

Package Ravages (RARE Variant Analysis and GENetic Simulation)

Herve Perdry and Ozvan Bocher

2021-08-17

Introduction

Ravages was developped to simulate genetic data and to perform rare variant association tests (burden tests and the variance-component test SKAT) on more than two groups of individuals (Bocher et al., 2019, doi:10.1002/gepi.22210). Ravages relies on the package Gaston developped by Herve Perdry and Claire Dandine-Roulland. 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 rare variant association tests on real data. A second vignette is available showing how to simulate genetic data and how to use them for power calculation. To learn more about all options of the functions, the reader is advised to look at the manual pages.

Example of analysis using LCT data

Below is an example of an association analysis and previous steps of data filtering using the dataset LCT available with the package Ravages. This dataset contains data from the 1000Genome project in the genomic region containing the Lactase gene. In this example, we look for an association between rare variants and the european populations of 1000Genomes. The population of each individual is available in the dataframe LCT.matrix.pop1000G. Details about each function is given right after this example.

```
#Import data in a bed matrix
x <- as.bed.matrix(x=LCT.matrix.bed, fam=LCT.matrix.fam, bim=LCT.snps)
#Add population
x@ped[,c("pop", "superpop")] <- LCT.matrix.pop1000G[,c("population", "super.population")]

#Select EUR superpopulation
x <- select.inds(x, superpop=="EUR")
x@ped$pop <- droplevels(x@ped$pop)

# Group variants within know genes by extending their positions
# 500bp upstream and downstream
x <- set.genomic.region(x, flank.width=500)
table(x@snps$genomic.region, useNA = "ifany")

##
## R3HDM1  UBXN4    LCT    MCM6    DARS    <NA>
##   2047   1207   1454   1149    924   1295

# Filter variants with maf in the entire sample lower than 1%
# And keep only genomic region with at least 10 SNPs
```

```

x1 <- filter.rare.variants(x, filter = "whole", maf.threshold = 0.01, min.nb.snps = 10)
table(x1@snps$genomic.region, useNA="ifany")

##
## R3HDM1  UBXN4    LCT    MCM6    DARS
##      268    172    208    163    136

# run burden test CAST, using the 1000Genome population as "outcome"
# Null model for CAST
x1.H0.burden <- NullObject.parameters(x1@ped$pop, ref.level = "CEU",
                                     RVAT = "burden", pheno.type = "categorical")
burden(x1, NullObject = x1.H0.burden, burden = "CAST", cores = 1)

## Categorical phenotype

##           p.value is.err
## R3HDM1 1.300274e-04      0
## UBXN4  1.993354e-05      0
## LCT    4.489474e-08      0
## MCM6   3.142808e-08      0
## DARS   2.064661e-03      0

# run SKAT, using the 1000Genome population as "outcome"
# Null model for SKAT with only few permutations
x1.H0.SKAT <- NullObject.parameters(x1@ped$pop, RVAT = "SKAT", pheno.type = "categorical")
SKAT(x1, x1.H0.SKAT, params.sampling=list(perm.target = 10, perm.max = 500))

## Categorical phenotype
## permutations

##           stat      p.perm      p.chi2      p.value
## R3HDM1 6.445786 0.001996008 1.042745e-05 1.042745e-05
## UBXN4  2.136390 0.011976048 9.329497e-03 9.329497e-03
## LCT    3.623174 0.001996008 4.413600e-04 4.413600e-04
## MCM6   2.923228 0.007984032 4.899532e-03 4.899532e-03
## DARS   2.132967 0.052884615 8.235545e-02 5.288462e-02

```

Defining genomic regions

For rare variant association tests, the unit of analysis is not a single variant but a genomic region, typically a gene. The first step of the analysis is therefore to group variants into genomic regions. This can be done using the function `set.genomic.region()` and known gene positions. It works on a `bed.matrix` (see Gaston) and simply adds a column “genomic.region” to the slot `x@snps` containing the genomic region (a factor) assigned to each variant. Gene positions should be given to `regions` as a dataframe containing the following columns: `Chr`, `Start`, `End`, `Gene_Name`. The dataframe should absolutely be ordered in the genome order, as well as the levels of `regions$Gene_Name`. By default, any variant being outside the gene positions won’t be annotated. Gene positions can be extended to annotate more variants using the argument `flank.width` corresponding to the number of base pairs upstream and downstream the gene. If `flank.width=Inf`, each variant will be assigned to the nearest gene. If two regions overlap, variants in the overlapping zone will be attributed to both regions, separated by a comma. By default, `split=TRUE` which means that variants attributed to multiple regions will be duplicated in the bed matrix. This is done by calling the function `bed.matrix.split.genomic.region()` which takes a bed matrix as argument (`x`) and duplicates variants being assigned to multiple regions separated by `split.pattern`. If `changeID=TRUE`, the id from the bedmatrix will be changed to the format `chr:pos:A1:A2:genomic.region` to distinguish the duplicated variants.

The files **genes.b37** and **genes.b38** available in Ravages which contain gene positions from ENSEMBL versions hg19 and hg38 can be used for *regions* to define gene positions.

An example is shown below:

```
#Example bed matrix with 4 variants
x.ex <- as.bed.matrix(x=matrix(0, ncol=4, nrow=10),
                     bim=data.frame(chr=1:4, id=paste("rs", 1:4, sep=""),
                                   dist = rep(0,4), pos=c(150,150,200,250),
                                   A1=rep("A", 4), A2=rep("T", 4)))

#Example genes dataframe
genes.ex <- data.frame(Chr=c(1,1,3,4), Start=c(10,110,190,220), End=c(170,180,250,260),
                      Gene_Name=factor(letters[1:4]))

#Attribute genomic regions without splitting the variants
#attributed to multiple genomic regions
x.ex <- set.genomic.region(x.ex, regions = genes.ex, split = F)
x.ex@snp$genomic.region

#Split genomic regions
x.ex.split <- bed.matrix.split.genomic.region(x.ex, split.pattern = ",")
x.ex.split@snp$genomic.region
```

CADD regions

In addition to the positions of the genes, Ravages offers the positions of CADD regions which correspond to non-overlapping regions defined genome-wide using the variant pathogenicity score CADD: regions are defined between variants observed at least two times in GnomAD with a high CADD score. This CADD score is not the original CADD score v1.4 published by Rentzsch et al., but an adjusted score that take into account three types of functional areas : coding, regulatory and intergenic areas. This enable to find the most important functional variants within each of those three areas. To assign variants to CADD regions and to the functional areas, the function **set.CADDregions()** can be used. It will add *genomic.region* to the bed matrix with the corresponding CADD regions, and *SubRegion* with the functional area. In addition, *adjCADD.Median* will be given which correspond to the median adjusted CADD score of variants observed at least two times in GnomAD and can be used for variant filtering as explained later.

Rare variant selection

Frequency filter

To perform rare variant analysis, it is also important to define what is a rare variant in order to leave out common ones. The function **filter.rare.variants()** enables to keep only variants with a MAF (Minor Allele Frequency) below a given threshold while leaving out monomorphic variants. This function uses and returns a bed.matrix which can be filtered in three different ways:

- If *filter*="whole", all the variants with a MAF lower than the threshold in the entire sample will be kept.
- If *filter*="controls", all the variants with a MAF lower than the threshold in the controls group will be kept. In this situation, the controls group needs to be specified to the argument *ref.level*.
- If *filter*="any", all the variants with a MAF lower than the threshold in any of the groups will be kept.

It is also possible to specify the minimum number of variants needed in a genomic region to keep it using the parameter *min.nb.snps*, as well as the minimum cumulative MAF using *min.cumulative.maf*.

For the filter *any* and *controls*, the group of each individual should be given as a factor to *group*.

CADD filter

In the analysis by CADD regions, we propose a new approach to select rare variants to include in the RVAT that keeps within each CADD region only the variants with an adjusted CADD score greater than the median observed in GnomAD. To this end, the function **filter.adjustedCADD()** can be used which relies on the same arguments than **filter.rare.variants()** for the filtering based on the frequency. Within this function, bi-allelic variants of the bed matrix will be directly annotated with the adjusted CADD scores using the corresponding file in the package *CADDRegions* and only those with a score greater than the median will be kept. It is necessary that the bed matrix is first annotated with **set.CADDregion()**.

Rare variant association tests

We have implemented two burden tests extensions (CAST and WSS) and an extension of the variance-component test SKAT to perform the association tests between a gene and more than two groups of individuals. The general idea of burden tests is to compute a genetic score per individual and per genomic region and to test if it differs between the different groups of individuals. To extend these tests to more than two groups of individuals, a non-ordinal multinomial regression is used. The independent variable in this regression is the genetic effect of the gene represented by the genetic score. Covariates can be added in the model. In addition to the genetic scores CAST and WSS directly implemented in the package, the user can specify another genetic score for the regression.

The variance-component test SKAT looks at the dispersion of genetics effects of rare variants. A geometrical interpretation of the test was used for its extension to more than two groups of individuals. Covariates can also be included in this model.

Genetic score for burden tests

We have implemented two functions to compute CAST and WSS scores respectively. These functions return a matrix with one row per individual and one column by genomic region. They are directly called in the function **burden()** if these scores are used to perform the association tests. It is also possible to compute genetic scores in a genomic region based on a vector of weights for each variant using the function **burden.weighted.matrix()**. It is important to note that all burden functions compute a score for the rare alleles. Therefore, if the reference allele is rare, the alleles will be flipped and it will be counted in the score instead of the alternative allele. This can be avoided only in the **CAST()** score with the argument *flip.rare.alleles* set at TRUE by default.

CAST

CAST is based on a binary score which has a value of one if an individual carries at least one variant in the considered genomic region, and 0 otherwise. A MAF threshold for the definition of a rare variant is therefore needed as an argument to *maf.threshold*. This score can be computed using the function **CAST()** as shown here on the LCT data:

```
#Calculation of the genetic score with a maf threshold of 1%
CAST.score <- CAST(x = x1, genomic.region = x1@snps$genomic.region, maf.threshold = 0.01)
head(CAST.score)
```

##	R3HDM1	UBXN4	LCT	MCM6	DARS
## HG00096	0	0	1	1	1
## HG00097	1	0	0	0	0
## HG00099	1	0	0	0	0
## HG00100	0	1	0	0	0
## HG00101	0	1	0	0	0
## HG00102	1	0	0	0	1

WSS

WSS (Weighted Sum Statistic) is based on a continuous score giving the highest weights to the rarest variants:

$$WSS_j = \sum_{i=1}^R I_{ij} \times w_i$$

with

$$w_i = \frac{1}{\sqrt{t_i \times q_i \times (1 - q_i)}}$$

and

$$q_i = \frac{n_i + 1}{2t_i + 1}$$

Where n_i is the total number of minor alleles genotyped for variant i , t_i is the total number of alleles genotyped for variant i and I_{ij} is the number of minor alleles of variant i for the individual j . In the original method, each variant is weighted according to its frequency in the controls group. In our version of WSS, the weights depend on allele frequencies calculated on the entire sample. The function **WSS()** can be used to compute the WSS score as shown on the LCT data:

```
WSS.score <- WSS(x = x1, genomic.region = x1@snps$genomic.region)
head(WSS.score)
```

##	R3HDM1	UBXN4	LCT	MCM6	DARS
## HG00096	0.0000000	0.0000000	0.8185268	1.26932	1.418436
## HG00097	0.8185268	0.0000000	0.0000000	0.000000	0.000000
## HG00099	1.0019881	0.0000000	0.0000000	0.000000	0.000000
## HG00100	0.0000000	1.001988	0.0000000	0.000000	0.000000
## HG00101	0.0000000	1.001988	0.0000000	0.000000	0.000000
## HG00102	1.0019881	0.0000000	0.0000000	0.000000	1.001988

Other genetic scores

It is also possible to compute other genetic scores based on variants weights using the function **burden.weighted.matrix()**. The weights should be given as a vector to *weights* having the same size as the number of variants. The genetic score will be compute as:

$$Score_j = \sum_{i=1}^R I_{ij} \times w_i$$

with w_i the weight of each variant in *weights*, and I_{ij} the number of minor alleles for individual j in variant i . Here is an example corresponding to a genetic score with all the weights at 1, i.e. counting the number of minor alleles:

```
Sum.score <- burden.weighted.matrix(x = x1, weights = rep(1, ncol(x1)))
head(Sum.score)
```

##	R3HDM1	UBXN4	LCT	MCM6	DARS
## HG00096	0	0	1	2	2
## HG00097	1	0	0	0	0
## HG00099	1	0	0	0	0
## HG00100	0	1	0	0	0
## HG00101	0	1	0	0	0
## HG00102	1	0	0	0	1

Regressions

We have extended CAST and WSS using non-ordinal multinomial regression models. Let consider C groups of individuals including a group of controls ($c = 1$) and $C - 1$ groups of cases with different sub-phenotypes of the disease. We can compute $C - 1$ probability ratios, one for each group of cases:

$$\ln \frac{P(Y_j = c)}{P(Y_j = 1)} = \beta_{0,c} + \beta_{G,c}X_G + \beta_{k1,c}K_1 + \dots + \beta_{kl,c}K_l$$

Where Y_j corresponds to the phenotype of the individual j and K_l is a vector for the l th covariate with the corresponding coefficient β_{kl} . The genetic effect is represented by X_G and correspond to the genetic score CAST or WSS with $\beta_{G,c}$ the log-odds ratio associated to this burden score.

The p-value associated to the genetic effect is calculated using a likelihood ratio test comparing this model to the same model without the genetic effect (null hypothesis). If only two groups are compared, a classical logistic regression is performed.

This regression can be performed on a `bed.matrix` using the function **burden()** which relies on the package `mlogit`. Parameters under the null model can be obtained using the function **NullObject.parameters()**. To generate this null model, the *pheno* of each individual should be given as a factor, and the potential covariates to include in the model should be given as a matrix to the argument *data* (one row per individual and one column by covariate). If only a subset of covariates from *data* are to be included in the model, a R formula should be given to *formula* with these covariates, otherwise all the covariates will be included. In addition, the reference group should be given to the argument *ref.level*, i.e. all odds ratios will be computed in comparison to this group of individuals in the regression on the score. The choice of the reference group won't affect the p-value. As the parameters needed to run the association tests depend on the type of test performed (burden tests or SKAT), and the type of phenotype (continuous or categorical), both arguments should be given to **NullObject.parameters()** (*RVAT* and *pheno.type* respectively).

NullObject.parameters() will return a list with the parameters to use in **burden()** to run the burden test, including the Log-Likelihood computed under the null model, the argument *data* with the covariates to include determined from the argument *formula*.

Once the null model has been run, the function **burden()** can be used to perform burden tests. To do so, the user needs to give the results from **NullObject.parameters()** to the argument *NullObject*, and needs to specify the genomic region associated to each variant (argument *genomic.region*).

The CAST or WSS genetic scores can be directly calculated in the regression (*burden*="CAST" or *burden*="WSS"). The user can also use another genetic score in the regression, which has to be specified as a matrix with one individual per row and one genomic region per column to *burden*. In this situation, no *bed* matrix is needed, and the result from **burden.weighted.matrix()** can be used directly.

To speed up the computation time, calculations can be parallelised using the argument *cores*, set at 10 by default.

burden() will return the p-value associated to the regression for each genomic region. If there is a convergence problem with the regression, the function will return 1 in the column *is.err*. The effect size (odds ratio for categorical phenotypes and beta value for continuous phenotypes) associated to each group of cases compared to the reference group (*NullObject\$ref.level*) with its confidence interval at a given alpha threshold (argument *alpha*) can also be obtained if *get.effect.size=TRUE*.

An example of the p-value and OR calculation with its 95% confidence interval using WSS on the LCT data is shown below with or without the inclusion of covariates. The outcome here corresponds to the population from 1000Genome.

```
#Null model
x1.H0 <- NullObject.parameters(x1@ped$pop, ref.level = "CEU",
                              RVAT = "burden", pheno.type = "categorical")

# WSS
burden(x = x1, NullObject = x1.H0, burden = "WSS",
       alpha=0.05, get.effect.size=TRUE, cores = 1)

#Sex + a simulated variable as covariates
sex <- x1@ped$sex
```

```

u <- runif(nrow(x1))
covar <- cbind(sex, u)
#Null model with the covariate "sex"
x1.H0.covar <- NullObject.parameters(x1@ped$pop, ref.level = "CEU",
                                     RVAT = "burden", pheno.type = "categorical",
                                     data = covar, formula = ~ sex)

#Regression with the covariate "sex" without OR values
#Using the score matrix WSS computed previously
burden(NullObject = x1.H0.covar, burden=WSS.score, cores = 1)

```

Finally, using Ravages, it is also possible to perform burden tests with a continuous phenotype by specifying *pheno.type* = “continuous”, and by giving a numeric vector to *pheno* as showed below:

```

#Random continuous phenotype
set.seed(1) ; pheno1 <- rnorm(nrow(x1))
#Null model
x1.H0.continuous <- NullObject.parameters(pheno1, RVAT = "burden",
                                          pheno.type = "continuous")

#Test CAST
burden(x1, NullObject = x1.H0.continuous, burden = "CAST", cores = 1)

```

Regressions with sub-scores

In the analysis by CADD regions, different types of functional areas can be found within CADD regions. To take them into account, we propose to use subscores in the regression, one for each functional area, within each CADD region. There will be at most three sub-scores in the regression. This can be performed using **burden.subscores()** which works similarly as **burden()** with the extra argument *SubRegion* corresponding to the vector with the subregions on which subscores should be computed. **burden.subscores()** can be directly applied on the bed matrix obtained from **set.CADDregions()** or **filter.adjustedCADD()**.

SKAT

We also extended the variance-component test SKAT using a geometric interpretation. Unlike the burden tests, there is no burden calculated in this test: the distribution of the genetic effects in the genomic region is compared to a null distribution. SKAT is based on a linear mixed model where the random effects correspond to the genetic effects.

Before running the SKAT association test, the function **NullObject.parameters()** first needs to be called as for the burden tests, by specifying *RVAT* = “SKAT”. The group of each individual should be given as a factor to *pheno* if a categorical phenotype is studied, and as a numeric vector if a continuous phenotype is studied. As before, potential covariates should be included as a matrix to *data*. If only some of them are to be included, they should be given as a R formula to *formula*. This function will compute parameters to use the **SKAT()** function, and should be given to this function using the argument *NullObject*.

To compute the p-values, a chi-square approximation is used based on the statistics’ moments. The moments can either be estimated using a sampling procedure, or be analytically computed using the method from Liu et al. 2008. The chi-square approximation can be based on the first three moments (*estimation.pvalue* = “skewness”), or on moments 1, 2 and 4 to have a more precise estimation of the tail distribution (*estimation.pvalue* = “kurtosis”). If *get.moments* = “theoretical” and *estimation.pvalue* = “skewness”, it is equivalent to the “liu” method in the SKAT package, and if *estimation.pvalue* = “kurtosis”, it corresponds to the “liu.mod” method in the SKAT package. If *debug* = *TRUE*, the statistics’ moments will be returned in addition to the p-values.

If the sample size is lower than 2000, we recommend to use the sampling procedure. If no covariates are present, a simple permutation procedure can be used (*get.moments* = “permutations”), otherwise, a bootstrap

sampling should be used (*get.moments* = “bootstrap”). For those two situations, a sequential procedure is used to compute the p-values: permuted statistics are computed and each one is compared to the observed statistics. The sampling procedure stops when either *perm.target* (the number of times a permuted statistics should be greater than the observed statistics) or *perm.max* (the maximum number of permutations to perform) is reached. P-values are then computed in two different ways: if *perm.target* is reached, the p-value is computed as *perm.target* divided by the number of permutations performed to reach this value; if *perm.max* is reached before *perm.target* (that is, for pretty small p-values), p-values are computed using the chi-square approximation based on moments estimated from the permuted statistics. *perm.target* and *perm.max* should be given as a list to the argument *params.sampling* of SKAT.

If the sample size is bigger than 2000, the analytical calculation from Liu et al. can be used to compute the theoretical moments. In this situation, it is possible to parallelise the calculations using the argument *cores*, set at 10 by default.

It is possible to clearly ask for a specific method to compute the moments using *get.moments* = “permutations”, “bootstrap” or “theoretical”. By default, *get.moments* = “size.based”, and the method will depend on the sample size.

An example of the SKAT function by specifying the “permutations” or “theoretical” method is shown.

```
#Null model
x1.null <- NullObject.parameters(x1@ped$pop, RVAT = "SKAT", pheno.type = "categorical")
#Permutations because no covariates
SKAT(x1, x1.null, get.moments = "permutations", debug = TRUE,
     params.sampling = list(perm.target = 100, perm.max = 5e4))
#Theoretical on 1 core
SKAT(x1, x1.null, get.moments = "theoretical", debug = TRUE, cores = 1)
```

It is also possible to perform the SKAT test on a continuous phenotype by using the argument *pheno.type* = “continuous” in **NullObject.parameters()** before the **SKAT()** function:

```
#Random continuous phenotype
set.seed(1) ; pheno1 <- rnorm(nrow(x1))
#Null Model with covariates
x1.H0.c <- NullObject.parameters(pheno1, RVAT = "SKAT", pheno.type = "continuous",
                                data = covar)

#Run SKAT
SKAT(x1, x1.H0.c)
```

Data management

Data in plink format or in vcf format can be loaded in R using the functions **read.bed.matrix()** and **read.vcf()** respectively from the package *gaston*.

If the data for the controls and the different groups of cases are in different files, they can be loaded separately and then combined using the function **gaston:::rbind()** as long as the same variants are present between the different groups of individuals.

An example is given below where the simulated data have been split according to the group of each individual, and then combined in a bed.matrix:

```
#Selection of each group of individuals
CEU <- select.inds(x1, pop=="CEU")
CEU

## A bed.matrix with 99 individuals and 947 markers.
## snps stats are set
## There are 748 monomorphic SNPs
## ped stats are set
```



```

FIN <- select.inds(x1, pop=="FIN")
FIN

## A bed.matrix with 99 individuals and 947 markers.
## snps stats are set
##   There are 771 monomorphic SNPs
## ped stats are set

GBR <- select.inds(x1, pop=="GBR")
GBR

## A bed.matrix with 91 individuals and 947 markers.
## snps stats are set
##   There are 792 monomorphic SNPs
## ped stats are set

#Combine in one file:
CEU.FIN.GBR <- rbind(CEU, FIN, GBR)
CEU.FIN.GBR

## A bed.matrix with 289 individuals and 947 markers.
## snps stats are set
##   There are 515 monomorphic SNPs
## ped stats are set

```

Whole-Genome analysis

RVAT can be run genome-wide with the approach by CADD regions. Due to time and memory management, we advise the user to import the data and apply the different functions of annotation, filtering and association chromosome by chromosome for large datasets.