Package 'MultiSTAAR'

January 11, 2023

Type Package

Title Multi-Trait STAAR (MultiSTAAR) Procedure for Dynamic Incorporation of Multiple Functional Annotations in Whole-Genome Sequencing Studies

Version 0.9.7

Date 2023-01-11

Author Xihao Li [aut, cre], Zilin Li [aut, cre], Han Chen [aut], Zhonghua Liu [aut]

Maintainer Xihao Li <xihaoli@g.harvard.edu>, Zilin Li <li@hsph.harvard.edu>

Description
An R package for performing MultiSTAAR procedure in whole-genome sequencing studies.

License GPL-3

Copyright See COPYRIGHTS for details.

Imports Rcpp, GMMAT, GENESIS, STAAR, Matrix, methods

Encoding UTF-8

LazyData true

Depends R (>= 3.2.0)

LinkingTo Rcpp, RcppArmadillo, STAAR

RoxygenNote 7.1.2

Suggests knitr, rmarkdown

VignetteBuilder knitr

R topics documented:

`fit_null_glmmkin_multi` *Fitting multivariate linear mixed model for multiple traits with known relationship matrices under the null hypothesis.*

Description

The `fit_null_glmmkin_multi` function is a wrapper of the `glmmkin` function from the `GMMAT` package that fits a regression model under the null hypothesis, which provides the preliminary step for subsequent variant-set tests in whole-genome sequencing data analysis. See `glmmkin` for more details.

Usage

```
fit_null_glmmkin_multi(
  fixed,
  data = parent.frame(),
  kins,
  use_sparse = NULL,
  kins_cutoff = 0.022,
  id,
  random.slope = NULL,
  groups = NULL,
  family = gaussian(link = "identity"),
  method = "REML",
  method.optim = "AI",
  maxiter = 500,
  tol = 1e-05,
  taumin = 1e-05,
  taumax = 1e+05,
  tauregion = 10,
  verbose = FALSE,
  ...
)
```

Arguments

`fixed`	an object of class `formula` (or one that can be coerced to that class): a symbolic description of the fixed effects model to be fitted. For multiple phenotype analysis, `formula` recognized by `lm`, such as `cbind(y1,y2,y3) ~ x1 + x2`, can be used in `fixed` as fixed effects.
`data`	a data frame or list (or object coercible by `as.data.frame` to a data frame) containing the variables in the model.
`kins`	a known positive semi-definite relationship matrix (e.g. kinship matrix in genetic association studies) or a list of known positive semi-definite relationship matrices. The rownames and colnames of these matrices must at least include all samples as specified in the `id` column of the data frame `data`. For each matrix in `kins`, variance components corresponding to each phenotype, as well as their covariance components, will be estimated. If `kins` is NULL, `fit_null_glmmkin_multi` will switch to the generalized linear model with no random effects.

use_sparse	a logical switch of whether the provided dense kins matrix should be transformed to a sparse matrix (default = NULL).
kins_cutoff	the cutoff value for clustering samples to make the output matrix sparse block-diagonal (default = 0.022).
id	a column in the data frame data, indicating the id of samples. When there are duplicates in id, the data is assumed to be longitudinal with repeated measures.
random.slope	an optional column indicating the random slope for time effect used in a mixed effects model for longitudinal data. It must be included in the names of data. There must be duplicates in id and method.optim must be "AI" (default = NULL).
groups	an optional categorical variable indicating the groups used in a heteroscedastic linear mixed model (allowing residual variances in different groups to be different). This variable must be included in the names of data, and family must be "gaussian" and method.optim must be "AI" (default = NULL).
family	a description of the error distribution and link function to be used in the model. This can be a character string naming a family function, a family function or the result of a call to a family function. (See family for details of family functions). Currently, family must be "gaussian".
method	method of fitting the generalized linear mixed model. Either "REML" or "ML" (default = "REML").
method.optim	optimization method of fitting the generalized linear mixed model. Currently, method.optim must be "AI".
maxiter	a positive integer specifying the maximum number of iterations when fitting the generalized linear mixed model (default = 500).
tol	a positive number specifying tolerance, the difference threshold for parameter estimates below which iterations should be stopped (default = 1e-5).
taumin	the lower bound of search space for the variance component parameter τ (default = 1e-5), used when method.optim = "Brent". See Details.
taumax	the upper bound of search space for the variance component parameter τ (default = 1e5), used when method.optim = "Brent". See Details.
tauregion	the number of search intervals for the REML or ML estimate of the variance component parameter τ (default = 10), used when method.optim = "Brent". See Details.
verbose	a logical switch for printing detailed information (parameter estimates in each iteration) for testing and debugging purpose (default = FALSE).
...	additional arguments that could be passed to glm.

Value

The function returns an object of the model fit from glmmkin (obj_nullmodel) and whether the kins matrix is sparse when fitting the null model. See glmmkin for more details.

References

Chen, H., et al. (2016). Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *The American Journal of Human Genetics*, *98*(4), 653-666. (pub)

Chen, H., et al. (2019). Efficient variant set mixed model association tests for continuous and binary traits in large-scale whole-genome sequencing studies. *The American Journal of Human Genetics*, *104*(2), 260-274. (pub)

Chen, H. (2021). GMMAT: Generalized linear Mixed Model Association Tests Version 1.3.2. (web)

`Indiv_Score_Test_Region_multi` *Multi-trait score test for individual variants in a given variant-set*

Description

The `Indiv_Score_Test_Region_multi` function takes in genotype and the object from fitting the null model to analyze the associations between multiple (quantitative) phenotypes and all individual variants in a given variant-set by using score test.

Usage

```
Indiv_Score_Test_Region_multi(
  genotype,
  obj_nullmodel,
  rare_maf_cutoff = 0.01,
  rv_num_cutoff = 2
)
```

Arguments

`genotype`	an n*p genotype matrix (dosage matrix) of the target sequence, where n is the sample size and p is the number of genetic variants.
`obj_nullmodel`	an object from fitting the null model, which is the output from either `fit_null_glmmkin_multi` function for unrelated or related samples. Note that `fit_null_glmmkin_multi` is a wrapper of the `glmmkin` function from the `GMMAT` package.
`rare_maf_cutoff`	the cutoff of maximum minor allele frequency in defining rare variants (default = 0.01).
`rv_num_cutoff`	the cutoff of minimum number of variants of analyzing a given variant-set (default = 2).

Value

a data frame with p rows corresponding to the p genetic variants in the given variant-set and (k+1) columns: `Score1`,...,`Scorek` (the score test statistics for each phenotype) and `pvalue` (the multivariate score test p-value). If a variant in the given variant-set has minor allele frequency = 0 or greater than `rare_maf_cutoff`, the corresponding row will be `NA`. If a variant in the given variant-set has degenerate covariance matrix across multiple phenotypes, the p-value will be set as 1.

References

Chen, H., et al. (2016). Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *The American Journal of Human Genetics*, *98*(4), 653-666. (pub)

`Indiv_Score_Test_Region_multi_cond` *Multi-trait conditional score test for individual variants in a given variant-set*

Description

The `Indiv_Score_Test_Region_multi_cond` function takes in genotype, the genotype of variants to be adjusted for in conditional analysis, and the object from fitting the null model to analyze the conditional associations between multiple (quantitative) phenotypes and all individual variants in a given variant-set by using score test, adjusting for a given list of variants.

Usage

```
Indiv_Score_Test_Region_multi_cond(
  genotype,
  genotype_adj,
  obj_nullmodel,
  rare_maf_cutoff = 0.01,
  rv_num_cutoff = 2,
  method_cond = c("optimal", "naive")
)
```

Arguments

`genotype`	an n*p genotype matrix (dosage matrix) of the target sequence, where n is the sample size and p is the number of genetic variants.
`genotype_adj`	an n*p_adj genotype matrix (dosage matrix) of the target sequence, where n is the sample size and p_adj is the number of genetic variants to be adjusted for in conditional analysis (or a vector of a single variant with length n if p_adj is 1).
`obj_nullmodel`	an object from fitting the null model, which is the output from either `fit_null_glmmkin_multi` function for unrelated or related samples. Note that `fit_null_glmmkin_multi` is a wrapper of the `glmmkin` function from the `GMMAT` package.
`rare_maf_cutoff`	the cutoff of maximum minor allele frequency in defining rare variants (default = 0.01).
`rv_num_cutoff`	the cutoff of minimum number of variants of analyzing a given variant-set (default = 2).
`method_cond`	a character value indicating the method for conditional analysis. `optimal` refers to regressing residuals from the null model on `genotype_adj` as well as all covariates used in fitting the null model (fully adjusted) and taking the residuals; `naive` refers to regressing residuals from the null model on `genotype_adj` and taking the residuals (default = `optimal`).

Value

a data frame with p rows corresponding to the p genetic variants in the given variant-set and (k+1) columns: `Score1_cond`,...,`Scorek_cond` (the conditional score test statistics adjusting for variants in `genotype_adj` for each phenotype) and `pvalue_cond` (the multivariate conditional score test p-value). If a variant in the given variant-set has minor allele frequency = 0 or greater than

rare_maf_cutoff, the corresponding row will be NA. If a variant in the given variant-set has degenerate covariance matrix across multiple phenotypes, the p-value will be set as 1.

References

Chen, H., et al. (2016). Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *The American Journal of Human Genetics*, *98*(4), 653-666. (pub)

Sofer, T., et al. (2019). A fully adjusted two-stage procedure for rank-normalization in genetic association studies. *Genetic Epidemiology*, *43*(3), 263-275. (pub)

MultiSTAAR *Multi-trait STAAR (MultiSTAAR) procedure using omnibus test*

Description

The MultiSTAAR function takes in genotype, the object from fitting the null model, and functional annotation data to analyze the association between multiple (quantitative) phenotypes and a variant-set by using MultiSTAAR procedure. For each variant-set, the multi-trait STAAR-O (MultiSTAAR-O) p-value is a p-value from an omnibus test that aggregated multi-trait SKAT(1,25), SKAT(1,1), Burden(1,25), Burden(1,1), ACAT-V(1,25), and ACAT-V(1,1) together with p-values of each test weighted by each annotation using Cauchy method.

Usage

```
MultiSTAAR(
  genotype,
  obj_nullmodel,
  annotation_phred = NULL,
  rare_maf_cutoff = 0.01,
  rv_num_cutoff = 2
)
```

Arguments

genotype	an n*p genotype matrix (dosage matrix) of the target sequence, where n is the sample size and p is the number of genetic variants.
obj_nullmodel	an object from fitting the null model, which is the output from fit_null_glmmkin_multi function for unrelated or related samples. Note that fit_null_glmmkin_multi is a wrapper of the glmmkin function from the GMMAT package.
annotation_phred	a data frame or matrix of functional annotation data of dimension p*q (or a vector of a single annotation score with length p). Continuous scores should be given in PHRED score scale, where the PHRED score of j-th variant is defined to be -10*log10(rank(-score_j)/total) across the genome. (Binary) categorical scores should be taking values 0 or 1, where 1 is functional and 0 is non-functional. If not provided, MultiSTAAR will perform the multi-trait SKAT(1,25), SKAT(1,1), Burden(1,25), Burden(1,1), ACAT-V(1,25), ACAT-V(1,1) and ACAT-O tests (default = NULL).

`rare_maf_cutoff`
: the cutoff of maximum minor allele frequency in defining rare variants (default = 0.01).

`rv_num_cutoff`
: the cutoff of minimum number of variants of analyzing a given variant-set (default = 2).

Value

a list with the following members:

`num_variant`: the number of variants with minor allele frequency > 0 and less than `rare_maf_cutoff` in the given variant-set that are used for performing the variant-set using MultiSTAAR.

`cMAC`: the cumulative minor allele count of variants with minor allele frequency > 0 and less than `rare_maf_cutoff` in the given variant-set.

`RV_label`: the boolean vector indicating whether each variant in the given variant-set has minor allele frequency > 0 and less than `rare_maf_cutoff`.

`results_STAAR_O`: the multi-trait STAAR-O (MultiSTAAR-O) p-value that aggregated multi-trait SKAT(1,25), SKAT(1,1), Burden(1,25), Burden(1,1), ACAT-V(1,25), and ACAT-V(1,1) together with p-values of each test weighted by each annotation using Cauchy method.

`results_ACAT_O`: the multi-trait ACAT-O p-value that aggregated multi-trait SKAT(1,25), SKAT(1,1), Burden(1,25), Burden(1,1), ACAT-V(1,25), and ACAT-V(1,1) using Cauchy method.

`results_STAAR_S_1_25`: a vector of multi-trait STAAR-S(1,25) p-values, including multi-trait SKAT(1,25) p-value weighted by MAF, the multi-trait SKAT(1,25) p-values weighted by each annotation, and a multi-trait STAAR-S(1,25) p-value by aggregating these p-values using Cauchy method.

`results_STAAR_S_1_1`: a vector of multi-trait STAAR-S(1,1) p-values, including multi-trait SKAT(1,1) p-value weighted by MAF, the multi-trait SKAT(1,1) p-values weighted by each annotation, and a multi-trait STAAR-S(1,1) p-value by aggregating these p-values using Cauchy method.

`results_STAAR_B_1_25`: a vector of multi-trait STAAR-B(1,25) p-values, including multi-trait Burden(1,25) p-value weighted by MAF, the multi-trait Burden(1,25) p-values weighted by each annotation, and a multi-trait STAAR-B(1,25) p-value by aggregating these p-values using Cauchy method.

`results_STAAR_B_1_1`: a vector of multi-trait STAAR-B(1,1) p-values, including multi-trait Burden(1,1) p-value weighted by MAF, the multi-trait Burden(1,1) p-values weighted by each annotation, and a multi-trait STAAR-B(1,1) p-value by aggregating these p-values using Cauchy method.

`results_STAAR_A_1_25`: a vector of multi-trait STAAR-A(1,25) p-values, including multi-trait ACAT-V(1,25) p-value weighted by MAF, the multi-trait ACAT-V(1,25) p-values weighted by each annotation, and a multi-trait STAAR-A(1,25) p-value by aggregating these p-values using Cauchy method.

`results_STAAR_A_1_1`: a vector of multi-trait STAAR-A(1,1) p-values, including multi-trait ACAT-V(1,1) p-value weighted by MAF, the multi-trait ACAT-V(1,1) p-values weighted by each annotation, and a multi-trait STAAR-A(1,1) p-value by aggregating these p-values using Cauchy method.

References

Li, X., Li, Z., et al. (2020). Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nature Genetics*, *52*(9), 969-983. (pub)

Li, Z., Li, X., et al. (2022). A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies. *Nature Methods*, *19*(12), 1599-1611. (pub)

Liu, Z. & Lin, X. (manuscript). Analyzing multiple quantitative traits in sequencing studies using multivariate linear mixed models.

Liu, Y., et al. (2019). Acat: A fast and powerful p value combination method for rare-variant analysis in sequencing studies. *The American Journal of Human Genetics*, *104*(3), 410-421. (pub)

Li, Z., Li, X., et al. (2020). Dynamic scan procedure for detecting rare-variant association regions in whole-genome sequencing studies. *The American Journal of Human Genetics*, *104*(5), 802-814. (pub)

`MultiSTAAR_cond` *Multi-trait STAAR (MultiSTAAR) procedure for conditional analysis using omnibus test*

Description

The `MultiSTAAR_cond` function takes in genotype, the genotype of variants to be adjusted for in conditional analysis, the object from fitting the null model, and functional annotation data to analyze the conditional association between multiple (quantitative) phenotypes and a variant-set by using MultiSTAAR procedure, adjusting for a given list of variants. For each variant-set, the conditional multi-trait STAAR-O (MultiSTAAR-O) p-value is a p-value from an omnibus test that aggregated conditional multi-trait SKAT(1,25), SKAT(1,1), Burden(1,25), Burden(1,1), ACAT-V(1,25), and ACAT-V(1,1) together with conditional p-values of each test weighted by each annotation using Cauchy method.

Usage

```
MultiSTAAR_cond(
  genotype,
  genotype_adj,
  obj_nullmodel,
  annotation_phred = NULL,
  rare_maf_cutoff = 0.01,
  rv_num_cutoff = 2,
  method_cond = c("optimal", "naive")
)
```

Arguments

`genotype`	an n*p genotype matrix (dosage matrix) of the target sequence, where n is the sample size and p is the number of genetic variants.
`genotype_adj`	an n*p_adj genotype matrix (dosage matrix) of the target sequence, where n is the sample size and p_adj is the number of genetic variants to be adjusted for in conditional analysis (or a vector of a single variant with length n if p_adj is 1).
`obj_nullmodel`	an object from fitting the null model, which is the output from `fit_null_glmmkin_multi` function for unrelated or related samples. Note that `fit_null_glmmkin_multi` is a wrapper of the `glmmkin` function from the `GMMAT` package.
`annotation_phred`	a data frame or matrix of functional annotation data of dimension p*q (or a vector of a single annotation score with length p). Continuous scores should

be given in PHRED score scale, where the PHRED score of j-th variant is defined to be -10*log10(rank(-score_j)/total) across the genome. (Binary) categorical scores should be taking values 0 or 1, where 1 is functional and 0 is non-functional. If not provided, MultiSTAAR will perform the multi-trait SKAT(1,25), SKAT(1,1), Burden(1,25), Burden(1,1), ACAT-V(1,25), ACAT-V(1,1) and ACAT-O tests (default = NULL).

`rare_maf_cutoff`
: the cutoff of maximum minor allele frequency in defining rare variants (default = 0.01).

`rv_num_cutoff`
: the cutoff of minimum number of variants of analyzing a given variant-set (default = 2).

`method_cond`
: a character value indicating the method for conditional analysis. `optimal` refers to regressing residuals from the null model on `genotype_adj` as well as all covariates used in fitting the null model (fully adjusted) and taking the residuals; `naive` refers to regressing residuals from the null model on `genotype_adj` and taking the residuals (default = `optimal`).

Value

a list with the following members:

`num_variant`: the number of variants with minor allele frequency > 0 and less than `rare_maf_cutoff` in the given variant-set that are used for performing the variant-set using MultiSTAAR.

`cMAC`: the cumulative minor allele count of variants with minor allele frequency > 0 and less than `rare_maf_cutoff` in the given variant-set.

`RV_label`: the boolean vector indicating whether each variant in the given variant-set has minor allele frequency > 0 and less than `rare_maf_cutoff`.

`results_STAAR_O_cond`: the conditional multi-trait STAAR-O (MultiSTAAR-O) p-value that aggregated conditional SKAT(1,25), SKAT(1,1), Burden(1,25), Burden(1,1), ACAT-V(1,25), and ACAT-V(1,1) together with conditional p-values of each test weighted by each annotation using Cauchy method.

`results_ACAT_O_cond`: the conditional multi-trait ACAT-O p-value that aggregated conditional multi-trait SKAT(1,25), SKAT(1,1), Burden(1,25), Burden(1,1), ACAT-V(1,25), and ACAT-V(1,1) using Cauchy method.

`results_STAAR_S_1_25_cond`: a vector of conditional multi-trait STAAR-S(1,25) p-values, including conditional multi-trait SKAT(1,25) p-value weighted by MAF, the conditional multi-trait SKAT(1,25) p-values weighted by each annotation, and a conditional multi-trait STAAR-S(1,25) p-value by aggregating these p-values using Cauchy method.

`results_STAAR_S_1_1_cond`: a vector of conditional multi-trait STAAR-S(1,1) p-values, including conditional multi-trait SKAT(1,1) p-value weighted by MAF, the conditional multi-trait SKAT(1,1) p-values weighted by each annotation, and a conditional multi-trait STAAR-S(1,1) p-value by aggregating these p-values using Cauchy method.

`results_STAAR_B_1_25_cond`: a vector of conditional multi-trait STAAR-B(1,25) p-values, including conditional multi-trait Burden(1,25) p-value weighted by MAF, the conditional multi-trait Burden(1,25) p-values weighted by each annotation, and a conditional multi-trait STAAR-B(1,25) p-value by aggregating these p-values using Cauchy method.

`results_STAAR_B_1_1_cond`: a vector of conditional multi-trait STAAR-B(1,1) p-values, including conditional multi-trait Burden(1,1) p-value weighted by MAF, the conditional multi-trait Burden(1,1) p-values weighted by each annotation, and a conditional multi-trait STAAR-B(1,1) p-value by aggregating these p-values using Cauchy method.

`results_STAAR_A_1_25_cond`: a vector of conditional multi-trait STAAR-A(1,25) p-values, including conditional multi-trait ACAT-V(1,25) p-value weighted by MAF, the conditional multi-trait ACAT-V(1,25) p-values weighted by each annotation, and a conditional multi-trait STAAR-A(1,25) p-value by aggregating these p-values using Cauchy method.

`results_STAAR_A_1_1_cond`: a vector of conditional multi-trait STAAR-A(1,1) p-values, including conditional multi-trait ACAT-V(1,1) p-value weighted by MAF, the conditional multi-trait ACAT-V(1,1) p-values weighted by each annotation, and a conditional multi-trait STAAR-A(1,1) p-value by aggregating these p-values using Cauchy method.

References

Li, X., Li, Z., et al. (2020). Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nature Genetics*, *52*(9), 969-983. (pub)

Li, Z., Li, X., et al. (2022). A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies. *Nature Methods*, *19*(12), 1599-1611. (pub)

Liu, Z. & Lin, X. (manuscript). Analyzing multiple quantitative traits in sequencing studies using multivariate linear mixed models.

Liu, Y., et al. (2019). Acat: A fast and powerful p value combination method for rare-variant analysis in sequencing studies. *The American Journal of Human Genetics*, *104*(3), 410-421. (pub)

Li, Z., Li, X., et al. (2020). Dynamic scan procedure for detecting rare-variant association regions in whole-genome sequencing studies. *The American Journal of Human Genetics*, *104*(5), 802-814. (pub)

Sofer, T., et al. (2019). A fully adjusted two-stage procedure for rank-normalization in genetic association studies. *Genetic Epidemiology*, *43*(3), 263-275. (pub)

Index