Package 'BinaryDosage'

January 13, 2020

Title Creates, Merges, and Reads Binary Dosage Files

Version 1.0.0

Description Tools to create binary dosage from either VCF or GEN files, merge binary dosage files, and read binary dosage files.

License GPL-3

Encoding UTF-8

LazyData true

Suggests knitr, rmarkdown, testthat (>= 2.1.0), covr

VignetteBuilder knitr

RoxygenNote 7.0.2

LinkingTo Rcpp

Imports Rcpp, digest, prodlim

NeedsCompilation yes

Author John Morrison [aut, cre],
NIEHS [fnd] (P01 CA196559),
NIEHS [fnd] (R01 CA201407),
NIEHS [fnd] (P30 ES007048),
NIEHS [fnd] (P01 HL115606)

Maintainer John Morrison <jmorr@usc.edu>

Repository CRAN

Date/Publication 2020-01-13 16:20:08 UTC

R topics documented:

---

bdapply *Apply a function to each SNP in a binary dosage file*

Description

A routine that reads in the SNP data serially from a binary dosage file and applies a user specified function to the data.

Usage

```
bdapply(bdinfo, func, ...)
```

Arguments

bdinfo	List with information about the binary dosage file returned from getbdinfo
func	A user supplied function to apply to the data for each snp. The function must be provide with the following parameters, dosage, p0, p1, and p2, where dosage is the dosage values for each subject and p0, p1, and p2 are the probabilities that a subject has zero, one, and two copies of the alternate allele, respectively.
...	Additional parameters needed by the user supplied function

Value

A list with length equal to the number of SNPs in the binary dosage file. Each element of the list is the value returned by the user supplied function

See Also

Other Iterating functions: genapply(), vcfapply()

Examples

```
# Get information about a binary dosage file

vcf1abdfile <- system.file("extdata", "vcf1a.bdose", package = "BinaryDosage")
bdinfo <- getbdinfo(bdfiles = vcf1abdfile)

# Apply the getaaf, get alternate allele frequency, function
# to all the SNPs in the binary dosage file
```

```
aaf <- bdapply(bdinfo = bdinfo,
               func = BinaryDosage:::getaaf)
```

---

bdmerge *Merge binary dosage files together*

Description

Routine to merge binary dosage files together. The files don't have to be in the same format. They will be merged into a file with the format specified. Information about the SNPs, aaf, maf, avgcall, rsq, can be maintained for each file, or recalculated for the merged set.

Usage

```
bdmerge(
  mergefiles,
  format = 4,
  subformat = 0L,
  bdfiles,
  famfiles = character(),
  mapfiles = character(),
  onegroup = TRUE,
  bdoptions = character(),
  snpjoin = "inner"
)
```

Arguments

mergefiles	Vector of file names for the merged binary files. The first is the binary dosage data containing the dosages and genetic probabilities. The second file name is the family information file. The third file name is the SNP information file. The family and SNP information files are not used if the binary dosage file is in format 4. For this format the family and SNP information are in the file with the dosages and genetic probabilities.
format	The format of the output binary dosage file. Allowed values are 1, 2, 3, and 4. The default value is 4. Using the default value is recommended.
subformat	The subformat of the format of the output binary dosage file. A value of 1 or 3 indicates that only the dosage value is saved. A value of 2 or 4 indicates the dosage and genetic probabilities will be output. Values of 3 or 4 are only allowed with formats 3 and 4. If a value of zero if provided, and genetic probabilities are in the vcf file, subformat 2 will be used for formats 1 and 2, and subformat 4 will be used for formats 3 and 4. If the vcf file does not contain genetic probabilities, subformat 1 will be used for formats 1 and 2, and subformat 3 will be used for formats 3 and 4. The default value is 0.
bdfiles	Vector of binary dosage file names to be merged.

famfiles	Vector of family file names that correspond to the names in bdfiles. If the binary dosage files are all in format 4, this may be an empty character array. Default value is character().
mapfiles	Vector of map file names that correspond to the names in bdfiles. If the binary dosage files are all in format 4, this may be an empty character array. Default value is character().
onegroup	Indicator to combine all the samples in one group. If this is FALSE, the groups in each binary dosage file are maintained and any binary dosage file with one group is made into its own group. Default value is TRUE.
bdoptions	Options indicating what information to calculate and store for each SNP. These can be aaf, maf, and rsq. This option is only available if format is equal to 4 and onegroup is TRUE. Default value is character().
snpjoin	Character value that can be either "inner" or "outer". This indicates whether to do an inner or outer join of the SNPs in each binary dosage file. Default value is "inner".

Value

None

Examples

```
bdvcf1afile <- system.file("extdata", "vcf1a.bdose", package = "BinaryDosage")
bdvcf1bfile <- system.file("extdata", "vcf1b.bdose", package = "BinaryDosage")
mergefiles <- tempfile()

BinaryDosage:::bdmerge(mergefiles = mergefiles,
                       bdfiles = c(bdvcf1afile, bdvcf1bfile),
                       bdoptions = "maf")
bdinfo <- getbdinfo(mergefiles)
```

---

genapply *Apply a function to each SNP in a gen, impute2, file*

Description

A routine that reads in the SNP data serially from a gen file and applies a user specified function to the data.

Usage

```
genapply(geninfo, func, ...)
```

Arguments

geninfo	List with information about the gen, impute2, file returned from getgeninfo
func	A user supplied function to apply to the data for each snp. The function must be provide with the following parameters, dosage, p0, p1, and p2, where dosage is the dosage values for each subject and p0, p1, and p2 are the probabilities that a subject has zero, one, and two copies of the alternate allele, respectively.
...	Additional parameters needed by the user supplied function

Value

A list with length equal to the number of SNPs in the vcf file. Each element of the list is the value returned by the user supplied function

See Also

Other Iterating functions: bdapply(), vcfapply()

Examples

```
# Get information about a gen, impute2, file

gen1afile <- system.file("extdata", "set1a.imp", package = "BinaryDosage")
geninfo <- getgeninfo(genfiles = gen1afile,
                      snpcolumns = c(1L, 3L, 2L, 4L, 5L),
                      header = TRUE)


aaf <- genapply(geninfo = geninfo,
                func = BinaryDosage:::getaaf)
```

---

gentobd *Convert a gen file to a binary dosage file*

Description

Routine to read information from a gen file and create a binary dosage file. Note: This routine can take a long time to run if the gen file is large.

Usage

```
gentobd(
  genfiles,
  snpcolumns = 1L:5L,
  startcolumn = 6L,
  impformat = 3L,
  chromosome = character(),
  header = c(FALSE, TRUE),
```

```
  gz = FALSE,
  sep = "\t",
  bdfiles,
  format = 4L,
  subformat = 0L,
  snpidformat = 0L,
  bdoptions = character(0)
)
```

Arguments

genfiles	A vector of file names. The first is the name of the gen file. The second is name of the sample file that contains the subject information.
snpcolumns	Column numbers containing chromosome, snpid, location, reference allele, alternate allele, respectively. This must be an integer vector. All values must be positive except for the chromosome. The value for the chromosome may be -1 or -0. -1 indicates that the chromosome value is passed to the routine using the chromosome parameter. 0 indicates that the chromosome value is in the snpid and that the snpid has the format chromosome:other_data. Default value is c(1L, 2L, 3L, 4L, 5L).
startcolumn	Column number of first column with genetic probabilities or dosages. Must be an integer value. Default value is 6L.
impformat	Number of genetic data values per subject. 1 indicates dosage only, 2 indicates P(g=0) and P(g=1) only, 3 indicates P(g=0), P(g=1), and P(g=2). Default value is 3L.
chromosome	Chromosome value to use if the first value of the snpcolumns is equal to 0. Default value is character().
header	Indicators if the gen and sample files have headers. If the gen file does not have a header. A sample file must be included. Default value is c(FALSE, TRUE).
gz	Indicator if file is compressed using gzip. Default value is FALSE.
sep	Separator used in the gen file. Default value is "\t"
bdfiles	Vector of names of the output files. The binary dosage file name is first. The family and map files follow. For format 4, no family and map file names are needed.
format	The format of the output binary dosage file. Allowed values are 1, 2, 3, and 4. The default value is 4. Using the default value is recommended.
subformat	The subformat of the format of the output binary dosage file. A value of 1 or 3 indicates that only the dosage value is saved. A value of 2 or 4 indicates the dosage and genetic probabilities will be output. Values of 3 or 4 are only allowed with formats 3 and 4. If a value of zero if provided, and genetic probabilities are in the vcf file, subformat 2 will be used for formats 1 and 2, and subformat 4 will be used for formats 3 and 4. If the vcf file does not contain genetic probabilities, subformat 1 will be used for formats 1 and 2, and subformat 3 will be used for formats 3 and 4. The default value is 0.

snpidformat	The format that the SNP ID will be saved as. -1 - SNP ID not written. 0 - same as in the VCF file. 1 - chromosome:location. 2 - chromosome:location:reference_allele:alternate_allele. If snpidformat is 1 and the VCF file uses format 2, an error is generated. Default value is 0.
bdoptions	Character array containing any of the following value, "aaf", "maf", "rsq". The presence of any of these values indicates that the specified values should be calculates and stored in the binary dosage file. These values only apply to format 4.

Value

None

Examples

```
# Find the gen file names
gen3afile <- system.file("extdata", "set3a.imp", package = "BinaryDosage")
gen3asample <- system.file("extdata", "set3a.sample", package = "BinaryDosage")
# Get temporary output file name
bdfiles <- tempfile()
# Convert the file
gentobd(genfiles = c(gen3afile, gen3asample),
        snpcolumns = c(0L, 2L:5L),
        bdfiles = bdfiles)
# Verify the file was written correctly
bdinfo <- getbdinfo(bdfiles = bdfiles)
```

---

getaaf *Calculate alternate allele frequency*

Description

Routine to calculate the alternate allele frequency given the dosages. Missing values for dosage ignored. This function is used internally and is exported for use in examples.

Usage

```
getaaf(dosage, p0, p1, p2)
```

Arguments

dosage	Dosage values
p0	Pr(g=0) - unused
p1	Pr(g=1) - unused
p2	Pr(g=2) - unused

Value

Alternate allele frequency

Examples

```
# Get information about binary dosage file
bdfile <- system.file("extdata", "vcf1a.bdose", package = "BinaryDosage")
bdinfo <- getbdinfo(bdfiles = bdfile)
snp1 <- getsnp(bdinfo = bdinfo, 1)
aaf <- getaaf(snp1$dosage)
```

`getbdinfo` *Get information about a binary dosage file*

Description

Routine to return information about a binary dosage file. This information is used by other routines to allow for quicker extraction of values from the file.

Usage

```
getbdinfo(bdfiles)
```

Arguments

`bdfiles`	Vector of file names. The first is the binary dosage data containing the dosages and genetic probabilities. The second file name is the family information file. The third file name is the SNP information file. The family and SNP information files are not used if the binary dosage file is in format 4. For this format the family and SNP information are in the file with the dosages and genetic probabilities.

Value

List with information about the binary dosage file. This includes family and subject IDs along with a list of the SNPs in the file. Other information needed to read the file is also included.

Examples

```
vcf1abdfile <- system.file("extdata", "vcf1a.bdose", package = "BinaryDosage")
bdinfo <- getbdinfo(bdfiles = vcf1abdfile)
```

getgeninfo *Get information about a gen, impute2, file*

Description

Routine to return information about a gen file. This information is used by other routines to allow for quicker extraction of values from the file.

Usage

```
getgeninfo(
  genfiles,
  snpcolumns = 1L:5L,
  startcolumn = 6L,
  impformat = 3L,
  chromosome = character(),
  header = c(FALSE, TRUE),
  gz = FALSE,
  index = TRUE,
  snpidformat = 0L,
  sep = c("\t", "\t")
)
```

Arguments

genfiles	A vector of file names. The first is the name of the gen file. The second is name of the sample file that contains the subject information.
snpcolumns	Column numbers containing chromosome, snpid, location, reference allele, alternate allele, respectively. This must be an integer vector. All values must be positive except for the chromosome. The value for the chromosome may be -1 or -0. -1 indicates that the chromosome value is passed to the routine using the chromosome parameter. 0 indicates that the chromosome value is in the snpid and that the snpid has the format chromosome:other_data. Default value is c(1L, 2L, 3L, 4L, 5L).
startcolumn	Column number of first column with genetic probabilities or dosages. Must be an integer value. Default value is 6L.
impformat	Number of genetic data values per subject. 1 indicates dosage only, 2 indicates P(g=0) and P(g=1) only, 3 indicates P(g=0), P(g=1), and P(g=2). Default value is 3L.
chromosome	Chromosome value to use if the first value of the snpcolumns is equal to 0. Default value is character().
header	Indicators if the gen and sample files have headers. If the gen file does not have a header. A sample file must be included. Default value is c(FALSE, TRUE).
gz	Indicator if file is compressed using gzip. Default value is FALSE.

index	Indicator if file should be indexed. This allows for faster reading of the file. Indexing a gzipped file is not supported. Default value is TRUE.
snpidformat	Format to change the snpid to. 0 indicates to use the snpid format in the file. 1 indicates to change the snpid into chromosome:location, 2 indicates to change the snpid into chromosome:location:referenceallele:alternateallele, 3 indicates to change the snpid into chromosome:location_referenceallele_alternateallele, Default value is 0.
sep	Separators used in the gen file and sample files, respectively. If only value is provided it is used for both files. Default value is c("\t", "\t")

Value

List with information about the gen file. This includes family and subject IDs along with a list of the SNPs in the file. Other information needed to read the file is also included.

Examples

```
# Get file names of th gen and sample file
gen3afile <- system.file("extdata", "set3a.imp", package = "BinaryDosage")
gen3ainfo <- system.file("extdata", "set3a.sample", package = "BinaryDosage")

# Get the information about the gen file
geninfo <- getgeninfo(genfiles = c(gen3afile, gen3ainfo),
                      snpcolumns = c(0L, 2L:5L))
```

---

getmaf *Calculate minor allele frequency*

Description

Routine to calculate the minor allele frequency given the dosages. Missing values for dosage ignored. This function is used internally and is exported for use in examples. Note: The minor allele in one data set may be different from another data set. This can make comparing minor allele frequencies between data sets nonsensical.

Usage

```
getmaf(dosage, p0, p1, p2)
```

Arguments

dosage	Dosage values
p0	Pr(g=0) - unused
p1	Pr(g=1) - unused
p2	Pr(g=2) - unused

Value

Minor allele frequency

Examples

```
# Get information about binary dosage file
bdfile <- system.file("extdata", "vcf1a.bdose", package = "BinaryDosage")
bdinfo <- getbdinfo(bdfiles = bdfile)
snp1 <- getsnp(bdinfo = bdinfo, 1)
maf <- getmaf(snp1$dosage)
```

---

getrsq *Calculate imputation r squared*

Description

Routine to calculate the imputation r squared given the dosages and Pr(g=2). This is an estimate for the imputation r squared returned from minimac and impute2. The r squared values are calculated slightly differently between the programs. This estimate is based on the method used by minimac. It does well for minor allele frequencies above 5%. This function is used internally and is exported for use in examples.

Usage

```
getrsq(dosage, p0, p1, p2)
```

Arguments

dosage	Dosage values
p0	Pr(g=0) - unused
p1	Pr(g=1) - unused
p2	Pr(g=2)

Value

Imputation r squared

Examples

```
# Get information about binary dosage file
bdfile <- system.file("extdata", "vcf1a.bdose", package = "BinaryDosage")
bdinfo <- getbdinfo(bdfiles = bdfile)
snp1 <- getsnp(bdinfo = bdinfo, 1, dosageonly = FALSE)
rsq <- BinaryDosage:::getrsq(snp1$dosage, p2 = snp1$p2)
```

getsnp *Read SNP data from a binary dosage file*

Description

Routine to read the dosage and genetic probabilities about a SNP from a binary dosage file

Usage

```
getsnp(bdinfo, snp, dosageonly = TRUE)
```

Arguments

bdinfo	Information about a binary dosage file return from getbdinfo
snp	The SNP to read the information about. This may be the SNP ID or the index of the SNP in the snps dataset in the bdinfo list
dosageonly	Indicator to return the dosages only or the dosages allowing with the genetic probabilities. Default value is TRUE

Value

A list with either the dosages or the dosages and the genetic probabilities.

Examples

```
# Get the information about the file
vcf1abdfile <- system.file("extdata", "vcf1a.bdose", package = "BinaryDosage")
bdinfo <- getbdinfo(bdfiles = vcf1abdfile)

# Read the first SNP
getsnp(bdinfo, 1, FALSE)
```

getvcfinfo *Get information about a vcf file*

Description

Routine to return information about a vcf file. This information is used by other routines to allow for quicker extraction of values from the file.

Usage

```
getvcfinfo(vcffiles, gz = FALSE, index = TRUE, snpidformat = 0L)
```

Arguments

vcffiles	A vector of file names. The first is the name of the vcf file. The second is name of the file that contains information about the imputation of the SNPs. This file is produced by minimac 3 and 4.
gz	Indicator if VCF file is compressed using gzip. Default value is FALSE.
index	Indicator if file should be indexed. This allows for faster reading of the file. Indexing a gzipped file is not supported. Default value is TRUE.
snpidformat	The format that the SNP ID will be saved as. 0 - same as in the VCF file 1 - chromosome:location 2 - chromosome:location:referenceallele:alternateallele If snpidformat is 1 and the VCF file uses format 2, an error is generated. Default value is 0.

Value

List containing information about the VCF file to include file name, subject IDs, and information about the SNPs. Indices for faster reading will be included if index is set to TRUE

Examples

```
# Get file names of th vcf and infromation file
vcf1afile <- system.file("extdata", "set1a.vcf", package = "BinaryDosage")
vcf1ainfo <- system.file("extdata", "set1a.info", package = "BinaryDosage")

# Get the information about the vcf file
vcf1ainfo <- getvcfinfo(vcffiles = c(vcf1afile, vcf1ainfo))
```

---

vcfapply *Apply a function to each SNP in a vcf file*

Description

A routine that reads in the SNP data serially from a vcf file and applies a user specified function to the data.

Usage

```
vcfapply(vcfinfo, func, ...)
```

Arguments

vcfinfo	List with information about the vcf file returned from getvcfinfo
func	A user supplied function to apply to the data for each snp. The function must be provide with the following parameters, dosage, p0, p1, and p2, where dosage is the dosage values for each subject and p0, p1, and p2 are the probabilities that a subject has zero, one, and two copies of the alternate allele, respectively.
...	Additional parameters needed by the user supplied function

Value

A list with length equal to the number of SNPs in the vcf file. Each element of the list is the value returned by the user supplied function

See Also

Other Iterating functions: `bdapply()`, `genapply()`

Examples

```
# Get information about a vcf file

vcf1afile <- system.file("extdata", "set1a.vcf", package = "BinaryDosage")
vcfinfo <- getvcfinfo(vcffiles = vcf1afile)

# Apply the getaaf, get alternate allele frequency, function
# to all the SNPs in the vcf file

aaf <- vcfapply(vcfinfo = vcfinfo,
                func = BinaryDosage:::getaaf)
```

`vcftobd` *Convert a VCF file to a binary dosage file*

Description

Routine to read information from a VCF file and create a binary dosage file. The function is designed to use files return from the Michigan Imputation Server but will run on other VCF files if they contain dosage and genetic probabilities. Note: This routine can take a long time to run if the VCF file is large.

Usage

```
vcftobd(
  vcffiles,
  gz = FALSE,
  bdfiles,
  format = 4L,
  subformat = 0L,
  snpidformat = 0,
  bdoptions = character(0)
)
```

Arguments

vcffiles	A vector of file names. The first is the name of the vcf file. The second is name of the file that contains information about the imputation of the SNPs. This file is produced by minimac 3 and 4.
gz	Indicator if VCF file is compressed using gzip. Default value is FALSE.
bdfiles	Vector of names of the output files. The binary dosage file name is first. The family and map files follow. For format 4, no family and map file names are needed.
format	The format of the output binary dosage file. Allowed values are 1, 2, 3, and 4. The default value is 4. Using the default value is recommended.
subformat	The subformat of the format of the output binary dosage file. A value of 1 or 3 indicates that only the dosage value is saved. A value of 2 or 4 indicates the dosage and genetic probabilities will be output. Values of 3 or 4 are only allowed with formats 3 and 4. If a value of zero if provided, and genetic probabilities are in the vcf file, subformat 2 will be used for formats 1 and 2, and subformat 4 will be used for formats 3 and 4. If the vcf file does not contain genetic probabilities, subformat 1 will be used for formats 1 and 2, and subformat 3 will be used for formats 3 and 4. The default value is 0.
snpidformat	The format that the SNP ID will be saved as. -1 SNP ID not written 0 - same as in the VCF file 1 - chromosome:location 2 - chromosome:location:reference_allele:alternate_allele If snpidformat is 1 and the VCF file uses format 2, an error is generated. Default value is 0.
bdoptions	Character array containing any of the following value, "aaf", "maf", "rsq". The presence of any of these values indicates that the specified values should be calculates and stored in the binary dosage file. These values only apply to format 4.

Value

None

Examples

```
# Find the vcf file names
vcf1afile <- system.file("extdata", "set1a.vcf", package = "BinaryDosage")
vcf1ainfo <- system.file("extdata", "set1a.info", package = "BinaryDosage")
bdfiles <- tempfile()
# Convert the file
vcftobd(vcffiles = c(vcf1afile, vcf1ainfo), bdfiles = bdfiles)
# Verify the file was written correctly
bdinfo <- getbdinfo(bdfiles)
```

Index