Package 'STAARpipelinePheWAS'

December 28, 2024

Type Package

Title STAARpipeline for Phenome-Wide Association Study of Whole-Genome/Whole-Exome Sequencing Data

Version 0.9.7

Date 2024-12-28

Author Xihao Li [aut, cre], Zilin Li [aut, cre]

Maintainer Xihao Li <xihaoli@unc.edu>, Zilin Li <lizl@nenu.edu.cn>

Description An R package for performing STAARpipeline in phenome-wide association study of whole-genome/whole-exome sequencing data.

License GPL-3

Copyright See COPYRIGHTS for details.

Imports STAAR, MultiSTAAR, STAARpipeline, dplyr, SeqArray, SeqVarTools, GenomicFeatures, TxDb.Hsapiens.UCSC.hg38.knownGene, Matrix, methods

Encoding UTF-8

LazyData true

Depends R (>= 3.2.0)

RoxygenNote 7.2.3

Suggests knitr, rmarkdown

VignetteBuilder knitr

R topics documented:

`Gene_Centric_Coding_PheWAS` *Gene-centric PheWAS analysis of coding functional categories using STAAR procedure*

Description

The `Gene_Centric_Coding_PheWAS` function takes in chromosome, gene name, functional category, the object of opened annotated GDS file, and the list of objects from fitting the null models to analyze the association between a series of quantitative/dichotomous phenotypes (including imbalanced case-control design) and coding functional categories of a gene by using STAAR procedure. For each coding functional category, the STAAR-O p-value is a p-value from an omnibus test that aggregated 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. For imbalance case-control setting, the results correspond to the STAAR-B p-value, which is a p-value from an omnibus test that aggregated Burden(1,25) and Burden(1,1) together with p-values of each test weighted by each annotation using Cauchy method. For multiple phenotype analysis (`obj_nullmodel$n.pheno > 1`), the results correspond to multi-trait association p-values (e.g. MultiSTAAR-O) by leveraging the correlation structure between multiple phenotypes.

Usage

```
Gene_Centric_Coding_PheWAS(
  chr,
  gene_name,
  category = c("all_categories", "plof", "plof_ds", "missense", "disruptive_missense",
    "synonymous", "ptv", "ptv_ds", "all_categories_incl_ptv"),
  genofile,
  obj_nullmodel_list,
  rare_maf_cutoff = 0.01,
  rv_num_cutoff = 2,
  rv_num_cutoff_max = 1e+09,
  rv_num_cutoff_max_prefilter = 1e+09,
  QC_label = "annotation/filter",
  variant_type = c("SNV", "Indel", "variant"),
  geno_missing_imputation = c("mean", "minor"),
  Annotation_dir = "annotation/info/FunctionalAnnotation",
  Annotation_name_catalog,
  Use_annotation_weights = c(TRUE, FALSE),
  Annotation_name = NULL,
  SPA_p_filter = TRUE,
  p_filter_cutoff = 0.05,
  silent = FALSE
)
```

Arguments

`chr`	chromosome.
`gene_name`	name of the gene to be analyzed using STAAR procedure.

`category` the coding functional category to be analyzed using STAAR procedure. Choices include `all_categories`, `plof`, `plof_ds`, `missense`, `disruptive_missense`, `synonymous`, `ptv`, `ptv_ds`, `all_categories_incl_ptv` (default = `all_categories`).

`genofile` an object of opened annotated GDS (aGDS) file.

`obj_nullmodel_list`
a list of objects from fitting the null models, which are either the output from `fit_nullmodel` function in the `STAARpipeline` package, or the output from `fitNullModel` function in the `GENESIS` package and transformed using the `genesis2staar_nullmodel` function in the `STAARpipeline` 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).

`rv_num_cutoff_max`
the cutoff of maximum number of variants of analyzing a given variant-set (default = 1e+09).

`rv_num_cutoff_max_prefilter`
the cutoff of maximum number of variants before extracting the genotype matrix (default = 1e+09).

`QC_label` channel name of the QC label in the GDS/aGDS file (default = "annotation/filter").

`variant_type` type of variant included in the analysis. Choices include "SNV", "Indel", or "variant" (default = "SNV").

`geno_missing_imputation`
method of handling missing genotypes. Either "mean" or "minor" (default = "mean").

`Annotation_dir` channel name of the annotations in the aGDS file (default = "annotation/info/FunctionalAnnotation").

`Annotation_name_catalog`
a data frame containing the name and the corresponding channel name in the aGDS file.

`Use_annotation_weights`
use annotations as weights or not (default = TRUE).

`Annotation_name`
a vector of annotation names used in STAAR (default = NULL).

`SPA_p_filter` logical: are only the variants with a normal approximation based p-value smaller than a pre-specified threshold use the SPA method to recalculate the p-value, only used for imbalanced case-control setting (default = TRUE).

`p_filter_cutoff`
threshold for the p-value recalculation using the SPA method, only used for imbalanced case-control setting (default = 0.05).

`silent` logical: should the report of error messages be suppressed (default = FALSE).

Value

A list of data frames containing the STAAR p-values (including STAAR-O or STAAR-B in imbalanced case-control setting) corresponding to each coding functional category of the given gene for each phenotype.

References

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)

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)

`Gene_Centric_Noncoding_PheWAS` *Gene-centric PheWAS analysis of noncoding functional categories using STAAR procedure*

Description

The `Gene_Centric_Noncoding_PheWAS` function takes in chromosome, gene name, functional category, the object of opened annotated GDS file, and the list of objects from fitting the null models to analyze the association between a series of quantitative/dichotomous phenotypes (including imbalanced case-control design) and noncoding functional categories of a gene by using STAAR procedure. For each noncoding functional category, the STAAR-O p-value is a p-value from an omnibus test that aggregated 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. For imbalance case-control setting, the results correspond to the STAAR-B p-value, which is a p-value from an omnibus test that aggregated Burden(1,25) and Burden(1,1) together with p-values of each test weighted by each annotation using Cauchy method. For multiple phenotype analysis (`obj_nullmodel$n.pheno > 1`), the results correspond to multi-trait association p-values (e.g. MultiSTAAR-O) by leveraging the correlation structure between multiple phenotypes.

Usage

```
Gene_Centric_Noncoding_PheWAS(
  chr,
  gene_name,
 category = c("all_categories", "downstream", "upstream", "UTR", "promoter_CAGE",
    "promoter_DHS", "enhancer_CAGE", "enhancer_DHS"),
  genofile,
  obj_nullmodel_list,
  rare_maf_cutoff = 0.01,
  rv_num_cutoff = 2,
  rv_num_cutoff_max = 1e+09,
  rv_num_cutoff_max_prefilter = 1e+09,
  QC_label = "annotation/filter",
  variant_type = c("SNV", "Indel", "variant"),
  geno_missing_imputation = c("mean", "minor"),
  Annotation_dir = "annotation/info/FunctionalAnnotation",
  Annotation_name_catalog,
  Use_annotation_weights = c(TRUE, FALSE),
  Annotation_name = NULL,
  SPA_p_filter = TRUE,
  p_filter_cutoff = 0.05,
  silent = FALSE
)
```

Arguments

chr	chromosome.
gene_name	name of the gene to be analyzed using STAAR procedure.
category	the noncoding functional category to be analyzed using STAAR procedure. Choices include all_categories, downstream, upstream, UTR, promoter_CAGE, promoter_DHS, enhancer_CAGE, enhancer_DHS (default = all_categories).
genofile	an object of opened annotated GDS (aGDS) file.
obj_nullmodel_list	a list of objects from fitting the null models, which are either the output from fit_nullmodel function in the STAARpipeline package, or the output from fitNullModel function in the GENESIS package and transformed using the genesis2staar_nullmodel function in the STAARpipeline 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).
rv_num_cutoff_max	the cutoff of maximum number of variants of analyzing a given variant-set (default = 1e+09).
rv_num_cutoff_max_prefilter	the cutoff of maximum number of variants before extracting the genotype matrix (default = 1e+09).
QC_label	channel name of the QC label in the GDS/aGDS file (default = "annotation/filter").
variant_type	type of variant included in the analysis. Choices include "SNV", "Indel", or "variant" (default = "SNV").
geno_missing_imputation	method of handling missing genotypes. Either "mean" or "minor" (default = "mean").
Annotation_dir	channel name of the annotations in the aGDS file (default = "annotation/info/FunctionalAnnotation").
Annotation_name_catalog	a data frame containing the name and the corresponding channel name in the aGDS file.
Use_annotation_weights	use annotations as weights or not (default = TRUE).
Annotation_name	a vector of annotation names used in STAAR (default = NULL).
SPA_p_filter	logical: are only the variants with a normal approximation based p-value smaller than a pre-specified threshold use the SPA method to recalculate the p-value, only used for imbalanced case-control setting (default = TRUE).
p_filter_cutoff	threshold for the p-value recalculation using the SPA method, only used for imbalanced case-control setting (default = 0.05).
silent	logical: should the report of error messages be suppressed (default = FALSE).

Value

A list of data frames containing the STAAR p-values (including STAAR-O or STAAR-B in imbalanced case-control setting) corresponding to each noncoding functional category of the given gene for each phenotype.

References

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)

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)

Individual_Analysis_PheWAS

Individual-variant PheWAS analysis using score test

Description

The `Individual_Analysis_PheWAS` function takes in chromosome, starting location, ending location, the object of opened annotated GDS file, and the list of objects from fitting the null models to analyze the association between a series of quantitative/dichotomous phenotypes (including imbalanced case-control design) and each individual variant in a genetic region by using score test. For multiple phenotype analysis (`obj_nullmodel$n.pheno > 1`), the results correspond to multi-trait score test p-values by leveraging the correlation structure between multiple phenotypes.

Usage

```
Individual_Analysis_PheWAS(
  chr,
  start_loc,
  end_loc,
  genofile,
  obj_nullmodel_list,
  mac_cutoff = 20,
  subset_variants_num = 5000,
  QC_label = "annotation/filter",
  variant_type = c("variant", "SNV", "Indel"),
  geno_missing_imputation = c("mean", "minor"),
  tol = .Machine$double.eps^0.25,
  max_iter = 1000,
  SPA_p_filter = TRUE,
  p_filter_cutoff = 0.05
)
```

Arguments

`chr`	chromosome.
`start_loc`	starting location (position) of the genetic region for each individual variant to be analyzed using score test.

`end_loc` ending location (position) of the genetic region for each individual variant to be analyzed using score test.

`genofile` an object of opened annotated GDS (aGDS) file.

`obj_nullmodel_list`
a list of objects from fitting the null models, which are either the output from `fit_nullmodel` function in the `STAARpipeline` package, or the output from `fitNullModel` function in the `GENESIS` package and transformed using the `genesis2staar_nullmodel` function in the `STAARpipeline` package.

`mac_cutoff` the cutoff of minimum minor allele count in defining individual variants (default = 20).

`subset_variants_num`
the number of variants to run per subset for each time (default = 5e3).

`QC_label` channel name of the QC label in the GDS/aGDS file (default = "annotation/filter").

`variant_type` type of variant included in the analysis. Choices include "variant", "SNV", or "Indel" (default = "variant").

`geno_missing_imputation`
method of handling missing genotypes. Either "mean" or "minor" (default = "mean").

`tol` a positive number specifying tolerance, the difference threshold for parameter estimates in saddlepoint approximation algorithm below which iterations should be stopped (default = ".Machine$double.eps^0.25").

`max_iter` a positive integer specifying the maximum number of iterations for applying the saddlepoint approximation algorithm (default = "1000").

`SPA_p_filter` logical: are only the variants with a score-test-based p-value smaller than a pre-specified threshold use the SPA method to recalculate the p-value, only used for imbalanced case-control setting (default = TRUE).

`p_filter_cutoff`
threshold for the p-value recalculation using the SPA method, only used for imbalanced case-control setting (default = 0.05)

Value

A list of data frames containing the score test p-value and the estimated effect size of the minor allele for each individual variant in the given genetic region for each phenotype. The first 4 columns of each data frame correspond to chromosome (CHR), position (POS), reference allele (REF), and alternative allele (ALT).

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)

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)

ncRNA_PheWAS *Gene-centric PheWAS analysis of long noncoding RNA (ncRNA) category using STAAR procedure*

Description

The ncRNA_PheWAS function takes in chromosome, gene name, the object of opened annotated GDS file, and the list of objects from fitting the null models to analyze the association between a series of quantitative/dichotomous phenotypes (including imbalanced case-control design) and the exonic and splicing category of an ncRNA gene by using STAAR procedure. For each ncRNA category, the STAAR-O p-value is a p-value from an omnibus test that aggregated 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. For imbalance case-control setting, the results correspond to the STAAR-B p-value, which is a p-value from an omnibus test that aggregated Burden(1,25) and Burden(1,1) together with p-values of each test weighted by each annotation using Cauchy method. For multiple phenotype analysis (obj_nullmodel$n.pheno > 1), the results correspond to multi-trait association p-values (e.g. MultiSTAAR-O) by leveraging the correlation structure between multiple phenotypes.

Usage

```
ncRNA_PheWAS(
  chr,
  gene_name,
  genofile,
  obj_nullmodel_list,
  rare_maf_cutoff = 0.01,
  rv_num_cutoff = 2,
  rv_num_cutoff_max = 1e+09,
  rv_num_cutoff_max_prefilter = 1e+09,
  QC_label = "annotation/filter",
  variant_type = c("SNV", "Indel", "variant"),
  geno_missing_imputation = c("mean", "minor"),
  Annotation_dir = "annotation/info/FunctionalAnnotation",
  Annotation_name_catalog,
  Use_annotation_weights = c(TRUE, FALSE),
  Annotation_name = NULL,
  SPA_p_filter = TRUE,
  p_filter_cutoff = 0.05,
  silent = FALSE
)
```

Arguments

chr	chromosome.
gene_name	name of the ncRNA gene to be analyzed using STAAR procedure.
genofile	an object of opened annotated GDS (aGDS) file.
obj_nullmodel_list	a list of objects from fitting the null models, which are either the output from fit_nullmodel function in the STAARpipeline package, or the output from

fitNullModel function in the GENESIS package and transformed using the genesis2staar_nullmodel function in the STAARpipeline 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).

rv_num_cutoff_max
: the cutoff of maximum number of variants of analyzing a given variant-set (default = 1e+09).

rv_num_cutoff_max_prefilter
: the cutoff of maximum number of variants before extracting the genotype matrix (default = 1e+09).

QC_label
: channel name of the QC label in the GDS/aGDS file (default = "annotation/filter").

variant_type
: type of variant included in the analysis. Choices include "SNV", "Indel", or "variant" (default = "SNV").

geno_missing_imputation
: method of handling missing genotypes. Either "mean" or "minor" (default = "mean").

Annotation_dir
: channel name of the annotations in the aGDS file (default = "annotation/info/FunctionalAnnotation").

Annotation_name_catalog
: a data frame containing the name and the corresponding channel name in the aGDS file.

Use_annotation_weights
: use annotations as weights or not (default = TRUE).

Annotation_name
: a vector of annotation names used in STAAR (default = NULL).

SPA_p_filter
: logical: are only the variants with a normal approximation based p-value smaller than a pre-specified threshold use the SPA method to recalculate the p-value, only used for imbalanced case-control setting (default = TRUE).

p_filter_cutoff
: threshold for the p-value recalculation using the SPA method, only used for imbalanced case-control setting (default = 0.05).

silent
: logical: should the report of error messages be suppressed (default = FALSE).

Value

A list of data frames containing the STAAR p-values (including STAAR-O or STAAR-B in imbalanced case-control setting) corresponding to the exonic and splicing category of the given ncRNA gene for each phenotype.

References

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)

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)

`Sliding_Window_PheWAS` *Genetic region PheWAS analysis of sliding windows using STAAR procedure*

Description

The `Sliding_Window_PheWAS` function takes in chromosome, starting location, ending location, sliding window length, the object of opened annotated GDS file, and the list of objects from fitting the null models to analyze the association between a series of quantitative/dichotomous phenotypes (including imbalanced case-control design) and variants in a genetic region by using STAAR procedure. For each sliding window, the STAAR-O p-value is a p-value from an omnibus test that aggregated 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. For imbalance case-control setting, the results correspond to the STAAR-B p-value, which is a p-value from an omnibus test that aggregated Burden(1,25) and Burden(1,1) together with p-values of each test weighted by each annotation using Cauchy method. For multiple phenotype analysis (`obj_nullmodel$n.pheno > 1`), the results correspond to multi-trait association p-values (e.g. MultiSTAAR-O) by leveraging the correlation structure between multiple phenotypes.

Usage

```
Sliding_Window_PheWAS(
  chr,
  start_loc,
  end_loc,
  sliding_window_length = 2000,
  type = c("single", "multiple"),
  genofile,
  obj_nullmodel_list,
  rare_maf_cutoff = 0.01,
  rv_num_cutoff = 2,
  rv_num_cutoff_max = 1e+09,
  rv_num_cutoff_max_prefilter = 1e+09,
  QC_label = "annotation/filter",
  variant_type = c("SNV", "Indel", "variant"),
  geno_missing_imputation = c("mean", "minor"),
  Annotation_dir = "annotation/info/FunctionalAnnotation",
  Annotation_name_catalog,
  Use_annotation_weights = c(TRUE, FALSE),
  Annotation_name = NULL,
  SPA_p_filter = TRUE,
  p_filter_cutoff = 0.05,
  silent = FALSE
)
```

Arguments

`chr`	chromosome.
`start_loc`	starting location (position) of the genetic region to be analyzed using STAAR procedure.

`end_loc` ending location (position) of the genetic region to be analyzed using STAAR procedure.

`sliding_window_length` the (fixed) length of the sliding window to be analyzed using STAAR procedure.

`type` the type of sliding window to be analyzed using STAAR procedure. Choices include `single`, `multiple` (default = `single`).

`genofile` an object of opened annotated GDS (aGDS) file.

`obj_nullmodel_list` a list of objects from fitting the null models, which are either the output from `fit_nullmodel` function in the `STAARpipeline` package, or the output from `fitNullModel` function in the `GENESIS` package and transformed using the `genesis2staar_nullmodel` function in the `STAARpipeline` 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).

`rv_num_cutoff_max` the cutoff of maximum number of variants of analyzing a given variant-set (default = 1e+09).

`rv_num_cutoff_max_prefilter` the cutoff of maximum number of variants before extracting the genotype matrix (default = 1e+09).

`QC_label` channel name of the QC label in the GDS/aGDS file (default = "annotation/filter").

`variant_type` type of variant included in the analysis. Choices include "SNV", "Indel", or "variant" (default = "SNV").

`geno_missing_imputation` method of handling missing genotypes. Either "mean" or "minor" (default = "mean").

`Annotation_dir` channel name of the annotations in the aGDS file (default = "annotation/info/FunctionalAnnotation").

`Annotation_name_catalog` a data frame containing the name and the corresponding channel name in the aGDS file.

`Use_annotation_weights` use annotations as weights or not (default = TRUE).

`Annotation_name` a vector of annotation names used in STAAR (default = NULL).

`SPA_p_filter` logical: are only the variants with a normal approximation based p-value smaller than a pre-specified threshold use the SPA method to recalculate the p-value, only used for imbalanced case-control setting (default = TRUE).

`p_filter_cutoff` threshold for the p-value recalculation using the SPA method, only used for imbalanced case-control setting (default = 0.05).

`silent` logical: should the report of error messages be suppressed (default = FALSE).

Value

A list of data frames containing the STAAR p-values (including STAAR-O or STAAR-B in imbalanced case-control setting) corresponding to each sliding window in the given genetic region for each phenotype.

References

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)

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)

Index