Package 'IN.power'

July 8, 2010

Title An R package (beta version) for computing the number of susceptibility SNPs

Version 0.0.1

Date 2010-07-08

Author Ju-Hyun Park

Maintainer Ju-Hyun Park <parkj3@mail.nih.gov>

Description An R package (beta version) for computing the number of susceptibility SNPs and power of future studies

Depends R (>= 2.10.1), mvtnorm

License GPL-2

R topics documented:

IN.power 1
INPower 2

Index **5**

`IN.power` *An R package (beta version) to estimate the number of susceptibility SNPs and power of future studies*

Description

This function uses the effect sizes for a set of known susceptibility SNPs and the power of detection of these SNPs from the original discovery samples to obtain an estimate of the total number of underlying susceptibility SNPs for that trait and the distribution of their effect sizes. The function can further use the estimated number of loci and distribution of effect sizes to evaluate the power for discovery of a future GWAS study (up to three-stage).

Author(s)

Ju-Hyun Park <prkj3@mail.nih.gov>

`INPower` *Estimate the number of susceptibility SNPs and power of future studies*

Description

This function uses the effect sizes for a set of known susceptibility SNPs and the power of detection of these SNPs from the original discovery samples to obtain an estimate of the total number of underlying susceptibility SNP for that trait and the distribution of their effect sizes. The function can further use the estimated number of loci and distribution of effect sizes to evaluate the power for discovery of a future GWAS study (up to three-stage).

Usage

```
INPower(MAFs, betas, pow, span=0.5, binary.outcome=TRUE,
         sample.size, signif.lvl, multi.stage.option=NULL, tgv=NULL, k)
```

Arguments

`MAFs`	Vector of minor allele frequencies associated with the set of known loci
`betas`	Vector of regression effects for the set of known loci under an additive genetic model. For a continuous phenotype analyzed with linear regression model, it is assumed that the outcome has been standardized so that the coefficients correspond to mean change in outcome per unit of s.d. for each copy of the given allele. For a binary outcome analyzed with logistic regression, the regression coefficients should correspond to change in log-odds-ratio per copy of the given allele.
`pow`	A vector representing the powers for the known loci in the original studies that led to their discoveries. Note these power calculations should be carefully done to avoid winner's curse (it is best to obtain effect size estimates from independent replication study) and to take into consideration all complexities of the designs of the original study. If the total SNP set is obtained from a group of studies for a given trait, then the power for an individual marker should reflect the probability of its detection in at least one of the studies.
`span`	The parameter which controls the degree of smoothing in Loess. It specifies the fraction of SNPs that are used in local linear regression to obtain the estimated number of loci at each effect size. The default is set at 0.5, but we recommend the user to set it at a value depending on the total size of the SNP set so that about 10-20 SNPs are used for local smoothing at each effect size. The total size of the SNP set should be reasonably large (e.g. at least 20 and preferably more) for application of Loess.
`binary.outcome`	TRUE/FALSE Is the outcome binary or continuous?
`sample.size`	Sample size for a future study for which integrated power calculation is desired. For case-control studies, half of the subjects are assumed to be cases and half to be controls. It can take a vector of several sample sizes for the same study as shown in the example below.
`signif.lvl`	The required genome-wide significance level for future study.

`multi.stage.option`
: This option allows to set-up design parameters for the future study if it would be done in multiple stages (up to three). The option has a list of two arguments alpha and pi, where alpha specifies the significance level(s) used for each stage to select markers for the subsequent stage and pi specifies the fraction of subjects who are included in the corresponding stages. The default for the option is NULL, that is, the study is assumed to be single-stage.

`tgv`
: An optional argument using which the user can input an estimate of the known total genetic variance (TGV) of the trait that may be available from familial aggregation studies. For a continuous outcome, this could be an estimate of the fraction of the total variance of the trait attributed to heritability. For a binary outcome, this could logarithm of squared sibling-relative-risk that is known to approximate total genetic variance under log-normal model for risk.

`k`
: A vector of integer values for which the user would like to calculate probabilities of the type Pr(X>= k) to evaluate the probability of detection of at least a specified number of loci in future studies. In addition, the function automatically finds nine values for "k", for which the probabilities are close to 0.1 to 0.9 with an increment of 0.1.

Details

The projections are only shown in the range of effect size for which the original studies had at least 1 percent power. The loess fitting procedure, however, may include additional SNPs with smaller effect sizes for local linear smoothing. The user is recommended to remove SNPs that may seem clearly outliers compared to the rest in terms of their effect sizes. By default the program currently removes all SNPs with power less than 0.1 percent from the analysis to avoid undue influence of potentially outlying observations.

Value

A list of two sublists with names `esdist.summary` and `future.study.summary`. The sublist `esdist.summary` contains the estimated number of loci (`t.n.loci`), the genetic variance explained by the estimated number of loci (`gve`), and the estimated number of loci at each different effect size (`es.dist`). Note for linear regression, `gve` is expressed as a percentage of the total variance of the outcome, since it assumed that outcome has been standardized. Further, if an estimate of total genetic variance (TGV) is provided by the user, then the estimate for GVE will be automatically expressed as a percentage of TGV. The sublist `future.study.summary` contains the expected number of loci to be discovered in the future study (`e.discov`), expected genetic variance explained (`e.gve`), and a table of probabilities of discovering at least `k` loci for the diffferent values of `k` (`prob.k`). Note that `e.gve` is defined similarly to `gve`.

References

Park et al. (2010). Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nature Genetics, 42:570-5.

Examples

```
# Analysis of height data used in Park et al. (2010)
MAFs<- c(0.13, 0.14, 0.50, 0.48, 0.28, 0.37, 0.22, 0.30, 0.24, 0.48, 0.14, 0.21,
         0.27, 0.44, 0.50, 0.48, 0.37, 0.33, 0.36, 0.23, 0.11, 0.31, 0.15, 0.50,
         0.28, 0.48, 0.38, 0.26, 0.44, 0.45)
betas<- c(0.0354, -0.0390,  0.0330, -0.0335, -0.0483,  0.0450, 0.0540, -0.0502,
```

```
          -0.0550, 0.0472, -0.0690, -0.0590, -0.0562, 0.0540, 0.0550, -0.0570,
          -0.0600, -0.0630, 0.0631, 0.0750,  -0.1050, 0.0715, 0.0950, 0.0679,
          -0.0770, -0.0720, -0.0750, -0.0830, 0.0950, 0.1350)
pow<- c(0.010, 0.028, 0.123, 0.137, 0.580, 0.584, 0.650, 0.710, 0.756, 0.765,
        0.793, 0.800, 0.863, 0.948, 0.966, 0.983, 0.988, 0.993, 0.996, 0.999,
        1.000, 1.000, 1.000, 1.000, 1.000, 1.000, 1.000, 1.000, 1.000, 1.000)

# Suppose that a one-stage design with a genome-wide significant level of 10^(-7)
# is considered for a future study. It is known that the heritability of height
# is ~0.8. It is of interest to predict the expected number of discoveries with
# sample sizes from 25,000 to 12,5000 with an increment of 25,000.
# In addition, for the given sample sizes,
# it is also of interest to find power to detect at least k loci, with k ranging
# from 25 to 125 with an increment of 25.
INPower(MAFs, betas, pow, span=0.5, binary.outcome=FALSE,
        sample.size=seq(25000,125000,by=25000),
        signif.lvl=10^(-7), tgv=0.8, k=seq(25,125,by=25))

# The function call below shows the same results as the above,
# but without the tgv argument. As a result, the genetic variance explained
# is expressed as a percentage of the total variance of the outcome,
# not of the total genetic variance.
INPower(MAFs, betas, pow, span=0.5, binary.outcome=FALSE,
        sample.size=seq(25000,125000,by=25000),
        signif.lvl=10^(-7), k=seq(25,125,by=25))

# Now a two-stage study is considered with all the other conditions remaining the
# same (including the estimate of total heritability).
# In addition, the selection criterion for SNPs
# taken toward the second stage is 5*10^(-5) and 30% of the given sample size is
# assigned to the first stage (and hence 70% to the second stage).
INPower(MAFs, betas, pow, span=0.5, binary.outcome=FALSE,
        sample.size=seq(25000,125000,by=25000),
        signif.lvl=10^(-7), multi.stage.option=list(al=5*10^(-5), pi=0.3),
        tgv=0.8 , k=seq(25,125,by=25))
```

Index

∗Topic **models**
    INPower, 2
∗Topic **package**
    IN.power, 1

IN.power, 1
INPower, 2