Package 'SNPMClust'

July 2, 2014

Type Package

Title A bivariate Gaussian genotype clustering and calling algorithm for Illumina microarrays

Version 1.0

Date 2013-03-29

Author Stephen W. Erickson, PhD, with contributions from Joshua Callaway, MPH

Maintainer Stephen W. Erickson, PhD <SErickson@uams.edu>

Description A bivariate Gaussian genotype clustering and calling
algorithm for Illumina microarrays, building on the package 'mclust'. Pronounced snip-em-clust.

Imports MASS, mclust

License GPL (>= 2)

NeedsCompilation no

Repository CRAN

Date/Publication 2013-05-02 08:11:50

R topics documented:

generatepriors *A function to create pseudodata for* snpmclust *algorithm*

Description

generatepriors creates bivariate normal pseudodata for the homozygous and heterozygous minor genotypes.

Usage

```
generatepriors(x, y, calls, priorpoints = length(x) * 0.2,
               xm1 = NA, xm2 = NA, xm3 = NA,
               ym1 = NA, ym2 = NA, ym3 = NA, ranseed = ranseed)
```

Arguments

x	x-vector of signal intensity data in transformed scale.
y	y-vector of signal intensity data in transformed scale.
calls	A priori genotype calls for intensity data.
priorpoints	The number of observations of pseudodata to be generated for the heterozygous and homozygous minor genotypes.
xm1, xm2, xm3, ym1, ym2, ym3	Pseudodata cluster means can be user-specified through these parameters. The ordered pair (xm1,ym1) gives the cluster mean for genotype AA; similarly for (xm2,ym2), (xm3,ym3) and AB, BB, respectively. Default values are NA, in which case cluster means are estimated from the data, conditional on the a priori genotypes passed via calls.
ranseed	Random seed for generation of pseudodata. The default is 1969.

Value

A priorpoints-by-2 matrix.

Author(s)

Stephen W. Erickson, PhD <SErickson@uams.edu> with Joshua Callaway, MPH <jcallaw3@utk.edu>

prepdata *Function to convert data epxorted from GenomeStudio into form usable by the function* snpmclust.

Description

prepdata converts and transforms data from GenomeStudio output into form that can be handled by the snpmclust() function.

Usage

```
prepdata(rawdata)
```

Arguments

rawdata	Data frame taken from an import of GenomeStudio full data table.

Details

prepdata expects a data frame that includes columns from an import of a GenomeStudio full data table. These columns include Name (the column of SNP rs-numbers) and the subcolumns Theta, R, GType, Score, X, Y, X.Raw, Y.Raw. Sample ID numbersare taken from the subcolumn prefixes. The data transformations in prepdata() are an integral part of the SNPMClust methodology.

Value

A list with the following components:

SNP	Character vector of SNP IDs ("rs numbers").
SampleID	Character vector of sample ID numbers, taken from subcolumn prefixes.
P	Length of SNP.
N	Length of SampleID.
Theta	Numeric PxN matrix of Theta subcolumns.
R	Numeric PxN matrix of R subcolumns.
GType	CharacterPxN matrix of GType subcolumns.
Score	Numeric PxN matrix of Score subcolumns.
X.Raw	Numeric PxN matrix of X.Raw subcolumns.
Y.Raw	Numeric PxN matrix of Y.Raw subcolumns.
X	Numeric PxN matrix of X subcolumns.
Y	Numeric PxN matrix of Y subcolumns.
logratio	Numeric PxN matrix of normalized signal intensity log-ratios.
R.trans	Numeric PxN matrix of Box-Cox-transformed signal magnitudes.

Author(s)

Stephen W. Erickson, PhD <SErickson@uams.edu> with Joshua Callaway, MPH <jcallaw3@utk.edu>

Examples

```
data(testset)
tmpfile = prepdata(testset)
```

---

snpmclust *Bivariate Gaussian genotype clustering and calling algorithm for Illumina microarrays.*

Description

snpmclust17 is the main function that develops the genotype clustering.

Usage

```
snpmclust(indata, p = 1, priorfrac = 0.2, uncertcutoff = 0.01, showplots = FALSE,
          xm1 = NA, xm2 = NA, xm3 = NA, ym1 = NA, ym2 = NA, ym3 = NA,
          ranseed = 1969, R.lowcutoff = 0.05)
```

Arguments

indata	A list containing input data on one or all SNPs, and would normally be produced by the function prepdata(). Details on the different components of indata are given below.
p	A positive integer specifying which SNP to cluster. The default is 1.
priorfrac	A non-negative scalar specifying the number of observations, as a fraction of the number of samples N, of pseudodata to be appended to the heterozygous and homozygous minor genotypes. The default is 0.2.
uncertcutoff	Genotype calls with uncertainty greater than uncertcutoff are set to "NC" (no call). The default is 0.01.
showplots	A logical value. If TRUE, the function will produce a series of plots. The default is FALSE.
xm1, xm2, xm3, ym1, ym2, ym3	Pseudodata cluster means can be user-specified through these parameters. The ordered pair (xm1,ym1) gives the cluster mean for genotype AA; similarly for (xm2,ym2), (xm3,ym3) and AB, BB, respectively. Default values are NA, in which case cluster means are estimated from the data, conditional on the a priori genotypes produced by GenomeStudio.
ranseed	Random seed for generation of pseudodata. The default is 1969.
R.lowcutoff	Genotypes for which R is less than R.lowcutoff are set to "NC" (no call). The default is 0.05.

Value

A data frame with N rows and three columns.

SNP	Locus (rs-number).
MClustCalls	Genotype call. Either "AA", "AB", "BB", or "NC" (no call).
Uncertainty	Uncertainty score for the corresponding genotype call.

Author(s)

Stephen W. Erickson, PhD <SErickson@uams.edu> with Joshua Callaway, MPH <jcallaw3@utk.edu>

Examples

```
data(testset)
tmpfile = prepdata(testset)
snpmclust(tmpfile, p=1, showplots=TRUE)
```

---

testset *De-identified test set*

Description

De-identified and scrambled test set to serve as the rawdata argument for prepdata. Five SNPs and 200 individuals.

Usage

```
data(testset)
```

Format

A data frame with 5 observations and 1801 variables.

Examples

```
data(testset)
```

Index