Package 'simplePHENOTYPES'

December 4, 2019

Date 2019-11-25

Type Package

Version 1.0.5

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 HapMap file as an input and numericalize it with GAPIT.numericalization Lipka et al. (2012) <doi:10.1093/bioinformatics/bts444>. The numericalized dataset is then used by a framework adapted from Rice and Lipka (2019) <doi:10.3835/plantgenome2018.07.0052> for simulating multiple traits.

License MIT + file LICENSE

Encoding UTF-8

LazyData true

biocViews

Depends R (>= 3.5.0)

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

Suggests knitr, rmarkdown

RoxygenNote 6.1.1

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 2019-12-04 10:40:02 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, rep = NULL, ntraits = 1, h2 = NULL,
  model = NULL, add_QTN_num = NULL, dom_QTN_num = NULL,
  epi_QTN_num = NULL, add_effect = NULL, same_add_dom_QTN = FALSE,
  dom_effect = NULL, degree_of_dom = 1, epi_effect = NULL,
  architecture = "pleiotropic", 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, ld = 0.5,
  sim_method = "geometric", vary_QTN = FALSE,
  big_add_QTN_effect = NULL, cor = NULL, seed = NULL,
  export_gt = FALSE, home_dir = NULL, output_dir = NULL,
  to_r = FALSE, output_format = "long", out_geno = NULL,
  gdsfile = NULL, constrains = list(maf_above = NULL, maf_below =
  NULL), prefix = NULL, maf_cutoff = NULL, nrows = Inf,
  na_string = "NA", SNP_effect = "Add", SNP_impute = "Middle",
  major_allele_zero = FALSE, quiet = FALSE, verbose = TRUE)
```

Arguments

`geno_obj`	Marker dataset loaded as an R object. Currently either HapMap or numericalized files (code as aa = 0, Aa = 1 and AA = 2, e.g. 'data("SNP55K_maize282_maf04")') are accepted. 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 file.
`geno_path`	Path to a folder containing the marker dataset file/files (e.g. separated by chromosome).
`rep`	Number of experiments to be simulated.
`ntraits`	Number of traits to be simulated under pleiotropic, partially and LD architectures (see below). If not assigned, a single trait will be simulated. Currently, the only option for the LD architecture is 'ntraits = 2'.

h2 Heritability of all 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 a 'h2 = 0'.

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".

add_QTN_num Number of additive quantitative trait nucleotide to be simulated.

dom_QTN_num Number of dominance quantitative trait nucleotide to be simulated.

epi_QTN_num Number of epistatic (Currently, only additive x additive epistasis are simulated) quantitative trait nucleotide to be simulated.

add_effect Additive effect size to be simulated. It may be either a vector (assuming 'ntraits' = 1 or one allelic effect per trait) or a list of length = 'ntraits', i.e., if 'ntraits' > 1, one set of additive effects should be provided for each trait. In that case, each component should be a vector of either length one, if 'sim_method = "geometric"' (see below), or length equal to the number of additive QTNs being simulated.

same_add_dom_QTN A boolean for having the same quantitative trait nucleotide having additive and dominance effects.

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 quantitative trait nucleotide 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.

architecture Genetic architecture to be simulated. Should be provided if 'ntraits' > 1. Possible options are: 'pleiotropic', 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 QTNs in linkage disequilibrium (controlled by parameter 'ld'). Currently the only option for 'architecture = "LD"' is 'ntraits = 2'.

pleio_a Number of pleiotropic additive QTNs to be used if 'architecture = "partially"'.

pleio_d Number of pleiotropic dominance QTNs to be used if 'architecture = "partially"'.

pleio_e Number of pleiotropic epistatic QTNs to be used if 'architecture = "partially"'.

trait_spec_a_QTN_num Number of trait specific additive QTNs if 'architecture = "partially"'. It should have length equals to 'ntraits'.

`trait_spec_d_QTN_num`
: Number of trait specific dominance QTNs if ‘architecture = "partially"‘. It should have length equals to ‘ntraits‘.

`trait_spec_e_QTN_num`
: Number of trait specific epistatic QTNs if ‘architecture = "partially"‘. It should have length equals to ‘ntraits‘.

`ld`
: Linkage disequilibrium between selected marker two adjacent markers to be used as QTN. Default is ‘ld = 0.5‘.

`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 the add_effect = 0.5, the effect size of the first QTNs is 0.2, and the effect size of the second is 0.5^2 and the effect of the n^th QTN is 0.5^n.

`vary_QTN`
: A boolean to determine 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 experiment (‘vary_QTN = TRUE‘).

`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, the fist QTN will have the large effect.

`cor`
: Option to simulate traits with a pre-defined cor. It should be a square matrix with number of rows = ‘ntraits‘.

`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 is 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 actual seed used in every simulation is exported along with simulated phenotypes.

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

`home_dir`
: Directory where files will be saved. It might be home_dir = getwd().

`output_dir`
: Name to be used to create a folder and save output files.

`to_r`
: Option for saving simulated results into R in addition to saving it to file. If TRUE, results need to be assigned to an R object (see vignette).

`output_format`
: Four options for saving outputs: ’multi-file’, saves one simulation setting 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 with plink bed files (renaming .fam file might be necessary).

`out_geno`
: Saves numericalized genotype either as "numeric", "plink" or "gds". Default is NULL.

`gdsfile`
: Points to a gds file (in case there is one already created) to be used with option architecture = "LD". Default is NULL.

constrains	Set constrains for QTN selection. Currently, only minor allelic frequency is implemented. Either one or both of the following options may be non-null: 'list(maf_above = NULL, maf_below = NULL)'.
prefix	If 'geno_path' points to a folder with files other than the marker dataset, a part of the dataset 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).
maf_cutoff	Optional filter for minor allele frequency (The dataset will be filtered. Not to be confounded with the constrain option which will only filter possible QTNs).
nrows	Option for loading only part of a dataset. Please see data.table::fread for details.
na_string	Sets missing data as "NA".
SNP_effect	Parameter used for numericalization. Following GAPIT implementation, the default is 'Add'.
SNP_impute	Parameter used for numericalization. Following GAPIT implementation, the default is 'Middle'.
major_allele_zero	Parameter used for numericalization. Following GAPIT implementation, the default is FALSE.
quiet	Whether or not the log file should be opened once the simulation is done.
verbose	if FALSE, suppress prints.

Value

Numericalized marker dataset, selected QTNs, phenotypes for 'ntraits' traits, log file.

Author(s)

Samuel B Fernandes and Alexander E Lipka Last update: Nov 14, 2019

References

Rice, B., Lipka, A. E. (2019). Evaluation of RR-BLUP genomic selection models that incorporate peak genome-wide association study signals in maize and sorghum. Plant Genome 12, 1–14.doi: 10.3835/plantgenome2018.07.0052

Alexander E. Lipka, Feng Tian, Qishan Wang, Jason Peiffer, Meng Li, Peter J. Bradbury, Michael A. Gore, Edward S. Buckler, Zhiwu Zhang, GAPIT: genome association and prediction integrated tool, Bioinformatics, Volume 28, Issue 18, 15 September 2012, Pages 2397–2399, doi: 10.1093/bioinformatics/bts444

Examples

```
# Simulate 50 replications of a single phenotype.
## Not run:
pheno <- create_phenotypes(
  geno_obj = SNP55K_maize282_maf04,
  add_QTN_num = 3,
  add_effect = 0.1,
```

```
  big_add_QTN_effect = 0.9,
  rep = 10,
  h2 = 0.7,
  to_r = TRUE,
  model = "A",
  home_dir = tempdir()
)

## End(Not run)
# For more examples, please run the following:
# vignette("simplePHENOTYPES")
```

Index