Package 'bisect'

April 16, 2018

Title Estimating Cell Type Composition from Methylation Sequencing Data

Version 0.9.0

Maintainer Eyal Fisher <ef388@cam.ac.uk>

Author Eyal Fisher [aut, cre]

Description An implementation of Bisect, a method for inferring cell type composition of samples based on methylation sequencing data (Whole Genome Bisulfite Sequencing and Reduced Representation Sequencing). The method is specifically tailored for sequencing data, and therefore works better than methods developed for methylation arrays. It contains a supervised mode that requires a reference (the methylation probabilities in the pure cell types), and a semi-supervised mode, that requires cell counts for a subset of the samples, but does not require a reference.

URL https://github.com/EyalFisher/BiSect

Depends R (>= 3.4.0)

License GPL-3

Encoding UTF-8

LazyData true

RoxygenNote 6.0.1

Suggests dplyr, knitr, rmarkdown, tidyr, sirt, ggplot2

VignetteBuilder knitr

NeedsCompilation no

Repository CRAN

Date/Publication 2018-04-16 16:34:49 UTC

R topics documented:

`alpha_blood` *Recommended values for a Dirichlet prior on cell composition in blood samples. Estimated by fitting Dirichlet distribution to cell counts data.*

Description

Recommended values for a Dirichlet prior on cell composition in blood samples. Estimated by fitting Dirichlet distribution to cell counts data.

Usage

```
alpha_blood
```

Format

A vector of length 6

`baseline_GSE40279` *Cell composition for 650 individuals from GSE40279. Estimated by running an array method on the original (array) data.*

Description

Cell composition for 650 individuals from GSE40279. Estimated by running an array method on the original (array) data.

Usage

```
baseline_GSE40279
```

Format

A data frame with 650 rows and 6 variables

Source

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE40279

`bisect_semi_suprevised` *Add together two numbers.*

Description

Add together two numbers.

Usage

```
bisect_semi_suprevised(methylation_unkown_samples, total_reads_unknown_samples,
  methylation_known_samples, total_reads_known_samples,
  cell_composition_known_samples, alpha = NA, iterations = 200)
```

Arguments

`methylation_unkown_samples`	a matrix of individuals (rows) on sites (columns), containing the number of methylated reads for each site, in each individual for the samples with unknown cell composition.
`total_reads_unknown_samples`	a matrix of individuals (rows) on sites (columns), containing the total number of reads for each site, in each individual for the samples with unknown cell composition.
`methylation_known_samples`	a matrix of individuals (rows) on sites (columns), containing the number of methylated reads for each site, in each individual for the samples with known cell composition.
`total_reads_known_samples`	a matrix of individuals (rows) on sites (columns), containing the total number of reads for each site, in each individual for the samples with known cell composition.
`cell_composition_known_samples`	a matrix of individuals (rows) on cell types (columns), containing the proportion of each cell type, in each known sample.
`alpha`	a vector containing the hyper-parameters for the dirichelt prior. One value for each cell type. If NA, it is initiallized to 1/(number of cell types).
`iterations`	the number of iterations to use in the EM algorithm.

Value

A list containing P, a matrix of estimated cell proportions for the unknown samples, and Pi, an estimated reference (the probability of methylation in each cell type).

Examples

```
## Randomly choose samples to be used as known
n_known_samples <- 50
known_samples_indices <- sample.int(nrow(baseline_GSE40279), size = n_known_samples)
known_samples <- as.matrix(baseline_GSE40279[known_samples_indices, ])

## Fit a dirichlet distribution to known samples to use as prior
fit_dirichlet <- sirt::dirichlet.mle(as.matrix(known_samples))
alpha <- fit_dirichlet$alpha

## Prepare the 4 needed matrices
methylation_known <- methylation_GSE40279[known_samples_indices, ]
methylation_unknown <-methylation_GSE40279[-known_samples_indices, ]
total_known <- total_reads_GSE40279[known_samples_indices, ]
total_unknown <- total_reads_GSE40279[-known_samples_indices, ]

## Run Bisect. You should use around 200 iterations. I choose than to accelarate the example.
results <- bisect_semi_suprevised(methylation_unknown, total_unknown,
                                  methylation_known, total_known,
                                  known_samples, alpha, iterations = 10)
```

`bisect_supervised` *Add together two numbers.*

Description

Add together two numbers.

Usage

```
bisect_supervised(methylation, total_reads, reference, alpha = NA,
  iterations = 200)
```

Arguments

`methylation`	a matrix of individuals (rows) on sites (columns), containing the number of methylated reads for each site, in each individual.
`total_reads`	a matrix of individuals (rows) on sites (columns), containing the total number of reads for each site, in each individual.
`reference`	a matrix of sites (rows) on cell types (columns), containing the probability for methylation in each site, in each cell type.
`alpha`	a vector containing the hyper-parameters for the dirichelt prior. One value for each cell type. If NA, it is initiallized to 1/(number of cell types).
`iterations`	the number of iterations to use in the EM algorithm.

Value

A matrix of individuals (rows) on cell types (columns) containing the estimated proportion of each cell type, in each individual.

Examples

```
## Prepare the methylation and total reads matrices
methylation <- as.matrix(methylation_GSE40279)
total_reads <- as.matrix(total_reads_GSE40279)
## Remove the IDs column from the reference
Pi <- as.matrix(reference_blood[,-1])

## Run Bisect. You should use around 200 iterations. I choose than to accelarate the example.
results <- bisect_supervised(methylation, total_reads, Pi, alpha_blood, iterations = 10)
```

`methylation_GSE40279` *Simulated amount of methylated reads for 650 individuals from GSE40279. The data was sub-sampled to simulate a 30X coverage. Only 241 sites that are known to differ substaintally between cell types are recorded.*

Description

Simulated amount of methylated reads for 650 individuals from GSE40279. The data was sub-sampled to simulate a 30X coverage. Only 241 sites that are known to differ substaintally between cell types are recorded.

Usage

```
methylation_GSE40279
```

Format

A data frame with 650 rows and 241 variables

Source

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE40279

`reference_blood`	*A reference containing methylation proportions for pure cell types in the 241 chosen sites.*

Description

A reference containing methylation proportions for pure cell types in the 241 chosen sites.

Usage

```
reference_blood
```

Format

A data frame with 241 rows and 7 variables

Source

https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-0943-7

`total_reads_GSE40279`	*Simulated amount of total reads reads for 650 individuals from GSE40279. The data was sub-sampled to simulate a 30X coverage. Only 241 sites that are known to differ substaintally between cell types are recorded.*

Description

Simulated amount of total reads reads for 650 individuals from GSE40279. The data was sub-sampled to simulate a 30X coverage. Only 241 sites that are known to differ substaintally between cell types are recorded.

Usage

```
total_reads_GSE40279
```

Format

A data frame with 650 rows and 241 variables

Source

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE40279

Index