Package 'LOCATOR'

August 26, 2024

Type Package

Title LOCATOR: Perform local-ancestry-aware association tests for local-ancestry-sensitive populations

Version 0.1.0

Author Hanxiao Sun [aut, cre]
Han Chen [aut, cre]

Maintainer Han Chen <Han.Chen.2@uth.tmc.edu>

Description An R package that provides local ancestry related data manipulation tools. Yield robust local-ancestry-effect-adjusted GWAS results and relevant plotting with computational efficiency. Also can be adapted to adjust local effects different from the overall global effects for other data needs.

Encoding UTF-8

Roxygen list(markdown=TRUE)

Depends R (>= 4.2.0)

Imports data.table, dplyr (>= 1.1.4), tidyr, tibble, ggplot2, scatterpie, MASS, Matrix, rje, rlang, magrittr

Suggests GMMAT, SeqArray, SeqVarTools

RoxygenNote 7.3.2

License GPL (>= 3)

NeedsCompilation no

Contents

`get.anchor` *Get the population anchor matrix*

Description

This function generates the population anchor matrix that illustrates the ancestry distribution among designated PCs.

Usage

```
get.anchor(input.ga, input.gc)
```

Arguments

`input.ga`	A numeric matrix of global ancestry. Please do not include any sample IDs.
`input.gc`	A numeric matrix of global PCs. Please do not include any sample IDs and ensure it corresponds to input.ga.

Value

a square matrix of population anchor matrix sharing the same number of columns as input.gc.

`get.breakpoints` *Identify local ancestry breakpoints and generate local-ancestry-based global ancestry proportions if applicable.*

Description

This function reads in local ancestry files in chunks, identifies local ancestry tracts, and yields local-ancestry-based global ancestry proportions if applicable.

Usage

```
get.breakpoints(
  input.la,
  input.pos,
  input.id,
  input.anc,
  input.sep,
  output.bp,
  start.pos = 0,
  if.yield.ga = F,
  output.ga = NULL,
  if.haplo.ga = F
)
```

Arguments

input.la	filename for pure local ancestry records. Numeric. Do not include any character description.
input.pos	a numeric vector of SNP positions in base pair. Ascending Order.
input.id	a vector of sample IDs.
input.anc	a character vector of reference ancestry labels.
input.sep	delimiters in the input.la.
output.bp	filename of identified local ancestry breakpoints.
start.pos	An integer indicating the starting position for counting on local ancestry (default = 0).
if.yield.ga	a logical switch indicating whether to generate global ancestry proportions or not (default = TRUE).
output.ga	filename of output global ancestry proportions. Only effective when if.yield.ga = TRUE (default = NULL).
if.haplo.ga	a logical switch indicating whether to generate haplotype-based global ancestry proportions (default = FALSE).

Value

a compressed binary file that documents sample information, genetic positions at breakpoints and the corresponding local ancestry references, with optional global ancestry proportions.

---

get.la.aware.score *Generate local-ancestry-aware GWAS results*

Description

This function uses a glmmkin class object from the null GLMM to perform score tests for local-ancestry-aware association with genotypes in a GDS file .gds file.

Usage

```
get.la.aware.score(
  null.obj,
  geno.file,
  outfile,
  LAC.file,
  n_PC_used = NULL,
  id.type = c("character", "integer", "numeric"),
  MAF.range = c(1e-07, 0.5),
  miss.cutoff = 1,
  missing.method = "impute2mean",
  is.dosage = F,
  n.batch = 100
)
```

Arguments

null.obj	a class glmmkin or class glmmkin.multi object, returned by fitting the null GLMMusing glmmkin.
geno.file	the full name of a GDS file (including the suffix .gds), or an object of class SeqVarGDSClass.
outfile	the output file name.
LAC.file	the full name of a binary LAC file.
n_PC_used	the dimension of PCs that will be included in analysis. Default is set to the number of dimension of PCs in the LAC.file.
id.type	the class of sample IDs. Possible values are "character", "integer", and "numeric".
MAF.range	a numeric vector of length 2 defining the minimum and maximum minor allele frequencies of variants that should be included in the analysis (default = c(1e-7, 0.5)).
miss.cutoff	the maximum missing rate allowed for a variant to be included (default = 1, including all variants).
missing.method	method of handling missing genotypes. Either "impute2mean" or "omit" (default = "impute2mean").
is.dosage	a logical switch for whether imputed dosage should be used from a GDS infile (default = FALSE).
n.batch	an integer for how many SNPs should be tested in a batch (default = 100). The computational time can increase dramatically if this value is either small or large. The optimal value for best performance depends on the user's system.

Value

@format a dataframe with following components:

SNP SNP name, as supplied in snps.

CHR Chromosome, copied from .gds file.

POS physical position in base pairs, copied from .gds file.

REF reference allele, copied from .gds file.

ALT alternate allele, copied from .gds file.

MISSRATE number of individuals with non-missing genotypes for each SNP.

AF ALT allele frequency for each SNP.

N total sample size.

SCORE the summary score of the effect allele.

VAR the variance of the summary score.

PVAL local-ancestry-aware GWAS p-values.

get.LACs *Generate Local Ancestry Coordinates (LACs)*

Description

This function uses the extracted local ancestry breakpoints and corresponding local ancestry inferences and converts them into binary LACs files.

Usage

```
get.LACs(
  input.bp,
  psi,
  output.la,
  id.type = c("character", "integer", "numeric"),
  select.PC = NULL,
  select.pos = NULL,
  select.id = NULL,
  if.haplo = F
)
```

Arguments

input.bp	filename of extracted local ancestry inference at breakpoints.
psi	the designated population anchor matrix.
output.la	filename of the output LACs files.
id.type	the class of sample IDs. Possible values are "character", "integer", and "numeric".
select.PC	a vector indicating PCs included in further analysis, e.g. c(1,2,3). Default is equivalent to the number of columns in the psi matrix.
select.pos	a vector indicating genetic positions included in further analysis. Ascending Order. Default is set to include all genetic positions in the input.bp.
select.id	a vector indicating sample IDs included in further analysis. Default is set to include all samples in the input.bp.
if.haplo	a logical switch indicating whether to generate haplotype-based LACs (default = FALSE).

Value

a compressed binary file that documents sample information, genetic positions at breakpoints and refined LACs given corresponding local ancestry inferences.

`plot.features` *generate local-ancestry-aware phenotype distribution plots*

Description

This function generates utilizes ggplot2 to generate phenotype or covariate distribution in the context of local ancestry at a specfic genetic position and global PCs as reference. The proportion of phenotype subtypes will be distributed in pie chart ratios, and the size of pie chart indicates the number of samples sharing that specific local ancestry combination. Only works for categorical variables.

Usage

```
## S3 method for class 'features'
plot(
  input.file,
  PC1,
  PC2,
  anc,
  PC1_la,
  PC2_la,
  pheno,
  pheno.name,
  phenotype.name,
  y.scale = 1e+06,
  if.jitter = T,
  y.jitter = 0.01,
  if.round = T,
  round.digit = 2,
  if.filter = F,
  n.filter = 5,
  r.x.axis = -0.025,
  r.y.axis = 0.025,
  r.n.level = 4
)
```

Arguments

`input.file`	a dataframe that includes information of global PCs, LACs at genetic positions and phenotype or covariate of interest.
`PC1, PC2`	a pair of strings indicating the columns contributed to the x-, and y-coordinate in the background, which are usually supposed to be from global PCs, e.g. PC1 and PC2. Please do not include any single quote nor double quote surrounding the string.
`anc`	a vector indicating reference ancestry populations.
`PC1_la, PC2_la`	a pair of string indicating the columns corresponding to x-, and y-coordinate in LACs, e.g. PC1_la and PC2_la, on which the combination of local ancestry and phenotype subtypes will be counted. Please do not include any single quote nor double quote surrounding the string.
`pheno`	a string indicating of the column of phenotype or covariate of interest.

pheno.name	a string as the label of the phenotype or covariate.
phenotype.name	a string vector as labels of phenotype subtypes.
y.scale	an integer or a positive value, usually extreme large, adjusting the radius of pie charts to match the coordinates of PCs (default = 1e6).
if.jitter	a logical switch indicating whether to slightly jitter the y-coordinates of pie charts to reduce overlapping between local ancestry patterns and pie charts (default = TRUE).
y.jitter	a positive number to adjust the y-coordinates of pie charts (default = 0.01). Only effective when if.jitter = TRUE.
if.round	a logical switch indicating whether to round LACs to reduce the impact of deviation in the LACs (default = TRUE).
round.digit	an integer indicating the number decimal places to be used (default = 2). Only effective when if.round = TRUE.
if.filter	a logical switch indicating whether to filter the number of local ancestry patterns to reduce the impact of deviation in the LACs (default = TRUE).
n.filter	an integer indicating the number of threshold for local ancestry patterns preserving for further analysis (default = 5). Only effective when if.filter = TRUE.
r.x.axis, r.y.axis	a pair of numerics indicating the coordinates of legend of pie chart sizes (default = c(-0.025,0.025)).
r.n.level	an integer indicating the layers of circles in the legend of pie chart sizes (default = 4).

Value

a ggplot2-based plots that illustrates the relationship between global PCs, LACs and phenotype distribution per local ancestry pattern accordingly.

---

read.binary *Retrieve relevant local ancestry information at breakpoints*

Description

This function reads and extracts compressed local ancestry information, either in local ancestry inferences or in LACs, given designated genetic positions.

Usage

```
read.binary(
  input.file,
  type = c("breakpoints", "LACs"),
  id.type = c("character", "integer", "numeric"),
  select.pos,
  select.id = NULL,
  if.haplo = F,
  col.names = NULL
)
```

Arguments

input.file	filename of relevant compressed binary local ancestry files.
type	file type of input.file. Either be "breakpoints" or "LACs".
id.type	the class of sample IDs. Possible values are "character", "integer", and "numeric".
select.pos	a vector indicating genetic positions included in further analysis. Ascending Order. Default is set to include all genetic positions in the input.file.
select.id	a vector indicating sample IDs included in further analysis. Default is set to include all samples in the input.file.
if.haplo	a logical switch indicating whether it is haplotype-based local ancestry information (default = FALSE).
col.names	a vector indicating column names for local ancestry in local ancestry inference and PCs in LACs. For LACs files, default is set from PC1 to PCn for all columns in the input.file.

Value

a list of dataframes including genetic-position-specfic local ancestry information in either local ancestry references or in LACs.

Index