Package 'GFcom'

November 30, 2016

Type Package

Title Efficient Estimation of Disease Odds Ratios for the Follow-up Genetic Association Studies

Version 1.0

Date 2016-11-29

Author Jiyuan Hu

Maintainer Jiyuan Hu <jiyuan_hu@fudan.edu.cn>

Description This package is developed to efficiently estimate the disease odds ratios of candidate SNPs for the follow-up genetic association studies. OR estimates from the initial GWAS scan suffers the well known selection bias, or the Winner's curse, which can be rather severe when the association tests for the candidate markers lacks statistical power. The newly developed OR estimator GFcom produces both point and CI estimates for the candidate SNPs. Both the point and CI estimate of GFcom shows efficient performance compared with other OR estimators.

Depends R (>= 2.14)

License GPL (>= 2)

Encoding UTF-8

LazyData true

URL https://github.com/JiyuanHu/GFcom

BugReports http://github.com/JiyuanHu/GFcom/issues

RoxygenNote 5.0.1

R topics documented:

GFcom-package 2
CD 3
data.preprocessing 4
GFcom.estimate 6
lung 8
simulated.data.ranking.bias 9
simulated.data.significance.bias 10

Index **13**

GFcom-package *Efficient Estimation of Disease Odds Ratios for the Follow-up Genetic Association Studies*

Description

This package is developed to efficiently estimate the disease odds ratios of candidate SNPs for the follow-up genetic association studies. OR estimates from the initial GWAS scan suffers the well known selection bias, or the "Winner's curse". The selection bias can become rather severe when the association tests for the candidate markers lacks statistical power. The newly developed OR estimator GFcom produces both point and CI estimates for the candidate SNPs. The point estimate of GFcom is as follows:

beta.GFcom = beta.cMLE - tau* (beta.cMLE - beta2)

where beta.cMLE is the conditional MLE utilizing data from both GWAS and the follow-up study, tau is a shrinkage estimator which reduces the estimation bias of beta.cMLE, beta2 is the MLE produced by the follow-up study. Both the point and CI estimate of GFcom shows efficient performance compared with other OR estimators.

Details

Package:	GFcom
Type:	Package
Version:	1.0
Date:	2016-11-29
License:	GPL (>= 2)

Author(s)

Jiyuan Hu

Maintainer: Jiyuan Hu <jiyuan_hu@fudan.edu.cn>

References

HU J, LI X, PAN D, LI Q.(2016) Efficient Estimation of Disease Odds Ratios for the Follow-up Genetic Association Studies.

Examples

```
################################################
#############simulating significance bias data###
beta = log(1.3); #log Odds Ratio
p =0.30;#Minor allele frequency
pi =0.01;#Disease prevalence rate
C = 5; #Threshold of Wald test or
#alpha = 5e-7; #Significance level
n.cases1=1000;
n.controls1=1000;
```

```
n.cases2= 2000;
n.controls2 = 2000;
M = 100;
model=1;
alpha.CI = 0.05;
############
# Not run:
# d = simulated.data.significance.bias(n.cases1,n.controls1,n.cases2,n.controls2,
#   beta= beta,p=p,pi=pi,C= C,M=M,model= model);
# d = GFcom.estimate (d,CI.estimation = TRUE,alpha.CI = alpha.CI);

##################################################
#############simulating ranking bias data#########
psudo = 1.1;
n.cases1 = n.controls1 = 1000;
n.cases2 = n.controls2 = 2000;
pi =0.01;
I = 100;
M = 1e5;
model = 1;
K = 10;
C0 = 12;
alpha.CI = 0.05;
############
# Not run:
# drank = simulated.data.ranking.bias(n.cases1,n.controls1,n.cases2,n.controls2,
#   pi=pi,I, M=M,model= model,K= K,C0 = C0,psudo= psudo);
# drank = GFcom.estimate (drank,CI.estimation = TRUE,alpha.CI = alpha.CI);

##################################################
#############Real data(I) with significance bias##
data(CD);
############
# Not run:
# d = data.preprocessing(genotypes.of.2stages = CD,psudo= 1.1);
# d = GFcom.estimate (d,CI.estimation = TRUE,alpha.CI = 0.05,reportOR = TRUE);

##################################################
#############Real data(II) with ranking bias##
data(lung)
############
# Not run:
#d = data.preprocessing(dat = lung,psudo= 1.1);
#d = GFcom.estimate (d,CI.estimation = TRUE,alpha.CI = 0.05,reportOR = TRUE);
```

CD *Genotype data from a follow-up genetic association study of Crohn's disease*

Description

This is a real follow-up genetic association data of Crohn's disease. This data contains the genotype frequencies for the initial GWAS scan and the follow-up study.

Usage

```
data(CD);
```

Format

A data frame containing 12 variables for 11 SNPs.

`DAA1` The frequency of homogeneous reference genotype AA in disease group at GWAS scan.

`DAa1` The frequency of heterogeneous genotype Aa in disease group at GWAS scan.

`Daa1` The frequency of homogeneous variant genotype aa in disease group at GWAS scan.

`NAA1` The frequency of homogeneous reference genotype AA in normal group at GWAS scan.

`NAa1` The frequency of heterogeneous genotype Aa in normal group at GWAS scan.

`Naa1` The frequency of homogeneous variant genotype aa in normal group at GWAS scan.

`DAA2` The frequency of homogeneous reference genotype AA in disease group at the follow-up study.

`DAa2` The frequency of heterogeneous genotype Aa in disease group at the follow-up study.

`Daa2` The frequency of homogeneous variant genotype aa in disease group at the follow-up study.

`NAA2` The frequency of homogeneous reference genotype AA in normal group at the follow-up study.

`NAa2` The frequency of heterogeneous genotype Aa in normal group at the follow-up study.

`Naa2` The frequency of homogeneous variant genotype aa in normal group at the follow-up study.

References

Parkes M, Barrett JC, Prescott NJ et al. Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. Nature Genetics 2007; 39(7): 830–832.

Burton PR, Clayton DG, Cardon LR et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007; 447(7145): 661–678.

`data.preprocessing` *Real data preprocessing used to calculate parameters for the OR estimation.*

Description

This function does the preprocessing steps, calculates parameters for the OR estimation of follow-up genetic association studies.

Usage

```
data.preprocessing(dat = NULL, genotypes.of.2stages = NULL, model = 1, psudo = 1.1)
```

Arguments

`dat` A data frame containing six variables. Each row represents one candidate SNP.

- `OR1` The OR estimates at the GWAS.
- `OR1.L` The lower bound of 95% CI of ORs at the GWAS.
- `OR1.U` The upper bound of 95% CI of ORs at the GWAS.
- `OR2` The OR estimates at the follow-up study.
- `OR2.L` The lower bound of 95% CI of ORs at the follow-up study.
- `OR2.U` The upper bound of 95% CI of ORs at the follow-up study.

`genotypes.of.2stages` A data frame containing 12 variables. Each row represents one candidate SNP. In the following A is the reference allele and a is the variant allele, respectively.

- `DAA1` The frequency of genotype AA in disease group at GWAS scan.
- `DAa1` The frequency of genotype Aa in disease group at GWAS scan.
- `Daa1` The frequency of genotype aa in disease group at GWAS scan.
- `NAA1` The frequency of genotype AA in normal group at GWAS scan.
- `NAa1` The frequency of genotype Aa in normal group at GWAS scan.
- `Naa1` The frequency of genotype aa in normal group at GWAS scan.
- `DAA2` The frequency of genotype AA in disease group at the follow-up study.
- `DAa2` The frequency of genotype Aa in disease group at the follow-up study.
- `Daa2` The frequency of genotype aa in disease group at the follow-up study.
- `NAA2` The frequency of genotype AA in normal group at the follow-up study.
- `NAa2` The frequency of Aa in normal group at the follow-up study.
- `Naa2` The frequency of genotype aa in normal group at the follow-up study.

`model` Genetic model. `model=1` by default, which indicates the additive genetic model.

`psudo` The multiplier used to calculated the psudo-significance level. `psudo=1.1` by default.

Value

`par` A list containing the following elements:

- **alpha** The psudo-significance level.
- **C** The psudo-threshold of Wald statistics for the association test.
- **K** Number of candidate SNPs selected at the GWAS. Only the top `K` most significant SNPs are selected and reported.

`est` A data frame containing 6 variables of `K` SNPs.

- **beta1** The estimated log ORs at GWAS.
- **se1** The estimated standard error of log ORs at GWAS.
- **beta2** The estimated log ORs at the follow-up study.
- **se2** The estimated standard error of log ORs at the follow-up study.
- **beta.com** The combined estimated log ORs for both studies.
- **se.com** The combined estimated standard error of log ORs for both studies.

`genotypes.of.2stages` A matrix of genotype data for both studies. The first 6 columns are the genotype frequency for disease and normal groups at GWAS. The last 6 conlumns are the genotype frequency for disease and normal groups at the follow-up study. This element might be missed if the input parameter `genotypes.of.2stages=NULL`.

Author(s)

Jiyuan Hu

References

HU J, LI X, PAN D, LI Q.(2016) Efficient Estimation of Disease Odds Ratios for the Follow-up Genetic Association Studies.

Examples

```
################################################
#############Real data(I) with significance bias##
data(CD);
############
# Not run:
# d = data.preprocessing(genotypes.of.2stages = CD,psudo= 1.1);
# d = GFcom.estimate (d,CI.estimation = TRUE,alpha.CI = 0.05,reportOR = TRUE);

################################################
#############Real data(II) with ranking bias##
data(lung);
############
# Not run:
# d = data.preprocessing(dat = lung,psudo= 1.1);
# d = GFcom.estimate (d,CI.estimation = TRUE,alpha.CI = 0.05,reportOR = TRUE);
```

`GFcom.estimate` *The point and CI estimates of odds ratios (log odds ratios) for the follow-up genetic association studies.*

Description

This function gives the point and CI estimates of odds ratios (log odds ratios) for follow-up genetic association studies.

Usage

```
GFcom.estimate(d, CI.estimation = TRUE, alpha.CI = 0.05, reportOR = FALSE)
```

Arguments

`d`	A list produced by functions `data.preprocessing()`, `simulated.data.significance.bias()` or `simulated.data.ranking.bias()`. The list contains the following elements: `pars`, `est` and `genotypes.of.2.stages`.
`CI.estimation`	If `CI.estimation == TRUE`, this function will produce CI estimates for log ORs.
`alpha.CI`	The significance level of the confidence interval.
`reportOR`	If `reportOR == TRUE`, this function will produce CI estimates for ORs as well.

Value

The returning object is still the input list d, with the following elements added to d itself:

point.est	The point estimate of GFcom for log ORs.
CI.est	The confidence interval estimate of GFcom for log ORs.
point.est.OR	The point estimate of GFcom for ORs.
CI.est.OR	The confidence interval estimate of GFcom for ORs.

Author(s)

Jiyuan Hu

References

HU J, LI X, PAN D, LI Q.(2016) Efficient Estimation of Disease Odds Ratios for the Follow-up Genetic Association Studies.

Examples

```
################################################
#############simulating significance bias data###
beta = log(1.3); #Log Odds Ratio
p =0.30;#Minor allele frequency
pi =0.01;#Disease prevalence rate
C = 5; #Threshold of Wald test or
#alpha = 5e-7; #Significance level
n.cases1=1000;
n.controls1=1000;
n.cases2= 2000;
n.controls2 = 2000;
M = 100;
model=1;
alpha.CI = 0.05;
############
# Not run:
# d = simulated.data.significance.bias(n.cases1,n.controls1,n.cases2,n.controls2,beta= beta,
#   p=p,pi=pi,C= C,M=M,model= model);
# d = GFcom.estimate (d,CI.estimation = TRUE,alpha.CI = alpha.CI)

################################################
#############simulating ranking bias data#########
psudo = 1.1;
n.cases1 = n.controls1 = 1000;
n.cases2 = n.controls2 = 2000;
pi =0.01;
I = 100;
M = 1e5;
model = 1;
K = 10;
C0 = 12;
alpha.CI = 0.05;
############
# Not run:
# drank = simulated.data.ranking.bias(n.cases1,n.controls1,n.cases2,n.controls2,
#  pi=pi,I, M=M,model= model,K= K,C0 = C0,psudo= psudo);
```

```
# drank = GFcom.estimate (drank,CI.estimation = TRUE,alpha.CI = alpha.CI);

################################################
#############Real data(I) with significance bias##
data(CD);
############
# Not run:
# d = data.preprocessing(genotypes.of.2stages = CD,psudo= 1.1);
# d = GFcom.estimate (d,CI.estimation = TRUE,alpha.CI = 0.05,reportOR = TRUE);

################################################
#############Real data(II) with ranking bias##
data(lung);
############
# Not run:
# d = data.preprocessing(dat = lung,psudo= 1.1);
# d = GFcom.estimate (d,CI.estimation = TRUE,alpha.CI = 0.05,reportOR = TRUE);
```

`lung` *Estimated Odds Ratios (OR) of candidate SNPs for lung cancer*

Description

This is a real follow-up study data of lung cancer. This data contains the estimated ORs and corresponding 95% confidence intervals of 10 candidate SNPs for the initial GWAS scan and the follow-up study.

Usage

```
data(lung)
```

Format

A data frame containing six variables for 10 SNPs.

`OR1` The OR estimates at the GWAS.

`OR1.L` The lower bound of 95% CI of ORs at the GWAS.

`OR1.U` The upper bound of 95% CI of ORs at the GWAS.

`OR2` The OR estimates at the follow-up study.

`OR2.L` The lower bound of 95% CI of ORs at the follow-up study.

`OR2.U` The upper bound of 95% CI of ORs at the follow-up study.

References

Amos CI, Wu X, Broderick P et al. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. Nature Genetics 2008; 40(5): 616–622.

simulated.data.ranking.bias
The Generation of Simulated Data with ranking bias at GWAS

Description

This function generates simulated two-stage genetic association study data with ranking bias at GWAS.

Usage

```
simulated.data.ranking.bias(n.cases1, n.controls1, n.cases2, n.controls2,
  pi, I, M, model, K, C0, psudo)
```

Arguments

n.cases1	Number of disease individuals at GWAS.
n.controls1	Number of normal individuals at GWAS.
n.cases2	Number of disease individuals at the follow-up study.
n.controls2	Number of normal individuals at the follow-up study.
pi	Disease prevalence rate.
I	Number of truely diease-associated SNPs among M SNPs
M	Number of simulated SNPs in the genome-wide scale.
model	Genetic model. model=1 indicates additive genetic model; model=0 indicates recessive genetic model; model=2 indicates dominant genetic model.
K	Number of selected SNPs at the GWAS. Only the top K most significant SNPs are selected and reported.
C0	The screening threshold of the Cochran-Armitage Trend Test (CATT). To speed up the calculation, CATT is undertook first to filter out most of the SNPs. SNPs passing this screening step will further fit the logistic model.
psudo	The multiplier used to calculated the psudo-significance level.

Value

par	A list containing the input parameters of the function as well as: **beta** The true log OR value of top K SNPs. **p** The minor allele frequencies of top K SNPs. **alpha0** The retrospective baseline log ORs of top K SNPs. **alpha** The psudo-significance level. **C** The psudo-threshold of Wald statistics for genetic association test.
est	A data frame containing 6 variables of K SNPs. **beta1** The estimated log ORs at GWAS. **se1** The estimated standard error of log ORs at GWAS. **beta2** The estimated log ORs at the follow-up study. **se2** The estimated standard error of log ORs at the follow-up study. **beta.com** The combined estimated log ORs for both studies.

se.com The combined estimated standard error of log ORs for both studies.

`genotypes.of.2stages`
A matrix of genotype data for both studies. The first 6 columns are the genotype frequency for disease and normal groups at GWAS. The last 6 conlumns are the genotype frequency for disease and normal groups at the follow-up study.

Author(s)

Jiyuan Hu

References

HU J, LI X, PAN D, LI Q.(2016) Efficient Estimation of Disease Odds Ratios for the Follow-up Genetic Association Studies.

Examples

```
psudo = 1.1;
n.cases1 = n.controls1 = 1000;
n.cases2 = n.controls2 = 2000;
pi =0.01;
I = 100;
M = 1e5;
model = 1;
K = 10;
C0 = 12;
alpha.CI = 0.05;
############
# Not run:
# drank = simulated.data.ranking.bias(n.cases1,n.controls1,n.cases2,n.controls2,
#   pi=pi,I, M=M,model= model,K= K,C0 = C0,psudo= psudo);
# drank = GFcom.estimate (drank,CI.estimation = TRUE,alpha.CI = alpha.CI);
```

---

simulated.data.significance.bias

The Generation of Simulated Data with significance bias at GWAS

Description

This function generates simulated two-stage association study data with significance bias at GWAS.

Usage

```
simulated.data.significance.bias(n.cases1, n.controls1, n.cases2, n.controls2,
  beta = NULL, OR = NULL, p, pi, C = NULL, alpha = NULL, M = 100, model = 1)
```

Arguments

`n.cases1`	Number of disease individuals at GWAS.
`n.controls1`	Number of normal individuals at GWAS.
`n.cases2`	Number of disease individuals at the follow-up study.
`n.controls2`	Number of normal individuals at the follow-up study.

beta	True log odds ratio.
OR	True odds ratio. Either beta or OR should be unNULL.
p	Minor allele frequency.
pi	Disease prevalence rate.
C	The threshold for Wald statistics used for the association test.
alpha	The significance level of association tests. Notice that C is the upper alpha/2 quantile of a standard normal distribution.
M	Number of significant repetitions.
model	Genetic model. model=1 indicates additive genetic model; model=0 indicates recessive genetic model; model=2 indicates dominant genetic model.

Value

par	A list containing the input parameters of the function as well as **alpha0** The retrospective baseline log OR. **ps** The genotype frequencies for genotypes AA, Aa and aa, respectively.
est	A data frame containing 6 variables of M observations. **beta1** The estimated log ORs at GWAS. **se1** The estimated standard error of log ORs at GWAS. **beta2** The estimated log ORs at the follow-up study. **se2** The estimated standard error of log ORs at the follow-up study. **beta.com** The combined estimated log ORs for both studies. **se.com** The combined estimated standard error of log ORs for both studies.
genotypes.of.2stages	A matrix of genotype data for both studies. The first 6 columns are the genotype frequency for disease and normal groups at GWAS. The last 6 conlumns are the genotype frequency for disease and normal groups at the follow-up study.

Author(s)

Jiyuan Hu

References

HU J, LI X, PAN D, LI Q.(2016) Efficient Estimation of Disease Odds Ratios for the Follow-up Genetic Association Studies.

Examples

```
beta = log(1.3); #Log Odds Ratio
p =0.30;#Minor allele frequency
pi =0.01;#Disease prevalence rate
C = 5; #Threshold of Wald test or
#alpha = 5e-7; #Significance level
n.cases1=1000;
n.controls1=1000;
n.cases2= 2000;
n.controls2 = 2000;
M = 100;
model=1;
```

```
alpha.CI = 0.05;
############
# Not run:
# d = simulated.data.significance.bias(n.cases1,n.controls1,n.cases2,n.controls2, beta= beta,
#   p=p,pi=pi,C= C,M=M,model= model);
# d = GFcom.estimate (d,CI.estimation = TRUE,alpha.CI = alpha.CI)
```

Index

∗Topic **GFcom**
GFcom.estimate, 6
∗Topic **Real data**
CD, 3
data.preprocessing, 4
lung, 8
∗Topic **package**
GFcom-package, 2
∗Topic **ranking bias**
lung, 8
simulated.data.ranking.bias, 9
∗Topic **significance bias**
CD, 3
simulated.data.significance.bias, 10
∗Topic **simulated data**
simulated.data.ranking.bias, 9
simulated.data.significance.bias, 10

CD, 3

data.preprocessing, 4

GFcom *(*GFcom-package*)*, 2
GFcom-package, 2
GFcom.estimate, 6

lung, 8

simulated.data.ranking.bias, 9
simulated.data.significance.bias, 10