Package ‘heritEWAS’

October 13, 2022

Title Identify Heritable Methylation Marks

Version 0.2.0

Description A novel statistical method based on expectation maximisation (EM) algorithm and genetic segregation analysis to identify heritable DNA methylation marks. Details about the method can be found in Joo et al. (2018) <doi:10.1038/s41467-018-03058-6>.

Depends R (>= 3.5.0)

License GPL-3

Encoding UTF-8

LazyData true

RoxygenNote 7.1.0

Suggests knitr, rmarkdown

VignetteBuilder knitr

NeedsCompilation no

Author James Dowty [aut],
Kevin Wong [aut, cre]

Maintainer Kevin Wong <wongck.kevin@gmail.com>

Repository CRAN

Date/Publication 2020-06-04 04:50:07 UTC

R topics documented:

`carrier_probabilities` *Calculate carrier probabilities for the most heritable methylation sites*

Description

For each person in `dat` and each methylation site in `top_probes`, this function calculates the probability that the person carries a rare mutation at a hypothetical genetic locus that affects methylation at the methylation site.

Usage

```
carrier_probabilities(dat, M_values, top_probes, ncores = 1)
```

Arguments

`dat` A data frame with rows corresponding to people and columns corresponding to (at least) the following variables, which will be coerced to `character` type:

- `family` (family ID), an identifier for each person's family, constant within families
- `indiv` (individual ID), an identifier for each person, with no duplicates across the dataset
- `mother` (mother ID), the individual ID of each person's mother, or missing (`NA`) for founders
- `father` (father ID), the individual ID of each person's father, or missing (`NA`) for founders
- `typed` (epi-genotyped), equal to `1` for people with methylation data and `0` for all others

`M_values` A matrix of M-values, with rows corresponding to methylation sites and columns corresponding to people.

`top_probes` A data frame, usually the output of `ML_estimates` restricted to the most heritable methylation sites (those with the highest values of Δl). See `ML_estimates` and the example below for more details.

`ncores` The number of cores to be used, with `ncores = 1` (the default) corresponding to non-parallel computing. When `ncores > 1`, the `parallel` package is used to parallelize the calculation.

Value

A data frame containing the carrier probabilities described in (Joo et al., 2018), with rows of the data frame corresponding to the people in `dat` and columns corresponding to the methylation sites in `top_probes`. This calculation is based on the Mendelian model of (Joo et al., 2018) with parameter values taken from `top_probes`.

References

Joo JE, Dowty JG, Milne RL, Wong EM, Dugué PA, English D, Hopper JL, Goldgar DE, Giles GG, Southey MC, kConFab. Heritable DNA methylation marks associated with susceptibility to breast cancer. Nat Commun. 2018 Feb 28;9(1):867. https://doi.org/10.1038/s41467-018-03058-6

Examples

```
str(ped)
str(M_values)

# Calculate genotype probabilities
typed_genos <- genotype_combinations(ped)
str(typed_genos)


# Compute Delta l
MLEs <- ML_estimates(typed_genos, M_values, ncores = 4)

# Select top probes
top_probes <- MLEs[MLEs$delta.l > 10, ]

# Calculate carrier probabilities
CP <- carrier_probabilities(ped, M_values, top_probes, ncores = 2)
str(CP)
```

genotype_combinations *Calculate genotype probabilities*

Description

This function computes the joint probabilities of the possible genotypes of selected family members within each family, as an intermediate calculation for use in ML_estimates.

Usage

```
genotype_combinations(dat, ncores = 1, verbose = TRUE)
```

Arguments

dat A data frame with rows corresponding to people and columns corresponding to (at least) the following variables, which will be coerced to character type:

- family (family ID), an identifier for each person's family, constant within families
- indiv (individual ID), an identifier for each person, with no duplicates across the dataset

- `mother` (mother ID), the individual ID of each person's mother, or missing (`NA`) for founders
- `father` (father ID), the individual ID of each person's father, or missing (`NA`) for founders
- `typed` (epi-genotyped), equal to `1` for people with methylation data and `0` for all others

`ncores` The number of cores to be used, with `ncores = 1` (the default) corresponding to non-parallel computing. When `ncores > 1`, the `parallel` package is used to parallelize the calculation of the joint probabilities of the possible genotype combinations.

`verbose` `FALSE` if user wants to suppress messages. Default is `TRUE`.

Details

Currently, there is a maximum of 20 people per family with `typed = 1`, due to the need to store all genotype combinations for the typed people. It is possible to break families with more than 20 typed people into smaller families, though this is not ideal.

Each family within `dat` should be a complete pedigree, meaning that each (non-missing) mother or father ID should correspond to a row, and each person should either have both parent IDs missing (if a founder) or non-missing (if a non-founder). No family should contain a pedigree loop, such as those caused by inbreeding or by two sisters having children with two brothers from an unrelated family.

Value

A named list, with names equal to the different family IDs and with each element of the list being a data frame specifying the possible genotypes of selected family members (those with `dat$typed = 1`) within each family, and the joint probability of each genotype combination. For each individual, the possible genotypes are `0`, corresponding to the wildtype, and `1`, for carriers of a rare genetic variant at an autosomal locus. The calculation makes the same assumptions as in (Joo et al., 2018), including that the genetic variant is so rare that at most one founder is a carrier.

References

Joo JE, Dowty JG, Milne RL, Wong EM, Dugué PA, English D, Hopper JL, Goldgar DE, Giles GG, Southey MC, kConFab. Heritable DNA methylation marks associated with susceptibility to breast cancer. Nat Commun. 2018 Feb 28;9(1):867. `https://doi.org/10.1038/s41467-018-03058-6`

Examples

```
# Load family data
data(ped)

# Calculate genotype probabilites
typed_genos <- genotype_combinations(ped)
str(typed_genos)
```

ML_estimates *Compute maximum likelihood estimates and Δl*

Description

For each methylation site, this function computes certain maximum likelihood estimates and a measure of heritability called Δl (with higher values corresponding to more highly heritable methylation sites), as described briefly below and fully in (Joo et al., 2018).

Usage

```
ML_estimates(typed_genos, M_values, sort = TRUE, na_omit = TRUE, ncores = 1)
```

Arguments

`typed_genos`	A named list, usually generated by `genotype_combinations`. Each element of the list is a data frame specifying all possible joint genotypes of selected family members within each family, and the joint probability of each genotype combination.
`M_values`	A matrix of M-values,with rows corresponding to the methylation sites and columns corresponding to people. The column names should correspond to the column names appearing in `typed_genos`.
`sort`	Re-order the methylation sites to have decreasing values of `delta.l` if `TRUE` (the default), or leave the sites in the original order if `FALSE`
`na_omit`	Remove any methylation sites with missing values (`NA`) of `delta.l` if `TRUE` (the default), or return the results for all sites if `FALSE`. Usually, missing values of `delta.l` are due to well-known singularities in the likelihood of the Gaussian mixtures model.
`ncores`	The number of cores to be used, with `ncores` = `1` (the default) corresponding to non-parallel computing. When `ncores` > `1`, the `parallel` package is used to parallelize the calculation, by dividing the methylation sites between the cores.

Value

A data frame with 15 columns. In the column names, the suffixes `.mendel` and `.mix` refer the Mendelian and mixture models of (Joo et al., 2018). Briefly, the mixture model is the standard Gaussian mixture model with two groups (group `0` and group `1`), so group memberships are independent and the M-values of each group are normally distributed. The Mendelian model is the same except that group memberships are dependent within families, and are modelled as the carrier status of a rare, autosomal genetic variant. In the column names, the prefixes `mu` and `sd` refer to the maximum likelihood estimates of the mean and standard deviation of each group's normal distribution, and the suffix `ll` refers to each model's maximised log-likelihood (i.e., the log-likelihood function evaluated at the maximum likelihood estimates). The suffix `.null` refers to the null model that is nested inside both the Mendelian and mixture models, in which the means and standard deviations for the two groups are equal (i.e., `mu0` = `mu1` and `sd0` = `sd1`). The column `delta.l` gives the difference between `ll.mendel` and `ll.mix`, and is the measure of heritability (Δl) that was introduced in (Joo et al., 2018).

References

Joo JE, Dowty JG, Milne RL, Wong EM, Dugué PA, English D, Hopper JL, Goldgar DE, Giles GG, Southey MC, kConFab. Heritable DNA methylation marks associated with susceptibility to breast cancer. Nat Commun. 2018 Feb 28;9(1):867. https://doi.org/10.1038/s41467-018-03058-6

Examples

```
# Example data
str(ped)
str(M_values)

# Calculate genotype probabilities
typed_genos <- genotype_combinations(ped)
str(typed_genos)


# Compute Delta l
MLEs <- ML_estimates(typed_genos, M_values, ncores = 4)
str(MLEs)
```

M_values *M-values for 1000 DNA methylation sites*

Description

A dataset containing simulated M-values for 1000 DNA methylation sites on selected people from the ped dataset (those with typed = 1).

Usage

```
M_values
```

Format

A data frame with 1000 rows (corresponding to DNA methyaltion sites) and 128 columns (corresponding to persons in ped with typed = 1).

Source

Simulated

ped *Simulated data on 20 families.*

Description

A dataset giving the relationship structure of 20 families and phenotypic data on the family members

Usage

```
ped
```

Format

A data frame with 288 rows (corresponding to persons) and 8 variables:

family an identifier for the person's family

indiv an identifier (ID) for the person

mother the individual ID of the person's mother

father the individual ID of the person's father

sex the person's sex (M = male, F = female)

aff the person's affected status (1 = case, 0 = control)

age the person's age, in years

typed a flag indicating if methylation data is available (1 = available, 0 = unavailable)

Source

Simulated

Index