# Package 'MethylGenotyper'

April 5, 2024

**Type** Package

**Title** MethylGenotyper: Call genotypes from SNP probes and CCS probes on Illumina methylation array

**Version** 1.0

**RoxygenNote** 7.2.3

**Imports** minfi, tidyverse, foreach, doParallel, HardyWeinberg, multimode, rlist, stats4, ggplot2, ggpubr, ggsci

**LazyData** false

**VignetteBuilder** knitr

**Suggests** knitr, rmarkdown

**Roxygen** list(markdown = TRUE)

**Depends** R (>= 2.10)

**Encoding** UTF-8

## R topics documented:

---

`backgroundCorrectionNoobFit` *Fit Normal and exponential distributions (adapted from* `SeSAMe`*)*

---

## Description

Fit Normal and exponential distributions (adapted from `SeSAMe`)

## Usage

```
backgroundCorrectionNoobFit(ib, bg)
```

## Arguments

| | |
|---|---|
| `ib` | Foreground signals. |
| `bg` | Background signals. |

## Value

mu and sigma by fitting background signals with normal distribution.

alpha by fitting foreground signals with exponential distribution.

callGeno_snp *Call genotypes for SNP probes*

## Description

Call genotypes for SNP probes

## Usage

```
callGeno_snp(
  inData,
  input = "raw",
  plotRAI = FALSE,
  vcf = FALSE,
  vcfName = "genotypes.snp_probe.vcf",
  GP_cutoff = 0.9,
  outlier_cutoff = "max",
  missing_cutoff = 0.1,
  R2_cutoff_up = 1.1,
  R2_cutoff_down = 0.75,
  MAF_cutoff = 0.01,
  HWE_cutoff = 1e-06,
  pop = "EAS",
  bayesian = FALSE,
  platform = "EPIC",
  verbose = 1
)
```

## Arguments

| | |
|---|---|
| inData | If input="raw", provide rgData here (Noob and dye-bias corrected signals produced by using `correct_noob_dye`). Otherwise, provide beta or M-value matrix here. |
| input | Input data types. One of "raw", "beta", and "mval". If input is "beta" or "mval", please use probes as rows and samples as columns. |
| plotRAI | If TRUE, plot distribution of RAIs. |
| vcf | If TRUE, will write a VCF file in the current directory. |
| vcfName | VCF file name. Only effective when vcf=TRUE. |
| GP_cutoff | When calculating missing rate, genotypes with the highest genotype probability < GP_cutoff will be treated as missing. |
| outlier_cutoff | "max" or a number ranging from 0 to 1. If outlier_cutoff="max", genotypes with outlier probability larger than all of the three genotype probabilities will be set as missing. If outlier_cutoff is a number, genotypes with outlier probability > outlier_cutoff will be set as missing. |
| missing_cutoff | Missing rate cutoff to filter variants. Note that for VCF output, variants with missing rate above the cutoff will be marked in the `FILTER` column. For the returned dosage matrix, variants with missing rate above the cutoff will be removed. |

`R2_cutoff_up, R2_cutoff_down`
: R-square cutoffs to filter variants (Variants with R-square > R2_cutoff_up or < R2_cutoff_down should be removed). Note that for VCF output, variants with R-square outside this range will be marked in the `FILTER` column. For the returned dosage matrix, variants with R-square outside this range will be removed.

`MAF_cutoff`
: A MAF cutoff to filter variants. Note that for VCF output, variants with MAF below the cutoff will be marked in the `FILTER` column. For the returned dosage matrix, variants with MAF below the cutoff will be removed.

`HWE_cutoff`
: HWE p value cutoff to filter variants. Note that for VCF output, variants with HWE p value below the cutoff will be marked in the `FILTER` column. For the returned dosage matrix, variants with HWE p value below the cutoff will be removed.

`pop`
: Population. One of EAS, AMR, AFR, EUR, SAS, and ALL.

`bayesian`
: Use the Bayesian approach to calculate posterior genotype probabilities.

`platform`
: EPIC or 450K.

`verbose`
: Verbose mode: 0/1/2.

## Value

A list containing

`dosage`
: A matrix of genotype calls. Variants with R2s, HWE p values, MAFs, or missing rates beyond the cutoffs are removed.

`genotypes`
: A list containing RAI, shapes of the mixed beta distributions, prior probabilities that the RAI values belong to one of the three genotypes, proportion of RAI values being outlier (U), and genotype probability (GP).

---

# `callGeno_typeI` *Call genotypes for Type I probes*

---

## Description

Call genotypes for Type I probes

## Usage

```
callGeno_typeI(
  rgData,
  plotRAI = FALSE,
  vcf = FALSE,
  vcfName = "genotypes.typeI_probe.vcf",
  bw = 0.04,
  minDens = 0.001,
  GP_cutoff = 0.9,
  outlier_cutoff = "max",
  missing_cutoff = 0.1,
  R2_cutoff_up = 1.1,
  R2_cutoff_down = 0.75,
```

```
  MAF_cutoff = 0.01,
  HWE_cutoff = 1e-06,
  cpu = 1,
  pop = "EAS",
  bayesian = FALSE,
  platform = "EPIC",
  verbose = 1
)
```

## Arguments

| | |
|---|---|
| rgData | Noob and dye-bias corrected signals produced by using correct_noob_dye. |
| plotRAI | If TRUE, plot distribution of RAIs. |
| vcf | If TRUE, will write a VCF file in the current directory. |
| vcfName | VCF file name. Only effective when vcf=TRUE. |
| bw | band width. |
| minDens | A parameter for mode test. Minimum density for a valid peak. |
| GP_cutoff | When calculating missing rate, genotypes with the highest genotype probability < GP_cutoff will be treated as missing. |
| outlier_cutoff | "max" or a number ranging from 0 to 1. If outlier_cutoff="max", genotypes with outlier probability larger than all of the three genotype probabilities will be set as missing. If outlier_cutoff is a number, genotypes with outlier probability > outlier_cutoff will be set as missing. |
| missing_cutoff | Missing rate cutoff to filter variants. Note that for VCF output, variants with missing rate above the cutoff will be marked in the FILTER column. For the returned dosage matrix, variants with missing rate above the cutoff will be removed. |
| R2_cutoff_up, R2_cutoff_down | R-square cutoffs to filter variants (Variants with R-square > R2_cutoff_up or < R2_cutoff_down should be removed). Note that for VCF output, variants with R-square outside this range will be marked in the FILTER column. For the returned dosage matrix, variants with R-square outside this range will be removed. |
| MAF_cutoff | A MAF cutoff to filter variants. Note that for VCF output, variants with MAF below the cutoff will be marked in the FILTER column. For the returned dosage matrix, variants with MAF below the cutoff will be removed. |
| HWE_cutoff | HWE p value cutoff to filter variants. Note that for VCF output, variants with HWE p value below the cutoff will be marked in the FILTER column. For the returned dosage matrix, variants with HWE p value below the cutoff will be removed. |
| cpu | Number of CPU cores. |
| pop | Population. One of EAS, AMR, AFR, EUR, SAS, and ALL. Only probes with MAF of matching population > 0.01 will be kept. |
| bayesian | Use the Bayesian approach to calculate posterior genotype probabilities. |
| platform | EPIC or 450K. |
| verbose | Verbose mode: 0/1/2. |

## Value

A list containing

| | |
|---|---|
| dosage | A matrix of genotype calls. Variants with R2s, HWE p values, MAFs, or missing rates beyond the cutoffs are removed. |
| genotypes | A list containing RAI, shapes of the mixed beta distributions, prior probabilities that the RAI values belong to one of the three genotypes, proportion of RAI values being outlier (U), and genotype probability (GP). |

---

`callGeno_typeII` *Call genotypes for Type II probes*

---

## Description

Call genotypes for Type II probes

## Usage

```
callGeno_typeII(
  inData,
  input = "raw",
  plotRAI = FALSE,
  vcf = FALSE,
  vcfName = "genotypes.typeII_probe.vcf",
  bw = 0.04,
  minDens = 0.001,
  GP_cutoff = 0.9,
  outlier_cutoff = "max",
  missing_cutoff = 0.1,
  R2_cutoff_up = 1.1,
  R2_cutoff_down = 0.75,
  MAF_cutoff = 0.01,
  HWE_cutoff = 1e-06,
  cpu = 1,
  pop = "EAS",
  maxiter = 50,
  bayesian = FALSE,
  platform = "EPIC",
  verbose = 1
)
```

## Arguments

| | |
|---|---|
| inData | If input="raw", provide rgData here (Noob and dye-bias corrected signals produced by using `correct_noob_dye`). Otherwise, provide beta or M-value matrix here. |
| input | Input data types. One of "raw", "beta", and "mval". If input is "beta" or "mval", please use probes as rows and samples as columns. |
| plotRAI | If TRUE, plot distribution of RAIs. |
| vcf | If TRUE, will write a VCF file in the current directory. |

| | |
|---|---|
| vcfName | VCF file name. Only effective when vcf=TRUE. |
| bw | band width. |
| minDens | A parameter for mode test. Minimum density for a valid peak. |
| GP_cutoff | When calculating missing rate, genotypes with the highest genotype probability < GP_cutoff will be treated as missing. |
| outlier_cutoff | "max" or a number ranging from 0 to 1. If outlier_cutoff="max", genotypes with outlier probability larger than all of the three genotype probabilities will be set as missing. If outlier_cutoff is a number, genotypes with outlier probability > outlier_cutoff will be set as missing. |
| missing_cutoff | Missing rate cutoff to filter variants. Note that for VCF output, variants with missing rate above the cutoff will be marked in the FILTER column. For the returned dosage matrix, variants with missing rate above the cutoff will be removed. |
| R2_cutoff_up, R2_cutoff_down | R-square cutoffs to filter variants (Variants with R-square > R2_cutoff_up or < R2_cutoff_down should be removed). Note that for VCF output, variants with R-square outside this range will be marked in the FILTER column. For the returned dosage matrix, variants with R-square outside this range will be removed. |
| MAF_cutoff | A MAF cutoff to filter variants. Note that for VCF output, variants with MAF below the cutoff will be marked in the FILTER column. For the returned dosage matrix, variants with MAF below the cutoff will be removed. |
| HWE_cutoff | HWE p value cutoff to filter variants. Note that for VCF output, variants with HWE p value below the cutoff will be marked in the FILTER column. For the returned dosage matrix, variants with HWE p value below the cutoff will be removed. |
| cpu | Number of CPU cores. |
| pop | Population. One of EAS, AMR, AFR, EUR, SAS, and ALL. Only probes with MAF of matching population > 0.01 will be kept. |
| maxiter | Maximal number of iterations for the EM algorithm. |
| bayesian | Use the Bayesian approach to calculate posterior genotype probabilities. |
| platform | EPIC or 450K. |
| verbose | Verbose mode: 0/1/2. |

## Value

A list containing

| | |
|---|---|
| dosage | A matrix of genotype calls. Variants with R2s, HWE p values, MAFs, or missing rates beyond the cutoffs are removed. |
| genotypes | A list containing RAI, shapes of the mixed beta distributions, prior probabilities that the RAI values belong to one of the three genotypes, proportion of RAI values being outlier (U), and genotype probability (GP). |
| methyl_recalc | Re-calculated methylation levels on reference alleles. A list containing shapes of the mixed beta distributions and true methylation level (pM) for each probe. |

`call_genotypes` *Call genotypes based on EM algorithm*

## Description

The Expectation–maximization (EM) algorithm is used to fit a mixture of three beta distributions representing the three genotypes (AA, AB, and BB) and one uniform distribution representing the outliers (adapted from ewastools). Probe-specific weights were used in the EM algorithm.

## Usage

```
call_genotypes(
  RAI,
  pop,
  type,
  maxiter = 50,
  bayesian = FALSE,
  platform = "EPIC",
  verbose = 1
)
```

## Arguments

| | |
|---|---|
| `RAI` | A matrix of RAI (Ratio of Alternative allele Intensity) for probes. Provide probes as rows and samples as columns. |
| `pop` | Population to be used to extract AFs. One of EAS, AMR, AFR, EUR, SAS, and ALL. |
| `type` | One of snp_probe, typeI_probe, and typeII_probe. |
| `maxiter` | Maximal number of iterations for the EM algorithm. |
| `bayesian` | Use the Bayesian approach to calculate posterior genotype probabilities. |
| `platform` | EPIC or 450K. |
| `verbose` | Verbose mode: 0/1/2. |

## Value

A list containing

| | |
|---|---|
| `RAI` | Ratio of Alternative allele Intensity |
| `shapes` | Shapes of the mixed beta distributions |
| `weights` | Prior probabilities that the RAI values belong to one of the three genotypes |
| `U` | Overall probability of RAI values being outlier |
| `outliers` | Probability of each RAI value being outlier |
| `logLik` | Log-likelihood |
| `GP` | Genotype probabilities of the three genotypes |

`constrain_R2` *Constrain R2*

## Description

R2 is calculated by var(G)/2p(1-p), where G is dosage genotype and p is allele frequency. Variants with 1<R2<=1.1 are constrained to 1. Variants with R2>1.1 (marked as .) are recommended to remove.

## Usage

```
constrain_R2(R2)
```

## Arguments

| | |
|---|---|
| `R2` | R-square |

## Value

Constrained R2

`correct_noob_dye` *Noob and dye-bias correction*

## Description

Noob and dye-bias correction

## Usage

```
correct_noob_dye(target, platform = "EPIC", cpu = 1)
```

## Arguments

| | |
|---|---|
| `target` | A data frame of two columns: Sample_Name, Basename, where Basename tells the location of IDAT files. |
| `platform` | EPIC or 450K. |
| `cpu` | Number of CPU. |

## Value

A list of noob and dye-bias corrected signals containing:

| | |
|---|---|
| `AR` | - A matrix of probeA signals in Red channel |
| `AG` | - A matrix of probeA signals in Green channel |
| `BR` | - A matrix of probeB signals in Red channel |
| `BG` | - A matrix of probeB signals in Green channel |

## dosage2hard *Get hard genotypes from genotype probabilities*

### Description

Hard genotype is defined as the genotype with highest genotype probability.

### Usage

```
dosage2hard(AA, AB, BB)
```

### Arguments

| | |
|---|---|
| AA | Genotype probability of AA or 0/0. |
| AB | Genotype probability of AB or 0/1. |
| BB | Genotype probability of BB or 1/1. |

### Value

A matrix of hard genotypes.

## eBeta *Moments estimator for beta distribution (adapted from ewastools)*

### Description

Moments estimator for beta distribution (adapted from ewastools)

### Usage

```
eBeta(x, w)
```

### Arguments

| | |
|---|---|
| x | A vector of RAI values. |
| w | Weights. |

### Value

A list of beta distribution shapes.

## filter_by_AF *Filter by AF*

### Description

Variants with MAF<0.01 are recommended to remove.

### Usage

```
filter_by_AF(AF, MAF_cutoff = 0.01)
```

### Arguments

| | |
|---|---|
| AF | Allele frequency |
| MAF_cutoff | An MAF (Minor allele frequency) cutoff to filter variants. |

### Value

Whether the variant passed filtering.

## filter_by_HWE *Filter by Hardy–Weinberg Equilibrium (HWE) p value*

### Description

Variants with Hardy–Weinberg Equilibrium (HWE) p value < HWE_cutoff are recommended to remove.

### Usage

```
filter_by_HWE(hwe_p, HWE_cutoff = 1e-06)
```

### Arguments

| | |
|---|---|
| hwe_p | Hardy–Weinberg Equilibrium (HWE) p values |
| HWE_cutoff | A HWE p value cutoff to filter variants. |

### Value

Whether the variant passed filtering.

## filter_by_missing *Filter by missing rate*

### Description

Variants with missing rate > missing_cutoff are recommended to remove.

### Usage

```
filter_by_missing(F_MISSING, missing_cutoff = 0.1)
```

### Arguments

| | |
|---|---|
| F_MISSING | Fraction of missing genotypes |
| missing_cutoff | Missing rate cutoff to filter variants. |

### Value

Whether the variant passed filtering.

## filter_by_R2 *Filter by R2*

### Description

Variants with R2>1.1 (marked as .) or R2<0.75 are recommended to remove.

### Usage

```
filter_by_R2(R2, R2_cutoff_up = 1.1, R2_cutoff_down = 0.75)
```

### Arguments

| | |
|---|---|
| R2 | R-square |
| R2_cutoff_up | Variants with R-square greater than this cutoff should be removed. |
| R2_cutoff_down | Variants with R-square less than this cutoff should be removed. |

### Value

Whether the variant passed filtering.

## fit_beta_em *Estimate mixed beta distribution parameters based on EM algorithm*

### Description

The Expectation–maximization (EM) algorithm is used to fit a mixture of three beta distributions representing the three genotypes (AA, AB, and BB) and one uniform distribution representing the outliers (adapted from ewastools).

### Usage

```
fit_beta_em(RAI, maxiter = 50, verbose = 1)
```

### Arguments

| | |
|---|---|
| RAI | A matrix of RAI (Ratio of Alternative allele Intensity) for probes. Provide probes as rows and samples as columns. |
| maxiter | Maximal number of iterations for the EM algorithm. |
| verbose | Verbose mode: 0/1/2. |

### Value

A list containing

| | |
|---|---|
| shapes | Shapes of the mixed beta distributions |
| weights | Prior probabilities that the RAI values belong to one of the three genotypes |
| U | Overall probability of RAI values being outlier |
| outliers | Probability of each RAI value being outlier |
| logLik | Log-likelihood |
| GP | Genotype probabilities of the three genotypes |

## format_genotypes *Format genotype calls*

### Description

Format genotype calls

### Usage

```
format_genotypes(
  genotypes,
  vcf = FALSE,
  vcfName,
  GP_cutoff = 0.9,
  outlier_cutoff = "max",
  missing_cutoff = 0.1,
  R2_cutoff_up = 1.1,
```

```
  R2_cutoff_down = 0.75,
  MAF_cutoff = 0.01,
  HWE_cutoff = 1e-06,
  pop = "ALL",
  type,
  plotAF = FALSE,
  platform = "EPIC"
)
```

## Arguments

| | |
|---|---|
| genotypes | Genotype calls. |
| vcf | If TRUE, will write a VCF file in the current directory. |
| vcfName | VCF file name. Only effective when vcf=TRUE. |
| GP_cutoff | When calculating missing rate, genotypes with the highest genotype probability < GP_cutoff will be treated as missing. |
| outlier_cutoff | "max" or a number ranging from 0 to 1. If outlier_cutoff="max", genotypes with outlier probability larger than all of the three genotype probabilities will be set as missing. If outlier_cutoff is a number, genotypes with outlier probability > outlier_cutoff will be set as missing. |
| missing_cutoff | Missing rate cutoff to filter variants. Note that for VCF output, variants with missing rate above the cutoff will be marked in the FILTER column. For the returned dosage matrix, variants with missing rate above the cutoff will be removed. |
| R2_cutoff_up, R2_cutoff_down | R-square cutoffs to filter variants (Variants with R-square > R2_cutoff_up or < R2_cutoff_down should be removed). Note that for VCF output, variants with R-square outside this range will be marked in the FILTER column. For the returned dosage matrix, variants with R-square outside this range will be removed. |
| MAF_cutoff | MAF cutoff to filter variants. Note that for VCF output, variants with MAF below the cutoff will be marked in the FILTER column. For the returned dosage matrix, variants with MAF below the cutoff will be removed. |
| HWE_cutoff | HWE p value cutoff to filter variants. Note that for VCF output, variants with HWE p value below the cutoff will be marked in the FILTER column. For the returned dosage matrix, variants with HWE p value below the cutoff will be removed. |
| pop | Population to be used to extract AFs. One of EAS, AMR, AFR, EUR, SAS, and ALL. |
| type | One of snp_probe, typeI_probe, and typeII_probe. |
| plotAF | To plot the distribution of AFs in 1KGP and input data. |
| platform | EPIC or 450K. |

## Value

A matrix of genotype calls. Variants with R2s, HWE p values, MAFs, or missing rates beyond the cutoffs are removed.

## getHWE *Calculate Hardy-Weinberg Equilibrium (HWE) p value*

### Description

Calculate Hardy-Weinberg Equilibrium (HWE) p value

### Usage

```
getHWE(hardgeno)
```

### Arguments

| | |
|---|---|
| hardgeno | A matrix of hard genotypes, with each row indicates a SNP and each column indicates a sample. |

### Value

HWE p values.

## getKinship *Get kinship coefficients and inbreeding coefficients using the SEEKIN estimator*

### Description

Only SNPs with missing rate < 10% were used.

### Usage

```
getKinship(dosage)
```

### Arguments

| | |
|---|---|
| dosage | A matrix of genotype calls. Provide probes as rows and samples as columns. |

### Value

A list containing

| | |
|---|---|
| kinship | A data frame containing kinship coefficient (Phi) and sample relationships between each two samples. |
| inbreed | A vector of inbreeding coefficients. |

## getKinship_het *Get kinship coefficients using the SEEKIN-het estimator*

### Description

Only SNPs with missing rate < 10% were used.

### Usage

```
getKinship_het(dosage, indAF)
```

### Arguments

| | |
|---|---|
| dosage | A matrix of genotype calls. Provide probes as rows and samples as columns. |
| indAF | A matrix of individual-specific AFs. Provide probes as rows and samples as columns. |

### Value

A data frame containing kinship coefficient (Phi) and sample relationships between each two samples.

## getMod *Estimate mode location for each probe*

### Description

Estimate mode location for each probe

### Usage

```
getMod(x, bw = 0.04, minDens = 0.001, cpu = 1)
```

### Arguments

| | |
|---|---|
| x | Matrix of beta or RAI values. Row names must be supplied. |
| bw | band width. |
| minDens | Minimum density for a valid peak. |
| cpu | Number of CPU. |

### Value

A data frame of mode locations.

getRAI_snp *Get RAI (Ratio of Alternative allele Intensity) for SNP probes*

## Description

Get RAI (Ratio of Alternative allele Intensity) for SNP probes

## Usage

```
getRAI_snp(inData, platform = "EPIC")
```

## Arguments

| | |
|---|---|
| inData | Noob and dye-bias corrected signals produced by using correct_noob_dye. |
| platform | EPIC or 450K. |

## Value

RAI (Ratio of Alternative allele Intensity).

getRelation *Get sample relationships*

## Description

Get sample relationships

## Usage

```
getRelation(phi)
```

## Arguments

| | |
|---|---|
| phi | A vector of kinship coefficient (Phi). |

## Value

A vector of sample relationships.

get_AF *Extract AFs from matching population in the 1000 Genomes Project (1KGP)*

## Description

Extract AFs from matching population in the 1000 Genomes Project (1KGP)

## Usage

```
get_AF(pop = "EAS", type, platform = "EPIC")
```

## Arguments

| | |
|---|---|
| pop | Population to be used to extract AFs. One of EAS, AMR, AFR, EUR, SAS, and ALL. |
| type | One of snp_probe, typeI_probe, and typeII_probe. |
| platform | EPIC or 450K. |

## Value

A vector of AFs

get_GP_bayesian *Infer posterior genotype probabilities based on the Bayesian approach*

## Description

Prior genotype probabilities were inferred from AFs. The AFs can be in population level or individual-specific level. For population level AFs, they can be extracted from the matched population in the 1000 Genomes Project (1KGP). For individual-specific AFs, they can be calculated according to the top four PCs.

## Usage

```
get_GP_bayesian(pD_AA, pD_AB, pD_BB, AF)
```

## Arguments

| | |
|---|---|
| pD_AA | A MxN matrix of AA genotype probabilities. Provide SNPs as rows and samples as columns. |
| pD_AB | A MxN matrix of AB genotype probabilities. Provide SNPs as rows and samples as columns. |
| pD_BB | A MxN matrix of BB genotype probabilities. Provide SNPs as rows and samples as columns. |
| AF | A MxN matrix of AFs. Provide SNPs as rows and samples as columns. |

### Value

A list containing

| | |
|---|---|
| pAA | Posterior genotype probability of AA |
| pAB | Posterior genotype probability of AB |
| pBB | Posterior genotype probability of BB |

---

## get_indAF *Calculate individual-specific AFs*

---

### Description

Calculate individual-specific AFs

### Usage

```
get_indAF(snpvec, refPC, studyPC)
```

### Arguments

| | |
|---|---|
| snpvec | A vector of SNP IDs. |
| refPC | Top PCs in the reference. |
| studyPC | Top PCs in study samples. |

### Value

A matrix of individual-specific AFs.

---

## get_target *Get example IDAT file list*

---

### Description

Get example IDAT file list

### Usage

```
get_target(platform = "EPIC")
```

### Arguments

| | |
|---|---|
| platform | One of "EPIC" and "450K" |

### Value

A data frame of the IDAT file list

## mnfst *EPIC manifest file*

### Description

A dataset containing all EPIC probes information.

### Usage

```
data(mnfst)
```

### Format

A data frame with 866554 rows and 5 columns:

**Name** CpG name
**AddressA_ID** AdressA ID
**AddressB_ID** AdressB ID
**Infinium_Design_Type** Infinium design type
**Color_Channel** Color channel

### Source

https://webdata.illumina.com/downloads/productfiles/methylationEPIC/infinium-methylationepic-v-1
zip

## mnfst_450K *450K manifest file*

### Description

A dataset containing all 450K probes information.

### Usage

```
data(mnfst_450K)
```

### Format

A data frame with 486428 rows and 5 columns:

**Name** CpG name
**AddressA_ID** AdressA ID
**AddressB_ID** AdressB ID
**Infinium_Design_Type** Infinium design type
**Color_Channel** Color channel

### Source

https://webdata.illumina.com/downloads/productfiles/humanmethylation450/humanmethylation450_
15017482_v1-2.csv

## mval2beta *Convert M values to beta values*

### Description

Convert M values to beta values

### Usage

```
mval2beta(mval)
```

### Arguments

| | |
|---|---|
| mval | M value matrix. |

### Value

Beta value matrix.

## normExpSignal *Normal-exponential deconvolution (adapted from* SeSAMe*)*

### Description

Normal-exponential deconvolution (adapted from SeSAMe)

### Usage

```
normExpSignal(mu, sigma, alpha, x)
```

### Arguments

| | |
|---|---|
| mu, sigma | Background signal parameters returned by backgroundCorrectionNoobFit. |
| alpha | Foreground signal parameters returned by backgroundCorrectionNoobFit. |
| x | Foreground signals to be corrected. |

### Value

The conditional expectation of the signal given the observed foreground and background.

## plot_AF — *Plot the distribution of AFs in 1KGP and input data.*

### Description

Plot the distribution of AFs in 1KGP and input data.

### Usage

```
plot_AF(AF_input, AF_1KGP, pop, type)
```

### Arguments

| | |
|---|---|
| AF_input | A vector. |
| AF_1KGP | A vector. |
| pop | Population. One of EAS, AMR, AFR, EUR, SAS, and ALL. |
| type | One of snp_probe, typeI_probe, and typeII_probe. |

## plot_PCA — *To plot the projection of study samples in reference ancestry space*

### Description

To plot the projection of study samples in reference ancestry space

### Usage

```
plot_PCA(refPC, studyPC)
```

### Arguments

| | |
|---|---|
| refPC | Top PCs in the reference |
| studyPC | Top PCs in study samples |

## plot_RAI_distribution — *Plot beta distributions for reference homozygous, heterozygous, and alternative homozygous*

### Description

Plot beta distributions for reference homozygous, heterozygous, and alternative homozygous

### Usage

```
plot_RAI_distribution(genotypes, type)
```

### Arguments

| | |
|---|---|
| genotypes | Genotype calls. |
| type | One of "snp_probe", "typeI_probe", and "typeII_probe". |

## pprocrustes *Projection Procrustes Analysis*

### Description

Adapted from http://csg.sph.umich.edu/chaolong/LASER

### Usage

```
pprocrustes(
  refPC_new,
  refPC,
  MAX_ITER = 10000,
  THRESHOLD = 1e-06,
  PROCRUSTES_SCALE = 0
)
```

### Arguments

| | |
|---|---|
| refPC_new | Top PCs in the combination of reference samples and one study sample |
| refPC | Top PCs in the reference samples |
| MAX_ITER | Maximum iterations for the projection Procrustes analysis |
| THRESHOLD | Convergence criterion for the projection Procrustes analysis |
| PROCRUSTES_SCALE | Fit the scaling parameter to maximize similarity |

### Value

Projection Procrustes Analysis results

## probeInfo_snp *SNP probe information for EPIC*

### Description

A dataset containing SNP probe information. Only autosome probes are included.

### Usage

```
data(probeInfo_snp)
```

### Format

A data frame with 53 rows and 14 columns:

**Chr** Chromosome ID

**Pos** Position

**SNP** SNP ID targeted by the CpG

**RefAllele** Reference allele

**AltAllele** Alternative allele

**CpG** CpG

**Color** Color channel

**Group** Probe types, color channel, and signal corresponds to alternative allele

**ALL_AF** Allele frequency of all population

**EAS_AF** Allele frequency of East Asian

**AMR_AF** Allele frequency of American

**AFR_AF** Allele frequency of African

**EUR_AF** Allele frequency of European

**SAS_AF** Allele frequency of South Asian

### Source

https://webdata.illumina.com/downloads/productfiles/methylationEPIC/infinium-methylationepic-v-1
zip

---

## probeInfo_snp_450K *SNP probe information for 450K*

---

### Description

A dataset containing SNP probe information. Only autosome probes are included.

### Usage

```
data(probeInfo_snp_450K)
```

### Format

A data frame with 57 rows and 14 columns:

**Chr** Chromosome ID

**Pos** Position

**SNP** SNP ID targeted by the CpG

**RefAllele** Reference allele

**AltAllele** Alternative allele

**CpG** CpG

**Color** Color channel

**Group** Probe types, color channel, and signal corresponds to alternative allele
**ALL_AF** Allele frequency of all population
**EAS_AF** Allele frequency of East Asian
**AMR_AF** Allele frequency of American
**AFR_AF** Allele frequency of African
**EUR_AF** Allele frequency of European
**SAS_AF** Allele frequency of South Asian

### Source

https://webdata.illumina.com/downloads/productfiles/methylationEPIC/infinium-methylationepic-v-1
zip

---

## probeInfo_typeI *Type I probe information for EPIC*

---

### Description

A dataset containing Type I probe information.

### Usage

```
data(probeInfo_typeI)
```

### Format

A data frame with 715 rows and 16 columns:

**Chr** Chromosome ID
**Pos** Position
**SNP** SNP ID targeted by the CpG
**RefAllele** Reference allele
**AltAllele** Alternative allele
**CpG** CpG
**Color** Color channel
**Group** NA
**ALL_AF** Allele frequency of all population
**EAS_AF** Allele frequency of East Asian
**AMR_AF** Allele frequency of American
**AFR_AF** Allele frequency of African
**EUR_AF** Allele frequency of European
**SAS_AF** Allele frequency of South Asian
**loc_pass** Passed peak position test or not

### Source

https://webdata.illumina.com/downloads/productfiles/methylationEPIC/infinium-methylationepic-v-1
zip

`probeInfo_typeII` *Type II probe information for EPIC*

## Description

A dataset containing information of Type II probes with SNPs at the extension bases. We only consider the situation that the alternative allele is A/T and the reference allele is C/G.

## Usage

```
data(probeInfo_typeII)
```

## Format

A data frame with 26420 rows and 16 columns:

**Chr** Chromosome ID

**Pos** Position

**SNP** SNP ID targeted by the CpG

**RefAllele** Reference allele

**AltAllele** Alternative allele

**CpG** CpG

**Color** Color channel

**Group** Probe types, color channel, and signal corresponds to alternative allele

**ALL_AF** Allele frequency of all population

**EAS_AF** Allele frequency of East Asian

**AMR_AF** Allele frequency of American

**AFR_AF** Allele frequency of African

**EUR_AF** Allele frequency of European

**SAS_AF** Allele frequency of South Asian

**loc_pass** Passed peak position test or not

## Source

https://webdata.illumina.com/downloads/productfiles/methylationEPIC/infinium-methylationepic-v-1
zip

probeInfo_typeII_450K *Type II probe information for 450K*

## Description

A dataset containing information of Type II probes with SNPs at the extension bases. We only consider the situation that the alternative allele is A/T and the reference allele is C/G.

## Usage

```
data(probeInfo_typeII_450K)
```

## Format

A data frame with 11875 rows and 14 columns:

**Chr** Chromosome ID

**Pos** Position

**SNP** SNP ID targeted by the CpG

**RefAllele** Reference allele

**AltAllele** Alternative allele

**CpG** CpG

**Color** Color channel

**Group** Probe types, color channel, and signal corresponds to alternative allele

**ALL_AF** Allele frequency of all population

**EAS_AF** Allele frequency of East Asian

**AMR_AF** Allele frequency of American

**AFR_AF** Allele frequency of African

**EUR_AF** Allele frequency of European

**SAS_AF** Allele frequency of South Asian

## Source

https://webdata.illumina.com/downloads/productfiles/methylationEPIC/infinium-methylationepic-v-1
zip

`probeInfo_typeI_450K` *Type I probe information for 450K*

## Description

A dataset containing Type I probe information.

## Usage

```
data(probeInfo_typeI_450K)
```

## Format

A data frame with 712 rows and 14 columns:

**Chr** Chromosome ID

**Pos** Position

**SNP** SNP ID targeted by the CpG

**RefAllele** Reference allele

**AltAllele** Alternative allele

**CpG** CpG

**Color** Color channel

**Group** NA

**ALL_AF** Allele frequency of all population

**EAS_AF** Allele frequency of East Asian

**AMR_AF** Allele frequency of American

**AFR_AF** Allele frequency of African

**EUR_AF** Allele frequency of European

**SAS_AF** Allele frequency of South Asian

## Source

https://webdata.illumina.com/downloads/productfiles/methylationEPIC/infinium-methylationepic-v-1
zip

## probelist *Probe list for EPIC*

### Description

A dataset containing the list of 53 SNP probes on autosomes, 715 Type I probes, and 26420 type II probes.

### Usage

```
data(probelist)
```

### Format

A data frame with 27188 rows and 2 columns:

**CpG** CpG list

**Type** Probe types

**A2** Alternative alleles

## probelist_450K *Probe list for 450K*

### Description

A dataset containing the list of 53 SNP probes on autosomes, 712 Type I probes, and 11875 type II probes.

### Usage

```
data(probelist_450K)
```

### Format

A data frame with 12644 rows and 2 columns:

**CpG** CpG list

**Type** Probe types

## procrustes *Standard Procrustes Analysis*

### Description

Adapted from http://csg.sph.umich.edu/chaolong/LASER

### Usage

```
procrustes(refPC_new, refPC, PROCRUSTES_SCALE = 0)
```

### Arguments

| | |
|---|---|
| refPC_new | Top PCs in the combination of reference samples and one study sample |
| refPC | Top PCs in the reference samples |
| PROCRUSTES_SCALE | Fit the scaling parameter to maximize similarity |

### Value

Procrustes Analysis results

## projection *PCA and Procrustes analysis*

### Description

PCA and Procrustes analysis

### Usage

```
projection(studyGeno, plotPCA = TRUE, cpu = 1, platform = "EPIC")
```

### Arguments

| | |
|---|---|
| studyGeno | A matrix of genotypes of study samples. Provide probes as rows and samples as columns. Include all SNP probes, type I probes, and type II probes if available. |
| plotPCA | To plot the projection of study samples in reference ancestry space. |
| cpu | Number of CPU. |
| platform | EPIC or 450K. |

### Value

A list containing

| | |
|---|---|
| refPC | Top PCs in the reference |
| studyPC | Top PCs in study samples |

recal_Geno *Recalibrate genotypes for samples of mixed population*

## Description

Recalibrate genotypes for samples of mixed population

## Usage

```
recal_Geno(
  genotypes,
  type,
  indAF,
  platform = "EPIC",
  GP_cutoff = 0.9,
  outlier_cutoff = "max",
  missing_cutoff = 0.1,
  R2_cutoff_up = 1.1,
  R2_cutoff_down = 0.75,
  MAF_cutoff = 0.01,
  HWE_cutoff = 1e-06
)
```

## Arguments

| | |
|---|---|
| genotypes | A list returned by either callGeno_snp, callGeno_typeI, or callGeno_typeII function. |
| type | One of snp_probe, typeI_probe, and typeII_probe. |
| indAF | A matrix of individual-specific AFs. Provide SNPs as rows and samples as columns. |
| platform | EPIC or 450K. |
| GP_cutoff | When calculating missing rate, genotypes with the highest genotype probability < GP_cutoff will be treated as missing. |
| outlier_cutoff | "max" or a number ranging from 0 to 1. If outlier_cutoff="max", genotypes with outlier probability larger than all of the three genotype probabilities will be set as missing. If outlier_cutoff is a number, genotypes with outlier probability > outlier_cutoff will be set as missing. |
| missing_cutoff | Missing rate cutoff to filter variants. Note that for VCF output, variants with missing rate above the cutoff will be marked in the FILTER column. For the returned dosage matrix, variants with missing rate above the cutoff will be removed. |
| R2_cutoff_up, R2_cutoff_down | R-square cutoffs to filter variants (Variants with R-square > R2_cutoff_up or < R2_cutoff_down should be removed). Note that for VCF output, variants with R-square outside this range will be marked in the FILTER column. For the returned dosage matrix, variants with R-square outside this range will be removed. |
| MAF_cutoff | A MAF cutoff to filter variants. Note that for VCF output, variants with MAF below the cutoff will be marked in the FILTER column. For the returned dosage matrix, variants with MAF below the cutoff will be removed. |

HWE_cutoff — HWE p value cutoff to filter variants. Note that for VCF output, variants with HWE p value below the cutoff will be marked in the FILTER column. For the returned dosage matrix, variants with HWE p value below the cutoff will be removed.

## Value

A list of recalibrated genotypes containing

dosage — A matrix of genotype calls. Variants with R2s, HWE p values, MAFs, or missing rates beyond the cutoffs are removed.

genotypes — A list containing RAI, shapes of the mixed beta distributions, prior probabilities that the RAI values belong to one of the three genotypes, proportion of RAI values being outlier (U), and genotype probability (GP)

indAF — A matrix of individual-specific AFs.

---

# refGeno_1KGP3 *Reference genotypes in the 1000 Genomes Project*

---

## Description

A matrix of reference genotypes in the 1000 Genomes Project (1KGP). It contains 2504 samples and 28,619 SNPs overlapping the methylation probes.

## Usage

```
data(refGeno_1KGP3)
```

## Format

A matrix with 28,619 rows and 2504 columns:

**Row** SNPs overlapping the methylation probes

**Column** Samples

---

# refGeno_1KGP3_SNP_failQC *SNPs in 1KGP with HWE<1e-20 or F_MISSING>=0.05*

---

## Description

SNPs to be removed in TRACE PCA

## Usage

```
data(refGeno_1KGP3_SNP_failQC)
```

## Format

A vector with 491 items:

**Value** SNP rs ID

## sam2pop *Population information for the 1KGP samples*

### Description

A vector of population information for the 2504 samples in 1KGP.

### Usage

```
data(sam2pop)
```

### Format

A vector with 2504 items:

**Name** Sample ID

**Value** Population

## TRACE *TRACE: fasT and Robust Ancestry Coordinate Estimation*

### Description

Adapted from http://csg.sph.umich.edu/chaolong/LASER

### Usage

```
TRACE(
  refGeno,
  studyGeno,
  MIN_LOCI = 100,
  DIM = 4,
  DIM_HIGH = 20,
  MAX_ITER = 10000,
  THRESHOLD = 1e-06,
  cpu = 1
)
```

### Arguments

| | |
|---|---|
| refGeno | A matrix of genotypes of reference individuals. Provide probes as rows and samples as columns. |
| studyGeno | A matrix of genotypes of study samples. Provide probes as rows and samples as columns. |
| MIN_LOCI | Minimum number of non-missing loci required |
| DIM | Number of PCs in the reference to match |
| DIM_HIGH | Number of PCs for sample-specific PCA |
| MAX_ITER | Maximum iterations for the projection Procrustes analysis |
| THRESHOLD | Convergence criterion for the projection Procrustes analysis |
| cpu | Number of CPU. |

## Value

A list containing

| | |
|---|---|
| `refPC` | Top PCs in the reference |
| `studyPC` | Top PCs in study samples |

# Index