Package 'simplePHENOTYPES'

October 14, 2022

Date 2021-01-19

Type Package

Version 1.3.0

Title Simulation of Pleiotropic, Linked and Epistatic Phenotypes

Description The number of studies involving correlated traits and the availability of tools to handle this type of data has increased considerably in the last decade. With such a demand, we need tools for testing hypotheses related to single and multi-trait (correlated) phenotypes based on many genetic settings. Thus, we implemented various options for simulation of pleiotropy and Linkage Disequilibrium under additive, dominance and epistatic models. The simulation currently takes a marker data set as an input and then uses it for simulating multiple traits as described in Fernandes and Lipka (2020) <doi:10.1186/s12859-020-03804-y>.

License MIT + file LICENSE

Encoding UTF-8

LazyData true

biocViews

Depends R (>= 3.5.0)

Imports data.table, mvtnorm, stats, utils, SNPRelate, gdsfmt

Suggests knitr, rmarkdown

RoxygenNote 7.1.1

VignetteBuilder knitr

URL https://github.com/samuelbfernandes/simplePHENOTYPES

BugReports https://github.com/samuelbfernandes/simplePHENOTYPES/issues

NeedsCompilation no

Author Samuel Fernandes [aut, cre] (<https://orcid.org/0000-0001-8269-535X>), Alexander Lipka [aut] (<https://orcid.org/0000-0003-1571-8528>)

Maintainer Samuel Fernandes <samuelf@illinois.edu>

Repository CRAN

Date/Publication 2021-01-20 16:30:02 UTC

R topics documented:

`create_phenotypes` *Simulation of single/multiple traits under different models and genetic architectures.*

Description

Simulation of single/multiple traits under different models and genetic architectures.

Usage

```
create_phenotypes(
  geno_obj = NULL,
  geno_file = NULL,
  geno_path = NULL,
  QTN_list = list(add = list(NULL), dom = list(NULL), epi = list(NULL)),
  prefix = NULL,
  rep = NULL,
  ntraits = 1,
  h2 = NULL,
  mean = NULL,
  model = NULL,
  architecture = "pleiotropic",
  add_QTN_num = NULL,
  dom_QTN_num = NULL,
  epi_QTN_num = NULL,
  epi_type = NULL,
  epi_interaction = 2,
  pleio_a = NULL,
  pleio_d = NULL,
  pleio_e = NULL,
  trait_spec_a_QTN_num = NULL,
  trait_spec_d_QTN_num = NULL,
  trait_spec_e_QTN_num = NULL,
  add_effect = NULL,
  big_add_QTN_effect = NULL,
  same_add_dom_QTN = FALSE,
  dom_effect = NULL,
  degree_of_dom = 1,
  epi_effect = NULL,
  type_of_ld = "indirect",
  ld_min = 0.2,
  ld_max = 0.8,
```

```
  ld_method = "composite",
  sim_method = "geometric",
  vary_QTN = FALSE,
  cor = NULL,
  cor_res = NULL,
  QTN_variance = FALSE,
  seed = NULL,
  home_dir = NULL,
  output_dir = NULL,
  export_gt = FALSE,
  output_format = "long",
  to_r = FALSE,
  out_geno = NULL,
  chr_prefix = "chr",
  remove_QTN = FALSE,
  warning_file_saver = TRUE,
  constraints = list(maf_above = NULL, maf_below = NULL, hets = NULL),
  maf_cutoff = NULL,
  nrows = Inf,
  na_string = "NA",
  SNP_effect = "Add",
  SNP_impute = "Middle",
  quiet = FALSE,
  verbose = TRUE,
  RNGversion = "3.5.1"
)
```

Arguments

geno_obj	Marker data set loaded as an R object. Currently either HapMap or numericalized files (code as aa = -1, Aa = 0 and AA = 1, e.g. 'data("SNP55K_maize282_maf04")') are accepted. These and other file formats (VCF, GDS, and Plink Bed/Ped files) may be read from file with 'geno_file' or 'geno_path'. Only one of 'geno_obj', 'geno_file' or 'geno_path' should be provided.
geno_file	Name of a marker data set to be read from a file. If in a different folder, the whole path should be provided. Formats accepted are Numeric, HapMap, VCF, GDS, and Plink Bed/Ped files. Notice that the major allele will always be 1 and the minor allele -1. Thus, when using Plink Bed files, the dosage information will be converted to the opposite value.
geno_path	Path to a folder containing the marker data set file/files (e.g., separated by chromosome). Formats accepted are: Numeric, HapMap, VCF, GDS, and Plink Bed/Ped files
QTN_list	A list of specific markers to be used as QTNs. If one wants to specify the QTNs instead of selecting them randomly, at least one of the following elements should be provided: 'QTN_list$add', 'QTN_list$dom', and/or 'QTN_list$epi'. The element '$add', '$dom', and '$epi' are lists containing a vector of markers for each of the traits to be simulated. For example, to simulate 2 traits controlled by 1 pleiotropic and 2 trait-specific additive QTNs, the user would create a list

of marker names ‘marker_list <- list(add = list(trait1 = c("marker1", "marker2", "marker3"), trait2 = c("marker1", "marker4", "marker5")))‘ and set ‘QTN_list = marker_list‘. On the other hand, to simulate a single trait controlled by 1 additive and 2 dominance QTNs, the marker list would be ‘marker_list <- list(add = list("marker1"), dom = list(c("marker2", "marker3")))‘. Notice that these vectors with maker names is used in the order they appear. For instance, in the list ‘marker_list <- list(add = list(trait9 = c("marker1"), trait4 = c("marker5")))‘, the vector names itself ("trait9" and "trait4") are ignored and "trait9" will be the vector of markers used to simulate the first trait and "trait4" will be the vector of markers used to simulate the second trait. Also, when using ‘QTN_list‘, many parameters used for selecting QTNs will be ignored (e.g., ‘constraints‘).

`prefix` If ‘geno_path‘ points to a folder with files other than the marker data set, a part of the data set name may be used to select the desired files (e.g., prefix = "Chr" would read files Chr1.hmp.txt, ..., Chr10.hmp.txt but not HapMap.hmp.txt).

`rep` The number of experiments (replicates of a trait with the same genetic architecture) to be simulated.

`ntraits` The number of multi-trait phenotypes to simulate under pleiotropic, partially [pleiotropic], and LD (spurious pleiotropy) architectures (see ‘architecture‘). If not assigned, a single trait will be simulated. Currently, the only option for the LD architecture is ‘ntraits = 2‘.

`h2` The heritability for each traits being simulated. It could be either a vector with length equals to ‘ntraits‘, or a matrix with ncol equals to ‘ntraits‘. If the later is used, the simulation will loop over the number of rows and will generate a result for each row. If a single trait is being simulated and h2 is a vector, one simulation of each heritability value will be conducted. Either none or all traits are expected to have ‘h2 = 0‘.

`mean` A vector with the mean (intercept) value for each of the simulated traits. If omitted, the simulated traits will be centered to zero.

`model` The genetic model to be assumed. The options are "A" (additive), "D" (dominance), "E" (epistatic) as well as any combination of those models such as "AE", "DE" or "ADE".

`architecture` The genetic architecture to be simulated. Should be provided if ‘ntraits‘ > 1. Possible options are: ’pleiotropic’ (default), for traits being controlled by the same QTNs; ’partially’, for traits being controlled by pleiotropic and trait-specific QTNs; ’LD’, for traits being exclusively controlled by different QTNs in "direct" or "indirect" (See ‘type_of_ld‘, ‘ld_min‘, and ‘ld_max‘ below) linkage disequilibrium. Currently the only option for ‘architecture = "LD"‘ is ‘ntraits = 2‘.

`add_QTN_num` The number of additive quantitative trait nucleotides (QTNs) to be simulated.

`dom_QTN_num` The number of dominance QTNs to be simulated.

`epi_QTN_num` The number of epistatic (Currently, only additive x additive epistasis are simulated) QTNs to be simulated.

`epi_type` to be implemented...

`epi_interaction`

Number of markers that compose an epistatic QTN. If ‘epi_interaction = 2‘ (default), a 2-way interaction (marker1 x marker2) will be used to simulate

epistatic QTNs. If 'epi_interaction = 3' a 3-way interaction (marker1 x marker2 x marker3) will be used instead.

`pleio_a` The number of pleiotropic additive QTNs to be used if 'architecture = "partially"'. When 'sim_method = custom' (see below), the first effects will be assigned to the pleiotropic QTNs and the last to the trait-specific ones. For instance, in a scenario where ntraits = 2, pleio_a = 2, trait_spec_a_QTN_num = 1, and add_effect = list(trait1 = c(0.1, 0.2, 0.3), trait2 = c(0.4, 0.5, 0.6)), the trait-specific QTNs for trait 1 and trait 2 will be 0.3 and 0.6, respectively. The first two allelic effects will be assigned to the pleiotropic QTNs.

`pleio_d` The number of pleiotropic dominance QTNs to be used if 'architecture = "partially"' (See pleio_a for details).

`pleio_e` The number of pleiotropic epistatic QTNs to be used if 'architecture = "partially"' (See pleio_a for details).

`trait_spec_a_QTN_num`
The number of trait-specific additive QTNs if 'architecture = "partially"'. It should be a vector of length equals to 'ntraits'.

`trait_spec_d_QTN_num`
The number of trait-specific dominance QTNs if 'architecture = "partially"'. It should be a vector of length equals to 'ntraits'.

`trait_spec_e_QTN_num`
The number of trait-specific epistatic QTNs if 'architecture = "partially"'. It should be a vector of length equals to 'ntraits'.

`add_effect` Additive effect size to be simulated. It may be either a vector (assuming 'ntraits' = 1 or one allelic effect per trait to create a geometric series ['sim_method = "geometric"']) or a list of length = 'ntraits', i.e., if 'ntraits' > 1, a list with one vector of additive effects should be provided for each trait. Unless 'big_add_QTN_effect' is provided, the length of each vector should be equal to the number of additive QTNs being simulated.

`big_add_QTN_effect`
Additive effect size for one possible major effect quantitative trait nucleotide. If 'ntraits' > 1, big_add_QTN_effect should have length equals 'ntraits'. If 'add_QTN_num' > 1, this large effect will be assigned to the fist QTN.

`same_add_dom_QTN`
A boolean for selecting markers to be both additive and dominance QTNs.

`dom_effect` Similar to the 'add_effect', it could be either a vector or a list. Optional if 'same_add_dom_QTN = TRUE'.

`degree_of_dom` If the same set of QTNs are being used for simulating additive and dominance effects, the dominance allelic effect could be a proportion of the additive allelic effect. In other words, 'degree_of_dom' equals to 0.5, 1, 1.5 will simulate, partial dominance, complete dominance and overdominance, respectively.

`epi_effect` Epistatic (additive x additive) effect size to be simulated. Similar to the 'add_effect', it could be either a vector or a list.

`type_of_ld` Type of LD used to simulate spurious pleiotropy. If "indirect" (default), an intermediate marker is selected from which two adjacent markers (one upstream and another downstream) will be chosen based on its LD with the intermediate

marker to be the QTNs. Optionally, in the "direct" method, one marker is selected to be a QTN for trait 1, and a second marker is selected based on its LD with the first selected marker to be the QTN for trait 2.

`ld_min` Minimum Linkage disequilibrium for selecting QTNs when 'architecture = LD'. The default is 'ld_min = 0.2' (markers should have a minimum LD of 0.2 to be used as QTNs).

`ld_max` Maximum Linkage disequilibrium for selecting QTNs when 'architecture = LD'. The default is 'ld_max = 0.8' (markers should have an LD of at maximum 0.8 to be used as QTNs).

`ld_method` Four methods can be used to calculate linkage disequilibrium values: "composite" for LD composite measure (Default), "r" for R coefficient (by EM algorithm assuming HWE, it could be negative), "dprime" for D', and "corr" for correlation coefficient (see snpgdsLDpair from package SNPRelate).

`sim_method` Provide the method of simulating allelic effects. The options available are "geometric" and "custom". For multiple QTNs, a geometric series may be simulated, i.e., if add_effect = 0.5, the effect size of the first QTNs will be 0.5, the effect size of the second QTN will be 0.5^2, and the effect of the n^th QTN will be 0.5^n.

`vary_QTN` A boolean that determines if the same set of quantitative trait nucleotide (QTN) should be used to generate genetic effects for each experiment ('vary_QTN = FALSE') or if a different set of QTNs should be used for each replication ('vary_QTN = TRUE').

`cor` Option to simulate traits with a predefined genetic correlation. It should be a correlation matrix with a number of rows = 'ntraits'. Default = NULL. Notice that when opting for controlling the correlation, the genetic effects are transformed using Cholesky decomposition. In this case, the correlation of genetic effects for different traits will be as provided, but due to the transformation, the actual allelic effects of correlated traits may be different than the input allelic effect.

`cor_res` Option to simulate traits with a predefined residual correlation. It should be a correlation matrix with number of rows = 'ntraits'. If NULL, an identity matrix (independent residuals) will be used.

`QTN_variance` Whether or not the percentage of the phenotypic variance explained by each QTN (QTN variance / phenotypic variance) should be exported. The default is FALSE. Notice that this is calculated prior to any transformation, such as the whitening/coloring transformation used to assign user-specified correlation to the genetic effect. In may not reflect the actual variance explained when the data is transformed.

`seed` Value to be used by set.seed. If NULL (default), runif(1, 0, 1000000) will be used. Notice that at each sampling step, a different seed generated based on the 'seed' parameter used. For example, if one uses 'seed = 123', when simulating the 10th replication of trait 1, the seed to be used is 'round((123 * 10 * 10) * 1)'. On the other hand, for simulating the 21st replication of trait 2, the seed to be used will be 'round((123 * 21 * 21) * 2)'. The master seed (unique value required to reproduce results) is saved at the top of the log file. Unless verbose = FALSE the actual seed used in every simulation is exported along with simulated phenotypes.

home_dir — Directory where files should be saved. It may be home_dir = getwd().

output_dir — Name to be used to create a folder inside ‘home_dir‘ and save output files.

export_gt — If TRUE genotypes of selected QTNs will be saved at file. If FALSE (default), only the QTN information will be saved.

output_format — Four options are available for saving simulated phenotypes: ’multi-file’, saves each simulation in a separate file; ’long’ (default for multiple traits), appends each experiment (rep) to the last one (by row); ’wide’, saves experiments by column (default for single trait) and ’gemma’, saves .fam files to be used by gemma or other software that uses plink bed files. (renaming .fam file with the same name of the bim and bed files is necessary).

to_r — Option for outputting the simulated results as an R data.frame in addition to saving it to file. If TRUE, results need to be assigned to an R object (see vignette).

out_geno — Optionally saves the numericalized genotype either as "numeric" (see vignettes for an example data), "BED" or "gds". The default is NULL.

chr_prefix — If input file format is VCF and out_geno = "BED", and a prefix is used in the chromosomes names, chr_prefix may be used to avoid issues in converting to bed files (e.g., chr_prefix = "chr" in "chr01").

remove_QTN — Whether or not a copy of the genotypic file should be saved without the simulated QTNs. The default is FALSE. If ‘vary_QTN = TRUE‘, the question "Are you sure that you want to save one genotypic file/rep (remove_QTN = TRUE and vary_QTN = TRUE) [type yes or no] ?" will pop up to avoid saving multiple large files unintentionally

warning_file_saver — Skips the interactive question and saves all files when ‘remove_QTN = TRUE‘ and ‘vary_QTN = TRUE‘.

constraints — Set constraints for QTN selection. Currently, the options are maf_above (the minimum value of minor allele frequency, a double between 0 - 0.5), maf_below (the maximum value of minor allele frequency, a double between 0 - 0.5),and hets (’include’ and ’remove’). All of these options are NULL by default (’list(maf_above = NULL, maf_below = NULL, hets = NULL)’). For instance, if the parameters used are ‘constraints = list(maf_above = 0.3, maf_below = 0.44, hets = "include")‘, only heterozygote markers with minor allele frequency between 0.3 and 0.44 will be selected to be QTNs. The option "remove" would only select homozygote markers to be QTNs.

maf_cutoff — Option for filtering the data set based on minor allele frequency (Not to be confounded with the constraints option which will only filter possible QTNs). It may be useful when outputting the genotypic data set.

nrows — Option for loading only part of a data set. Used when marker data is in numeric or HapMap format. Please see data.table::fread for details.

na_string — Tell create_phenotypes what character represents missing data (default is "NA"). Used when the input marker data is numeric or HapMap.

SNP_effect — Parameter used for numericalization. The options are: Add (AA = 1, Aa = 0, aa = -1), Dom (AA = -1, Aa = 0, aa = -1), Left (AA = 1, Aa = -1, aa = -1), Right (AA = 1, Aa = 1, aa = -1). The default option is Add.

SNP_impute	Naive imputation for HapMap numericalization. The options are: Major (NA <- 1), Middle (NA <- 0), and Minor (NA <- -1).
quiet	Whether or not the log file should pop up into R once the simulation is done.
verbose	If FALSE, suppress all prints and suppress individual seed numbers from being saved to file. The master seed (unique value required to reproduce results) is saved at the top of the log file.
RNGversion	Parameter to set the random number generator. Different R versions may be selected, the default value is '3.5.1'.

Value

Single or multi-trait phenotypes in one of many formats. Numericalized marker data set with or without the selected QTNs. Diagnostic files (log, QTN information, summary of LD between QTNs, proportion of phenotypic variance explained by each QTN).

Author(s)

Samuel B Fernandes and Alexander E Lipka Last update: Jan 19, 2021

References

Fernandes, S.B., and Lipka, A.E., 2020 simplePHENOTYPES: SIMulation of pleiotropic, linked and epistatic SIMulation of Pleiotropic, Linked and Epistatic PHENOTYPES. BMC Bioinformatics 21(1):491, doi: 10.1186/s1285902003804y

Examples

```
# Simulate 50 replications of a single phenotype.
data("SNP55K_maize282_maf04")
pheno <-
  create_phenotypes(
    geno_obj = SNP55K_maize282_maf04,
    add_QTN_num = 3,
    add_effect = 0.2,
    big_add_QTN_effect = 0.9,
    rep = 10,
    h2 = 0.7,
    model = "A",
    to_r = TRUE,
    home_dir = tempdir(),
    quiet = T
    )
# For more examples, please run the following:
# vignette("simplePHENOTYPES")
```

Index