Package 'RRate'

June 22, 2016

Type Package

Title RRate: Estimating Repliation Rate for GWAS

Version 1.0

Date 2016-06-13

Author Wei Jiang, Jing-Hao Xue and Weichuan Yu

Maintainer Wei Jiang <wjiangaa@connect.ust.hk>

Description Replication Rate (RR) is the probability of replicating a statistically significant association in GWAS. This R-package provide the estimation method for replication rate which makes use of the summary statistics from the primary study. We can use the estimated RR to determine the sample size of the replication study, and to check the consistency between the results of the primary study and those of the replication study.

License GPL-3

Depends splines

R topics documented:

`RRate-package` *RRate: Estimating Repliation Rate for GWAS*

Description

Replication Rate (RR) is the probability of replicating a statistically significant association in GWAS. This R-package provide the estimation method for replication rate which makes use of the summary statistics from the primary study. We can use the estimated RR to determine the sample size of the replication study, and to check the consistency between the results of the primary study and those of the replication study.

Details

Package:	RRate
Type:	Package
Version:	1.0
Date:	2016-06-13
License:	GPL-3

The goal of genome-wide association studies (GWAS) is to discover genetic variants associated with diseases/traits. Replication is a common validation method in GWAS. We regard an association as true finding when it shows significance in both the primary and replication studies. A worth pondering question is: what is the probability of a primary association (i.e. statistically significant association in the primary study) being validated in the replication study?

We refer the Bayesian replication probability as the replication rate (RR). Here we implement the estimation method for RR which makes use of the summary statistics from the primary study. We can use the estimated RR to determine the sample size of the replication study, and to check the consistency between the results of the primary study and those of the replication study.

The principal component of RRate package is `repRateEst`. Also we implement sample size determination method (`repSampleSizeRR` and `repSampleSizeRR2`) and consistency checking method (Hosmer-Lemeshow test, `HLtest`).

1. To estimate the RR, we need obtain the summary statistics of each genotyped SNPs in the primary study. We have put a example summary statistics (smryStats1) in the package. You can use `data(smryStats1)` to load the example data. You can also obtain the ground-truth parameters (allele frequencies, odds ratios) of the example data using `data(param)`. We also put the corresponding summary statistics of the replicaition study in the package (smryStats2).

2. You can use `SEest` to estimate the standard error of the observed log-odds ratio.

```
SEest(n0,n1,fU,fA)
```

Details about the function can be seen using `help(SEest)`.

3. You can use `repRateEst` to estimate the RR for each associations discovered from the primary study (i.e. primary associations).

```
repRateEst(MUhat,SE, SE2,zalpha2,zalphaR2,boot=100,output=TRUE,idx=TRUE,dir=output,info=T)
```

Details about the function can be seen using `help(repRateEst)`.

4. You can use `repSampleSizeRR` and `repSampleSizeRR2` to determine the sample size of the replication study.

```
repSampleSizeRR(RR, n, MUhat,SE,zalpha2,zalphaR2,idx=TRUE)
```

```
repSampleSizeRR2(RR,CCR2, MUhat,SE,fU,fA,zalpha2,zalphaR2, idx=TRUE)
```

Details about these functions can be seen using `help(repSampleSizeRR)` and `help(repSampleSizeRR2)`.

5. You can use `HLtest` to check the consistency between the results of the primary study and those of the replication study.

```
HLtest(x,p,g=10,null=all,boot=1000,info=T,dir=.)
```

Details about the function can be seen using `help(HLtest)`

Author(s)

Wei Jiang, Jing-Hao Xue and Weichuan Yu

Maintainer: Wei Jiang <wjiangaa@connect.ust.hk>

References

Jiang, W., Xue, J-H, and Yu, W. What is the probability of replicating a statistically significant association in genome-wide association studies?. *Submitted*.

See Also

repRateEst, SEest, repSampleSizeRR, repSampleSizeRR2, HLtest

Examples

```
alpha<-5e-6               #Significance level in the primary study
alphaR<-5e-3              #Significance level in the replication study
zalpha2<-qnorm(1-alpha/2)
zalphaR2<-qnorm(1-alphaR/2)

##Load data
data(smryStats1)        #Example of summary statistics in 1st study
z1<-smryStats1$Z          #Z values in 1st study
n2.0<-2000                #Number of individuals in control group
n2.1<-2000                #Number of individuals in case group

SE2<-SEest(n2.0, n2.1, smryStats1$F_U, smryStats1$F_A) #SE in replication study
######  RR estimation  ######
RRresult<-repRateEst(log(smryStats1$OR),smryStats1$SE, SE2,zalpha2,zalphaR2, output=TRUE,dir=.)
RR<-RRresult$RR           #Estimated RR
```

HLtest *Hosmer-Lemeshow test*

Description

Test whether each element of x is sampled with the probability specified by the corrsponding element in p.

Usage

```
HLtest(x, p, g = 10, null = "all", boot = 1000, info = T, dir = ".")
```

Arguments

x	A boolean vector.
p	A probability vector having the same length with x.
g	The group number used in the test.
null	a character in c('all', 'chi2','boot'). If null=='chi2', then we use (g-1) degree of freedom chi2 distribution to approximately compute p value. If null=='boot', then we use parametric bootstrap to compute p value. If null=='all', then both methods are used. This is the default option.
boot	The resampling times to compute p value. Only effective when null=='boot' or 'all'
info	Draw the null distribution of the test statistic.
dir	The directory to save the plot of the null distribution.

Details

Null Hypothesis: Each element of x is sampled with a probability which is the corresponding element of p. We group x to g groups according to p. Then we compare the success proportion with the mean value of p in each group.

Value

A list is returned:

`H`	The test statistic.
`pval_chi2`	The p value approximated by using chi2 distribution.
`pval_boot`	The p value computed by using parametric bootstrap.

Author(s)

Wei Jiang, Jing-Hao Xue and Weichuan Yu

Maintainer: Wei Jiang <wjiangaa@connect.ust.hk>

References

Hosmer, D. W., & Lemesbow, S. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in statistics-Theory and Methods*, 9(10), 1043-1069.

Jiang, W., Xue, J-H, and Yu, W. What is the probability of replicating a statistically significant association in genome-wide association studies?. *Submitted*.

See Also

RRate repRateEst, SEest, repSampleSizeRR, repSampleSizeRR2,

Examples

```
alpha<-5e-6               #Significance level in the primary study
alphaR<-5e-3              #Significance level in the replication study
zalpha2<-qnorm(1-alpha/2)
zalphaR2<-qnorm(1-alphaR/2)

##Load data
data(smryStats1)        #Example of summary statistics in 1st study
n2.0<-2000                #Number of individuals in control group
n2.1<-2000                #Number of individuals in case group

SE2<-SEest(n2.0, n2.1, smryStats1$F_U, smryStats1$F_A) #SE in replication study
######  RR estimation  ######
RRresult<-repRateEst(log(smryStats1$OR),smryStats1$SE, SE2,zalpha2,zalphaR2, output=TRUE,dir=.)

#### Hosmer-Lemeshow test  ####
data(smryStats2)        #Example of summary statistics in 2nd study
sigIdx<-(smryStats1$P<alpha)
repIdx<-(sign(smryStats1$Z[sigIdx])*smryStats2$Z[sigIdx]>zalphaR2)
groupNum<-10
HLresult<-HLtest(repIdx,RRresult$RR,g=groupNum,dir=.)
```

repSampleSizeRR *Sample size determination for the replication study based on RR*

Description

repSampleSizeRR and repSampleSizeRR2 implement the RR-based sample size determination method for the replication study. If the replication study has the same control-to-case ratio with the primary study, then repSampleSizeRR can be used. Otherwise, repSampleSize2 is more suitable.

Usage

```
repSampleSizeRR(GRR, n, MUhat, SE, zalpha2, zalphaR2, idx = TRUE)

repSampleSizeRR2(GRR,CCR2, MUhat,SE,fU,fA,zalpha2,zalphaR2, idx=TRUE)
```

Arguments

GRR	The desired global replication rate.
n	Sample size in the primary study.
MUhat	The observed effect size (log-odds ratio).
SE	The standard error of MUhat.
zalpha2	The critical value of z-values in the primary study, i.e. z_alpha/2.
zalphaR2	The critical value of z-values in the replication study, i.e. z_alphaR/2.
idx	The indexes of the SNPs having been further inverstigated in the replication study. We only calculate RR for primary associations with indexes in idx.
CCR2	The control-to-case ratio of the replication study.
fU	The allele frequency in the control group.
fA	The allele frequency in the case group.

Value

The determined sample size of the replication study is returned.

Author(s)

Wei Jiang, Jing-Hao Xue and Weichuan Yu

Maintainer: Wei Jiang <wjiangaa@connect.ust.hk>

References

Jiang, W., Xue, J-H, and Yu, W. What is the probability of replicating a statistically significant association in genome-wide association studies?. *Submitted.*

See Also

RRate repRateEst, SEest, HLtest

Examples

```
alpha<-5e-6               #Significance level in the primary study
alphaR<-5e-3              #Significance level in the replication study
zalpha2<-qnorm(1-alpha/2)
zalphaR2<-qnorm(1-alphaR/2)

##Load data
data(smryStats1)        #Example of summary statistics in 1st study
#### Sample size determination ###
n1<-4000                  #Sample size of the primary study
n2_1<-repSampleSizeRR(0.8, n1, log(smryStats1$OR),smryStats1$SE,zalpha2,zalphaR2)

CCR2<-2                   #Control-to-case ration in the replication study
n2_2<-repSampleSizeRR2(0.8, CCR2, log(smryStats1$OR),smryStats1$SE,smryStats1$F_U,
smryStats1$F_A,zalpha2,zalphaR2)
```

RRate-functions *Estimating Replication Rate for primary associations*

Description

repRateEst implements a replication rate estimation method. Two-component mixture prior is used in the estimation.

Usage

```
repRateEst(MUhat, SE, SE2, zalpha2, zalphaR2, boot = 100, output = TRUE,
idx = TRUE, dir = "output", info = TRUE)
```

Arguments

MUhat	The observed effect size (log-odds ratio) in the primary study.
SE	The standard error of the observed log-odds ratio in the primary study.
SE2	The standard error of the observed log-odds ratio in the replication study.
zalpha2	The critical value of z-values in the primary study, i.e. z_alpha/2.
zalphaR2	The critical value of z-values in the replication study, i.e. z_alphaR/2.
boot	The resampling number of bootstrop used for estimating the credible interval of the RR.
output	Bool value. To determine whether to output the estimated results in the dir or not.
idx	The indexes of the SNPs having been further inverstigated in the replication study. We only calculate RR for primary associations with indexes in idx.
dir	The directory to save the estimated results. It has effect when output=T
info	Bool value. To determine whether to show the parematers inference results in the terminal or not.

Details

The RR estimation is based on the following two-component mixture model: mu=pi_0 delta_0+(1-pi_0) N(0, sigma_0^2).

Details can be seen the following reference paper.

Value

`repRateEst` returns the RR, lfdr, prediction power and infered parameters. The returened value is a LIST:

`idx`	The index of the SNPs which RR are estimated.
`pi0`	The proportion of nonassociated SNPs.
`sigma02`	The variance of the associated SNPs' effect sizes
`RR`	Estimated replication rate.
`RRlow`	The lower limit of the 95% CI for RR.
`RRhigh`	The upper limit of the 95% CI for RR.
`lfdr`	Estimated local false discovery rate of the primary study
`lfdrLow`	The lower limit of the 95% CI for lfdr.
`lfdrHigh`	The upper limit of the 95% CI for lfdr.
`predPower`	The Bayesian predictive power of the replication study.
`predPowerLow`	The lower limit of the 95% CI for predPower.
`perdPowerHigh`	The upper limit of the 95% CI for predPower.
`GRR`	The Global Replication Rate (Mean value of RR)
`GRRlow`	The lower limit of the 95% CI for GRR.
`GRRhigh`	The upper limit of the 95% CI for GRR.

Author(s)

Wei Jiang, Jing-Hao Xue and Weichuan Yu

Maintainer: Wei Jiang <wjiangaa@connect.ust.hk>

References

Jiang, W., Xue, J-H, and Yu, W. What is the probability of replicating a statistically significant association in genome-wide association studies?. *Submitted*.

See Also

`RRate`, `SEest`, `repSampleSizeRR`, `repSampleSizeRR2`, `HLtest`

Examples

```
alpha<-5e-6               #Significance level in the primary study
alphaR<-5e-3              #Significance level in the replication study
zalpha2<-qnorm(1-alpha/2)
zalphaR2<-qnorm(1-alphaR/2)

##Load data
data(smryStats1)        #Example of summary statistics in 1st study
n2.0<-2000                #Number of individuals in control group
n2.1<-2000                #Number of individuals in case group

SE2<-SEest(n2.0, n2.1, smryStats1$F_U, smryStats1$F_A) #SE in replication study
######  RR estimation  ######
RRresult<-repRateEst(log(smryStats1$OR),smryStats1$SE, SE2,zalpha2,zalphaR2, output=TRUE,dir=.)
RR<-RRresult$RR           #Estimated RR
```

SEest *Estimating standard error of the observed log-odds ratio*

Description

SEest implements the Woolf's method to estimate the standard error of the observed log-odds ratio.

Usage

```
SEest(n0,n1,fU,fA)
```

Arguments

n0	Sample size in the control group.
n1	Sample size in the case group.
fU	Allele frequency in the control group.
fA	Allele frequency in the case group.

Details

The Woolf's method to estimate the standard error of log(OR) is based on the following formula: se(log(OR))=sqrt(1/(n0 fU(1-fU))+1/(n1 fA(1-fA))).

Value

The estimated standard error is returned.

Author(s)

Wei Jiang, Jing-Hao Xue and Weichuan Yu

Maintainer: Wei Jiang <wjiangaa@connect.ust.hk>

References

Woolf, B. (1955). On estimating the relation between blood group and disease. *Ann Hum Genet*, 19(4), 251-253.

Jiang, W., Xue, J-H, and Yu, W. What is the probability of replicating a statistically significant association in genome-wide association studies?. *Submitted*.

See Also

RRate, repRateEst, repSampleSizeRR, repSampleSizeRR2, HLtest

Examples

```
##Load data
data(smryStats1)        #Example of summary statistics in 1st study
n2.0<-2000                #Number of individuals in control group
n2.1<-2000                #Number of individuals in case group

SE2<-SEest(n2.0, n2.1, smryStats1$F_U, smryStats1$F_A) #SE in replication study
```

Index