Package ‘extraBinomial’

February 19, 2015

Type Package

Title Extra-binomial approach for pooled sequencing data

Version 2.1

Date 2012-07-09

Author Xin Yang, Chris Wallace

Maintainer Xin Yang <xin.yang@cimr.cam.ac.uk>

Description This package tests for differences in minor allele frequency between groups and is based on an extra-binomial variation model for pooled sequencing data.

License GPL-3

Repository CRAN

Date/Publication 2012-07-09 15:39:12

NeedsCompilation no

R topics documented:

`extraBinomial-package` *Extra-binomial approach for pooled sequencing data*

Description

This package tests for differences in minor allele frequency between groups and is based on extra-binomial variation model for pooled sequencing data.

Details

Package:	extraBinomial
Type:	Package
Version:	2.1
Date:	2012-07-09
License:	GPL-3

To use the function exbio, simply define two matrices R, R.alt with the same dimensions (rows index SNPs and columns index pools), a vector cc indicating the case and control status, number of chromosomes (n) and then do: exbio(R, R.alt, cc, n) to yield the estimated allele frequencies and p-value based on extra-binomial model.

Author(s)

Xin Yang, Chris Wallace

Maintainer: Xin Yang <xin.yang@cimr.cam.ac.uk>

References

Yang et al. "Extra-binomial variation approach for analysis of pooled DNA sequencing data", under review.

exbio *Extra-binomial approach for pooled sequencing data*

Description

This funtion tests for differences in minor allele frequency between groups and is based on extra-binomial variation model for pooled sequencing data.

Usage

```
exbio(R, R.alt, cc, n, tol = 0.001, a.start = 1, b.start = 1, max.it = 1000, digits = NULL, model.maf =
```

Arguments

`R`	A matrix with rows indexed by SNPs and columns by pools. The entries are counts of allele 1.
`R.alt`	A similarly formatted matrix containing the counts of allele 2.
`cc`	A case/control indicator vector with length = number of pools containing 0s (control pool) and 1s (case pool).
`n`	Number of chromosomes (twice the number of subjects) in each pooled sample.
`tol`	Maximum difference between coefficient values in successive glm before we can stop, the default=0.001.

`a.start` An intial value for the parameter a in linear regression, the default=1.

`b.start` An intial value for the parameter b in linear regression, the default=1.

`max.it` Maximum iterations, the default=1000.

`digits` How many significant digits are to be used for allele frequency and p-value. The default, 'NULL', uses 'getOption(digits)'.

`model.maf` A logical value indicating whether to allow the modelled error structure to depend on allele frequency (the default) or just read depth. The default=TRUE.

Details

R and R.alt contain the read counts for the major allele and the alternative allele respectively and are required to have the same dimension.

The extra-binomial model defined: E(R/N)=p, Var(R/N)=p(1-p)(a/n+b/N) when N=R+R.alt

We denote: W=1/(a/n+b/N), which may be interpreted as the adjusted depth of pool j for SNP i. Given the expected quantities: E(r2)=1/W=a/n+b/N, the parameters a and b can be estimated by linear regression of r2 on 1/N, giving a/n as the intercept and b as the slope. If model.maf=TRUE, W=1/(a/n+b/N+b2*p+b3*p^2) and two additional parameters (b2 and b3) are estimated. This regression is carried out using generalized linear model (GLM) by first adopting Gaussian errors to estimate a relatively good start value of a and b, and then using these start values to do GLM with gamma errors and identity link because both a and b are positive.

Since the estimated allele frequency p depends on a and b, the calculations are carried out iteratively.

A chi-square test is performed on a 2*2 table using the weighted allele counts to calculate the p-value.

Value

A list containing the following components:

`result` a data.frame with three columns: the first shows the minor allele frequency of controls; the second shows the minor allele freqeuncy of cases; the third shows the p-value. Each row stands for a SNP.

`parameters` a character vector indicating the values of the parameters a and b (and b2, b3 if model.maf=TRUE) in the linear regression and and the times of iteration.

Author(s)

Xin Yang, Chris Wallace

References

Yang et al. "Extra-binomial variation approach for analysis of pooled DNA sequencing data", under review.

Examples

```
R<-matrix(c(1409,1530,1490,1630,924,998,1000,1012),nrow=2,ncol=4,byrow=TRUE)
R.alt<-matrix(c(170,210,192,209,13,14,30,38),nrow=2,ncol=4,byrow=TRUE)
cc<-c(0,0,1,1)
n=96
exbio(R, R.alt, cc, n, max.it = 100, digits=3)
##=> p.value = 9.91e-01 for SNP1 and 4.01e-11 for SNP2,
##so association for SNP2 is established, but not for SNP1.
```

Index