Package ‘SNPMClust’

July 27, 2016

Type Package

Title Bivariate Gaussian Genotype Clustering and Calling for Illumina Microarrays

Version 1.3

Date 2016-07-26

Author Stephen W. Erickson and Joshua C. Callaway

Maintainer Joshua C. Callaway <joshcllw@gmail.com>

Description Bivariate Gaussian genotype clustering and calling for Illumina microarrays, building on the package ‘mclust’. Pronounced snip-em-clust.

Depends R (>= 3.1.0), MASS, mclust

License GPL (>= 2)

RoxygenNote 5.0.1

NeedsCompilation no

Repository CRAN

Date/Publication 2016-07-27 03:37:28

R topics documented:

generatepriors ................................................................. 2
prepdata ......................................................................... 3
snpmclust ................................................................. 4
testset ................................................................. 5

Index 7
generatepriors  

*Generate pseudodata for snpMClust*

**Description**

Generates bivariate normal pseudodata for the homozygous and heterozygous minor genotypes.

**Usage**

```r
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 <serickson@rti.org> with Joshua C. Callaway <joshc1lw@gmail.com>

**References**

Description

Converts and transforms data from GenomeStudio output into form that can be handled by the function snpmclust.

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 IDs are 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 <serickson@rti.org> with Joshua C. Callaway <joshcllw@gmail.com>

References


Examples

data(testset)
tmppfile = prepdata(testset)

snpmclust

Genotype clustering and calling

Description

Genotype clustering and calling for Illumina microarrays.

Usage

snpmclust(indata, p = 1L, priorfrac = 0.2L, uncertcutoff = 0.01L, q_cutoff = 0L, showplots = FALSEL, xm1 = NA, xm2 = NA, xm3 = NA, ym1 = NA, ym2 = NA, ym3 = NA, ranseed = 1969, rNlowcutoff = PNPUI)

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 can be seen in help(prepdata).

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.

cutoff Uncertainty scores lower than the q_cutoff’th quantile are reset to that value. When used with R.lowcutoff, this is equivalent to requiring a SNP-specific call rate of q_cutoff or higher.

showplots A logical value. If TRUE, the function will produce a series of plots. The default is FALSE.
Pseudodata cluster means can be user-specified through these parameters. The ordered pair \((xm_1, ym_1)\) gives the cluster mean for genotype AA; similarly for \((xm_2, ym_2)\), \((xm_3, ym_3)\) 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 list with the following components:

- **calls**: A data frame with N rows and 4 columns, namely, SNP, SampleID, MClustCalls (the genotype call), and Uncertainty.
- **snp**: The SNP name (i.e. rs-number).
- **callrate**: Call rate for the SNP.
- **priorfrac**: Value of argument in function call.
- **uncertcutoff**: Value of argument in function call.
- **qcutoff**: Value of argument in function call.

### Author(s)

Stephen W. Erickson <serickson@rti.org> with Joshua C. Callaway <joshc11w@gmail.com>

### References


### Examples

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

### 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.

References


Examples

data(testset)
Index

*Topic SNPs
   snpmclust, 4
*Topic cluster
   snpmclust, 4
*Topic convert
   prepdata, 3
*Topic priors
   generatepriors, 2
*Topic pseudodata
   generatepriors, 2
*Topic rawdata
   prepdata, 3
*Topic testdata
   testset, 5

generatepriors, 2
prepdata, 3
snpmclust, 4
testset, 5