Package Ravages (RAre Variant Analysis and GEnetic Simulation), Simulations

Herve Perdry and Ozvan Bocher

2021-10-18

Introduction

Ravages was developed to simulate genetic data and to perform rare variant association tests (burden tests and the variance-component test SKAT) on different types of phenotypes (Bocher et al., 2019, doi: 10.1002/gepi.22210, Bocher et al., 2020, doi:10.1038/s41431-020-00792-8) at a genome-wide scale. Ravages relies on the package Gaston. Most functions are written in C++ thanks to the packages Rcpp, RcppParallel and RcppEigen. Functions of Ravages use bed.matrix to manipulate genetic data as in the package Gaston (see documentation of this package for more details).

In this vignette, we illustrate how to perform genetic simulations and how to compute power of rare variant association tests. See the main vignette for more details about rare variant association tests. To learn more about all options of the functions, the reader is advised to look at the manual pages.

We developed two main simulations procedures, one based on allelic frequencies and genetic relative risks (GRR), and the other one based on haplotypes. All the functions can be used to simulate more than two groups of individuals.

Global parameters of Ravages

Functions in Ravages are parallelised either using RcppParallel when functions are implemented in C++ or using parallel otherwise. When the argument *cores* is present in the function, it can be directly changed to fix the number of cores to use. Otherwise, the number of threads used by multithreaded functions can be modified through RcppParallel function **setThreadOptions()**. It is advised to try several values for the number of threads, as using too many threads might be counterproductive due to an important overhead. The default value set by RcppParallel is generally too high.

Information on progression is provided if $verbose = T$ for multiple functions in Ravages.

Simulations based on allelic frequencies and GRR

The main function to use to simulate data based on this method is **rbm.GRR()** which is detailed below.

Calculation of frequencies in each group of individuals

The first step to simulate genetic data is to compute genotypic frequencies in each group of individuals based on frequencies in the general population and on genetic relative risk (GRR) values. GRR correspond to the increased risk of a disease for a given genotype compared to a reference genotype, here the homozygous

genotype for the reference allele. More precisely, the GRR associated to the heterozygous genotype in the group c can be calculated as follow:

$$GRR_{Aa} = \frac{P(Y = c|Aa)}{P(Y = c|AA)}$$

With Y the phenotype (1 for the controls, and c going from 2 to C for the subgroups of cases), A the reference allele, and a the alternate allele. The frequency of each genotype in each group of cases c can be calculated using Bayes theorem:

$$P(Aa|Y = c) = \frac{P(Y = c|Aa)P(Aa)}{\sum_{Geno=AA,Aa,aa} P(Y = c|Geno)P(Geno)} = \frac{GRR_{Aa} \times P(Aa)}{P(AA) + GRR_{Aa} \times P(Aa) + GRR_{aa} \times P(aa)}$$

P(AA), P(Aa) and P(aa) corresponding to the genotypic probabilities in the general population. The three genotypic frequencies can then be calculated in the controls group using the rule of total probability:

$$P(Geno|Y = 1) = P(Geno) - \sum_{c=2}^{c=C} P(Geno|Y = c)P(Y = c)$$

The function **genotypic.freq()** performs these calculations to obtain the three genotypic frequencies in the different groups of individuals. To do so, the user needs to give *P(Y=c)*, the prevalence of each group of cases (argument *prev*), and the GRR values. GRR values need to be in a matrix form with one row per cases group and one column per variant. If there is no supposed link between the GRR associated to the heterozygous genotype and the GRR associated to the homozygous alternative genotype (general model of the disease, *genetic.model* = "*general*"), the user needs to specify two GRR matrices: one for GRR_{Aa} (argument *GRR.het*) and one for GRR_{aa} (argument *GRR.homo.alt*). If *genetic.model=“recessive”*, *“multiplicative”* or *“dominant”*, only one GRR matrix is needed (*GRR.het*). The function **genotypic.freq()** will return a list with three matrices, one for each genotype containing the genotypic frequencies, with one row per group of individuals and one column per variant.

To help the user with the construction of the GRR matrix for **genotypic.freq()**, we implemented the function **GRR.matrix()**.To use this function, the user needs to specify how the GRR should be calculated (argument *GRR*):

- the same GRR is given to all the variants (*GRR=“constant”*), its value being specified to *GRR.value*;
- the GRR is computed using the same formula as in SKAT: $w = -0.4 * |log_{10}(MAF)|$ (*GRR=“SKAT”*);
- the GRR is computed using another function depending on MAF in the general population (*GRR=“variable”*), this function being specified to *GRR.function*.

In the last two situations, a file containing the MAF in the general population with at least a column “maf” and a column “gene” should be given to the argument *genes.maf*. Two such files are available in Ravages: the file *Kryukov* containing MAF simulated under the demographic model of Kryukov and the file *GnomADgenes* containing MAF from the NFE population in GnomAD. As these files contain MAF for multiple genes, the user needs to specify which gene to choose to simulate the data with the argument *select.gene*. If this argument is empty, only the first gene will be kept in the simulation procedure. Finally, the multiplicative factor of the GRR between each group of cases compared to the first group of cases needs to be specified to the argument *GRR.multiplicative.factor* (number of values = number of cases groups - 1).

The function **GRR.matrix()** will return a GRR matrix in the appropriate format for the function **genotypic.freq()**. Examples of these two functions are shown below:

```
# GRR calculated using the same formula as in SKAT,
# with values in the second group of cases being twice the values from the first one

GRR.del <- GRR.matrix(GRR = "SKAT", genes.maf = Kryukov, n.case.groups = 2,
                      GRR.multiplicative.factor=2, select.gene = "R1")
GRR.del[,1:5]
```

```
##           [,1]      [,2]     [,3]      [,4]      [,5]
## [1,]  5.037728  6.222656 12.34472  9.042877  8.133663
## [2,] 10.075455 12.445313 24.68944 18.085755 16.267327
```

```
# Calculation of genotype frequencies in the two groups of cases and the controls group
# The previous GRR matrix is used with a multiplicative model of the disease
# The prevalence in each group of cases is 0.001

geno.freq.groups <- genotypic.freq(genes.maf = Kryukov, select.gene="R1",
                                   GRR.het = GRR.del, prev = c(0.001, 0.001),
                                   genetic.model = "multiplicative")
str(geno.freq.groups)
```

```
## List of 3
##  $ freq.homo.ref: num [1:3, 1:383] 1 0.999 0.998 1 1 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:3] "controls" "cases_1" "cases_2"
##   .. ..$ : NULL
##  $ freq.het     : num [1:3, 1:383] 1.87e-04 9.56e-04 1.91e-03 5.57e-05 3.52e-04 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:3] "controls" "cases_1" "cases_2"
##   .. ..$ : NULL
##  $ freq.homo.alt: num [1:3, 1:383] 7.90e-09 2.29e-07 9.15e-07 6.49e-10 3.11e-08 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:3] "controls" "cases_1" "cases_2"
##   .. ..$ : NULL
```

```
# frequencies of the homozygous alternative genotype in the different groups
geno.freq.groups$freq.homo.alt[,1:5]
```

```
##                  [,1]         [,2]         [,3]         [,4]         [,5]
## controls 7.897334e-09 6.485888e-10 7.480447e-14 6.568600e-12 2.503543e-11
## cases_1  2.288672e-07 3.106821e-08 4.778956e-11 9.067311e-10 2.469545e-09
## cases_2  9.145933e-07 1.242291e-07 1.911556e-10 3.626706e-09 9.877199e-09
```

It is also possible to calculate the MAF in each group of individuals as follow:

```
#MAF calculation for the first five variants
geno.freq.groups$freq.homo.alt[,1:5] + 0.5*geno.freq.groups$freq.het[,1:5]
```

```
##                  [,1]         [,2]         [,3]         [,4]         [,5]
## controls 9.375276e-05 2.785699e-05 5.403418e-07 3.246158e-06 5.972867e-06
## cases_1  4.784006e-04 1.762618e-04 6.912999e-06 3.011198e-05 4.969452e-05
## cases_2  9.563437e-04 3.524614e-04 1.382590e-05 6.022214e-05 9.938410e-05
```

Simulation of genotypes

In addition to compute the genotypic frequencies in each group of individuals, it is possible to directly simulate genotypes for a group of controls and more than two groups of cases. This can be done using the function **rbm.GRR()** which relies on the function **genotypic.freq()** explained previously. The arguments *genes.maf*, *select.gene*, *prev* and *genetic.model* are the same as in **genotypic.freq()**.

In **rbm.GRR()**, the parameter *maf.threshold* corresponds to the maximum maf used to define a rare variants, i.e. among which causal variants will be sampled. The proportion of causal variants (among those rare variants) and the proportion of protective variants (among causal variants) simulated in the genomic region should be specified to *p.causal* and *p.protect* respectively. The argument *GRR.matrix.del* should contain a matrix with GRR values as if all the variants were deleterious. This matrix can be obtained using the

function **GRR.matrix()** previously explained. If *genetic.model="general"*, two GRR matrices need to be given as a list to the argument *GRR.matrix.del* (one for the heterozygous genotype and the other for the homozygous alternative genotype).

If the user wants to simulate protective variants in addition to deleterious variants, a similar argument to *GRR.matrix* should be given to *GRR.matrix.pro* with GRR values as if all variants were protective. If the argument *GRR.matrix.pro* is empty and *p.protect*>0, *GRR.matrix.pro* will correspond to $1/GRR.matrix$. These deleterious and protective GRR values will then be assigned to the sampled deleterious and protective variants in the simulations, the non-causal variants having GRR values of 1.

The size of the different groups of individuals should be a vector specified to *size*, and the user has to choose whether the causal variants will be the same between the different groups of cases with the argument *same.variant*. Finally, the number of simulations need to be specified via *replicates*.

The function **rbm.GRR()** will return a bed matrix with the group of each individual in the field *@ped$pheno*, the first one being considered by default as the controls group, and the replicate number corresponding to the genomic region in the field *@snps$genomic.region*.

The example below shows how to simulate a group of 1 000 controls and two groups of 500 cases with 50% of deleterious variants having GRR values from the previous example. The deleterious variants are different between the two groups of cases and 5 genomic regions are simulated.

```
x <- rbm.GRR(genes.maf = Kryukov, size = c(1000, 500, 500), replicates = 5,
             prev = c(0.001, 0.001), GRR.matrix.del = GRR.del, p.causal = 0.5,
             p.protect = 0, same.variant = FALSE,
             genetic.model = "multiplicative", select.gene = "R1")
x
```

```
## A bed.matrix with 2000 individuals and 1915 markers.
## snps stats are set
##   There are  1639  monomorphic SNPs
## ped stats are set
```

```
table(x@ped$pheno)
```

```
##
##    0    1    2
## 1000  500  500
```

```
table(x@snps$genomic.region)
```

```
##
##  R1  R2  R3  R4  R5
## 383 383 383 383 383
```

Simulations based on haplotypes

Ravages also offers the possibility to perform genetic simulations based on real haplotypes data to mimic linkage desequlibrium pattern observed in these data.
Two options are available to do so: genetic data are simulated using either haplotypes and their frequencies in different groups of individuals (function **rbm.haplos.freqs()**), or using haplotypes and a liability model on their burdens (function **rbm.haplos.thresholds()**).

For the first situation, **rbm.haplos.freqs()** requires a matrix of haplotypes (argument *haplos* with one row per haplotype and one column per variant), and a matrix with each haplotype frequency in each simulated group of individuals (argument *freqs* with one row per haplotype and one column per group of individuals). Observed haplotypes are then sampled in each group of individuals in respect to their frequencies in each one of them.

For the second situation, **rbm.haplos.thresholds()** requires a matrix of haplotypes (argument *haplos* as well), and a number of other arguments to draw the liability model. The idea of these simulations is to draw a liability distribution of the burden computed on causal variants for each haplotype, to compute the probability of each haplotype in each group of individuals based on this distribution, and finally to sample haplotypes in each group based on those probabilities. The parameter *weights* indicates how to compute the weights into burdens computation: either all variants have the same weights (*weights="constant"*), or higher weights are given to rare variants using the same formula as in SKAT : $w = -0.4 * |log_{10}(MAF)|$). The parameter *maf.threshold* corresponds to the maf threshold used to define a rare variants, i.e. all monomorphic variants and variants with a MAF upper this threshold will have a weight of 0. In addition, the proportion of causal variants among those variants needs to be given to *p.causal*, and the proportion of protective variants among causal variants needs to be given to *p.protect*. Causal variants will be sampled among variants with a positive weight, and haplotypes' burden will be computed on these variants.

The parameter *h2* corresponds to the phenotypic variance explained by the gene: haplotypes burdens will be adjusted on this value; and *prev* corresponds to the prevalence of each group of individuals and will be used to compute the liability threshold for each group of individuals (liabilities of individuals in group i will fall between the quantile 1-$prev_i$ and +Inf). This quantile is computed using a standard normal distribution if *normal.approx*=TRUE (applicable for low *h2* value and pretty large *prev* values), otherwise, quantiles will be computed from a sample of 10^6 observed burdens.

If one wants to simulate a group of unselected controls, *prev* needs to be set at 1, regardless of the other arguments: haplotypes are sampled uniformly in the whole liability distribution. The arguments *p.causal*, *p.protect*, *h2* and *prev* are vectors of same length as the number of groups. If they are of size 1, their value will be repeated.

As for the previous functions, the number of simulations to perform should be given to *replicates*. As computing haplotypes' probabilities for a set of causal variants can be lengthy, it is possible to run several multiple replicates on a given set. Indeed, even if the same causal variants are taken, different haplotypes (with same probabilities) will be sampled in each individual for each replicate. To ensure a minimal variability, we yet recommend to resample the set of causal variants several times. For example, for 1000 replicates, we recommend to resample causal variants 20 times, and therefore to perform 50 replicates by set of causal variants. The number of replicates to perform by set of causal variants should be given to *rep.by.causal*.

As for the genetic simulations based on GRR, the sizes of each group should be given to *size*, and the number of simulations to *replicates*. As before, the replicate number will be in the slot *@snps$genomic.region* and the group of individuals in *@ped$pheno*, both being factors. If **rbm.haplos.freqs()** is used, phenotype values will correspond to the colnames of *freqs*.

Two examples these functions are presented below to illustrate these functions. The first one simulates 5 groups of individuals, with haplotype frequencies from the five european populations for the LCT gene. The second one simulates one group of 100 controls and two groups of 50 cases based on the liability model with different genetic effects.

```
#Load LCT dataset for haplotype matrix
data(LCT.haplotypes)

#Simulation based on haplotypes frequencies
#for the variants in the LCT gene in the EUR population
LCT.gene.hap <- LCT.hap[which(LCT.sample$super.population=="EUR"),
                        which(LCT.snps$pos>=136545410 & LCT.snps$pos<=136594750)]
#Individuals from EUR
LCT.sample.EUR <- subset(LCT.sample, super.population=="EUR")
#Matrix of haplotype frequencies
LCT.freqs <- sapply(unique(LCT.sample.EUR$population), function(z)
                    ifelse(LCT.sample.EUR$population==z,
                           1/table(LCT.sample.EUR$population)[z], 0))
#Simulation of genetic data for five groups of 50 individuals
```

```
x <- rbm.haplos.freqs(haplos=LCT.gene.hap, freqs=LCT.freqs, size=rep(50,5), replicates=5)

#Simulation of 100 controls, and two groups of 50 cases with 30 causal variants
#and with the second group having half h2 and twice the prevalence
#compared to the first one
#5 replicates are performed and causal variants are sampled once
x <- rbm.haplos.thresholds(haplos=LCT.gene.hap, p.causal = 0.3, h2=c(0.01,0.01,0.02),
                           prev=c(1,0.01,0.005), size=c(100, 50, 50), replicates=5,
                           rep.by.causal = 5)
```

```
## 3  groups of individuals will be simulated
```

Power calculation

Power of the RVAT (burden tests and SKAT) can be directly computed using **Ravages**. To do so, the functions *rbm.GRR.power()* and *rbm.haplos.power()* which rely on simulations based on GRR and on haplotypes respectively can be used. Those two functions use the same arguments as the simulation functions previously presented. Power calculations are available for *CAST*, *WSS* and *SKAT*. In addition, using *rbm.GRR.power()*, theoretical power can be rapidly computed for CAST by using **power.type="theoretical"** and **RVAT="CAST"**.

Below is an example of CAST power based on theoretical calculations or simulations with haplotypes.

```
#CAST theoretical power
GRR.del <- GRR.matrix(GRR = "SKAT", genes.maf = Kryukov, n.case.groups = 2,
                      GRR.multiplicative.factor=2, select.gene = "R1")
rbm.GRR.power(genes.maf = Kryukov, size = c(1000, 500, 500),
              prev = c(0.001, 0.001), GRR.matrix.del = GRR.del,
              p.causal = 0.5, p.protect = 0, select.gene="R1",
              same.variant = FALSE, genetic.model = "multiplicative",
                        power.type="theoretical")
```

```
##   2.5e-06
## 0.5979181
```

```
#CAST power based on simulations on haplotypes (20 replicates)
rbm.haplos.power(haplos=LCT.gene.hap, freqs=LCT.freqs, size=rep(50,5),
                 replicates=20, rep.by.causal = 10,  RVAT = "CAST", cores = 1)
```

```
## Categorial phenotype
```

```
##      2.5e-06
## CAST    0.35
```