm reveals substantial contamination by immune cells in saliva, buccal and cervix.* Epigenomics (2018) 10: 925-940. doi:10.2217/epi-2018-0037.

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.

---

`centDHSbloodDMC.m` *Whole blood reference of 333 tsDHS-DMCs and 7 blood cell subtypes*

---

## Description

Reference-based cell-type fraction 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 whole blood reference DNA methylation dataset, as described in Teschendorff et al (2017). It contains 333 tsDHS-DMCs of 7 blood cell subtypes(*As the fractions of eosinophils are usually small, you could add the estimated fractions of neutrophils and eosinophils togetther as the estimations of granulocytes.*):

### Usage

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

### Format

A matrix with 333 rows and 7 columns

### Details

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

### 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.

---

## centEpiFibFatIC.m *Reference for breast tissue*

---

### Description

This reference was designed for estimating fractions of epithelial cells, fibroblasts, fat cells and total immune cells in breast tissue.

### Usage

```
data(centEpiFibFatIC.m)
```

### Format

A matrix with 491 rows and 4 columns

### Details

- Epi
- Fib
- Fat
- IC

## References

Zheng SC, Webster AP, Dong D, Feber A, Graham DG, Sullivan R, Jevons S, Lovat LB, Beck S, Widschwendter M, Teschendorff AE *A novel cell-type deconvolution algorithm reveals substantial contamination by immune cells in saliva, buccal and cervix.* Epigenomics (2018) 10: 925-940. doi:10.2217/epi-2018-0037.

---

`centEpiFibIC.m` *Reference for genenric epithelial tissue*

---

## Description

This reference could be used to estimate pproportions of epithelial cells, fibroblasts, and total immune cells in epithelial tissues.

## Usage

```
data(centEpiFibIC.m)
```

## Format

A matrix with 716 rows and 3 columns

## Details

- Epi
- Fib
- IC

## References

Zheng SC, Webster AP, Dong D, Feber A, Graham DG, Sullivan R, Jevons S, Lovat LB, Beck S, Widschwendter M, Teschendorff AE *A novel cell-type deconvolution algorithm reveals substantial contamination by immune cells in saliva, buccal and cervix.* Epigenomics (2018) 10: 925-940. doi:10.2217/epi-2018-0037.

---

`DoMetaEfron` *An R-function to perform a meta-analysis over multiple studies using an empirical Bayes procedure by Efron followed by Stouffer method.*

---

## Description

An R-function to perform a meta-analysis over multiple studies using an empirical Bayes procedure by Efron followed by Stouffer method.

## Usage

```
DoMetaEfron(stat.m, pval.m, bre = 120, df = 15, pct0 = 0.25, plotlocfdr = 0)
```

### Arguments

| | |
|---|---|
| stat.m | A matrix of signed statistics (e.g. t-statistics) with rows labeling genomic features (e.g. CpGs or genes) and columns labeling studies. rownames must be provided. |
| pval.m | A matrix matched to stat.m containing the associated P-values, with rows labeling genomic features (e.g. CpGs or genes) and columns labeling studies. |
| bre | The number of breakpoints to divide statistics per study into bins. By default this is 120. See input argument for locfdr function from locfdr package. |
| df | The number of degrees of freedom for fitting spline. By default this is 15. See input argument for locfdr function from locfdr package. |
| pct0 | Percentage of statistics to use for fitting null. By default this is 0.25 (i.e. 25%). |
| plotlocfdr | Determines whether to plot output or not. By default this is set to 0 meaning no plot. See input argument for locfdr function from locfdr package. |

### Value

meta.m A matrix with rows as in stat.m, and with 3 columns labeling Stouffer's z-statistic, P-value and Benjamini-Hochberg adjusted P-value.

---

## DummyBeta.m *Dummy beta value matrix*

---

### Description

A faked beta value matrix of 2000 CpGs and 10 samples. Only used for demostration purpose..

### Usage

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

### Format

A matrix with 2000 CpGs and 10 columns

### Details

- beta value matrix of 2000 CpGs and 10 samples

epidish *Epigenetic Dissection of Intra-Sample-Heterogeneity*

## Description

A reference-based function to infer the fractions 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 the user.

## Usage

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

## Arguments

| | |
|---|---|
| beta.m | A data matrix with rows labeling the molecular features (should use same ID as in ref.m) and columns labeling samples (e.g. primary tumour specimens). Missing value is not 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. Missing value is not 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 | Only used in RPC mode, the limit of the number of IWLS iterations |
| nu.v | Only used in CBS mode. It is a vector of several candidate nu values. nu is parameter needed for nu-classification, nu-regression, and one-classification in svm. The best estimation results among all candidate nu will be automatically returned. |
| constraint | Only used in 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: a matrix of the estimated fractions; ref: the reference centroid matrix used; dataREF: the subset of the input data matrix with only the probes defined in the reference matrix.

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

RPC-mode A list with the following entries: estF: a matrix of the estimated fractions; ref: the reference centroid matrix used; dataREF: the subset of the input data matrix with only 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')
frac.m <- out.l$estF
```

---

`hepidish` *Hierarchical EpiDISH (HEpiDISH)*

---

## Description

HEpiDISH is an iterative hierarchical procedure of EpiDISH. HEpiDISH uses two distinct DNAm references, a primary reference for the estimation of several cell-types fractions, and a separate secondary non-overlapping DNAm reference for the estimation of underlying subtype fractions of one of the cell-type in the primary reference.

## Usage

```
hepidish(
  beta.m,
  ref1.m,
  ref2.m,
  h.CT.idx,
  method = c("RPC", "CBS", "CP"),
  maxit = 50,
  nu.v = c(0.25, 0.5, 0.75),
  constraint = c("inequality", "equality")
)
```

## Arguments

| | |
|---|---|
| `beta.m` | A data matrix with rows labeling the molecular features (should use same ID as in reference matrices) and columns labeling samples (e.g. primary tumour specimens). Missing value is not allowed and all values should be positive or zero. In the case of DNA methylation, these are beta-values. |
| `ref1.m` | A matrix of **primary** 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. Missing value is not allowed, and all values in this matrix should be positive or zero. For DNAm data, values should be beta-values. |
| `ref2.m` | Similar to `ref1.m`, but now a A matrix of **secondary** reference. For example, `ref1.m` contains reference centroids for epithelial cells, fibroblasts and total immune cells. `ref2.m` can be subtypes of immune cells, such as B-cells, NK cells, monocytes and etc. |
| `h.CT.idx` | A index tells which cell-type in `ref1.m` is the higher order cell-types in `ref2.m`. For example, `ref1.m` contains reference centroids for epithelial cells, fibroblasts and total immune cells. `ref2.m` contains subtypes of immune cells, the `h.CT.idx` should be 3, corresponding to immune cells in `ref1.m`. |
| `method` | Chioce of a reference-based method ('RPC','CBS','CP') |
| `maxit` | Only used in RPC mode, the limit of the number of IWLS iterations |
| `nu.v` | Only used in CBS mode. It is a vector of several candidate nu values. nu is parameter needed for nu-classification, nu-regression, and one-classification in svm. The best estimation results among all candidate nu will be automatically returned. |
| `constraint` | Only used in 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

A matrix of the estimated fractions

## References

Zheng SC, Webster AP, Dong D, Feber A, Graham DG, Sullivan R, Jevons S, Lovat LB, Beck S, Widschwendter M, Teschendorff AE *A novel cell-type deconvolution algorithm reveals substantial contamination by immune cells in saliva, buccal and cervix.* Epigenomics (2018) 10: 925-940. doi:10.2217/epi-2018-0037.

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(centEpiFibIC.m)
data(centBloodSub.m)
data(DummyBeta.m)
frac.m <- hepidish(beta.m = DummyBeta.m, ref1.m = centEpiFibIC.m,
ref2.m = centBloodSub.m, h.CT.idx = 3, method = 'RPC')
```

---

# LiuDataSub.m *Whole blood example beta value matrix*

---

## Description

This beta value matrix is a subset matrix of Liu et al data(GSE42861). Beta values of 326 CpGs in the centDHSbloodDMC reference matrix and other randomly chosen 174 CpGs are included for 50 randomly chosen samples.

## Usage

```
data(LiuDataSub.m)
```

## Format

A matrix with 500 CpGs and 50 columns

## Details

- beta value matrix of 500 CpGs and 50 samples

# Index