

Package ‘gJLS2’

October 13, 2022

Title A Generalized Joint Location and Scale Framework for Association Testing

Version 0.2.0

Description An update to the Joint Location-Scale (JLS) testing framework that identifies associated SNPs, gene-sets and pathways with main and/or interaction effects on quantitative traits (Soave et al., 2015; <[doi:10.1016/j.ajhg.2015.05.015](https://doi.org/10.1016/j.ajhg.2015.05.015)>). The JLS method simultaneously tests the null hypothesis of equal mean and equal variance across genotypes, by aggregating association evidence from the individual location/mean-only and scale/variance-only tests using Fisher's method. The generalized joint location-scale (gJLS) framework has been developed to deal specifically with sample correlation and group uncertainty (Soave and Sun, 2017; <[doi:10.1111/biom.12651](https://doi.org/10.1111/biom.12651)>). The current release: gJLS2, include additional functionalities that enable analyses of X-chromosome genotype data through novel methods for location (Chen et al., 2021; <[doi:10.1002/gepi.22422](https://doi.org/10.1002/gepi.22422)>) and scale (Deng et al., 2019; <[doi:10.1002/gepi.22247](https://doi.org/10.1002/gepi.22247)>).

License GPL (>= 3)

Encoding UTF-8

LazyData true

RoxygenNote 7.1.1

Suggests knitr, markdown

VignetteBuilder knitr

Imports methods, nlme, quantreg, MCMCpack, MASS, plyr, ggplot2, moments

Depends R (>= 3.6.0)

NeedsCompilation no

Author Wei Deng [aut, cre],
Lei Sun [aut]

Maintainer Wei Deng <deng@utstat.toronto.edu>

Repository CRAN

Date/Publication 2021-09-30 09:00:05 UTC

R topics documented:

chrXdat	2
gJLS2	3
gJLS2s	5
levneRegA_per_SNP	7
levneRegX_per_SNP	9
levneTests_per_SNP	11
locReg	13
scaleReg	15

Index	19
--------------	-----------

chrXdat	<i>X-chromosomal example using the 1000 Genomes Project data</i>
---------	--

Description

A dataset containing real genotypes of 5 X-chromosomal SNPs and simulated phenotype of 473 unrelated samples from the 1000 Genomes Project.

Usage

```
chrXdat
```

Format

A data frame with 473 rows and 11 variables:

FID Family ID.

IID Individual ID.

PAT Paternal individual ID.

MAT Maternal individual ID.

SEX The genetic sex of a individual.

PHENOTYPE The quantitative trait value, simulated from a standard normal distribution.

rs5983012_A The genotype of SNP with MAF roughly equal to 0.1 in females.

rs4119090_G The genotype of SNP with MAF roughly equal to 0.2 in females.

rs5911042_T The genotype of SNP with MAF roughly equal to 0.3 in females.

rs986810_C The genotype of SNP with MAF roughly equal to 0.4 in females.

rs180495_G The genotype of SNP with MAF roughly equal to 0.45 in females.

Source

<https://www.internationalgenome.org/data/>

Description

This function takes as input the genotype of a SNP (GENO), the SEX (SEX), and a quantitative trait (Y) in a sample population, and possibly additional covariates, such as principal components. The function returns the location and scale association p -values for each SNP, as well as the gJLS p -value, which provides the combined evidence via Fisher's method (Soave et al., 2015, 2017). To perform this analysis genome-wide, we recommend to use the R-plugin written for PLINK, see gJLSPLINK for more details.

Usage

```
gJLS2(  
  GENO,  
  Y,  
  COVAR = NULL,  
  SEX = NULL,  
  Xchr = FALSE,  
  transformed = TRUE,  
  loc_alg = "LAD",  
  related = FALSE,  
  cov.structure = "corCompSymm",  
  clust = NULL,  
  genotypic = FALSE,  
  origLev = FALSE,  
  centre = "median",  
  XchrMethod = 3  
)
```

Arguments

GENO	a list of a genotype matrix/vector of SNPs, must contain values 0, 1, 2's coded for the number of reference allele. Alternatively, for imputed genotypes, it could either be a vector of dosage values between 0 and 2, or a list of matrix of genotype probabilities, numerically between 0 and 1 for each genotype. The length/dimension of GENO should match that of Y, and/or SEX and COVAR.
Y	a vector of quantitative traits, such as human height.
COVAR	optional: a vector or matrix of covariates that are used to reduce bias due to confounding, such as age.
SEX	optional: the genetic sex of individuals in the sample population, must be a vector of 1 and 2 following the default sex code is 1 for males and 2 for females in PLINK.
Xchr	a logical indicator for whether the analysis is for X-chromosome SNPs.

transformed	a logical indicating whether the quantitative response Y should be transformed using a rank-based method to resemble a normal distribution; recommended for traits with non-symmetric distribution. The default option is TRUE.
loc_alg	a character indicating the type of algorithm to compute the centre in stage 1; the value is either "OLS", corresponding to an ordinary linear regression under Gaussian assumptions to compute the mean, or "LAD", corresponding to a quantile regression to compute the median. The recommended default option is "LAD". For the quantile regression, the function calls <code>quantreg::rq</code> and the median is estimated using either the "br" (smaller samples) or "sfn" (larger samples and sparse problems) algorithm depending the sample size, for more details see <code>?quantreg::rq</code> .
related	optional: a logical indicating whether the samples should be treated as related; if TRUE while no relatedness covariance information is given, it is then estimated under a <code>cov.structure</code> and assumes this structure among all within-group errors pertaining to the same pair/cluster if specified using <code>clust</code> . This option currently only applies to autosomal SNPs.
cov.structure	optional: should be one of standard classes of correlation structures listed in <code>corClasses</code> from R package nlme . See <code>?corClasses</code> . The most commonly used option is <code>corCompSymm</code> for a compound symmetric correlation structure. This option currently only applies to autosomal SNPs.
clust	optional: a factor indicating the grouping of samples; it should have at least two distinct values. It could be the family ID (FID) for family studies. This option currently only applies to autosomal SNPs.
genotypic	a logical indicating whether the variance homogeneity should be tested with respect to an additively (linearly) coded or non-additively coded <code>geno_one</code> . The former has one less degree of freedom than the latter and is the default option. For dosage genotypes without genotypic probabilities, <code>genotypic</code> is forced to be FALSE.
origLev	a logical indicator for whether the reported p-values should also include original Levene's test.
centre	a character indicating whether the absolute deviation should be calculated with respect to "median" or "mean" in the traditional sex-specific and Fisher combined Levene's test p-values (three tests) for X-chromosome. The default value is "median". This option applies to sex-specific analysis using original Levene's test (i.e. when <code>regression\$\$=\$\$TRUE</code>).
XchrMethod	an integer taking values 0 (reports all models), 1,1, 1,2, 2, 3, for the choice of X-chromosome location association testing models; for more details, see locReg .

Value

a vector of location, scale and combined gJLS p-values for each SNP.

Note

For a genome-scan, we recommend to run this in PLINK via the plugin function `gJLSPLINK`, especially for large datasets and those with more than 20 covariates.

We highly recommend to quantile-normally transform Y for non-symmetrically distributed traits. This is typically done to avoid ‘scale-effect’ when the variance values tend to be proportional to mean values when stratified by $GEN0$, as observed by Pare et al. (2010) and Yang et al. (2011).

For the moment, only quantitative trait Y is accepted as the subsequent generalized joint location scale (gJLS) analyses require the variance be calculated on quantitative traits. However, we are working on to include binary response for the generalized JLS analyses in the next update of gJLS.

Author(s)

Wei Q. Deng <deng@utstat.toronto.edu>, Lei Sun <sun@utstat.toronto.edu>

References

Soave D, Corvol H, Panjwani N, Gong J, Li W, Boëlle PY, Durie PR, Paterson AD, Rommens JM, Strug LJ, Sun L. (2015). A Joint Location-Scale Test Improves Power to Detect Associated SNPs, Gene Sets, and Pathways. *American Journal of Human Genetics*. 2015 Jul 2;**97**(1):125-38. doi: [10.1016/j.ajhg.2015.05.015](https://doi.org/10.1016/j.ajhg.2015.05.015). PMID: 26140448; PMCID: PMC4572492.

Examples

```
N <- 1000
genDAT <- rbinom(N, 2, 0.3)
sex <- rbinom(N, 1, 0.5)+1
y <- rnorm(N)
covar <- matrix(rnorm(N*10), ncol=10)

gJLS2(GENO=data.frame("SNP1" = genDAT, "aSNP1" = genDAT), SEX=sex, Y=y, COVAR=covar)

gJLS2(GENO=genDAT, SEX=sex, Y=y, COVAR=covar , Xchr=TRUE)
```

gJLS2s

generalized Joint-Location-Scale (gJLS) test with summary statistics

Description

This function takes as input the gL and gS p-values for each SNP and combine to produce the gJLS p-values. It is used for genome-wide analysis where only the gL or gS p-values are available, caution should be exercised when combing gL and gS p-values obtained from separate datasets.

Usage

```
gJLS2s(gL, gS)
```

Arguments

gL	a vector of location p-values or a data.frame containing column names "SNP" and "gL".
gS	a vector of scale p-values or a data.frame containing column names "SNP" and "gS".

Value

a vector of combined gJLS p-values for each SNP.

Note

For a genome-scan, we recommend to run this in PLINK via the plugin function gJLSPLINK, especially for large datasets and those with more than 20 covariates.

We highly recommend to quantile-normally transform Y for non-symmetrically distributed traits. This is typically done to avoid ‘scale-effect’ when the variance values tend to be proportional to mean values when stratified by GEN0, as observed by Pare et al. (2010) and Yang et al. (2011).

For the moment, only quantitative trait Y is accepted as the subsequent generalized joint location scale (gJLS) analyses require the variance be calculated on quantitative traits. However, we are working on to include binary response for the generalized JLS analyses in the next update of gJLS.

Author(s)

Wei Q. Deng <deng@utstat.toronto.edu>, Lei Sun <sun@utstat.toronto.edu>

References

Soave D, Corvol H, Panjwani N, Gong J, Li W, Boëlle PY, Durie PR, Paterson AD, Rommens JM, Strug LJ, Sun L. (2015). A Joint Location-Scale Test Improves Power to Detect Associated SNPs, Gene Sets, and Pathways. *American Journal of Human Genetics*. 2015 Jul 2;**97**(1):125-38. doi: [10.1016/j.ajhg.2015.05.015](https://doi.org/10.1016/j.ajhg.2015.05.015). PMID: 26140448; PMCID: PMC4572492.

Examples

```
gL <- data.frame("SNP" = paste("rs", 1:100, sep=""), "gL"=runif(100))
gS <- runif(100)

gJLS2s(gL = gL, gS=gS)
```

leveneRegA_per_SNP	<i>The generalized Levene's test via a two-stage regression for variance homogeneity by SNP genotype (autosomes)</i>
--------------------	--

Description

This function takes as input the genotype of a SNP (GENO), a quantitative trait (Y) in a sample population, and possibly additional covariates, such as principal components. The function returns the scale association p -values for each autosomal SNP using the generalized Levene's test.

Usage

```
leveneRegA_per_SNP(
  geno_one,
  Y,
  COVAR = NULL,
  transformed = TRUE,
  loc_alg = "LAD",
  related = FALSE,
  cov.structure = "corCompSymm",
  clust = NULL,
  genotypic = FALSE
)
```

Arguments

geno_one	the genotype of a biallelic SNP, must be a vector of 0, 1, 2's coded for the number of reference allele. Alternatively, for imputed genotypes, it could be a matrix/vector of dosage values, numerically between 0 and 2. The length/dimension of geno_one should match that of Y, and/or SEX and COVAR.
Y	a vector of quantitative traits, such as human height.
COVAR	optional: a vector or matrix of covariates that are used to reduce bias due to confounding, such as age.
transformed	a logical indicating whether the quantitative response Y should be transformed using a rank-based method to resemble a normal distribution; recommended for traits with non-symmetric distribution. The default option is TRUE.
loc_alg	a character indicating the type of algorithm to compute the centre in stage 1; the value is either "OLS", corresponding to an ordinary linear regression under Gaussian assumptions to compute the mean, or "LAD", corresponding to a quantile regression to compute the median. The recommended default option is "LAD". For the quantile regression, the function calls <code>quantreg::rq</code> and the median is estimated using either the "fn" (smaller samples) or "sfn" (larger samples and sparse problems) algorithm depending the sample size, for more details see <code>?quantreg::rq</code> .

related	optional: a logical indicating whether the samples should be treated as related; if TRUE while no relatedness covariance information is given, it is then estimated under a <code>cov.structure</code> and assumes this structure among all within-group errors pertaining to the same pair/cluster if specified using <code>clust</code> . This option currently only applies to autosomal SNPs.
<code>cov.structure</code>	optional: should be one of standard classes of correlation structures listed in <code>corClasses</code> from R package nlme . See <code>?corClasses</code> . The most commonly used option is <code>corCompSymm</code> for a compound symmetric correlation structure. This option currently only applies to autosomal SNPs.
<code>clust</code>	optional: a factor indicating the grouping of samples; it should have at least two distinct values. It could be the family ID (FID) for family studies. This option currently only applies to autosomal SNPs.
genotypic	optional: a logical indicating whether the variance homogeneity should be tested with respect to an additively (linearly) coded or non-additively coded <code>geno_one</code> . The former has one less degree of freedom than the latter and is the default option. For dosage data without genotypic probabilities, <code>genotypic</code> is forced to be FALSE.

Value

Levene's test regression p-values for autosomal SNPs according to the model specified.

Note

We recommend to quantile-normally transform Y to avoid 'scale-effect' where the variance values tend to be proportional to mean values when stratified by `geno_one`.

When the `relatedness` option is used, the computational time is expected to be longer for larger sample size ($n > 1000$), thus we recommend this option for smaller studies rather than large population based studies.

There is no explicit argument to supply `SEX` for autosomal SNPs, the user can choose to include the genetic sex of individuals as a column of the `COVAR` argument.

Author(s)

Wei Q. Deng <deng@utstat.toronto.edu>, Lei Sun <sun@utstat.toronto.edu>

References

Soave D, Sun L. (2017). A generalized Levene's scale test for variance heterogeneity in the presence of sample correlation and group uncertainty. *Biometrics*. **73**(3):960-971. doi: [10.1111/biom.12651](https://doi.org/10.1111/biom.12651). PMID: 28099998.

Gastwirth JL, Gel YR, Miao W. (2009). The Impact of Levene's Test of Equality of Variances on Statistical Theory and Practice." *Statistical Science*. **24**(3) 343 - 360, doi: [10.1214/09STS301](https://doi.org/10.1214/09STS301)

Examples

```

N <- 100
genDAT <- rbinom(N, 2, 0.3)
Y <- rnorm(N)
covar <- matrix(rnorm(N*10), ncol=10)

# vanilla example:
leveneRegA_per_SNP(geno_one=genDAT, Y=Y, COVAR=covar)

# relatedness samples:
leveneRegA_per_SNP(geno_one=genDAT, Y=Y, COVAR=covar,
related=TRUE)
leveneRegA_per_SNP(geno_one=genDAT, Y=Y, COVAR=covar,
related=TRUE, clust = factor(rbinom(N, 2, 0.6)))

# dosage genotypes example:
library("MCMCpack")
a <- 0.3
geno <- rbinom(N, 2, 0.3)
a <- 0.3 ## uncertainty
genPP <- rbind(rdirichlet(sum(geno==0),c(a,(1-a)/2,(1-a)/2)),
               rdirichlet(sum(geno==1),c((1-a)/2,a,(1-a)/2)),
               rdirichlet(sum(geno==2),c((1-a)/2,(1-a)/2,a)))

leveneRegA_per_SNP(geno_one=genPP, Y=Y, COVAR=covar)
leveneRegA_per_SNP(geno_one=genPP, Y=Y, COVAR=covar,
genotypic=TRUE)

# dosage and related samples:
leveneRegA_per_SNP(geno_one=genPP, Y=Y, COVAR=covar,
related=TRUE, clust = factor(rbinom(N, 1, 0.3)))
leveneRegA_per_SNP(geno_one=genPP, Y=Y, COVAR=covar,
related=TRUE, clust = factor(rbinom(N, 1, 0.3)), genotypic=TRUE)

```

leveneRegX_per_SNP	<i>Levene's regression tests for variance homogeneity by SNP genotype (X-chromosome specific)</i>
--------------------	---

Description

This function takes as input the genotype of a SNP (`geno_one`), the genetic sex (`SEX`), a quantitative trait (`Y`) in a sample population, and possibly additional covariates, such as principal components. The function returns the scale association p -values for each X-chromosome SNP using the generalized Levene's test designed for X-chromosome biallelic markers.

Usage

```
leveneRegX_per_SNP(
  geno_one,
  SEX,
  Y,
  COVAR = NULL,
  genotypic = FALSE,
  transformed = TRUE,
  loc_alg = "LAD"
)
```

Arguments

geno_one	the genotype of a biallelic SNP, must be a vector of 0, 1, 2's coded for the number of reference allele. Alternatively, for imputed genotypes, it could be a matrix/vector of dosage values, numerically between 0 and 2. The length/dimension of geno_one should match that of Y, and/or SEX and COVAR.
SEX	optional: the genetic sex of individuals in the sample population, must be a vector of 1 and 2 following the default sex code is 1 for males and 2 for females in PLINK.
Y	a vector of quantitative traits, such as human height.
COVAR	optional: a vector or matrix of covariates that are used to reduce bias due to confounding, such as age.
genotypic	optional: a logical indicating whether the variance homogeneity should be tested with respect to an additively (linearly) coded or non-additively coded geno_one. The former has one less degree of freedom than the latter and is the default option. For dosage genotypes without genotypic probabilities, genotypic is forced to be FALSE.
transformed	a logical indicating whether the quantitative response Y should be transformed using a rank-based method to resemble a normal distribution; recommended for traits with non-symmetric distribution. The default option is TRUE.
loc_alg	a character indicating the type of algorithm to compute the centre in stage 1; the value is either "OLS", corresponding to an ordinary linear regression under Gaussian assumptions to compute the mean, or "LAD", corresponding to a quantile regression to compute the median. The recommended default option is "LAD". For the quantile regression, the function calls <code>quantreg::rq</code> and the median is estimated using either the "fn" (smaller samples) or "sfn" (larger samples and sparse problems) algorithm depending the sample size, for more details see <code>?quantreg::rq</code> .

Value

the Levene's test regression p-value according to the model specified.

Note

We recommend to quantile-normally transform Y to avoid 'scale-effect' where the variance values tend to be proportional to mean values when stratified by `geno_one`.

Author(s)

Wei Q. Deng <deng@utstat.toronto.edu>, Lei Sun <sun@utstat.toronto.edu>

References

Deng WQ, Mao S, Kalnapekis A, Esko T, Magi R, Pare G, Sun L. (2019) Analytical strategies to include the X-chromosome in variance heterogeneity analyses: Evidence for trait-specific polygenic variance structure. *Genet Epidemiol.* **43**(7):815-830. doi: [10.1002/gepi.22247](https://doi.org/10.1002/gepi.22247). PMID:31332826.

Gastwirth JL, Gel YR, Miao W. (2009). The Impact of Levene's Test of Equality of Variances on Statistical Theory and Practice. *Statistical Science.* **24**(3) 343-360, doi: [10.1214/09STS301](https://doi.org/10.1214/09STS301).

Examples

```
N <- 1000
sex <- rbinom(N, 1, 0.5)+1
Y <- rnorm(N)
genDAT <- NA
genDAT[sex==2] <- rbinom(sum(sex==2), 2, 0.3)
table(genDAT, sex)
genDAT[sex==1] <- rbinom(sum(sex==1), 1, 0.3)
table(genDAT, sex)

leveneRegX_per_SNP(geno_one=genDAT, SEX=sex, Y=Y)
leveneRegX_per_SNP(geno_one=genDAT, SEX=sex, Y=Y, genotypic=TRUE)
leveneRegX_per_SNP(geno_one=genDAT, SEX=sex, Y=Y, loc_alg="OLS")
```

leveneTests_per_SNP	<i>Levene's test for variance homogeneity by SNP genotypes (sex-specific p-values)</i>
---------------------	--

Description

This function takes as input the genotype of a SNP (`geno_one`), the genetic sex (`SEX`), a quantitative trait (`Y`) in a sample population. The function then returns the variance heterogeneity p -values for each sex and the overall variance heterogeneity signal using Fisher's method by combining the sex-specific results.

Usage

```
leveneTests_per_SNP(
  geno_one,
  SEX = NULL,
  Y,
  centre = "median",
  transformed = TRUE
)
```

Arguments

geno_one	the genotype of a bi-allelic SNP, must be a vector of 0, 1, 2's coded for the number of reference allele. Alternatively, for imputed genotypes, it could be a matrix/vector of dosage values, numerically between 0 and 2. The length/dimension of geno_one should match that of Y, and/or SEX and COVAR.
SEX	optional: the genetic sex of individuals in the sample population, must be a vector of 1 and 2 following the default sex code is 1 for males and 2 for females in PLINK.
Y	a vector of quantitative traits, such as human height.
centre	a character indicating whether the absolute deviation should be calculated with respect to "median" or "mean", the default option is "median".
transformed	a logical indicating whether the quantitative response Y should be transformed using a rank-based method to resemble a normal distribution; recommended for traits with non-symmetric distribution. The default option is TRUE.

Value

a vector of Levene's test p-values according to levels specified by geno_one in each sex and the Fisher's method to combine the sex-specific Levene's test p-values.

Note

We recommend to quantile-normally transform Y to avoid 'scale-effect' where the variance values tend to be proportional to mean values when stratified by G.

Author(s)

Wei Q. Deng <deng@utstat.toronto.edu>, Lei Sun <sun@utstat.toronto.edu>

References

Levene H. (1960) Robust tests for equality of variances. In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling* eds:I. Olkin, S.G. Ghurye, W. Hoeffding, W.G. Madow & H.B.Mann, pp.278-292. Stanford: Stanford University Press.

Examples

```
N <- 5000
sex <- rbinom(N, 1, 0.5)+1
genDAT <- rbinom(N, 2, 0.3)
y <- rnorm(N);

genDAT[sex==2] <- rbinom(sum(sex==2), 1, 0.3)
table(genDAT, sex)
leveneTests_per_SNP(geno_one=genDAT, SEX=sex, Y=y^2, transform=TRUE)

genDAT[sex==2] <- rbinom(sum(sex==2), 1, 0.01)
table(genDAT, sex)
leveneTests_per_SNP(geno_one=genDAT, SEX=sex, Y=y^2, transform=FALSE)
```

```
leveneTests_per_SNP(geno_one=rep(0, N), SEX=sex, Y=y^2, transform=TRUE)
leveneTests_per_SNP(geno_one=rep(0, N), Y=y^2, transform=TRUE)
```

locReg

*Location (mean-based association) test***Description**

This function takes as input the genotype of SNPs (GENO), the SEX (SEX), and a quantitative trait (Y) in a sample population, and possibly additional covariates, such as principal components. The function returns the location association p -values for each SNP.

Usage

```
locReg(
  GENO,
  Y,
  SEX = NULL,
  COVAR = NULL,
  Xchr = FALSE,
  XchrMethod = 3,
  transformed = FALSE,
  related = FALSE,
  cov.structure = "corCompSymm",
  clust = NULL
)
```

Arguments

GENO	a list of a genotype matrix/vector of SNPs, must contain values 0, 1, 2's coded for the number of reference allele. Alternatively, for imputed genotypes, it could either be a vector of dosage values between 0 and 2, or a list of matrix of genotype probabilities, numerically between 0 and 1 for each genotype. The length/dimension of GENO should match that of Y, and/or SEX and COVAR.
Y	a numeric vector of quantitative trait, such as human height.
SEX	the genetic sex of individuals in the sample population, must be a vector of 1's and 2's following PLINK default coding, where males are coded as 1 and females 2. Optional for analysis of autosomal SNPs, but required for X-chromosome.
COVAR	optional: a vector or a matrix of covariates, such as age or principal components.
Xchr	a logical indicator for whether the analysis is for X-chromosome SNPs, if TRUE then the following association testing model is used: $Y \sim G + G_D + S + G \times S$; with p -value given by comparing $Y \sim G + S + G \times S$ vs. $Y \sim S$ (G is the additive coded genotype; G_D is an indicator for female heterozygotes).

XchrMethod	an integer taking values 0 (reports all models), 1.1, 1.2, 2, 3, for the choice of X-chromosome association testing models: model 1.1: $Y \sim G$ (females only) model 1.2: $Y \sim G$ (males only) model 2: $Y \sim G + S + G \times \text{Sex}$; with p-value given by comparing $Y \sim G + \text{Sex} + G \times \text{Sex}$ vs. $Y \sim \text{Sex}$ (the additively coded G is robust to X-chromosome inactivation uncertainty). This is also the option for dosage genotypes. model 3 (recommended): $Y \sim G + G_D + S + G \times \text{Sex}$; with p-value given by comparing $Y \sim G + G_D + \text{Sex} + G \times \text{Sex}$ vs. $Y \sim \text{Sex}$ (G_D is an indicator for female heterozygotes, this model is robust to X-chromosome inactivation uncertainty and skewed inactivation). For imputed data in the form of genotypic probabilities, the model becomes: $Y \sim G1 + G2 + G1 \times \text{Sex} + \text{Sex}$, where G1 and G2 are the genotypic probabilities for the heterozygote and alternative allele homozygote.
transformed	a logical indicating whether the quantitative response Y should be transformed using a rank-based method to resemble a normal distribution; recommended for traits with non-symmetric distribution. The default option is FALSE.
related	optional: a logical indicating whether the samples should be treated as related; if TRUE while no relatedness covariance information is given, it is then estimated under a cov.structure and assumes this structure among all within-group errors pertaining to the same pair/cluster if specified using clust. This option currently only applies to autosomal SNPs.
cov.structure	optional: should be one of standard classes of correlation structures listed in corClasses from R package nlme . See ?corClasses. The most commonly used option is corCompSymm for a compound symmetric correlation structure. This option currently only applies to autosomal SNPs.
clust	optional: a factor indicating the grouping of samples; it should have at least two distinct values. It could be the family ID (FID) for family studies. This option currently only applies to autosomal SNPs.

Value

a vector of location association p -values for each SNP.

Note

The choice to use a rank-based inverse normal transformation is left to the user's discretion. See XXX for a discussion on the pros and cons of quantile transformation with respect to location association.

For X-chromosome markers, when the samples consist entirely of females or males, we report only results from model 1, regardless of the XchrMethod option.

Author(s)

Wei Q. Deng <deng@utstat.toronto.edu>, Lei Sun <sun@utstat.toronto.edu>

References

Chen B, Craiu RV, Sun L. (2020) Bayesian model averaging for the X-chromosome inactivation dilemma in genetic association study. *Biostatistics*. **21**(2):319-335. doi: [10.1093/biostatistics/kxy049](https://doi.org/10.1093/biostatistics/kxy049). PMID: 30247537.

Chen B, Craiu RV, Strug LJ, Sun L. (2021) The X factor: A robust and powerful approach to X-chromosome-inclusive whole-genome association studies. *Genetic Epidemiology*. doi: [10.1002/gepi.22422](https://doi.org/10.1002/gepi.22422). PMID: 34224641.

Examples

```
N <- 100
genDAT <- rbinom(N, 2, 0.3)
sex <- rbinom(N, 1, 0.5)+1
y <- rnorm(N)
COVAR <- matrix(rnorm(N*10), ncol=10)

locReg(GENO=genDAT, SEX=sex, Y=y, COVAR=COVAR)

# correlated example:
library("MASS")
yy <- mvrnorm(1, mu= rep(0, N), Sigma = matrix(0.3, N, N) + diag(0.7, N))
locReg(GENO=list("SNP1"= genDAT, "SNP2" = genDAT[sample(1:100)]),
SEX=sex, Y=as.numeric(yy), COVAR=COVAR, related = TRUE,
clust = rep(1, 100))

# sibpair example:
pairedY <- mvrnorm(N/2,rep(0,2),matrix(c(1,0.2,0.2,1), 2))
yy <- c(pairedY[,1], pairedY[,2])
locReg(GENO=list("SNP1"= genDAT, "SNP2" = genDAT[sample(1:100)]),
SEX=sex, Y=as.numeric(yy), COVAR=COVAR, related = TRUE,
clust = rep(c(1:50), 2))

# Xchr data example:
genDAT1 <- rep(NA, N)
genDAT1[sex==1] <- rbinom(sum(sex==1), 1, 0.5)
genDAT1[sex==2] <-rbinom(sum(sex==2), 2, 0.5)
locReg(GENO=genDAT1, SEX=sex, Y=y, COVAR=COVAR, Xchr=TRUE)
```

scaleReg

Scale (variance-based association) test

Description

This function takes as input the genotype of SNPs (GENO), the SEX (SEX), and a quantitative trait (Y) in a sample population, and possibly additional covariates, such as principal components. The function returns the scale association p -values for each SNP.

Usage

```
scaleReg(
  GENO,
  Y,
```

```

COVAR = NULL,
SEX = NULL,
Xchr = FALSE,
transformed = FALSE,
loc_alg = "LAD",
related = FALSE,
cov.structure = "corCompSymm",
clust = NULL,
genotypic = FALSE,
origLev = FALSE,
centre = "median"
)

```

Arguments

GENO	a list of a genotype matrix/vector of SNPs, must contain values 0, 1, 2's coded for the number of reference allele. Alternatively, for imputed genotypes, it could either be a vector of dosage values between 0 and 2, or a list of matrix of genotype probabilities, numerically between 0 and 1 for each genotype. The length/dimension of GENO should match that of Y, and/or SEX and COVAR.
Y	a vector of quantitative traits, such as human height.
COVAR	optional: a vector or matrix of covariates that are used to reduce bias due to confounding, such as age.
SEX	optional: the genetic sex of individuals in the sample population, must be a vector of 1 and 2 following the default sex code is 1 for males and 2 for females in PLINK.
Xchr	a logical indicator for whether the analysis is for X-chromosome SNPs.
transformed	a logical indicating whether the quantitative response Y should be transformed using a rank-based method to resemble a normal distribution; recommended for traits with non-symmetric distribution. The default option is FALSE.
loc_alg	a character indicating the type of algorithm to compute the centre in stage 1; the value is either "OLS", corresponding to an ordinary linear regression under Gaussian assumptions to compute the mean, or "LAD", corresponding to a quantile regression to compute the median. The recommended default option is "LAD". For the quantile regression, the function calls <code>quantreg::rq</code> and the median is estimated using either the "br" (smaller samples) or "sfn" (larger samples and sparse problems) algorithm depending the sample size, for more details see <code>?quantreg::rq</code> .
related	optional: a logical indicating whether the samples should be treated as related; if TRUE while no relatedness covariance information is given, it is then estimated under a <code>cov.structure</code> and assumes this structure among all within-group errors pertaining to the same pair/cluster if specified using <code>clust</code> . This option currently only applies to autosomal SNPs.
<code>cov.structure</code>	optional: should be one of standard classes of correlation structures listed in <code>corClasses</code> from R package nlme . See <code>?corClasses</code> . The most commonly used option is <code>corCompSymm</code> for a compound symmetric correlation structure. This option currently only applies to autosomal SNPs.

clust	optional: a factor indicating the grouping of samples; it should have at least two distinct values. It could be the family ID (FID) for family studies. This option currently only applies to autosomal SNPs.
genotypic	a logical indicating whether the variance homogeneity should be tested with respect to an additively (linearly) coded or non-additively coded geno_one. The former has one less degree of freedom than the latter and is the default option. For dosage genotypes without genotypic probabilities, genotypic is forced to be FALSE.
origLev	a logical indicator for whether the reported p-values should also include original Levene's test.
centre	a character indicating whether the absolute deviation should be calculated with respect to "median" or "mean" in the traditional sex-specific and Fisher combined Levene's test p-values (three tests) for X-chromosome. The default value is "median". This option applies to sex-specific analysis using original Levene's test (i.e. when regression $\$$ =\$\$TRUE).

Value

a vector of Levene's test regression p-values according to the models specified.

Note

We recommend to quantile-normally transform Y to avoid 'scale-effect' where the variance values tend to be proportional to mean values when stratified by GENO.

Author(s)

Wei Q. Deng <deng@utstat.toronto.edu>, Lei Sun <sun@utstat.toronto.edu>

References

- Deng WQ, Mao S, Kalnapekns A, Esko T, Magi R, Pare G, Sun L. (2019) Analytical strategies to include the X-chromosome in variance heterogeneity analyses: Evidence for trait-specific polygenic variance structure. *Genet Epidemiol.* **43**(7):815-830. doi: [10.1002/gepi.22247](https://doi.org/10.1002/gepi.22247). PMID:31332826.
- Gastwirth JL, Gel YR, Miao W. (2009). The Impact of Levene's Test of Equality of Variances on Statistical Theory and Practice." *Statistical Science.* **24**(3) 343-360, doi: [10.1214/09STS301](https://doi.org/10.1214/09STS301).
- Soave D, Sun L. (2017). A generalized Levene's scale test for variance heterogeneity in the presence of sample correlation and group uncertainty. *Biometrics.* **73**(3):960-971. doi: [10.1111/biom.12651](https://doi.org/10.1111/biom.12651). PMID:28099998.

Examples

```
N <- 1000
genoDAT <- rbinom(N, 2, 0.3)
sex <- rbinom(N, 1, 0.5)+1
Y <- rnorm(N)
covar <- matrix(rnorm(N*10), ncol=10)

# vanilla example:
```

```
scaleReg(GENO=list(genoDAT, genoDAT), Y=Y, COVAR=covar)
scaleReg(GENO=list(genoDAT, genoDAT), Y=Y, COVAR=covar, genotypic=TRUE)
scaleReg(GENO=list(genoDAT, genoDAT), Y=Y, COVAR=covar, origLev = TRUE)
scaleReg(GENO=list(genoDAT, genoDAT), Y=Y, COVAR=covar, origLev = TRUE, SEX=sex)
```

Index

* datasets

chrXdat, [2](#)

chrXdat, [2](#)

gJLS2, [3](#)

gJLS2s, [5](#)

leveneRegA_per_SNP, [7](#)

leveneRegX_per_SNP, [9](#)

leveneTests_per_SNP, [11](#)

locReg, [4](#), [13](#)

scaleReg, [15](#)