Package 'powerEQTL'

October 14, 2022

Type Package

Title Power and Sample Size Calculation for Bulk Tissue and Single-Cell eQTL Analysis

Version 0.3.4

Date 2021-07-21

Author Xianjun Dong [aut, ctb],
Xiaoqi Li [aut, ctb],
Tzuu-Wang Chang [aut, ctb],
Scott T. Weiss [aut, ctb],
Weiliang Qiu [aut, cre]

Maintainer Weiliang Qiu <weiliang.qiu@gmail.com>

Depends R (>= 3.6.0)

Imports stats, nlme, GLMMadaptive, parallel

Description Power and sample size calculation for bulk tissue and single-cell eQTL analysis based on ANOVA, simple linear regression, or linear mixed effects model. It can also calculate power/sample size
for testing the association of a SNP to a continuous type phenotype.
Please see the reference: Dong X, Li X, Chang T-
W, Scherzer CR, Weiss ST, Qiu W. (2021) <doi:10.1093/bioinformatics/btab385>.

License GPL (>= 2)

URL https://github.com/sterding/powerEQTL and
https://bwhbioinfo.shinyapps.io/powerEQTL/

NeedsCompilation no

Repository CRAN

Date/Publication 2021-07-22 08:30:02 UTC

R topics documented:

`powerEQTL.ANOVA` *Power Calculation for EQTL Analysis Based on Un-Balanced One-Way ANOVA*

Description

Power calculation for eQTL analysis that tests if a SNP is associated to a gene probe by using unbalanced one-way ANOVA. This function can be used to calculate one of the 3 parameters (power, sample size, and minimum allowable MAF) by setting the corresponding parameter as NULL and providing values for the other 2 parameters.

Usage

```
powerEQTL.ANOVA(MAF,
                deltaVec=c(-0.13, 0.13),
                n=200,
                power = NULL,
                sigma = 0.13,
                FWER = 0.05,
                nTests = 200000,
                n.lower = 4,
                n.upper = 1e+30)
```

Arguments

`MAF`	numeric. Minor allele frequency.
`deltaVec`	numeric. A vector having 2 elements. The first element is equal to $\mu_2 - \mu_1$ and the second element is equal to $\mu_3 - \mu_2$, where μ_1 is the mean gene expression level for the mutation homozygotes, μ_2 is the mean gene expression level for the heterozygotes, and μ_3 is the mean gene expression level for the wild-type gene expression level.
`n`	integer. Total number of subjects.
`power`	numeric. Power for testing if 3 genotypes have the same mean gene expression levels.
`sigma`	numeric. Standard deviation of the random error.
`FWER`	numeric. Family-wise Type I error rate.
`nTests`	integer. Number of tests (i.e., number of all (SNP, gene) pairs) in eQTL analysis.
`n.lower`	numeric. Lower bound of the total number of subjects. Only used when calculating total number of subjects.
`n.upper`	numeric. Upper bound of the total number of subjects. Only used when calculating total number of subjects.

Details

If we would like to test potential non-linear relationship between genotype of a SNP and expression of a gene, we can use un-balanced one-way ANOVA. Actually, an article published by the GTEx Consortium in 2013 used this approach.

Suppose there are $k = 3$ groups of subjects: (1) mutation homozygotes; (2) heterozygotes; and (3) wildtype homozygotes. We would like to test if the mean expression μ_i, $i = 1, \ldots, k$, of the gene is the same among the k groups of subjects. We can use the following one-way ANOVA model to characterize the relationship between observed gene expression level y_{ij} and the population mean expression level μ_i:

$$y_{ij} = \mu_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N\left(0, \sigma^2\right),$$

where $i = 1, \ldots, k$, $j = 1, \ldots, n_i$, y_{ij} is the observed gene expression level for the j-th subject in the i-th group, μ_i is the mean gene expression level of the i-th group, ϵ_{ij} is the random error, k is the number of groups, n_i is the number of subjects in the i-th group. Denote the total number of subjects as $N = \sum_{i=1}^{k} n_i$. That is, we have n_1 mutation homozygotes, n_2 heterozygotes, and n_3 wildtype homozygotes.

We would like to test the null hypothesis H_0 and alternative hypothesis H_1:

$$H_0 : \mu_1 = \mu_2 = \mu_3,$$

$$H_1 : \text{not all means are the same.}$$

According to O'Brien and Muller (1993), the power calculation formula for unbalanced one-way ANOVA is

$$power = Pr\left(F \geq F_{1-\alpha}\left(k-1, N-k\right) \middle| F \sim F_{k-1,N-k,\lambda}\right),$$

where $k = 3$ is the number of groups of subjects, N is the total number of subjects, $F_{1-\alpha}\left(k-1, N-k\right)$ is the $100(1-\alpha)$-th percentile of central F distribution with degrees of freedoms $k-1$ and $N-k$, and $F_{k-1,N-k,\lambda}$ is the non-central F distribution with degrees of freedoms $k-1$ and $N-k$ and non-central parameter (ncp) λ. The ncp λ is equal to

$$\lambda = \frac{N}{\sigma^2} \sum_{i=1}^{k} w_i \left(\mu_i - \mu\right)^2,$$

where μ_i is the mean gene expression level for the i-th group of subjects, w_i is the weight for the i-th group of subjects, σ^2 is the variance of the random errors in ANOVA (assuming each group has equal variance), and μ is the weighted mean gene expression level

$$\mu = \sum_{i=1}^{k} w_i \mu_i.$$

The weights $w_i = n_i/N$ are the sample proportions for the 3 groups of subjects, where $N = n_1 + n_2 + n_3$ is the total number of subjects. Hence, $\sum_{i=1}^{3} w_i = 1$. Based on Hardy-Weinberg Equilibrium, we have $w_1 = \theta^2$, $w_2 = 2\theta(1-\theta)$, and $w_3 = (1-\theta)^2$, where θ is MAF.

Without loss of generality, we set $\mu_1 = -\delta_1$, $\mu_2 = 0$, and $\mu_3 = \delta_2$.

We adopted the parameters from the GTEx cohort (see the Power analysis" section of Nature Genetics, 2013; https://www.nature.com/articles/ng.2653), where they modeled the expression data as having a log-normal distribution with a log standard deviation of 0.13 within each genotype class (AA, AB, BB). This level of noise is based on estimates from initial GTEx data. In their power analysis, they assumed the across-genotype difference delta = 0.13 (i.e., equivalent to detecting a log expression change similar to the standard deviation within a single genotype class).

Value

power if the input parameter `power = NULL`.

sample size (total number of subjects) if the input parameter `n = NULL`;

minimum detectable slope if the input parameter `slope = NULL`.

Author(s)

Xianjun Dong <XDONG@rics.bwh.harvard.edu>, Xiaoqi Li<xli85@bwh.harvard.edu>, Tzuu-Wang Chang <Chang.Tzuu-Wang@mgh.harvard.edu>, Scott T. Weiss <restw@channing.harvard.edu>, Weiliang Qiu <weiliang.qiu@gmail.com>

References

The GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nature Genetics, 45:580-585, 2013.

O'Brien, RG and Muller, KE. Unified power analysis for t-tests through multivariate hypotheses. In LK Edwards, editor, Applied Analysis of Variance in Behavioral Science, pages 297-344. New York: Dekker, 1993.

Dong X, Li X, Chang T-W, Scherzer CR, Weiss ST, and Qiu W. powerEQTL: An R package and shiny application for sample size and power calculation of bulk tissue and single-cell eQTL analysis. Bioinformatics, 2021;, btab385

Examples

```
# calculate power
powerEQTL.ANOVA(MAF = 0.1,
                deltaVec = c(-0.13, 0.13),
                n = 282,
                power = NULL,
                sigma = 0.13,
                FWER = 0.05,
                nTests = 200000)

# calculate sample size (total number of subjects)
powerEQTL.ANOVA(MAF = 0.1,
                deltaVec = c(-0.13, 0.13),
                n = NULL,
                power = 0.8,
                sigma = 0.13,
                FWER = 0.05,
                nTests = 200000)

# calculate minimum allowable MAF
powerEQTL.ANOVA(MAF = NULL,
                deltaVec = c(-0.13, 0.13),
                n = 282,
                power = 0.8,
                sigma = 0.13,
```

```
                FWER = 0.05,
                nTests = 200000)
```

`powerEQTL.scRNAseq` *Power Calculation for Association Between Genotype and Gene Expression Based on Single Cell RNAseq Data*

Description

Power calculation for association between genotype and gene expression based on single cell RNAseq data. This function can be used to calculate one of the 4 parameters (power, sample size, minimum detectable slope, and minimum allowable MAF) by setting the corresponding parameter as NULL and providing values for the other 3 parameters.

Usage

```
powerEQTL.scRNAseq(
  slope,
  n,
  m,
  power = NULL,
  sigma.y,
  MAF = 0.2,
  rho = 0.8,
  FWER = 0.05,
  nTests = 1,
  n.lower = 2.01,
  n.upper = 1e+30)
```

Arguments

`slope`	numeric. Slope under alternative hypothesis.
`n`	integer. Total number of subjects.
`m`	integer. Number of cells per subject.
`power`	numeric. Power for testing if the slope is equal to zero.
`sigma.y`	numeric. Standard deviation of the gene expression.
`MAF`	numeric. Minor allele frequency (between 0 and 0.5).
`rho`	numeric. Intra-class correlation (i.e., correlation between y_{ij} and y_{ik} for the j-th and k-th cells of the i-th subject).
`FWER`	numeric. Family-wise Type I error rate for one pair (SNP, gene).
`nTests`	integer. Number of tests (i.e., number of all (SNP, gene) pairs) in eQTL analysis.
`n.lower`	numeric. Lower bound of the total number of subjects. Only used when calculating total number of subjects.
`n.upper`	numeric. Upper bound of the total number of subjects. Only used when calculating total number of subjects.

Details

We assume the following simple linear mixed effects model for each (SNP, gene) pair to characterize the association between genotype and gene expression:

$$y_{ij} = \beta_{0i} + \beta_1 * x_i + \epsilon_{ij},$$

where

$$\beta_{0i} \sim N\left(\beta_0, \sigma_\beta^2\right),$$

and

$$\epsilon_{ij} \sim N\left(0, \sigma^2\right),$$

$i = 1, \ldots, n$, $j = 1, \ldots, m$, n is the number of subjects, m is the number of cells per subject, y_{ij} is the gene expression level for the j-th cell of the i-th subject, x_i is the genotype for the i-th subject using additive coding. That is, $x_i = 0$ indicates the i-th subject is a wildtype homozygote, $x_i = 1$ indicates the i-th subject is a heterozygote, and $x_i = 2$ indicates the i-th subject is a mutation homozygote.

We would like to test the following hypotheses:

$$H_0 : \beta_1 = 0,$$

and

$$H_1 : \beta_1 = \delta,$$

where $\delta \neq 0$.

For a given SNP, we assume Hardy-Weinberg Equilibrium and denote the minor allele frequency of the SNP as θ.

We can derive the power calculation formula is

$$power = 1 - \Phi\left(z_{\alpha^*/2} - a \times b\right) + \Phi\left(-z_{\alpha^*/2} - a \times b\right),$$

where

$$a = \frac{\sqrt{2\theta\left(1-\theta\right)}}{\sigma_y}$$

and

$$b = \frac{\delta\sqrt{m(n-1)}}{\sqrt{1+(m-1)\rho}}$$

and $z_{\alpha^*/2}$ is the upper $100\alpha^*/2$ percentile of the standard normal distribution, $\alpha^* = FWER/nTests$, nTests is the number of (SNP, gene) pairs, $\sigma_y = \sqrt{\sigma_\beta^2 + \sigma^2}$, and $\rho = \sigma_\beta^2 / \left(\sigma_\beta^2 + \sigma^2\right)$ is the intra-class correlation.

Value

power if the input parameter `power = NULL`.

sample size (total number of subjects) if the input parameter `n = NULL`;

minimum detectable slope if the input parameter `slope = NULL`;

minimum allowable MAF if the input parameter `MAF = NULL`.

Author(s)

Xianjun Dong <XDONG@rics.bwh.harvard.edu>, Xiaoqi Li<xli85@bwh.harvard.edu>, Tzuu-Wang Chang <Chang.Tzuu-Wang@mgh.harvard.edu>, Scott T. Weiss <restw@channing.harvard.edu>, Weiliang Qiu <weiliang.qiu@gmail.com>

References

Dong X, Li X, Chang T-W, Scherzer CR, Weiss ST, and Qiu W. powerEQTL: An R package and shiny application for sample size and power calculation of bulk tissue and single-cell eQTL analysis. Bioinformatics, 2021;, btab385

Examples

```
n = 102
m = 227868

# calculate power
power = powerEQTL.scRNAseq(
  slope = 0.6,
  n = n,
  m = m,
  power = NULL,
  sigma.y = 0.29,
  MAF = 0.05,
  rho = 0.8,
  nTests = 1e+6)

print(power)

# calculate sample size (total number of subjects)
n = powerEQTL.scRNAseq(
  slope = 0.6,
  n = NULL,
  m = m,
  power = 0.9567288,
  sigma.y = 0.29,
  MAF = 0.05,
  rho = 0.8,
  nTests = 1e+6)

print(n)

# calculate slope
slope = powerEQTL.scRNAseq(
  slope = NULL,
  n = n,
  m = m,
  power = 0.9567288,
  sigma.y = 0.29,
  MAF = 0.05,
  rho = 0.8,
  nTests = 1e+6)
```

```
print(slope)

# calculate MAF
MAF = powerEQTL.scRNAseq(
  slope = 0.6,
  n = n,
  m = m,
  power = 0.9567288,
  sigma.y = 0.29,
  MAF = NULL,
  rho = 0.8,
  nTests = 1e+6)
print(MAF)
```

`powerEQTL.scRNAseq.sim` *Power Calculation for Association Between Genotype and Gene Expression Based on Single Cell RNAseq Data via Simulation from ZINB Mixed Effects Regression Model*

Description

Power calculation for association between genotype and gene expression based on single cell RNAseq data via simulation from ZINB mixed effects regression model. This function can be used to calculate one of the 4 parameters (power, sample size, minimum detectable slope, and minimum allowable MAF) by setting the corresponding parameter as NULL and providing values for the other 3 parameters.

Usage

```
powerEQTL.scRNAseq.sim(slope,
       n,
       m,
                      power = NULL,
       m.int = -1,
       sigma.int = 1,
       zero.p = 0.1,
       theta = 1,
       MAF = 0.2,
       FWER = 0.05,
       nTests = 1,
       nSim = 1000,
       estMethod = "GLMMadaptive",
       nCores = 1,
                      n.lower = 2.01,
                      n.upper = 1e+4,
```

```
                        slope.lower =  1e-6,
                        slope.upper = log(1.0e+6),
                        MAF.lower = 0.05,
                        MAF.upper = 0.49
)
```

Arguments

slope	numeric. Slope (see details).
n	integer. Total number of subjects.
m	integer. Number of cells per subject.
power	numeric. Power for testing if the slope is equal to zero.
m.int	numeric. Mean of random intercept (see details).
sigma.int	numeric. Standard deviation of random intercept (see details).
zero.p	numeric. Probability that an excess zero occurs.
theta	numeric. Dispersion parameter of negative binomial distribution. The smaller theta is, the larger variance of NB random variable is.
MAF	numeric. Minor allele frequency of the SNP.
FWER	numeric. Family-wise Type I error rate.
nTests	integer. Number of tests (i.e., number of all (SNP, gene) pairs) in eQTL analysis.
nSim	integer. Number of simulated datasets to be generated.
estMethod	character. Indicates which method would be used to fit zero inflated negative binomial mixed effects model. Currently, the possible choice is "GLMMadaptive".
nCores	integer. Number of computer cores used by mclapply for parallel computing. For Windows, nCores=1.
n.lower	numeric. Lower bound of the total number of subjects. Only used when calculating total number of subjects.
n.upper	numeric. Upper bound of the total number of subjects. Only used when calculating total number of subjects.
slope.lower	numeric. Lower bound of the slope. Only used when calculating minimum slope.
slope.upper	numeric. Upper bound of the slope Only used when calculating minimum slope.
MAF.lower	numeric. Lower bound of the MAF. Only used when calculating minimum MAF.
MAF.upper	numeric. Upper bound of the MAF Only used when calculating minimum MAF.

Details

This function calculates the power for testing if genotypes of a SNP is associated with the expression of a gene via nSim datasets generated from zero-inflated negative binomial (ZINB) regression model.

Each dataset is generated from zero-inflated negative binomial mixed effects regression model with only one covariate: genotype. That is, the read counts of a gene follows a mixture of 2-component distributions. One component takes only one value: zero. The other component is negative binomial distribution, which takes non-negative values 0, 1, 2, ... The log mean of the negative binomial distribution is linear function of the genotype.

For each dataset, the p-value for testing if the slope for genotype is equal to zero will be calculated.

The proportion of p-values $< \alpha$ is the estimated power, where $\alpha = FWER/nTests$.

Each simulated dataset contains gene expression levels of one gene and genotypes of one SNP for subjects with multiple cells. The gene expression levels (read counts) follow zero-inflated negative binomial distribution. Denote Y_{ij} as the read counts for the j-th cell of the i-th subject, $i = 1, \ldots, n$, $j = 1, \ldots, m$, n is the number of subjects, and m is the number of cells per subject. Denote p as the probability that $Y_{ij} = 0$ is an excess zero. With probability $1 - p$, Y_{ij} follows a negative binomial distribution $NB(\mu, \theta)$, where μ is the mean (i.e., $\mu = E(Y_{ij})$) and θ is the dispersion parameter. The variance of the NB distribution is $\mu + \mu^2/\theta$. The relationship between gene expression and genotype for the i-th subject is characterized by the equation

$$\mu_i = \exp(\beta_{0i} + \beta_1 x_i),$$

where β_{0i} is the random intercept following a normal distribution $N(\beta_0, \sigma^2)$ to account for within-subject correlation of gene expression, β_0 is the mean of the random intercept, σ is the standard deviation of the random intercept, β_1 is the slope, and x_i is the additive-coded genotype for the SNP with minor allele frequency MAF.

We assume that the SNP satisfies the Hardy-Weinberg Equilibrium. That is, the probabilities of the 3 genotypes $(0, 1, 2)$ are $(1 - MAF)^2, 2MAF(1 - MAF), MAF^2$, respectively.

For simplicity, we assume that excess zeros are caused by technical issues, hence are not related to genotypes.

Value

power if the input parameter `power = NULL`.

sample size (total number of subjects) if the input parameter `n = NULL`;

minimum detectable slope if the input parameter `slope = NULL`;

minimum allowable MAF if the input parameter `MAF = NULL`.

Note

The speed of simulation approach is slow. It is recommended to run the function `powerEQTL.scRNAseq.sim` in parallel computing environment (e.g., Unix/Linux) and set `nCores > 1`. Also, it is important to set appropriate ranges to search sample size, minimum detectable slope, or minimum allowable MAF. If the function (e.g., GLMMadaptive) to fit data returns an error message for a simulated dataset, then the function `powerEQTL.scRNAseq.sim` will try a few more runs with different simulated datasets. If all the runs failed, zero (for power estimation) or `NA` or `NaN` (for estimating slope, sample size, or MAF) would be output. It is recommended to set `nSim >= 1000` to get stable results.

Author(s)

Xianjun Dong <XDONG@rics.bwh.harvard.edu>, Xiaoqi Li<xli85@bwh.harvard.edu>, Tzuu-Wang Chang <Chang.Tzuu-Wang@mgh.harvard.edu>, Scott T. Weiss <restw@channing.harvard.edu>, Weiliang Qiu <weiliang.qiu@gmail.com>

References

Dong X, Li X, Chang T-W, Scherzer CR, Weiss ST, and Qiu W. powerEQTL: An R package and shiny application for sample size and power calculation of bulk tissue and single-cell eQTL analysis. Bioinformatics, 2021;, btab385

Examples

```
nSubj = 10
nCellPerSubj = 10

# calculate power
power = powerEQTL.scRNAseq.sim(
  slope = 1.62, # slope
  n = nSubj, # total number of subjects
  m = nCellPerSubj, # number of cells per subject
  power = NULL, # power to be estimated
  m.int = -1, # mean of random intercept
  sigma.int = 1, # SD of the random intercept
  zero.p = 0.01, # probability that an excess zero occurs
  theta = 1, # dispersion parameter of NB distribution NB(mu, theta)
  MAF = 0.45,
  FWER = 0.05,
  nTests = 1,
  nSim  =  5, # number of simulations
  estMethod = "GLMMadaptive", # parameter estimation method for ZINB
  nCores = 1 # number of computer cores used by 'mclapply'
)

print(power)
```

---

powerEQTL.SLR *Power Calculation for EQTL Analysis Based on Simple Linear Regression*

Description

Power calculation for eQTL analysis that tests if a SNP is associated to a gene probe by using simple linear regression. This function can be used to calculate one of the 4 parameters (power, sample size, minimum detectable slope, and minimum allowable MAF) by setting the corresponding parameter as NULL and providing values for the other 3 parameters.

Usage

```
powerEQTL.SLR(
  MAF,
  slope = 0.13,
  n = 200,
  power = NULL,
  sigma.y = 0.13,
  FWER = 0.05,
  nTests = 2e+05,
  n.lower = 2.01,
  n.upper = 1e+30)
```

Arguments

MAF	numeric. Minor allele frequency.
slope	numeric. Slope of the simple linear regression.
n	integer. Total number of subjects.
power	numeric. Power for testing if the slope is equal to zero.
sigma.y	numeric. Standard deviation of the outcome y_i in simple linear regression.
FWER	numeric. Family-wise Type I error rate.
nTests	integer. Number of tests (i.e., number of all (SNP, gene) pairs) in eQTL analysis.
n.lower	numeric. Lower bound of the total number of subjects. Only used when calculating total number of subjects.
n.upper	numeric. Upper bound of the total number of subjects. Only used when calculating total number of subjects.

Details

To test if a SNP is associated with a gene probe, we use the simple linear regression

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where y_i is the gene expression level of the i-th subject, x_i is the genotype of the i-th subject, and ϵ_i is the random error term with mean zero and standard deviation σ. Additive coding for genotype is used. That is, $x_i = 0$ indicates wildtype homozygotes; $x_i = 1$ indicates heterozygotes; and $x_i = 2$ indicates mutation heterozygotes.

To test if the SNP is associated with the gene probe, we test the null hypothesis $H_0 : \beta_1 = 0$ versus the alternative hypothesis $H_1 : \beta_1 = \delta$, where $\delta \neq 0$.

Denote θ as the minor allele frequency (MAF) of the SNP. Under Hardy-Weinberg equilibrium, we can calculate the variance of genotype of the SNP: $\sigma_x^2 = 2\theta(1-\theta)$, where σ_x^2 is the variance of the predictor (i.e. the SNP) x_i.

The exact power calculation formula can be derived as

$$1 - T_{n-2,\lambda}(t_{n-2}(\alpha/2)) + T_{n-2,\lambda}(-t_{n-2}(\alpha/2)),$$

where $T_{n-2,\lambda}(a)$ is the value at a of cumulative distribution function of non-central t distribution with $n-2$ degrees of freedom and non-centrality parameter $\lambda = \delta/\sqrt{\sigma^2/[(n-1)\tilde{\sigma}_x^2]}$. And $\tilde{\sigma}_x^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)$.

Dupont and Plummer (1998) mentioned the following relationship:

$$\sigma^2 = \sigma_y^2 - \beta_1^2\sigma_x^2.$$

So we can plug in the above equation to the power calculation formula.

Under Hardy-Weinberg equilibrium, we have $\sigma_x^2 = 2\theta(1-\theta)$, where θ is the minor allele frequency (MAF).

Hence, the non-centrality parameter can be rewritten as

$$\lambda = \frac{\delta}{\sqrt{\left(\sigma_y^2 - \delta^2 2\left(1-\hat{\theta}\right)\hat{\theta}\right) / \left[(n-1)2\left(1-\hat{\theta}\right)\hat{\theta}\right]}}$$

We adopted the parameters from the GTEx cohort (see the Power analysis" section of Nature Genetics, 2013; https://www.nature.com/articles/ng.2653), where they modeled the expression data as having a log-normal distribution with a log standard deviation of 0.13 within each genotype class (AA, AB, BB). This level of noise is based on estimates from initial GTEx data. In their power analysis, they assumed the across-genotype difference delta = 0.13 (i.e., equivalent to detecting a log expression change similar to the standard deviation within a single genotype class).

Value

power if the input parameter `power = NULL`.

sample size (total number of subjects) if the input parameter `n = NULL`;

minimum detectable slope if the input parameter `slope = NULL`;

minimum allowable MAF if the input parameter `MAF = NULL`.

Author(s)

Xianjun Dong <XDONG@rics.bwh.harvard.edu>, Xiaoqi Li<xli85@bwh.harvard.edu>, Tzuu-Wang Chang <Chang.Tzuu-Wang@mgh.harvard.edu>, Scott T. Weiss <restw@channing.harvard.edu>, Weiliang Qiu <weiliang.qiu@gmail.com>

References

Dupont, W.D. and Plummer, W.D.. Power and Sample Size Calculations for Studies Involving Linear Regression. Controlled Clinical Trials. 1998;19:589-601.

Dong X, Li X, Chang T-W, Scherzer CR, Weiss ST, and Qiu W. powerEQTL: An R package and shiny application for sample size and power calculation of bulk tissue and single-cell eQTL analysis. Bioinformatics, 2021;, btab385

Examples

```
# calculate power
powerEQTL.SLR(
  MAF = 0.1,
  slope = 0.13,
  n = 179,
  power = NULL,
  sigma.y = 0.13,
  FWER = 0.05,
  nTests = 2e+05)

# calculate sample size (total number of subjects)
powerEQTL.SLR(
  MAF = 0.1,
  slope = 0.13,
  n = NULL,
  power = 0.8,
  sigma.y = 0.13,
  FWER = 0.05,
  nTests = 2e+05)

# calculate minimum detectable slope
powerEQTL.SLR(
  MAF = 0.1,
  slope = NULL,
  n = 179,
  power = 0.8,
  sigma.y = 0.13,
  FWER = 0.05,
  nTests = 2e+05)

# calculate minimum allowable MAF
powerEQTL.SLR(
  MAF = NULL,
  slope = 0.13,
  n = 179,
  power = 0.8,
  sigma.y = 0.13,
  FWER = 0.05,
  nTests = 2e+05)
```

---

powerLME *Power Calculation for Simple Linear Mixed Effects Model*

Description

Power calculation for simple linear mixed effects model. This function can be used to calculate one of the 3 parameters (power, sample size, and minimum detectable slope) by setting the corresponding parameter as NULL and providing values for the other 2 parameters.

Usage

```
powerLME(
  slope,
  n,
  m,
  sigma.y,
  sigma.x,
  power = NULL,
  rho = 0.8,
  FWER = 0.05,
  nTests = 1,
  n.lower = 2.01,
  n.upper = 1e+30)
```

Arguments

slope	numeric. Slope under alternative hypothesis.
n	integer. Total number of subjects.
m	integer. Number of observations per subject.
sigma.y	numeric. Standard deviation of the outcome y.
sigma.x	numeric. Standard deviation of the predictor x.
power	numeric. Desired power.
rho	numeric. Intra-class correlation (i.e., correlation between y_{ij} and y_{ik} for the j-th and k-th observations of the i-th subject).
FWER	numeric. Family-wise Type I error rate.
nTests	integer. Number of tests (e.g., number of genes in differential expression analysis based on scRNAseq to compare gene expression between diseased subjects and healthy subjects).
n.lower	numeric. Lower bound of the total number of subjects. Only used when calculating toal number of subjects.
n.upper	numeric. Upper bound of the total number of subjects. Only used when calculating total number of subjects.

Details

We assume the following simple linear mixed effects model to characterize the association between the predictor x and the outcome y:

$$y_{ij} = \beta_{0i} + \beta_1 * x_i + \epsilon_{ij},$$

where

$$\beta_{0i} \sim N\left(\beta_0, \sigma^2_\beta\right),$$

and

$$\epsilon_{ij} \sim N\left(0, \sigma^2\right),$$

$i = 1, \ldots, n$, $j = 1, \ldots, m$, n is the number of subjects, m is the number of observations per subject, y_{ij} is the outcome value for the j-th observation of the i-th subject, x_i is the predictor value for the i-th subject. For example, x_i is the binary variable indicating if the i-th subject is a diseased subject or not.

We would like to test the following hypotheses:

$$H_0 : \beta_1 = 0,$$

and

$$H_1 : \beta_1 = \delta,$$

where $\delta \neq 0$.

We can derive the power calculation formula is

$$power = 1 - \Phi\left(z_{\alpha^*/2} - a \times b\right) + \Phi\left(-z_{\alpha^*/2} - a \times b\right),$$

where

$$a = \frac{\hat{\sigma}_x}{\sigma_y}$$

and

$$b = \frac{\delta\sqrt{m(n-1)}}{\sqrt{1 + (m-1)\rho}}$$

and $z_{\alpha^*/2}$ is the upper $100\alpha^*/2$ percentile of the standard normal distribution, $\alpha^* = \alpha/nTests$, nTests is the number of tests, $\sigma_y = \sqrt{\sigma_\beta^2 + \sigma^2}$, $\hat{\sigma}_x = \sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2/(n-1)}$, and $\rho = \sigma_\beta^2/\left(\sigma_\beta^2 + \sigma^2\right)$ is the intra-class correlation.

Value

power if the input parameter `power = NULL`.

sample size (total number of subjects) if the input parameter `n = NULL`;

minimum detectable slope if the input parameter `slope = NULL`.

Author(s)

Xianjun Dong <XDONG@rics.bwh.harvard.edu>, Xiaoqi Li<xli85@bwh.harvard.edu>, Tzuu-Wang Chang <Chang.Tzuu-Wang@mgh.harvard.edu>, Scott T. Weiss <restw@channing.harvard.edu>, Weiliang Qiu <weiliang.qiu@gmail.com>

References

Dong X, Li X, Chang T-W, Scherzer CR, Weiss ST, and Qiu W. powerEQTL: An R package and shiny application for sample size and power calculation of bulk tissue and single-cell eQTL analysis. Bioinformatics, 2021;, btab385

Examples

```
n = 102
m = 227868

# calculate power
power = powerLME(
  slope = 0.6,
  n = n,
  m = m,
  sigma.y = 0.29,
  sigma.x = 0.308,
  power = NULL,
  rho = 0.8,
  FWER = 0.05,
  nTests = 1e+6)

print(power)

# calculate sample size (total number of subjects)
n = powerLME(
  slope = 0.6,
  n = NULL,
  m = m,
  sigma.y = 0.29,
  sigma.x = 0.308,
  power = 0.9562555,
  rho = 0.8,
  FWER = 0.05,
  nTests = 1e+6)

print(n)

# calculate slope
slope = powerLME(
  slope = NULL,
  n = n,
  m = m,
  sigma.y = 0.29,
  sigma.x = 0.308,
  power = 0.9562555,
  rho = 0.8,
  FWER = 0.05,
  nTests = 1e+6)

print(slope)
```

`powerLMEnoCov` *Power Calculation for Simple Linear Mixed Effects Model Without Covariate*

Description

Power calculation for simple linear mixed effects model without covariate. This function can be used to calculate one of the 3 parameters (power, sample size, and minimum detectable slope) by setting the corresponding parameter as NULL and providing values for the other 2 parameters.

Usage

```
powerLMEnoCov(
  slope,
  n,
  m,
  sigma.y,
  power = NULL,
  rho = 0.8,
  FWER = 0.05,
  nTests = 1,
  n.lower = 2.01,
  n.upper = 1e+30)
```

Arguments

`slope`	numeric. Slope under alternative hypothesis.
`n`	integer. Total number of subjects.
`m`	integer. Number of pairs of replicates per subject.
`sigma.y`	numeric. Standard deviation of the outcome y.
`power`	numeric. Desired power.
`rho`	numeric. Intra-class correlation (i.e., correlation between y_{ij} and y_{ik} for the j-th and k-th observations of the i-th subject).
`FWER`	numeric. Family-wise Type I error rate.
`nTests`	integer. Number of tests (e.g., number of genes in differential expression analysis based on scRNAseq to compare gene expression before and after treatment).
`n.lower`	numeric. Lower bound of the total number of subjects. Only used when calculating total number of subjects.
`n.upper`	numeric. Upper bound of the total number of subjects. Only used when calculating total number of subjects.

Details

In an experiment, there are n samples. For each sample, we get m pairs of replicates. For each pair, one replicate will receive no treatment. The other replicate will receive treatment. The outcome is the expression of a gene. The overall goal of the experiment is to check if the treatment affects gene

expression level or not. Or equivalently, the overall goal of the experiment is to test if the mean within-pair difference of gene expression is equal to zero or not. In the design stage, we would like to calculate the power/sample size of the experiment for testing if the mean within-pair difference of gene expression is equal to zero or not.

We assume the following linear mixed effects model to characterize the relationship between the within-pair difference of gene expression y_{ij} and the mean of the within-pair difference β_{0i}:

$$y_{ij} = \beta_{0i} + \epsilon_{ij},$$

where

$$\beta_{0i} \sim N\left(\beta_0, \sigma_\beta^2\right),$$

and

$$\epsilon_{ij} \sim N\left(0, \sigma^2\right),$$

$i = 1, \ldots, n$, $j = 1, \ldots, m$, n is the number of subjects, m is the number of pairs of replicates per subject, y_{ij} is the within-pair difference of outcome value for the j-th pair of the i-th subject.

We would like to test the following hypotheses:

$$H_0 : \beta_0 = 0,$$

and

$$H_1 : \beta_0 = \delta,$$

where $\delta \neq 0$. If we reject the null hypothesis H_0 based on a sample, we then get evidence that the treatment affects the gene expression level.

We can derive the power calculation formula:

$$power = 1 - \Phi\left(z_{\alpha^*/2} - a \times b\right) + \Phi\left(-z_{\alpha^*/2} - a \times b\right),$$

where

$$a = \frac{1}{\sigma_y}$$

and

$$b = \frac{\delta\sqrt{mn}}{\sqrt{1 + (m-1)\rho}}$$

and $z_{\alpha^*/2}$ is the upper $100\alpha^*/2$ percentile of the standard normal distribution, $\alpha^* = \alpha/nTests$, nTests is the number of tests, $\sigma_y = \sqrt{\sigma_\beta^2 + \sigma^2}$ and $\rho = \sigma_\beta^2 / \left(\sigma_\beta^2 + \sigma^2\right)$ is the intra-class correlation.

Value

power if the input parameter `power = NULL`.

sample size (total number of subjects) if the input parameter `n = NULL`;

minimum detectable slope if the input parameter `slope = NULL`.

Author(s)

Xianjun Dong <XDONG@rics.bwh.harvard.edu>, Xiaoqi Li<xli85@bwh.harvard.edu>, Tzuu-Wang Chang <Chang.Tzuu-Wang@mgh.harvard.edu>, Scott T. Weiss <restw@channing.harvard.edu>, Weiliang Qiu <weiliang.qiu@gmail.com>

References

Dong X, Li X, Chang T-W, Scherzer CR, Weiss ST, and Qiu W. powerEQTL: An R package and shiny application for sample size and power calculation of bulk tissue and single-cell eQTL analysis. Bioinformatics, 2021;, btab385

Examples

```
n = 17
m = 5
sigma.y = 0.68
slope = 1.3*sigma.y
print(slope)

# estimate power
power = powerLMEnoCov(
  slope = slope,
  n = n,
  m = m,
  sigma.y = sigma.y,
  power = NULL,
  rho = 0.8,
  FWER = 0.05,
  nTests = 20345)

print(power)

# estimate sample size (total number of subjects)
n = powerLMEnoCov(
  slope = slope,
  n = NULL,
  m = m,
  sigma.y = sigma.y,
  power = 0.8721607,
  rho = 0.8,
  FWER = 0.05,
  nTests = 20345)

print(n)

# estimate slope
slope = powerLMEnoCov(
  slope = NULL,
  n = n,
  m = m,
  sigma.y = sigma.y,
  power = 0.8721607,
  rho = 0.8,
  FWER = 0.05,
  nTests = 20345)

print(slope)
```

simDat.eQTL.scRNAseq *Generate Gene Expression Levels Of One Gene And Genotypes Of One SNP For Subjects With Multiple Cells Based On ZINB Mixed Effects Regression Model*

Description

Generate gene expression levels of one gene and genotypes of one SNP for subjects with multiple cells based on ZINB mixed effects regression model.

Usage

```
simDat.eQTL.scRNAseq(nSubj = 50,
     nCellPerSubj = 100,
     zero.p = 0.01,
     m.int = 0,
     sigma.int = 1,
     slope = 1,
     theta = 1,
     MAF = 0.45)
```

Arguments

nSubj	integer. Total number of subjects.
nCellPerSubj	integer. Number of cells per subject.
zero.p	numeric. Probability that an excess zero occurs.
m.int	numeric. Mean of random intercept (see details).
sigma.int	numeric. Standard deviation of random intercept (see details).
slope	numeric. Slope (see details).
theta	numeric. dispersion parameter of negative binomial distribution. The smaller theta is, the larger variance of NB random variable is.
MAF	numeric. Minor allele frequency of the SNP.

Details

This function simulates gene expression levels of one gene and genotypes of one SNP for subjects with multiple cells based on zero-inflated negative binomial (ZINB) regression model with only one covariate: genotype. That is, the read counts of a gene follows a mixture of 2-component distributions. One component takes only one value: zero. The other component is negative binomial distribution, which takes non-negative values 0, 1, 2, ... The log mean of the negative binomial distribution is linear function of the genotype.

Denote Y_{ij} as the read counts for the j-th cell of the i-th subject, $i = 1, \ldots, n$, $j = 1, \ldots, m$, n is the number of subjects, and m is the number of cells per subject.

Denote p as the probability that $Y_{ij} = 0$ is an excess zero. With probability $1 - p$, Y_{ij} follows a negative binomial distribution $NB(\mu, \theta)$, where μ is the mean (i.e., $\mu = E(Y_{ij})$) and θ is the

dispersion parameter. The variance of the NB distribution is $\mu + \mu^2/\theta$. The relationship between gene expression and genotype for the i-th subject is characterized by the equation

$$\mu_i = \exp(\beta_{0i} + \beta_1 x_i),$$

where β_{0i} is the random intercept following a normal distribution $N(\beta_0, \sigma^2)$ to account for within-subject correlation of gene expression, β_0 is the mean of the random intercept, σ is the standard deviation of the random intercept, β_1 is the slope, and x_i is the additive-coded genotype for the SNP with minor allele frequency MAF.

We assume that the SNP satisfies the Hardy-Weinberg Equilibrium. That is, the probabilities of the 3 genotypes $(0, 1, 2)$ are $(1 - MAF)^2, 2MAF(1 - MAF), MAF^2$, respectively.

For simplicity, we assume that excess zeros are caused by technical issues, hence are not related to genotypes.

Value

A data frame with 3 columns:

`id`	subject id
`geno`	additive-coded genotype of the SNP
`counts`	gene expression of the gene

Author(s)

Xianjun Dong <XDONG@rics.bwh.harvard.edu>, Xiaoqi Li<xli85@bwh.harvard.edu>, Tzuu-Wang Chang <Chang.Tzuu-Wang@mgh.harvard.edu>, Scott T. Weiss <restw@channing.harvard.edu>, Weiliang Qiu <weiliang.qiu@gmail.com>

References

Dong X, Li X, Chang T-W, Scherzer CR, Weiss ST, and Qiu W. powerEQTL: An R package and shiny application for sample size and power calculation of bulk tissue and single-cell eQTL analysis. Bioinformatics, 2021;, btab385

Examples

```
frame = simDat.eQTL.scRNAseq(nSubj = 5,
     nCellPerSubj = 3,
     zero.p = 0.01,
     m.int = 0,
     sigma.int = 1,
     slope = 1,
     theta = 1,
     MAF = 0.45)
print(dim(frame))
print(frame[1:10,])
```

Index