Package 'MetaSTAARlite'

April 17, 2025

Type Package

Title MetaSTAARlite Pipeline for Meta-Analysis of Whole-Genome/Whole-Exome Sequencing Data

Version 0.9.6.1

Date 2025-04-14

Author Xihao Li [aut, cre], Yohhan Kumarasinghe [aut, cre], Zilin Li [aut, cre], Yuxin Yuan [aut, cre]

Maintainer Xihao Li <xihaoli@unc.edu>, Yohhan Kumarasinghe <yohhan@unc.edu>, Zilin Li <lizl@nenu.edu.cn>, Yuxin Yuan <yuxinyuan@nenu.edu.cn>

Description A lightweight R package for using the MetaSTAARlite pipeline in meta-analysis of whole-genome/whole-exome sequencing data.

License GPL-3

Copyright See COPYRIGHTS for details.

Imports Rcpp, STAAR, MetaSTAAR, STAARpipeline, STAARpipelineSummary, dplyr, SeqArray, SeqVarTools, GenomicFeatures, TxDb.Hsapiens.UCSC.hg38.knownGene, Matrix, methods, expm, MASS

Encoding UTF-8

LazyData true

Depends R (>= 3.2.0)

RoxygenNote 7.3.2

Suggests knitr, rmarkdown

VignetteBuilder knitr

Contents

`coding_MetaSTAARlite` *Performs meta-analysis of coding functional categories using MetaSTAARlite*

Description

This function performs meta-analysis to detect associations between a quantitative/dichotomous phenotype and coding functional categories of a gene by using the MetaSTAARlite pipeline. For each coding functional category, the MetaSTAAR-O p-value is a p-value from an omnibus test that aggregated SKAT-MS(1,25), SKAT-MS(1,1), Burden-MS(1,25), Burden-MS(1,1), ACAT-V-MS(1,25), and ACAT-V-MS(1,1) together with p-values of each test weighted by each annotation using Cauchy method.

Usage

```
coding_MetaSTAARlite(
  chr,
  gene_name,
  genes,
  sample.sizes,
  coding_sumstat_gene_list,
  coding_cov_gene_list,
  cov_maf_cutoff,
  rare_maf_cutoff = 0.01,
  rv_num_cutoff = 2,
  check_qc_label = FALSE,
  variant_type = c("SNV", "Indel", "variant"),
  Use_annotation_weights = TRUE,
  Annotation_name = NULL,
  silent = FALSE
)
```

Arguments

chr	an integer which specifies the chromosome number.
gene_name	a character which specifies the name of the gene to be meta-analyzed using MetaSTAARlite.
genes	the genes_info object from the STAARpipeline package.
sample.sizes	a numeric vector with the length of study.names indicating the sample size of each study.
coding_sumstat_gene_list	a list containing study-specific summary statistics corresponding to the specified gene.
coding_cov_gene_list	a list containing study-specific sparse weighted covariance matrices corresponding to the specified gene.
cov_maf_cutoff	a numeric vector with the length of study.names indicating the maximum minor allele frequency cutoffs under which the sparse weighted covariance files between variants are stored.
rare_maf_cutoff	a numeric value specifying the cutoff of maximum minor allele frequency in defining rare variants (default = 0.01).
rv_num_cutoff	an integer specifying the cutoff of minimum number of variants of meta-analyzing a given variant-set (default = 2).
check_qc_label	a logical value indicating whether variants need to be dropped according to qc_label specified in coding_MetaSTAARlite_worker (default = FALSE).
variant_type	a character value specifying the type of variant included in the analysis. Choices include "SNV", "Indel", or "variant" (default = "SNV").
Use_annotation_weights	a logical value which determines if annotations will be used as weights or not (default = TRUE).
Annotation_name	a character vector of annotation names used in MetaSTAARlite (default = NULL).
silent	a logical value which determines if the report of error messages will be suppressed (default = FALSE).

Value

a list of data frames containing the MetaSTAAR p-values (including MetaSTAAR-O) corresponding to each coding functional category of the given gene.

References

Li, X., et al. (2023). Powerful, scalable and resource-efficient meta-analysis of rare variant associations in large whole genome sequencing studies. *Nature Genetics*, *55*(1), 154-164. (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*. (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)

`coding_MetaSTAARlite_cond` *Performs conditional meta-analysis of coding functional categories using MetaSTAARlite*

Description

This function performs meta-analysis to detect conditional associations between a quantitative/dichotomous phenotype and coding functional categories of a gene adjusting for set of known variants by using the MetaSTAARlite pipeline. For each coding functional category, the conditional MetaSTAAR-O p-value is a p-value from an omnibus test that aggregated conditional SKAT-MS(1,25), SKAT-MS(1,1), Burden-MS(1,25), Burden-MS(1,1), ACAT-V-MS(1,25), and ACAT-V-MS(1,1) together with conditional p-values of each test weighted by each annotation using Cauchy method.

Usage

```
coding_MetaSTAARlite_cond(
  chr,
  gene_name,
  genes,
  sample.sizes,
  coding_sumstat_gene_list,
  coding_cov_gene_list,
  coding_cov_cond_gene_list,
  cov_maf_cutoff,
  rare_maf_cutoff = 0.01,
  rv_num_cutoff = 2,
  effect.cond = c("homogeneous", "heterogeneous"),
  check_qc_label = FALSE,
  variant_type = c("SNV", "Indel", "variant"),
  Use_annotation_weights = TRUE,
  Annotation_name = NULL,
  silent = FALSE
)
```

Arguments

`chr`	an integer which specifies the chromosome number.
`gene_name`	a character which specifies the name of the gene to be meta-analyzed using MetaSTAARlite.
`genes`	the `genes_info` object from the `STAARpipeline` package.
`sample.sizes`	a numeric vector with the length of `study.names` indicating the sample size of each study.
`coding_sumstat_gene_list`	a list containing study-specific summary statistics corresponding to the specified gene.
`coding_cov_gene_list`	a list containing study-specific sparse weighted covariance matrices corresponding to the specified gene.

`coding_cov_cond_gene_list`
: a list containing study-specific summary statistics and covariance matrices corresponding to the specified gene for variants to be conditioned on.

`cov_maf_cutoff`
: a numeric vector with the length of `study.names` indicating the maximum minor allele frequency cutoffs under which the sparse weighted covariance files between variants are stored.

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

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

`effect.cond`
: a character value indicating the effects of variants to be adjusted for in conditional analysis are "homogeneous" or "heterogeneous" (default = "homogeneous").

`check_qc_label`
: a logical value indicating whether variants need to be dropped according to `qc_label` specified in `coding_MetaSTAARlite_worker` (default = FALSE).

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

`Use_annotation_weights`
: a logical value which determines if annotations will be used as weights or not (default = TRUE).

`Annotation_name`
: a character vector of annotation names used in MetaSTAARlite (default = NULL).

`silent`
: a logical value which determines if the report of error messages will be suppressed (default = FALSE).

Value

a list of data frames containing the conditional MetaSTAAR p-values (including MetaSTAAR-O) corresponding to each coding functional category of the given gene.

References

Li, X., et al. (2023). Powerful, scalable and resource-efficient meta-analysis of rare variant associations in large whole genome sequencing studies. *Nature Genetics*, *55*(1), 154-164. (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*. (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)

`coding_MetaSTAARlite_worker`

Generates summary statistics of coding functional categories using MetaSTAARlite

---

Description

This function uses MetaSTAARlite to generate variant-level summary statistics and sparse covariance matrices for coding functional categories of a gene.

Usage

```
coding_MetaSTAARlite_worker(
  chr,
  gene_name,
  genofile,
  obj_nullmodel,
  genes,
  known_loci = NULL,
  cov_maf_cutoff = 0.05,
  signif.digits = NULL,
  QC_label = "annotation/filter",
  check_qc_label = FALSE,
  variant_type = c("SNV", "Indel", "variant"),
  Annotation_dir = "annotation/info/FunctionalAnnotation",
  Annotation_name_catalog,
  Use_annotation_weights = TRUE,
  Annotation_name = NULL,
  silent = FALSE
)
```

Arguments

chr	an integer which specifies the chromosome number.
gene_name	a character which specifies the name of the gene to be meta-analyzed using MetaSTAARlite.
genofile	an object of opened annotated GDS (aGDS) file.
obj_nullmodel	an object from fitting the null model, which is 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.
genes	the genes_info object from the STAARpipeline package.
known_loci	the data frame of variants to be adjusted for in conditional analysis. Should contain four columns in the following order: chromosome (CHR), position (POS), reference allele (REF), and alternative allele (ALT) (default = NULL).
cov_maf_cutoff	a numeric value indicating the maximum minor allele frequency cutoff under which the sparse weighted covariance file between variants is stored (default = 0.05).
signif.digits	an integer indicating the number of significant digits to be used for storing the sparse weighted covariance file. If signif.digits is NULL, it is assumed that no rounding will be performed (default = NULL).
QC_label	a character specifying the channel name of the QC label in the GDS/aGDS file (default = "annotation/filter").
check_qc_label	a logical value indicating whether variants need to be dropped according to qc_label. If check_qc_label is FALSE, then the summary statistics will be stored for PASS variants from the study. If check_qc_label is TRUE, then the summary statistics will be stored for all variants from the study, together will an additional column of qc_label (default = FALSE).
variant_type	a character value specifying the type of variant included in the analysis. Choices include "SNV", "Indel", or "variant" (default = "SNV").

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

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

`Use_annotation_weights` a logical value which specifies if annotations will be used as weights or not (default = TRUE).

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

`silent` a logical value which determines if the report of error messages will be suppressed (default = FALSE).

Value

a list of the following objects corresponding to each coding functional category of the given gene: (1) the data frame of all variants in the variant-set (the variant-level summary statistics file), including the following information: chromosome (chr), position (pos), reference allele (ref), alternative allele (alt), quality control status (qc_label, optional), alternative allele count (alt_AC), minor allele count (MAC), minor allele frequency (MAF), study sample size (N), score statistic (U), variance (V), variant annotations specified in `Annotation_name`, and the low-rank decomposed component of the covariance file; (2) the sparse matrix of all variants in the variant-set whose minor allele frequency is below `cov_maf_cutoff` (the sparse weighted covariance file); (3) the summary statistics and covariance matrices corresponding to the specified gene for variants to be conditioned on in `known_loci`.

References

Li, X., et al. (2023). Powerful, scalable and resource-efficient meta-analysis of rare variant associations in large whole genome sequencing studies. *Nature Genetics*, *55*(1), 154-164. (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*. (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)

`custom_MetaSTAARlite` *Performs meta-analysis of custom variant-sets (masks) using MetaSTAARlite*

Description

This function performs meta-analysis to detect associations between a quantitative/dichotomous phenotype and a user-specified custom mask by using the MetaSTAARlite pipeline. For each custom mask, the MetaSTAAR-O p-value is a p-value from an omnibus test that aggregated SKAT-MS(1,25), SKAT-MS(1,1), Burden-MS(1,25), Burden-MS(1,1), ACAT-V-MS(1,25), and ACAT-V-MS(1,1) together with p-values of each test weighted by each annotation using Cauchy method.

Usage

```
custom_MetaSTAARlite(
  chr,
  mask_name,
  sample.sizes,
  custom_sumstat_mask_list,
  custom_cov_mask_list,
  cov_maf_cutoff,
  rare_maf_cutoff = 0.01,
  rv_num_cutoff = 2,
  check_qc_label = FALSE,
  variant_type = c("SNV", "Indel", "variant"),
  Use_annotation_weights = TRUE,
  Annotation_name = NULL,
  silent = FALSE
)
```

Arguments

`chr`	an integer which specifies the chromosome number.
`mask_name`	a character which specifies the name of the mask to be meta-analyzed using MetaSTAARlite.
`sample.sizes`	a numeric vector with the length of `study.names` indicating the sample size of each study.
`custom_sumstat_mask_list`	a list containing study-specific summary statistics corresponding to the custom mask.
`custom_cov_mask_list`	a list containing study-specific sparse weighted covariance matrices corresponding to the custom mask.
`cov_maf_cutoff`	a numeric vector with the length of `study.names` indicating the maximum minor allele frequency cutoffs under which the sparse weighted covariance files between variants are stored.
`rare_maf_cutoff`	a numeric value specifying the cutoff of maximum minor allele frequency in defining rare variants (default = 0.01).
`rv_num_cutoff`	an integer specifying the cutoff of minimum number of variants of meta-analyzing a given variant-set (default = 2).
`check_qc_label`	a logical value indicating whether variants need to be dropped according to `qc_label` specified in `custom_MetaSTAARlite_worker` (default = FALSE).
`variant_type`	a character value specifying the type of variant included in the analysis. Choices include "SNV", "Indel", or "variant" (default = "SNV").
`Use_annotation_weights`	a logical value which determines if annotations will be used as weights or not (default = TRUE).
`Annotation_name`	a character vector of annotation names used in MetaSTAARlite (default = NULL).
`silent`	a logical value which determines if the report of error messages will be suppressed (default = FALSE).

Value

a list of data frames containing the MetaSTAAR p-values (including MetaSTAAR-O) corresponding to the custom mask.

References

Li, X., et al. (2023). Powerful, scalable and resource-efficient meta-analysis of rare variant associations in large whole genome sequencing studies. *Nature Genetics*, *55*(1), 154-164. (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*. (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)

`custom_MetaSTAARlite_cond`

Performs conditional meta-analysis of custom variant-sets (masks) using MetaSTAARlite

---

Description

This function performs meta-analysis to detect conditional associations between a quantitative/dichotomous phenotype and a user-specified custom mask adjusting for set of known variants by using the MetaSTAARlite pipeline. For each custom mask, the conditional MetaSTAAR-O p-value is a p-value from an omnibus test that aggregated conditional SKAT-MS(1,25), SKAT-MS(1,1), Burden-MS(1,25), Burden-MS(1,1), ACAT-V-MS(1,25), and ACAT-V-MS(1,1) together with conditional p-values of each test weighted by each annotation using Cauchy method.

Usage

```
custom_MetaSTAARlite_cond(
  chr,
  mask_name,
  sample.sizes,
  custom_sumstat_mask_list,
  custom_cov_mask_list,
  custom_cov_cond_mask_list,
  cov_maf_cutoff,
  rare_maf_cutoff = 0.01,
  rv_num_cutoff = 2,
  effect.cond = c("homogeneous", "heterogeneous"),
  check_qc_label = FALSE,
  variant_type = c("SNV", "Indel", "variant"),
  Use_annotation_weights = TRUE,
  Annotation_name = NULL,
  silent = FALSE
)
```

Arguments

`chr`	an integer which specifies the chromosome number.
`mask_name`	a character which specifies the name of the mask to be meta-analyzed using MetaSTAARlite.
`sample.sizes`	a numeric vector with the length of `study.names` indicating the sample size of each study.
`custom_sumstat_mask_list`	a list containing study-specific summary statistics corresponding to the custom mask.
`custom_cov_mask_list`	a list containing study-specific sparse weighted covariance matrices corresponding to the custom mask.
`custom_cov_cond_mask_list`	a list containing study-specific summary statistics and covariance matrices corresponding to the custom mask for variants to be conditioned on.
`cov_maf_cutoff`	a numeric vector with the length of `study.names` indicating the maximum minor allele frequency cutoffs under which the sparse weighted covariance files between variants are stored.
`rare_maf_cutoff`	a numeric value specifying the cutoff of maximum minor allele frequency in defining rare variants (default = 0.01).
`rv_num_cutoff`	an integer specifying the cutoff of minimum number of variants of meta-analyzing a given variant-set (default = 2).
`effect.cond`	a character value indicating the effects of variants to be adjusted for in conditional analysis are "homogeneous" or "heterogeneous" (default = "homogeneous").
`check_qc_label`	a logical value indicating whether variants need to be dropped according to `qc_label` specified in `custom_MetaSTAARlite_worker` (default = FALSE).
`variant_type`	a character value specifying the type of variant included in the analysis. Choices include "SNV", "Indel", or "variant" (default = "SNV").
`Use_annotation_weights`	a logical value which determines if annotations will be used as weights or not (default = TRUE).
`Annotation_name`	a character vector of annotation names used in MetaSTAARlite (default = NULL).
`silent`	a logical value which determines if the report of error messages will be suppressed (default = FALSE).

Value

a list of data frames containing the conditional MetaSTAAR p-values (including MetaSTAAR-O) corresponding to the custom mask.

References

Li, X., et al. (2023). Powerful, scalable and resource-efficient meta-analysis of rare variant associations in large whole genome sequencing studies. *Nature Genetics*, *55*(1), 154-164. (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*. (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)

`custom_MetaSTAARlite_worker` *Generates summary statistics of custom variant-sets (masks) using MetaSTAARlite*

Description

This function uses MetaSTAARlite to generate variant-level summary statistics and sparse covariance matrices for a user-specified custom mask.

Usage

```
custom_MetaSTAARlite_worker(
  chr,
  variant_list,
  agds_variant_list = NULL,
  genofile,
  obj_nullmodel,
  known_loci = NULL,
  cov_maf_cutoff = 0.05,
  signif.digits = NULL,
  QC_label = "annotation/filter",
  check_qc_label = FALSE,
  variant_type = c("SNV", "Indel", "variant"),
  Annotation_dir = "annotation/info/FunctionalAnnotation",
  Annotation_name_catalog,
  Use_annotation_weights = TRUE,
  Annotation_name = NULL,
  silent = FALSE
)
```

Arguments

`chr`	an integer which specifies the chromosome number.
`variant_list`	the data frame of variants in the custom mask to be included. Should contain five columns in the following order: chromosome (CHR), position (POS), reference allele (REF), alternative allele (ALT), and mask name (MaskName). An example is given (here)
`agds_variant_list`	the pre-loaded data frame of all variants in the aGDS for faster computation. Should contain four columns in the following order: chromosome (CHR), position (POS), reference allele (REF), and alternative allele (ALT),
`genofile`	an object of opened annotated GDS (aGDS) file.
`obj_nullmodel`	an object from fitting the null model, which is 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.

known_loci — the data frame of variants to be adjusted for in conditional analysis. Should contain four columns in the following order: chromosome (CHR), position (POS), reference allele (REF), and alternative allele (ALT) (default = NULL).

cov_maf_cutoff — a numeric value indicating the maximum minor allele frequency cutoff under which the sparse weighted covariance file between variants is stored (default = 0.05).

signif.digits — an integer indicating the number of significant digits to be used for storing the sparse weighted covariance file. If signif.digits is NULL, it is assumed that no rounding will be performed (default = NULL).

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

check_qc_label — a logical value indicating whether variants need to be dropped according to qc_label. If check_qc_label is FALSE, then the summary statistics will be stored for PASS variants from the study. If check_qc_label is TRUE, then the summary statistics will be stored for all variants from the study, together will an additional column of qc_label (default = FALSE).

variant_type — a character value specifying the type of variant included in the analysis. Choices include "SNV", "Indel", or "variant" (default = "SNV").

Annotation_dir — a character specifying the channel name of the annotations in the aGDS file (default = "annotation/info/FunctionalAnnotation").

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

Use_annotation_weights — a logical value which specifies if annotations will be used as weights or not (default = TRUE).

Annotation_name — a character vector of annotation names used in MetaSTAARlite (default = NULL).

silent — a logical value which determines if the report of error messages will be suppressed (default = FALSE).

Value

a list of the following objects corresponding to the custom mask: (1) the data frame of all variants in the variant-set (the variant-level summary statistics file), including the following information: chromosome (chr), position (pos), reference allele (ref), alternative allele (alt), quality control status (qc_label, optional), alternative allele count (alt_AC), minor allele count (MAC), minor allele frequency (MAF), study sample size (N), score statistic (U), variance (V), variant annotations specified in Annotation_name, and the low-rank decomposed component of the covariance file; (2) the sparse matrix of all variants in the variant-set whose minor allele frequency is below cov_maf_cutoff (the sparse weighted covariance file); (3) the summary statistics and covariance matrices corresponding to the specified gene for variants to be conditioned on in known_loci.

References

Li, X., et al. (2023). Powerful, scalable and resource-efficient meta-analysis of rare variant associations in large whole genome sequencing studies. *Nature Genetics*, *55*(1), 154-164. (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*. (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_Coding_Results_Summary_meta` *Generates summary table and visualization for the meta-analysis of coding functional categories that was conducted using MetaSTAARlite*

Description

This function takes in objects of gene-centric coding meta-analysis results from MetaSTAARlite and generates a summary table, Manhattan plot, and QQ plot for the meta-analysis of coding functional categories that was conducted based on the parameters provided by the user.

Usage

```
Gene_Centric_Coding_Results_Summary_meta(
  gene_centric_coding_jobs_num,
  input_path,
  output_path,
  gene_centric_results_name,
  alpha = 2.5e-06,
  manhattan_plot = FALSE,
  QQ_plot = FALSE
)
```

Arguments

`gene_centric_coding_jobs_num`	an integer which specifies the number of jobs done in the gene-centric coding meta-analysis.
`input_path`	a character which specifies the file path to the gene-centric coding meta-analysis results files.
`output_path`	a character which specifies the file path to the desired location of the produced summary table and visualizations for gene-centric coding meta-analysis.
`gene_centric_results_name`	a character which specifies the name (excluding the jobs number) of the gene-centric coding meta-analysis results files.
`alpha`	a numeric value which specifies the desired significance threshold for the gene-centric coding meta-analysis (default = 2.5E-06).
`manhattan_plot`	a logical value which determines if a Manhattan plot is generated (default = FALSE).
`QQ_plot`	a logical value which determines if a QQ plot is generated (default = FALSE).

`Gene_Centric_Noncoding_Results_Summary_meta`
Generates summary table and visualization for the meta-analysis of noncoding functional categories that was conducted using MetaSTAARlite

Description

This function takes in objects of gene-centric noncoding meta-analysis results from MetaSTAARlite and generates a summary table, Manhattan plot, and QQ plot for the meta-analysis of noncoding functional categories that was conducted based on the parameters provided by the user.

Usage

```
Gene_Centric_Noncoding_Results_Summary_meta(
  gene_centric_noncoding_jobs_num,
  input_path,
  output_path,
  gene_centric_results_name,
  ncRNA_jobs_num,
  ncRNA_input_path,
  ncRNA_output_path,
  ncRNA_results_name,
  alpha = 2.5e-06,
  alpha_ncRNA = 2.5e-06,
  ncRNA_pos = NULL,
  manhattan_plot = FALSE,
  QQ_plot = FALSE
)
```

Arguments

`gene_centric_noncoding_jobs_num`
an integer which specifies the number of jobs done in the gene-centric noncoding meta-analysis.

`input_path` a character which specifies the file path to the gene-centric noncoding meta-analysis results files.

`output_path` a character which specifies the file path to the desired location of the produced summary table and visualizations for gene-centric noncoding meta-analysis.

`gene_centric_results_name`
a character which specifies the name (excluding the jobs number) of the gene-centric noncoding meta-analysis results files.

`ncRNA_jobs_num` an integer which specifies the number of jobs done in the ncRNA meta-analysis.

`ncRNA_input_path`
a character which specifies the file path to the ncRNA meta-analysis results files.

`ncRNA_output_path`
a character which specifies the file path to the desired location of the produced summary table and visualizations for the ncRNA meta-analysis.

`ncRNA_results_name`
: a character which specifies the name (excluding the jobs number) of the ncRNA meta-analysis results files.

`alpha`
: a numeric value which specifies the desired significance threshold for the gene-centric noncoding meta-analysis (default = 2.5E-06).

`alpha_ncRNA`
: a numeric value which specifies the desired significance threshold for the ncRNA meta-analysis (default = 2.5E-06).

`ncRNA_pos`
: positions of ncRNA genes, required for generating the Manhattan plot and Q-Q plot of the results of ncRNA genes (default = NULL).

`manhattan_plot`
: a logical value which determines if a Manhattan plot is generated (default = FALSE).

`QQ_plot`
: a logical value which determines if a QQ plot is generated (default = FALSE).

`Genotype_flip_sp_extraction` *Genotype extraction and filtering for association analysis*

Description

The `Genotype_flip_sp_extraction` function extracts genotype data for a given set of variants and samples from a GDS file. It applies multiple filters based on allele frequency, missingness rate, and user-defined thresholds to categorize variants into different dosage groups. The function returns a sparse genotype matrix (`"dgCMatrix"`), variant-level information, and optional annotation data.

Usage

```
Genotype_flip_sp_extraction(
  genofile,
  variant.id,
  sample.id,
  REF_AF,
  rare_maf_cutoff = 0.01,
  variant_maf_cutoff_filter = 1,
  Missing_rate,
  Missing_cutoff = 0.01,
  subset_variants_num = 1000,
  rv_num_cutoff_max_prefilter = 1e+09,
  annotation_phred = NULL,
  QC_label = "annotation/filter"
)
```

Arguments

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

`variant.id`
: ID of selected variants.

`sample.id`
: ID of selected samples.

`REF_AF`
: a numeric vector of reference allele frequencies for each variant.

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

`variant_maf_cutoff_filter`
: a numeric value specifying the MAF threshold for excluding common variants in variant-set analysis. Default is 1 for individual-variant analysis.

`Missing_rate`
: a numeric vector of missing rates for each variant.

`Missing_cutoff`
: a numeric value specifying the maximum missing rate threshold for defining high-missing variants (default = 0.01).

`subset_variants_num`
: the number of variants to extract per subset for each time in Case 1 and Case 2 (default = 2e3).

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

`annotation_phred`
: a optional data frame or matrix of functional annotation data of dimension p*q (or a vector of a single annotation score with length p). See `STAAR::STAAR` for more details.

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

Details

The extracted variants are processed based on the following three cases: (1) Case 1: `ALT_AF > 0.5` - Use `"$dosage"` to extract Geno (dosages of the reference allele). - Convert the extracted data into the `"dgCMatrix"` format (sparse matrix).

(2) Case 2: `ALT_AF <= 0.5` and (`MAF >= rare_maf_cutoff` or `Missing_rate >= Missing_cutoff`) - Use `"$dosage_alt"` to extract Geno (dosages of the alternative allele). - Convert the extracted data into the `"dgCMatrix"` format (sparse matrix).

(3) Case 3: `ALT_AF <= 0.5` and (`MAF < rare_maf_cutoff` and `Missing_rate < Missing_cutoff`) - Use `"$dosage_sp"` to directly extract Geno in the `"dgCMatrix"` format.

Note: - `REF_AF` and `Missing_rate` can be efficiently computed using `SeqArray::seqGetAF_AC_Missing(genofile, minor=FALSE)`.

Value

A list containing:

`Geno`
: A sparse matrix of genotypes.

`results_information`
: A data frame with variant-level details, including chromosome, position, reference and alternative alleles, MAF, ALT_AF, missing rate, and variant ID.

`annotation_phred`
: A data frame containing filtered functional annotation data if provided; otherwise, NULL.

individual_analysis_MetaSTAARlite

Performs meta-analysis of individual variants using MetaSTAARlite

Description

This function performs meta-analysis to detect associations between a quantitative/dichotomous phenotype and each individual variant in a genetic region by using score test.

Usage

```
individual_analysis_MetaSTAARlite(
  sample.sizes,
  sumstat.list,
  mac_cutoff = 20,
  check_qc_label = FALSE
)
```

Arguments

sample.sizes	a numeric vector with the length of study.names indicating the sample size of each study.
sumstat.list	a list containing study-specific summary statistics from all participating studies.
mac_cutoff	an integer specifying the cutoff of minimum combined minor allele count in defining individual variants (default = 20).
check_qc_label	a logical value indicating whether variants need to be dropped according to qc_label specified in individual_analysis_MetaSTAARlite_worker (default = FALSE).

Value

a data frame containing the meta-analysis score test p-value and the estimated effect size of the alternative allele for each individual variant in the given genetic region.

individual_analysis_MetaSTAARlite_cond

Performs conditional meta-analysis of individual variants using MetaSTAARlite

Description

This function performs conditional meta-analysis to detect associations between a quantitative/dichotomous phenotype and each (significant) individual variant by using conditional score test.

Usage

```
individual_analysis_MetaSTAARlite_cond(
  individual_results,
  sample.sizes,
  sumstat.list,
  covcond.list,
  mac_cutoff = 20,
  effect.cond = c("homogeneous", "heterogeneous"),
  check_qc_label = FALSE
)
```

Arguments

individual_results	a dataframe containing the (significant) results of the individual variant meta-analysis.
sample.sizes	a numeric vector with the length of study.names indicating the sample size of each study.
sumstat.list	a list containing study-specific summary statistics from all participating studies.
covcond.list	a list containing study-specific summary statistics and covariance matrices for variants to be conditioned on from all participating studies.
mac_cutoff	an integer specifying the cutoff of minimum combined minor allele count in defining individual variants (default = 20).
effect.cond	a character value indicating the effects of variants to be adjusted for in conditional analysis are "homogeneous" or "heterogeneous" (default = "homogeneous").
check_qc_label	a logical value indicating whether variants need to be dropped according to qc_label specified in coding_MetaSTAARlite_worker (default = FALSE).

Value

a data frame containing the conditional meta-analysis score test p-value and the estimated effect size of the alternative allele for each (significant) individual variant in individual_results.

---

individual_analysis_MetaSTAARlite_worker

Generates summary statistics of individual variants using MetaSTAARlite

---

Description

This function uses MetaSTAARlite to generate variant-level summary statistics for individual variants of interest.

Usage

```
individual_analysis_MetaSTAARlite_worker(
  chr,
  start_loc,
  end_loc,
  genofile,
  obj_nullmodel,
  known_loci = NULL,
  subsegment.size = 50000,
  QC_label = "annotation/filter",
  check_qc_label = FALSE,
  variant_type = c("variant", "SNV", "Indel"),
  Annotation_dir = "annotation/info/FunctionalAnnotation",
  Annotation_name_catalog,
  Use_annotation_weights = c(FALSE, TRUE),
  Annotation_name = NULL,
  silent = FALSE
)
```

Arguments

chr	an integer which specifies the chromosome number.
start_loc	an integer which specifies the starting location of variants to be analyzed.
end_loc	an integer which specifies the end location of variants to be analyzed.
genofile	an object of opened annotated GDS (aGDS) file.
obj_nullmodel	an object from fitting the null model, which is 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.
known_loci	the data frame of variants to be adjusted for in conditional analysis. Should contain four columns in the following order: chromosome (CHR), position (POS), reference allele (REF), and alternative allele (ALT) (default = NULL).
subsegment.size	a numeric value which specifies the size of each subsegment for computation (default = 5e4).
QC_label	a character specifying the channel name of the QC label in the GDS/aGDS file (default = "annotation/filter").
check_qc_label	a logical value indicating whether variants need to be dropped according to qc_label. If check_qc_label is FALSE, then the summary statistics will be stored for PASS variants from the study. If check_qc_label is TRUE, then the summary statistics will be stored for all variants from the study, together will an additional column of qc_label (default = FALSE).
variant_type	a character value specifying the type of variant included in the analysis. Choices include "SNV", "Indel", or "variant" (default = "variant").
Annotation_dir	a character specifying the channel name of the annotations in the aGDS file (default = "annotation/info/FunctionalAnnotation").
Annotation_name_catalog	a data frame containing the annotation name and the corresponding channel name in the aGDS file.

`Use_annotation_weights`
: a logical value which specifies if annotations will be used as weights or not (default = FALSE).

`Annotation_name`
: a character vector of annotation names used in MetaSTAARlite (default = NULL).

`silent`
: a logical value which determines if the report of error messages will be suppressed (default = FALSE).

Value

a list of the following objects corresponding to each individual variant in the given genetic region: (1) the data frame of all variants in the variant-set (the variant-level summary statistics file), including the following information: chromosome (chr), position (pos), reference allele (ref), alternative allele (alt), quality control status (qc_label, optional), alternative allele count (alt_AC), minor allele count (MAC), minor allele frequency (MAF), study sample size (N), score statistic (U), variance (V), variant annotations specified in `Annotation_name` (optional), and the low-rank decomposed component of the covariance file; (2) the summary statistics and covariance matrices corresponding to the specified gene for variants to be conditioned on in `known_loci`.

References

Li, X., et al. (2023). Powerful, scalable and resource-efficient meta-analysis of rare variant associations in large whole genome sequencing studies. *Nature Genetics*, *55*(1), 154-164. (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*. (pub)

`Individual_Analysis_Results_Summary_meta` *Generates summary table and visualization for the meta-analysis of individual variants that was conducted using MetaSTAARlite*

Description

This function takes in objects of individual meta-analysis results and generates a summary table, Manhattan plot, and QQ plot for the meta-analysis of individual variants that was conducted based on the parameters provided by the user.

Usage

```
Individual_Analysis_Results_Summary_meta(
  jobs_num,
  input_path,
  output_path,
  individual_results_name,
  alpha = 5e-09,
  manhattan_plot = FALSE,
  QQ_plot = FALSE
)
```

Arguments

jobs_num	an integer which specifies the number of jobs done in the individual meta-analysis.
input_path	a character which specifies the file path to the individual meta-analysis results files.
output_path	a character which specifies the file path to the desired location of the produced summary table and visualizations for individual meta-analysis.
individual_results_name	a character which specifies the name (excluding the jobs number) of the individual meta-analysis results files.
alpha	a numeric value which specifies the desired significance threshold (default = 5E-09).
manhattan_plot	a logical value which determines if a Manhattan plot is generated (default = FALSE).
QQ_plot	a logical value which determines if a QQ plot is generated (default = FALSE).

MetaSTAARlite_merge *Merges the generated summary statistics and sparse covariance matrices of different studies*

Description

This function merges the generated variant-level summary statistics and sparse covariance matrices of all participating studies in preparation for the meta-analysis step of MetaSTAARlite.

Usage

```
MetaSTAARlite_merge(
  chr,
  sample.sizes,
  sumstat.list,
  cov.list,
  rare_maf_cutoff = 0.01,
  cov_maf_cutoff,
  check_qc_label = FALSE,
  variant_type = c("SNV", "Indel", "variant"),
  Use_annotation_weights = TRUE,
  Annotation_name = NULL
)
```

Arguments

chr	an integer which specifies the chromosome number.
sample.sizes	a numeric vector with the length of `study.names` indicating the sample size of each study.
sumstat.list	a list containing study-specific summary statistics from all participating studies.
cov.list	a list containing study-specific sparse weighted covariance matrices from all participating studies.

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

`cov_maf_cutoff`
: a numeric vector with the length of `study.names` indicating the maximum minor allele frequency cutoffs under which the sparse weighted covariance files between variants are stored.

`check_qc_label`
: a logical value indicating whether variants need to be dropped according to `qc_label` (default = FALSE).

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

`Use_annotation_weights`
: a logical value which determines if annotations will be used as weights or not (default = TRUE).

`Annotation_name`
: a character vector of annotation names used in MetaSTAARlite (default = NULL).

Value

a list with the following members:

`info`: the merged data frame of all variants in the variant-set of interest whose combined minor allele frequency is below `rare_maf_cutoff`, including the following information (listed in the same order as `U` and the rows/columns of `cov`): chromosome (chr), position (pos), reference allele (ref), alternative allele (alt), combined minor allele count (MAC), and combined minor allele frequency (MAF).

`U` the merged score statistics vector of all variants in the variant-set of interest whose combined minor allele frequency is below `rare_maf_cutoff`.

`cov` the merged covariance matrix of all variants in the variant-set of interest whose combined minor allele frequency is below `rare_maf_cutoff`.

`annotation_phred` the merged functional annotation data in PHRED score scale of all variants in the variant-set of interest whose combined minor allele frequency is below `rare_maf_cutoff`.

References

Li, X., et al. (2023). Powerful, scalable and resource-efficient meta-analysis of rare variant associations in large whole genome sequencing studies. *Nature Genetics*, *55*(1), 154-164. (pub)

`MetaSTAARlite_merge_cond`
: *Merges the generated summary statistics and covariance matrices of different studies for conditional analysis*

Description

This function merges the generated variant-level summary statistics and covariance matrices of all participating studies in preparation for the conditional meta-analysis step of MetaSTAARlite.

Usage

```
MetaSTAARlite_merge_cond(
  chr,
  sample.sizes,
  sumstat.list,
  cov.list,
  covcond.list,
  rare_maf_cutoff = 0.01,
  cov_maf_cutoff,
  effect.cond = c("homogeneous", "heterogeneous"),
  check_qc_label = FALSE,
  variant_type = c("SNV", "Indel", "variant"),
  Use_annotation_weights = TRUE,
  Annotation_name = NULL
)
```

Arguments

chr	an integer which specifies the chromosome number.
sample.sizes	a numeric vector with the length of study.names indicating the sample size of each study.
sumstat.list	a list containing study-specific summary statistics from all participating studies.
cov.list	a list containing study-specific sparse weighted covariance matrices from all participating studies.
covcond.list	a list containing study-specific summary statistics and covariance matrices for variants to be conditioned on from all participating studies.
rare_maf_cutoff	the cutoff of maximum minor allele frequency in defining rare variants (default = 0.01).
cov_maf_cutoff	a numeric vector with the length of study.names indicating the maximum minor allele frequency cutoffs under which the sparse weighted covariance files between variants are stored.
effect.cond	a character value indicating the effects of variants to be adjusted for in conditional analysis are "homogeneous" or "heterogeneous" (default = "homogeneous").
check_qc_label	a logical value indicating whether variants need to be dropped according to qc_label (default = FALSE).
variant_type	a character value specifying the type of variant included in the analysis. Choices include "SNV", "Indel", or "variant" (default = "SNV").
Use_annotation_weights	a logical value which determines if annotations will be used as weights or not (default = TRUE).
Annotation_name	a character vector of annotation names used in MetaSTAARlite (default = NULL).

Value

a list with the following members:

`info`: the merged data frame of all variants in the variant-set of interest whose combined minor allele frequency is below `rare_maf_cutoff`, including the following information (listed in the same order as `U` and the rows/columns of `cov`): chromosome (chr), position (pos), reference allele (ref), alternative allele (alt), combined minor allele count (MAC), and combined minor allele frequency (MAF).

`U_cond`: the merged conditional score statistics vector of all variants in the variant-set of interest whose combined minor allele frequency is below `rare_maf_cutoff`, adjusting for a given list of variants.

`cov_cond`: the merged conditional covariance matrix of all variants in the variant-set of interest whose combined minor allele frequency is below `rare_maf_cutoff`, adjusting for a given list of variants.

`annotation_phred` the merged functional annotation data in PHRED score scale of all variants in the variant-set of interest whose combined minor allele frequency is below `rare_maf_cutoff`.

References

Li, X., et al. (2023). Powerful, scalable and resource-efficient meta-analysis of rare variant associations in large whole genome sequencing studies. *Nature Genetics*, *55*(1), 154-164. (pub)

`MetaSTAARlite_worker_cov` *Generating sparse weighted covariance file using MetaSTAARlite (the "worker" step)*

Description

The `MetaSTAARlite_worker_cov` function takes in genotype and the object from fitting the null model to generate the sparse weighted covariance file for the given variant-set.

Usage

```
MetaSTAARlite_worker_cov(
  genotype,
  obj_nullmodel,
  cov_maf_cutoff = 0.05,
  qc_label = NULL,
  signif.digits = 3
)
```

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. If the input genotype matrix is sparse (e.g. `dgCMatrix` format), it is assumed that it has been flipped to represent minor allele coding.
`obj_nullmodel`	an object from fitting the null model, which is 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.

cov_maf_cutoff	a numeric value indicating the maximum minor allele frequency cutoff under which the sparse weighted covariance file between variants is stored (default = 0.05).
qc_label	a vector of quality control status for each variant in genotype, where a PASS variant is labeled as "PASS". If qc_label is NULL, it is assumed that all variants are PASS variants in the study (default = NULL).
signif.digits	an integer indicating the number of significant digits to be used for storing the sparse weighted covariance file. If signif.digits is NULL, it is assumed that no rounding will be performed (default = 3).

Value

GTSinvG_rare: the sparse matrix of all variants in the variant-set whose minor allele frequency is below cov_maf_cutoff (the sparse weighted covariance file).

References

Li, X., et al. (2023). Powerful, scalable and resource-efficient meta-analysis of rare variant associations in large whole genome sequencing studies. *Nature Genetics*, *55*(1), 154-164. (pub)

MetaSTAARlite_worker_cov_cond

Generating summary statistics file for conditional analysis using MetaSTAARlite (the "worker" step)

Description

The MetaSTAARlite_worker_cov_cond function takes in genotype, the genotype of variants to be adjusted for in conditional analysis, the object from fitting the null model, variant information and adjusted variant information (unique identifier) to generate the summary statistics file for the given variant-set, adjusting for a given list of variants.

Usage

```
MetaSTAARlite_worker_cov_cond(
  genotype,
  genotype_adj,
  obj_nullmodel,
  variant_info,
  variant_adj_info
)
```

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. If the input genotype matrix is sparse (e.g. dgCMatrix format), it is assumed that it has been flipped to represent minor allele coding.
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 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.

`variant_info` a data frame of variant information (unique identifier) with p rows (listed in the same order as the columns of `genotype`) and should contain the following 4 columns: chromosome (chr), position (pos), reference allele (ref), and alternative allele (alt).

`variant_adj_info` a data frame of adjusted variant information (unique identifier) with p_adj rows (listed in the same order as the rows of `genotype_adj`) and should contain the following 4 columns: chromosome (chr), position (pos), reference allele (ref), and alternative allele (alt).

Value

a list with the following members:

`GTPG_cond`: the covariance matrix between all variants in the variant-set (rows) and all variants in the conditional variant-set (columns) (the covariance file for conditional analysis).

`variant_info`: the data frame of variant information (unique identifier) with p rows (listed in the same order as the rows of `GTPG_cond`) and 4 columns: chromosome (chr), position (pos), reference allele (ref), and alternative allele (alt).

`variant_adj_info`: the data frame of adjusted variant information (unique identifier) with p_adj rows (listed in the same order as the columns of `GTPG_cond`) and 4 columns: chromosome (chr), position (pos), reference allele (ref), alternative allele (alt), score statistic (U), and variance (V).

References

Li, X., et al. (2023). Powerful, scalable and resource-efficient meta-analysis of rare variant associations in large whole genome sequencing studies. *Nature Genetics*, *55*(1), 154-164. (pub)

`MetaSTAARlite_worker_sumstat` *Generating variant-level summary statistics file using MetaSTAARlite (the "worker" step)*

Description

The `MetaSTAARlite_worker_sumstat` function takes in genotype, the object from fitting the null model, and variant information (unique identifier) to generate the variant-level summary statistics file for the given variant-set.

Usage

```
MetaSTAARlite_worker_sumstat(
  genotype,
  ALT_AF = NULL,
  obj_nullmodel,
  variant_info,
  qc_label = NULL,
```

```
  annotation_phred = NULL,
  for_individual_analysis = FALSE
)
```

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. If the input genotype matrix is sparse (e.g. dgCMatrix format), it is assumed that it has been flipped to represent minor allele coding.
ALT_AF	a numeric vector of alternate allele frequencies for each variant in genotype. This is required when genotype is in sparse format (e.g., dgCMatrix). This vector is the ALT_AF column from the results_info data frame returned by the Genotype_flip_sp_extraction function (default = NULL).
obj_nullmodel	an object from fitting the null model, which is 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.
variant_info	a data frame of variant information (unique identifier) with p rows (listed in the same order as the columns of genotype) and should contain the following 4 columns: chromosome (chr), position (pos), reference allele (ref), and alternative allele (alt).
qc_label	a vector of quality control status for each variant in variant_info, where a PASS variant is labeled as "PASS". If qc_label is NULL, it is assumed that all variants are PASS variants in the study (default = NULL).
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, STAAR will perform the 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).
for_individual_analysis	a logical value indicating whether it is used for individual (single-variant) meta-analysis (default = FALSE).

Value

sumstat: the data frame of all variants in the variant-set or the list of individual variants (the variant-level summary statistics file), including the following information: chromosome (chr), position (pos), reference allele (ref), alternative allele (alt), quality control status (qc_label, optional), alternative allele count (alt_AC), minor allele count (MAC), minor allele frequency (MAF), study sample size (N), score statistic (U), variance (V), variant annotations provided in annotation_phred, and the low-rank decomposed component of the covariance file.

References

Li, X., et al. (2023). Powerful, scalable and resource-efficient meta-analysis of rare variant associations in large whole genome sequencing studies. *Nature Genetics*, *55*(1), 154-164. (pub)

na.replace.sp *Replace Missing Values in a Sparse Genotype Matrix*

Description

The na.replace.sp function replaces missing values (NA) in a sparse genotype matrix ('dgCMatrix' format). If 'is_NA_to_Zero = TRUE', NA values are replaced with 0. Otherwise, NA values in each column are replaced with the corresponding entries in 'm'. This function is inspired by glmnet::na.replace for sparse matrices.

Usage

```
na.replace.sp(genotype_sp, m, is_NA_to_Zero = FALSE)
```

Arguments

genotype_sp	A sparse genotype matrix of class 'dgCMatrix' from the Matrix package.
m	A numeric vector specifying the replacement values for each column.
is_NA_to_Zero	A logical value indicating whether NA values should be replaced with 0 (default: FALSE). If FALSE, NA values are replaced column-wise using 'm'.

Value

A 'dgCMatrix' object with missing values replaced accordingly.

Examples

```
library(Matrix)
set.seed(123)
# Create a sparse matrix with some NA values
mat <- Matrix(c(1, NA, 3, 0, NA, 2, 4, 5, NA), nrow = 3, sparse = TRUE)
print(mat)

# Replace NA values with 0
mat_imputed <- na.replace.sp(mat, m = c(0.5, 1, 1.5), is_NA_to_Zero = TRUE)
print(mat_imputed)

# Replace NA values with values from m
mat_imputed_m <- na.replace.sp(mat, m = c(0.5, 1, 1.5), is_NA_to_Zero = FALSE)
print(mat_imputed_m)
```

ncRNA_MetaSTAARlite *Performs meta-analysis of long noncoding RNA (ncRNA) category using MetaSTAARlite*

Description

This function performs meta-analysis to detect associations between a quantitative/dichotomous phenotype and the exonic and splicing category of an ncRNA gene by using the MetaSTAARlite pipeline. For each coding functional category, the MetaSTAAR-O p-value is a p-value from an omnibus test that aggregated SKAT-MS(1,25), SKAT-MS(1,1), Burden-MS(1,25), Burden-MS(1,1), ACAT-V-MS(1,25), and ACAT-V-MS(1,1) together with p-values of each test weighted by each annotation using Cauchy method.

Usage

```
ncRNA_MetaSTAARlite(
  chr,
  gene_name,
  sample.sizes,
  ncRNA_sumstat_gene_list,
  ncRNA_cov_gene_list,
  cov_maf_cutoff,
  rare_maf_cutoff = 0.01,
  rv_num_cutoff = 2,
  check_qc_label = FALSE,
  variant_type = c("SNV", "Indel", "variant"),
  Use_annotation_weights = TRUE,
  Annotation_name = NULL,
  silent = FALSE
)
```

Arguments

`chr`	an integer which specifies the chromosome number.
`gene_name`	a character which specifies the name of the ncRNA gene to be meta-analyzed using MetaSTAARlite.
`sample.sizes`	a numeric vector with the length of `study.names` indicating the sample size of each study.
`ncRNA_sumstat_gene_list`	a list containing study-specific summary statistics corresponding to the specified gene.
`ncRNA_cov_gene_list`	a list containing study-specific sparse weighted covariance matrices corresponding to the specified gene.
`cov_maf_cutoff`	a numeric vector with the length of `study.names` indicating the maximum minor allele frequency cutoffs under which the sparse weighted covariance files between variants are stored.
`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 meta-analyzing a given variant-set (default = 2).
`check_qc_label`	a logical value indicating whether variants need to be dropped according to `qc_label` specified in `ncRNA_MetaSTAARlite_worker` (default = FALSE).
`variant_type`	a character value specifying the type of variant included in the analysis. Choices include "SNV", "Indel", or "variant" (default = "SNV").

`Use_annotation_weights`
: a logical value which determines if annotations will be used as weights or not (default = TRUE).

`Annotation_name`
: a character vector of annotation names used in MetaSTAARlite (default = NULL).

`silent`
: a logical value which determines if the report of error messages will be suppressed (default = FALSE).

Value

a list of data frames containing the MetaSTAAR p-values (including MetaSTAAR-O) corresponding to the exonic and splicing category of the given ncRNA gene.

References

Li, X., et al. (2023). Powerful, scalable and resource-efficient meta-analysis of rare variant associations in large whole genome sequencing studies. *Nature Genetics*, *55*(1), 154-164. (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*. (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)

`ncRNA_MetaSTAARlite_cond`

Performs conditional meta-analysis of long noncoding RNA (ncRNA) category using MetaSTAARlite

---

Description

This function performs meta-analysis to detect conditional associations between a quantitative/dichotomous phenotype and the exonic and splicing category of an ncRNA gene adjusting for set of known variants by using the MetaSTAARlite pipeline. For each coding functional category, the conditional MetaSTAAR-O p-value is a p-value from an omnibus test that aggregated conditional SKAT-MS(1,25), SKAT-MS(1,1), Burden-MS(1,25), Burden-MS(1,1), ACAT-V-MS(1,25), and ACAT-V-MS(1,1) together with conditional p-values of each test weighted by each annotation using Cauchy method.

Usage

```
ncRNA_MetaSTAARlite_cond(
  chr,
  gene_name,
  sample.sizes,
  ncRNA_sumstat_gene_list,
  ncRNA_cov_gene_list,
  ncRNA_cov_cond_gene_list,
  cov_maf_cutoff,
  rare_maf_cutoff = 0.01,
  rv_num_cutoff = 2,
```

```
  effect.cond = c("homogeneous", "heterogeneous"),
  check_qc_label = FALSE,
  variant_type = c("SNV", "Indel", "variant"),
  Use_annotation_weights = TRUE,
  Annotation_name = NULL,
  silent = FALSE
)
```

Arguments

chr	an integer which specifies the chromosome number.
gene_name	a character which specifies the name of the ncRNA gene to be meta-analyzed using MetaSTAARlite.
sample.sizes	a numeric vector with the length of study.names indicating the sample size of each study.
ncRNA_sumstat_gene_list	a list containing study-specific summary statistics corresponding to the specified gene.
ncRNA_cov_gene_list	a list containing study-specific sparse weighted covariance matrices corresponding to the specified gene.
ncRNA_cov_cond_gene_list	a list containing study-specific summary statistics and covariance matrices corresponding to the specified gene for variants to be conditioned on.
cov_maf_cutoff	a numeric vector with the length of study.names indicating the maximum minor allele frequency cutoffs under which the sparse weighted covariance files between variants are stored.
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 meta-analyzing a given variant-set (default = 2).
effect.cond	a character value indicating the effects of variants to be adjusted for in conditional analysis are "homogeneous" or "heterogeneous" (default = "homogeneous").
check_qc_label	a logical value indicating whether variants need to be dropped according to qc_label specified in ncRNA_MetaSTAARlite_worker (default = FALSE).
variant_type	a character value specifying the type of variant included in the analysis. Choices include "SNV", "Indel", or "variant" (default = "SNV").
Use_annotation_weights	a logical value which determines if annotations will be used as weights or not (default = TRUE).
Annotation_name	a character vector of annotation names used in MetaSTAARlite (default = NULL).
silent	a logical value which determines if the report of error messages will be suppressed (default = FALSE).

Value

a list of data frames containing the conditional MetaSTAAR p-values (including MetaSTAAR-O) corresponding to the exonic and splicing category of the given ncRNA gene.

References

Li, X., et al. (2023). Powerful, scalable and resource-efficient meta-analysis of rare variant associations in large whole genome sequencing studies. *Nature Genetics*, *55*(1), 154-164. (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*. (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)

`ncRNA_MetaSTAARlite_worker`

Generates summary statistics of long noncoding RNA (ncRNA) category using MetaSTAARlite

---

Description

This function uses MetaSTAARlite to generate variant-level summary statistics and sparse covariance matrices the exonic and splicing category of an ncRNA gene.

Usage

```
ncRNA_MetaSTAARlite_worker(
  chr,
  gene_name,
  genofile,
  obj_nullmodel,
  known_loci = NULL,
  cov_maf_cutoff = 0.05,
  signif.digits = NULL,
  QC_label = "annotation/filter",
  check_qc_label = FALSE,
  variant_type = c("SNV", "Indel", "variant"),
  Annotation_dir = "annotation/info/FunctionalAnnotation",
  Annotation_name_catalog,
  Use_annotation_weights = TRUE,
  Annotation_name = NULL,
  silent = FALSE
)
```

Arguments

`chr`	an integer which specifies the chromosome number.
`gene_name`	a character which specifies the name of the gene to be meta-analyzed using MetaSTAARlite.
`genofile`	an object of opened annotated GDS (aGDS) file.
`obj_nullmodel`	an object from fitting the null model, which is 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.

known_loci — the data frame of variants to be adjusted for in conditional analysis. Should contain four columns in the following order: chromosome (CHR), position (POS), reference allele (REF), and alternative allele (ALT) (default = NULL).

cov_maf_cutoff — a numeric value indicating the maximum minor allele frequency cutoff under which the sparse weighted covariance file between variants is stored (default = 0.05).

signif.digits — an integer indicating the number of significant digits to be used for storing the sparse weighted covariance file. If signif.digits is NULL, it is assumed that no rounding will be performed (default = NULL).

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

check_qc_label — a logical value indicating whether variants need to be dropped according to qc_label. If check_qc_label is FALSE, then the summary statistics will be stored for PASS variants from the study. If check_qc_label is TRUE, then the summary statistics will be stored for all variants from the study, together will an additional column of qc_label (default = FALSE).

variant_type — a character value specifying the type of variant included in the analysis. Choices include "SNV", "Indel", or "variant" (default = "SNV").

Annotation_dir — a character specifying the channel name of the annotations in the aGDS file (default = "annotation/info/FunctionalAnnotation").

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

Use_annotation_weights — a logical value which specifies if annotations will be used as weights or not (default = TRUE).

Annotation_name — a character vector of annotation names used in MetaSTAARlite (default = NULL).

silent — a logical value which determines if the report of error messages will be suppressed (default = FALSE).

Value

a list of the following objects corresponding to the exonic and splicing category of the given ncRNA gene: (1) the data frame of all variants in the variant-set (the variant-level summary statistics file), including the following information: chromosome (chr), position (pos), reference allele (ref), alternative allele (alt), quality control status (qc_label, optional), alternative allele count (alt_AC), minor allele count (MAC), minor allele frequency (MAF), study sample size (N), score statistic (U), variance (V), variant annotations specified in Annotation_name, and the low-rank decomposed component of the covariance file; (2) the sparse matrix of all variants in the variant-set whose minor allele frequency is below cov_maf_cutoff (the sparse weighted covariance file); (3) the summary statistics and covariance matrices corresponding to the specified gene for variants to be conditioned on in known_loci.

References

Li, X., et al. (2023). Powerful, scalable and resource-efficient meta-analysis of rare variant associations in large whole genome sequencing studies. *Nature Genetics*, *55*(1), 154-164. (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*. (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)

`noncoding_MetaSTAARlite` *Performs meta-analysis of noncoding functional categories using MetaSTAARlite*

Description

This function performs meta-analysis to detect associations between a quantitative/dichotomous phenotype and noncoding functional categories of a gene by using the MetaSTAARlite pipeline. For each coding functional category, the MetaSTAAR-O p-value is a p-value from an omnibus test that aggregated SKAT-MS(1,25), SKAT-MS(1,1), Burden-MS(1,25), Burden-MS(1,1), ACAT-V-MS(1,25), and ACAT-V-MS(1,1) together with p-values of each test weighted by each annotation using Cauchy method.

Usage

```
noncoding_MetaSTAARlite(
  chr,
  gene_name,
  sample.sizes,
  noncoding_sumstat_gene_list,
  noncoding_cov_gene_list,
  cov_maf_cutoff,
  rare_maf_cutoff = 0.01,
  rv_num_cutoff = 2,
  check_qc_label = FALSE,
  variant_type = c("SNV", "Indel", "variant"),
  Use_annotation_weights = TRUE,
  Annotation_name = NULL,
  silent = FALSE
)
```

Arguments

`chr`	an integer which specifies the chromosome number.
`gene_name`	a character which specifies the name of the gene to be meta-analyzed using MetaSTAARlite.
`sample.sizes`	a numeric vector with the length of `study.names` indicating the sample size of each study.
`noncoding_sumstat_gene_list`	a list containing study-specific summary statistics corresponding to the specified gene.
`noncoding_cov_gene_list`	a list containing study-specific sparse weighted covariance matrices corresponding to the specified gene.

`cov_maf_cutoff` a numeric vector with the length of `study.names` indicating the maximum minor allele frequency cutoffs under which the sparse weighted covariance files between variants are stored.

`rare_maf_cutoff` a numeric value specifying the cutoff of maximum minor allele frequency in defining rare variants (default = 0.01).

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

`check_qc_label` a logical value indicating whether variants need to be dropped according to `qc_label` specified in `noncoding_MetaSTAARlite_worker` (default = FALSE).

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

`Use_annotation_weights` a logical value which determines if annotations will be used as weights or not (default = TRUE).

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

`silent` a logical value which determines if the report of error messages will be suppressed (default = FALSE).

Value

a list of data frames containing the MetaSTAAR p-values (including MetaSTAAR-O) corresponding to each noncoding functional category of the given gene.

References

Li, X., et al. (2023). Powerful, scalable and resource-efficient meta-analysis of rare variant associations in large whole genome sequencing studies. *Nature Genetics*, *55*(1), 154-164. (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*. (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)

`noncoding_MetaSTAARlite_cond` *Performs conditional meta-analysis of noncoding functional categories using MetaSTAARlite*

Description

This function performs meta-analysis to detect conditional associations between a quantitative/dichotomous phenotype and noncoding functional categories of a gene adjusting for set of known variants by using the MetaSTAARlite pipeline. For each coding functional category, the conditional MetaSTAAR-O p-value is a p-value from an omnibus test that aggregated conditional SKAT-MS(1,25), SKAT-MS(1,1), Burden-MS(1,25), Burden-MS(1,1), ACAT-V-MS(1,25), and ACAT-V-MS(1,1) together with conditional p-values of each test weighted by each annotation using Cauchy method.

Usage

```
noncoding_MetaSTAARlite_cond(
  chr,
  gene_name,
  sample.sizes,
  noncoding_sumstat_gene_list,
  noncoding_cov_gene_list,
  noncoding_cov_cond_gene_list,
  cov_maf_cutoff,
  rare_maf_cutoff = 0.01,
  rv_num_cutoff = 2,
  effect.cond = c("homogeneous", "heterogeneous"),
  check_qc_label = FALSE,
  variant_type = c("SNV", "Indel", "variant"),
  Use_annotation_weights = TRUE,
  Annotation_name = NULL,
  silent = FALSE
)
```

Arguments

chr	an integer which specifies the chromosome number.
gene_name	a character which specifies the name of the gene to be meta-analyzed using MetaSTAARlite.
sample.sizes	a numeric vector with the length of study.names indicating the sample size of each study.
noncoding_sumstat_gene_list	a list containing study-specific summary statistics corresponding to the specified gene.
noncoding_cov_gene_list	a list containing study-specific sparse weighted covariance matrices corresponding to the specified gene.
noncoding_cov_cond_gene_list	a list containing study-specific summary statistics and covariance matrices corresponding to the specified gene for variants to be conditioned on.
cov_maf_cutoff	a numeric vector with the length of study.names indicating the maximum minor allele frequency cutoffs under which the sparse weighted covariance files between variants are stored.
rare_maf_cutoff	a numeric value specifying the cutoff of maximum minor allele frequency in defining rare variants (default = 0.01).
rv_num_cutoff	an integer specifying the cutoff of minimum number of variants of meta-analyzing a given variant-set (default = 2).
effect.cond	a character value indicating the effects of variants to be adjusted for in conditional analysis are "homogeneous" or "heterogeneous" (default = "homogeneous").
check_qc_label	a logical value indicating whether variants need to be dropped according to qc_label specified in noncoding_MetaSTAARlite_worker (default = FALSE).
variant_type	a character value specifying the type of variant included in the analysis. Choices include "SNV", "Indel", or "variant" (default = "SNV").

`Use_annotation_weights`
: a logical value which determines if annotations will be used as weights or not (default = TRUE).

`Annotation_name`
: a character vector of annotation names used in MetaSTAARlite (default = NULL).

`silent`
: a logical value which determines if the report of error messages will be suppressed (default = FALSE).

Value

a list of data frames containing the conditional MetaSTAAR p-values (including MetaSTAAR-O) corresponding to each noncoding functional category of the given gene.

References

Li, X., et al. (2023). Powerful, scalable and resource-efficient meta-analysis of rare variant associations in large whole genome sequencing studies. *Nature Genetics*, *55*(1), 154-164. (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*. (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)

---

noncoding_MetaSTAARlite_worker

Generates summary statistics of noncoding functional categories using MetaSTAARlite

---

Description

This function uses MetaSTAARlite to generate variant-level summary statistics and sparse covariance matrices for noncoding functional categories of a gene.

Usage

```
noncoding_MetaSTAARlite_worker(
  chr,
  gene_name,
  genofile,
  obj_nullmodel,
  known_loci = NULL,
  cov_maf_cutoff = 0.05,
  signif.digits = NULL,
  QC_label = "annotation/filter",
  check_qc_label = FALSE,
  variant_type = c("SNV", "Indel", "variant"),
  Annotation_dir = "annotation/info/FunctionalAnnotation",
  Annotation_name_catalog,
  Use_annotation_weights = TRUE,
  Annotation_name = NULL,
  silent = FALSE
)
```

Arguments

chr	an integer which specifies the chromosome number.
gene_name	a character which specifies the name of the gene to be meta-analyzed using MetaSTAARlite.
genofile	an object of opened annotated GDS (aGDS) file.
obj_nullmodel	an object from fitting the null model, which is 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.
known_loci	the data frame of variants to be adjusted for in conditional analysis. Should contain four columns in the following order: chromosome (CHR), position (POS), reference allele (REF), and alternative allele (ALT) (default = NULL).
cov_maf_cutoff	a numeric value indicating the maximum minor allele frequency cutoff under which the sparse weighted covariance file between variants is stored (default = 0.05).
signif.digits	an integer indicating the number of significant digits to be used for storing the sparse weighted covariance file. If signif.digits is NULL, it is assumed that no rounding will be performed (default = NULL).
QC_label	a character specifying the channel name of the QC label in the GDS/aGDS file (default = "annotation/filter").
check_qc_label	a logical value indicating whether variants need to be dropped according to qc_label. If check_qc_label is FALSE, then the summary statistics will be stored for PASS variants from the study. If check_qc_label is TRUE, then the summary statistics will be stored for all variants from the study, together will an additional column of qc_label (default = FALSE).
variant_type	a character value specifying the type of variant included in the analysis. Choices include "SNV", "Indel", or "variant" (default = "SNV").
Annotation_dir	a character specifying the channel name of the annotations in the aGDS file (default = "annotation/info/FunctionalAnnotation").
Annotation_name_catalog	a data frame containing the annotation name and the corresponding channel name in the aGDS file.
Use_annotation_weights	a logical value which specifies if annotations will be used as weights or not (default = TRUE).
Annotation_name	a character vector of annotation names used in MetaSTAARlite (default = NULL).
silent	a logical value which determines if the report of error messages will be suppressed (default = FALSE).

Value

a list of the following objects corresponding to each noncoding functional category of the given gene: (1) the data frame of all variants in the variant-set (the variant-level summary statistics file), including the following information: chromosome (chr), position (pos), reference allele (ref), alternative allele (alt), quality control status (qc_label, optional), alternative allele count (alt_AC), minor allele count (MAC), minor allele frequency (MAF), study sample size (N), score statistic (U), variance (V), variant annotations specified in Annotation_name, and the low-rank decomposed component of the covariance file; (2) the sparse matrix of all variants in the variant-set whose minor

allele frequency is below `cov_maf_cutoff` (the sparse weighted covariance file); (3) the summary statistics and covariance matrices corresponding to the specified gene for variants to be conditioned on in `known_loci`.

References

Li, X., et al. (2023). Powerful, scalable and resource-efficient meta-analysis of rare variant associations in large whole genome sequencing studies. *Nature Genetics*, *55*(1), 154-164. (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*. (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