sambadaOnR

January 21, 2019

Type Package

Title Processing pipeline for Sambada from pre to post processing

Version 1.0

Date 2019-01-16

Author Solange Gaillard, Sylvie Stucki, Oliver Selmoni, Elia Vajana

Maintainer Your Name <your@email.com>

Description Processing pipeline for Sambada from pre to post processing

License GPL (>= 2)

Imports SNPRelate

LinkingTo

RoxygenNote 6.1.0

Suggests knitr,
rmarkdown

VignetteBuilder knitr

R topics documented:

`sambadaOnR-package` *sambadaOnR: A package for running sambada within R with pipeline from pre to post-processing*

Description

The sambadaOnR package provides four categories of important functions: Install sambada, Preprocessing, Running sambada and Post-processing.

Install sambada functions

You can download sambada (if not already on your computer) from GitHub using the function `downloadSambada`

Preprocessing functions

The Preprocessing functions contain three functions:

- `prepareGeno`: translate genomic file to sambada's input file while applying genomic filters
- `createEnv`: create your environmental file from file location from local raster or global worldclim database
- `prepareEnv`: reduce envrionmental file with correlated variables and analyse population structure

Running sambada function

To run sambada, you will want to use the function: `sambadaParallel`

Postprocessing functions

The Postprocessing functions contain three functions:

- `prepareOutput`: calculate p and q-values from sambada output
- `plotResultInteractive`: start an interactive local web page to query a manhattan plot with maps, plots and ensembl query result
- `plotMap`: create a map of marker, population structure or environmental variable distribution

`changePath` *Adds folder to the PATH environmental variable*

Description

Adds the `directory` path to the environmental PATH variable. This operation is only valid for the current R session. You must run change_path for every new R session. Alternatively, you can permantently edit your "PATH" environmental variable on your OS so that it entails the path to the binaries folder of sambada.

Usage

```
changePath(directory)
```

Arguments

directory	character The path to samBada binaries folder

Author(s)

Solange Gaillard

---

createEnv *Create env file from raster file(s) and/or glabal database present in the raster r package*

Description

Create env file as an input for SamBada (it is recommended to run prepare_env function before running samBada) raster file(s) and/or glabal database present in the raster r package

Usage

```
createEnv(locationFileName, x, y, locationProj, separator = ",",
  rasterName, rasterProj, directory = FALSE, worldclim = TRUE,
  srtm = FALSE, interactiveChecks, verbose = TRUE)
```

Arguments

locationFileName	!! char Name of the file containing location of individuals. Must be in the active directory. Supported extension are .csv, .shp. All columns present in this file will also be present in the output file
x	char Name of the x (or longitude if not projected coordinate system) column in the locationFileName. Required if locationFileName extension is .csv
locationProj	integer Coordinate system EPSG code of the locationFileName. If locationFileName is already georeferenced, this argument will be skipped. Required if locationFileName extension is csv.
rasterName	char or list Name or list of name of raster files to import. Supported format are the one of raster package. If directory is TRUE then the path to the directory
rasterProj	integer or list of integer Coordinate system EPSG code of the rasterlayer. If rasterlyer is already georeferenced, this argument will be skipped. If rasterName is a list, can be either a single number if all projections are the same or a list of projection for all files if different. If directory is TRUE, can only contain one number (all projections must be equal or rasters must be georeferenced)
directory	logical If true, all .tif, .gtiff, .img, .sdat, . present in rasterName will be loaded
worldclim	logical If TRUE worldclim bio, tmin, tmax and prec variables will be downloaded at a resolution of 0.5 minutes of degree (the finest resolution). Rely rgdal and gdalUtils R package to merge the tiles. The downloaded tiles will be stored in the (new) wc0.5 directory of the active directory
srtm	logical If TRUE the SRTM (altitude) variables will be downloaded at a resolution ... Rely rgdal and gdalUtils R package to merge the tiles. The downloaded tiles will be stored in the (new) wc0.5 directory of the active directory

`interactiveChecks`
: logical If TRUE, shows loaded rasters and point locations

`verbose`
: logical If TRUE, indication on process will be shown

`x`
: char Name of the y (or latitude if not projected coordinate system) column in the `locationFileName`. Required if `locationFileName` extension is .csv

Value

None

Author(s)

Solange Gaillard

Examples

```
#Own raster + worldclim download
createEnv(rasterName=c('prec.tif','tmin.sdat'),locationFileName='MyFile.shp',
      rasterProj=c(4326,21781), worldclim=TRUE,interactiveChecks=TRUE)

#Worldclim download only
createEnv(locationFileName='MyFile.csv',x='Longitude',y='Latitude',locationProj=4326,
      worldclim=TRUE,interactiveChecks=FALSE)
```

`downloadSambada` *Download samBada*

Description

Downloads from GitHub the version of samBada that corresponds to your OS. Unzips the folder and adds the path to the binary folder to the environmental path variable. This operation is only valid for the current R session. You must run change_path for every new R session. Alternatively, you can manually edit your "PATH" environmental variable permanently on your OS so that it entails the path to the binaries folder of sambada (this procedure different for every OS).

Usage

```
downloadSambada(directory = NULL)
```

Arguments

`directory`
: character The directory where sambada should be downloaded. If null, downloads in active directory.

Author(s)

Solange Gaillard

Examples

```
downloadSamada('D:/Sambada')
```

plotMap *Plotting of maps*

Description

Plots several kinds of maps (environmental variable distribution, population structure, marker absence or presence, autocorrelation of marker). Unlike plotResultInteractive, the resulting maps are non-interactive. The function can handle several marker/variables at once and create separate outputfiles.

Usage

```
plotMap(envFile, x, y, locationProj, popStrCol, gdsFile, markerName,
  mapType, varEnvName, SAMethod = NULL, SAThreshold = NULL,
  saveType = NULL, rasterName = NULL, simultaneous = FALSE)
```

Arguments

envFile	char The file containing the input environmental variable of sambada.
x	char The name of the column corresponding to the x-coordinate in the envFile. Can be set to null if unknown, in this case the maps will not be available
y	char The name of the column corresponding to the y-coordinate in the env file. Can be set to null if x is null.
locationProj	integer EPSG code of the geographical projection in the envFile
popStrCol	char The name or vector of name of column(s) in envFile describing population structure. If provided, additional layers on the map will be available reprenting population structure.
gdsFile	char The GDS file created in the preprocessing of sambada. If null, will try with envFile(without -env.csv) and .gds
markerName	name of the marker to be plotter if mapType is 'marker' or 'AS'. markerName can be found in preparedOutput$sambadaOutput[,"] where preparedOutput would be the result of the function prepareOutput
mapType	char A string or vector of string containing one or several of 'marker' (presence/absence of marker), 'env' (envrionmental variable distribution), 'popStr' (appartnance to a population in pie charts), 'AS' (autocorrelation of the marker). Note that the background of all maps, if found, will be the raster of the environmental variable. Thus the 'env' mapType is preferred when no raster is provided. For the 'AS' type, it is calculated on the fly for the markers provided and not the one possibly calculated by sambada.
varEnvName	char Name of the environmental variable. If a raster of the variable is located in your working directory, you can provide varEnvName even for mapType such as 'marker' or 'AS'. The function will scan the folder of your working directory for raster with the same name as varEnvName (and commonly used extension for raster) and put it as background.
SAMethod	char If mapType contains 'AS', then you must specify the method for setting the weights of neighbours. Can be one of 'knn' (k-nearest neighbours) or 'distance'
saveType	char One of NULL, 'png' or 'pdf'. To be implemented... If NULL is set, the maps will be shown in the R plotting window. Otherwise, it will be saved in the specified format in your working directory.

rasterName	char If a raster file with the environmental variable distribution exists with a different name than varEnvName, provide it here (including extension)
simultaneous	boolean If TRUE and mapType contains several kinds of maps, all maps corresponding to the same marker will be plotted on the same window. The resulting maps can be very small.
SAMThreshold	char If mapType contains 'AS' and SAMethod id 'knn' then the number of neighbours. If SAThreshold is 'distance' then the distance in map-unit (unless you use a spherical projection (latitude/longitude), in which case you should use km)

Value

None

Author(s)

Solange Gaillard

Examples

```
# Map of marker
plotMap('EnvFile.csv','longitude','latitude', locationProj=4326,  popStrCol='pop1',
     gdsFile='GDSFile.gds', markerName='ARS-BFGL-NGS-106879_AA',
     mapType=c('marker'), varEnvName='bio1')

# Maps of marker and population structure (two subplot)
plotMap('EnvFile.csv','longitude','latitude', locationProj=4326,  popStrCol='pop1',
     gdsFile='GDSFile.gds', markerName='ARS-BFGL-NGS-106879_AA',
     mapType=c('marker', 'popStr'), varEnvName='bio1', simultaneous=TRUE)
```

---

plotResultInteractive *Interactive plotting of results*

Description

Plots the manhattan plot for a given environmental variable. The plot is interactive and a map of the distribution of the marker can be retrieved as well as nearby genes listed in Ensembl.

Usage

```
plotResultInteractive(preparedOutput, varEnv, envFile, species = NULL,
  pass = NULL, x = NULL, y = NULL, valueName = "pvalueG",
  chromo = "all", gdsFile = NULL, IDCol = NULL, popstrCol = NULL)
```

Arguments

preparedOutput	char The prepared output list from prepare_output function
varEnv	char The name of the environmental variable one wish to study (as in the header of envFIle)
envFile	char The file containing the input environmental variable of sambada.
species	char The abbreviated latin name of the species without capitals nor punctuation (e.g. btaurus, chircus,...). Can be set to null if species not present in ensembl database

pass	integer Number of BP around a SNP in which to look for an annotation in Ensembl. Set to null if species is null
x	char The name of the column corresponding to the x-coordinate in the envFile. Can be set to null if unknown, in this case the maps will not be available
y	char The name of the column corresponding to the y-coordinate in the env file. Can be set to null if x is null.
valueName	char Name of the p- or q-value one wish to plot the manhattan on. This can be either pvalueG, pvalueW, qvalueG, qvalueW for G- or Waldscore respectively.
chromo	char/integer Name or vector of name of the chromosome to investigate. If all is chosen (default), all numerical chromosome will be mapped. If your sambada output is large (typically if you are working with more than 50K genomic file), you should probably map a subset of your dataset (e.g. chr=1)
gdsFile	char The GDS file created in the preprocessing of sambada. If null, will try with envFile(without -env.csv) and .gds
IDCol	char The name of the column in envFile corresponding to the ID of the individuall. If provided, hover on the output map will give the id of the animal
popStrCol	char The name or vector of name of column(s) in envFile describing population structure. If provided, additional layers on the map will be available reprenting population structure.

Value

None

Author(s)

Solange Gaillard

Examples

```
plotResultInteractive('myFile','chircus',1,'Longitude','Latitude','bio1',c('1','2'))
```

---

prepareEnv *Prepare environmental input*

Description

Writes a new environmental file that sambada can work with after having removed too correlated variables. Also calculates population structure from a PCA in SNPRelate and add it at the end of the environmental file

Usage

```
prepareEnv(envFile, maxCorr, idName, separator = ",", genoFile = NULL,
  numPc = NULL, mafThresh = NULL, missingnessThresh = NULL,
  ldThresh = NULL, numPop = -1, clustMethod = "kmeans",
  includeCol = NULL, excludeCol = NULL, popStrCol = NULL, x, y,
  locationProj, interactiveChecks = FALSE, verbose = TRUE)
```

Arguments

envFile	char Name of the input environmental file (must be in active directory). Can be .csv or .shp
maxCorr	double A number between 0 and 1 specifying the maximum allowable correlation coefficient between environmental files. If above, one of the variables will be deleted
idName	char Name of the id in the environmental file matching the one of genoFile
separator	char If envFile is .csv, the separator character. If file created with create_env, separator is ' '
genoFile	char (optional) Name of the input genomic file (must be in active directory). If not null, population variable will be calculated from a PCA relying on the SNPRelate package. Can be .gds, .ped, .bed, .vcf. If different from .gds, a gds file (SNPrelate specific format) will be created
numPc	double If above 1, number of principal components to analyze. If between 0 and 1, automatic detection of number of PC. If 0, PCA and population structure will not be computed: in that case, the genoFile will only be used to make the sample order in the envFile match the one of the envFile (necessary for sambada's computation). Set it to null if genoFile is null
mafThresh	double A number between 0 and 1 specifying the Major Allele Frequency (MAF) filtering when computing PCA (if null no filtering on MAF will be computed)
missingnessThresh	double A number between 0 and 1 specifying the missing rate filtering when computing PCS(if null no filtering on missing rate will be computed)
ldThresh	double A number between 0 and 1 specifying the linkage desiquilibrium (LD) rate filtering before computing the PCA (if null no filtereing on LD will be computed)
numPop	integer If not null, clustering based on numPc first PC will be computed to devide into numPop populations. If -1 automatic detection of number of cluster (elbow method if clustMethod='kmeans', maximise branch length if clustMethod='hclust'). If null, no clustering will be computed: if genoFile is set, principal component scores will be included as population information in the final file.
clustMethod	char One of 'kmeans' or 'hclust' for K-means and hierarchical clustering respectively. Default 'kmeans'
includeCol	character vector Columns in the envrionmental file to be considered as variables. If none specified, all numeric variables will be considered as env var except for the id
excludeCol	character vector Columns in the envrionmental file to exclude in the output (non-variable column). If none specified, all numeric variables will be considered as env var except for the id
popStrCol	character vector Columns in the envrionmental file describing population structure (ran elsewhere). Those columns won't be excluded when correlated with environmental files
x	character Name of the column corresponding to the x coordinate (or longitude if sperical coordinate). If not null, x column won't be removed even if correlated with other variable. This parameter is also used to display the map of the population structure.

y	character Name of the column corresponding to the y coordinate (or latitude if sperical coordinate). If not null, y column won't be removed even if correlated with other variable. This parameter is also used to display the map of the population structure.
locationProj	integer EPSG code of the projection of x-y coordinate
interactiveChecks	logical If TRUE, plots will show up showing number of populations chosen, and correlation between variables and the user can interactively change the chosen threshold for maxCorr and numPop (optional, default value=FALSE)
verbose	boolean If true show information about progress of the process

Value

None

Author(s)

Solange Gaillard, Oliver Selmoni

Examples

```
#Calculating PCA-based population structure
prepareEnv('myFile-env.csv',0.8,'Nom',' ','myFile.gds', numPc=0.2,
     mafThresh=0.05, missingnessThresh=0.1, ldThresh=0.2, numPop=NULL,
     x='Longitude', y='Latitude', locationProj=4326, interactiveChecks = TRUE)

#Calculating structure membership coefficient based on kmeans clustering
prepareEnv('myFile-env.csv',0.8,'Nom',' ','myFile.gds', numPc=0.2,
     mafThresh=0.05, missingnessThresh=0.1, ldThresh=0.2, numPop=NULL,
     x='Longitude', y='Latitude', locationProj=4326, interactiveChecks = TRUE)

#Without calculating population structure.
prepareEnv('myFile-env.csv',0.8,'Nom',' ', x='Longitude',y='Latitude',
     locationProj=4326, interactiveChecks = TRUE)
```

---

prepareGeno *Prepare genomic input*

Description

Writes a new genomic file that sambada can work with after having applied the selected genomic filtering options. The output file has the same name as the input file but with a .csv extension

Usage

```
prepareGeno(fileName, mafThresh = NULL, missingnessThresh = NULL,
  ldThresh = NULL, mgfThresh = NULL, directory = NULL,
  interactiveChecks = FALSE, verbose = FALSE)
```

Arguments

`fileName`	char Name of the input file (must be in active directory). Can be .gds, .ped, .bed, .vcf. If different from .gds, a gds file (SNPrelate specific format) will be created unless no filtering options are chosen
`mafThresh`	double A number between 0 and 1 specifying the Major Allele Frequency (MAF) filtering (if null no filtering on MAF will be computed)
`missingnessThresh`	double A number between 0 and 1 specifying the missing rate filtering (if null no filtering on missing rate will be computed)
`ldThresh`	double A number between 0 and 1 specifying the linkage desiquilibrium (LD) rate filtering (if null no filtering on LD will be computed)
`mgfThresh`	double A number between 0 and 1 specifying the Major Genotype Frequency (MGF) rate filtering (if null no filtering on MGF will be computed). NB: sambada computations relie on gentoypes
`directory`	char The directory where binaries of sambada are saved. This parameter is not necessary if directoy path is permanently stored in the PATH environmental variable or if a function invoking samabada executable (prepareGeno or sambada_parallel) has been already run in the R active session.
`interactiveChecks`	logical If TRUE, plots will show up showing distribution of allele frequency etc... and the user can interactively change the chosen threshold for mafThresh, missingnessThresh, mgfThresh (optional, default value=FALSE)

Value

None

Author(s)

Solange Gaillard, Oliver Selmoni

Examples

```
#With ped input file
prepareGeno('myPlinkFile.ped',mafThresh=0.05, missingnessThresh=0.05,
     mgfThresh=0.8,interactiveChecks=TRUE)

#With gds input file
prepareGeno('myGDSFile.gds',mafThresh=0.05, missingnessThresh=0.05,
     mgfThresh=0.8,interactiveChecks=FALSE)
```

`prepareOutput` *Prepare output (usefull for all postprocessing analysis)*

Description

Read sambada's output and prepare it by retrieving the snp position and chromosome (usefull for plotting manhattan)

Usage

```
prepareOutput(sambadaname, dimMax, gdsFile = NULL, popStr = FALSE,
  nrows = NULL, interactiveChecks = TRUE)
```

Arguments

sambadaname	char The name of the genofile without extension name given to sambada (or outputfile of sambada without the ending -Out-Dim.csv)
dimMax	integer The maximum number of dimension given in sambada
gdsFile	char Name of the gds file associated with sambada's input file. If null, will try with sambadaname.gds
popStr	logical Indicates whether sambada was run using the POPSTRVAR parameter (i.e. population structure was taken into account). Default false
nrows	integer Specifies the number of line to read from the input file. Useful if saveType END ALL was used in sambada and that the number of models run is large so that the reading and processing is too slow. The saveType END parameter ensures that most significant models are located at the top of the file.
interactiveChecks	logical

Value

a list containing a) $sambadaOutput a matrix containing the output from sambada with 3 additional column: corresponding snp, chromosome and position of the marker b) chrSNPNum The total number of SNPs in each chromosme c) $chrMaxPos The highest position found in each chromosome

Examples

```
prepare_output('myFile',1)
```

---

sambadaParallel *Run sambada on parallel cores*

Description

Read sambadas input file to retrieve necessary information (num indiv etc...), split the dataset using SamBada's Supervision tool, run sambada on the splitted dataset and merge all using Supervision. See sambada's documentation for more information.

Usage

```
sambadaParallel(genoFile, envFile, idGeno, idEnv, dimMax = 1,
  cores = NULL, wordDelim = " ", saveType = "END BEST 0.05",
  populationVar = NULL, spatial = NULL, autoCorr = NULL,
  shapeFile = NULL, outputFile = NULL, colSupEnv = NULL,
  colSupMark = NULL, subsetVarEnv = NULL, subsetVarMark = NULL,
  headers = TRUE, directory = NULL, keepAllFiles = FALSE)
```

Arguments

genoFile	The name of the file in the current directory of genetic information, complient with samBada's format (use prepareGeno to transform it)
envFile	The name of the file in the current directory of environmental information (use link{createEnv} to create it and link{prepareEnv} to reduce the correlated dataset and check order)
idGeno	Name of the column in the genoFile corresponding to the id of the animals
idEnv	Name of the column in the envFile corresponding to the id of the animals
dimMax	Maximum number of environmental variables included in the logistic models. Use 1 for univariate models, 2 for univariate and bivariates models
cores	Number of cores to use. If NULL, the #cores-1 will be used where #cores corresponds to all cores available on your computer.
wordDelim	char Word delimiter of input file(s). Default ' ',
saveType	composed of three words 1) one of 'end' or 'real' to save the result during the analysis or at the end (allows sorting of result) 2) one of 'all' or 'best' to save all models or only significant models 3) If 'best' specify the threshold of significance (before applying Bonferroni's correction). Default 'END BEST 0.05',
populationVar	one of 'first' or 'last'. This option indicates whether any explanatory variables represent the population structure. If present, the said population variables must be gathered in the input file, either on the left or on the right side of the group of environmental variables. Default null.
spatial	composed of 5 words 1) Column name (or number) for longitude 2) Column name (or number) for latitude 3) one of 'sperical' or 'cartesian': to indicate the type of coordinate 4) one of 'distance', 'gaussian', bisquare' or 'nearest': type of weighting scheme (see sambadoc) 4) Number bandwith of weighting function Input type is (double). Units are in [m] for spherical coordinates; for cartesian coordinates, units match those of the samples' positions. Case nearest: Input type is (int)
autoCorr	composed of 3 words. 1) one of global, local or both: to indicate the type of spatial autocorrelation to compute. 2) one of env, mark or both: to inidicate the variables on which to compute the analysis 3) integer The number of permutation to compute the pseudo p-value. Ex 'global both 999'
shapeFile	one of yes or no. With this option, the LISA are saved as a shapefile (in addition to the usual output)
outputFile	char Base name(s) for the results file(s). Default: construction from input file with suffixes (e.g. -Out-)
colSupEnv	char or vector of char Name(s) of the column(s) in the environmental data to be excluded from the analysis. Default NULL
colSupMark	char or vector of char Name(s) of the column(s) in the molecular data to be excluded from the analysis. Default NULL
subsetVarEnv	char or vector of char Name(s) of the column(s) in the environmental data to be included in the analysis while the other columns are set as inactive. Default NULL
subsetVarMark	char or vector of char Name(s) of the column(s) in the molecular data to be included in the analysis while the other columns are set as inactive. Default NULL
headers	logical Presence or absence of variable names in input files Default TRUE

directory	char The directory where binaries of sambada are saved. This parameter is not necessary if directoy path is permanently stored in the PATH environmental variable or if a function invoking samabada executable (prepareGeno or sambadaParallel) has been already run in the R active session.
keepAllFiles	logical If TRUE, all parameter files and splitted genofile and log-files are not removed. Default FALSE
All	additional parameters in samBada: see documentation. In case you have to specify several words, you can either specify them in one string and separate them with a space or add a vector string

Author(s)

Solange Gaillard, Sylvie Stucki

Examples

```
#With all default parameter
sambadaParallel('File-molecular.csv','File-.env.csv','ID_indiv','sampleID')

#With population structure
sambadaParallel('File-molecular.csv','File-.env.csv','ID_indiv','sampleID'
     dimMax=2, saveType='END ALL', populationVar='pop1')
```

Index