Package: BinaryDosage (via r-universe)

October 28, 2026

Title Creates, Merges, and Reads Binary Dosage Files

Version 2.0.0

Description Tools to create binary dosage files from either VCF or GEN files, merge binary dosage files, and read binary dosage files. Binary dosage files tend to have quicker read times than VCF and GEN formats. There is a small increase in size compared to compressed VCF and GEN files.

License GPL-3

Encoding UTF-8

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

VignetteBuilder knitr

RoxygenNote 7.3.3

LinkingTo Rcpp

Imports Rcpp, digest, prodlim

NeedsCompilation yes

Author John Morrison [aut, cre], NCI [fnd] (CA196559), NCI [fnd] (CA201407), NIEHS [fnd] (ES007048), NHLBI [fnd] (HL115606)

Maintainer John Morrison <jmorr@usc.edu>

Repository https://cran.r-universe.dev

Date/Publication 2026-04-29 18:52:28 UTC

RemoteUrl https://github.com/cran/BinaryDosage

RemoteRef HEAD

RemoteSha 40f112c2b3a31b7d67a5ed4a8d9fef4dc8018581

Contents

bdapply 2
bdmerge 3
closebd5con 5
genapply 5

gentobd 6
getaaf 8
getbd5snp 9
getbd5snp_buf 9
getbd5snp_con 10
getbdinfo 11
getgeninfo 11
getmaf 13
getrsq 14
getsnp 14
getvcfinfo 15
mergebd 16
openbd5con 17
subsetbd 18
updatebd 19
vcfapply 20
vcftobd 21
vcftobdlegacy 22

Index **24**

`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)
```

closebd5con *Close a persistent Format 5 connection*

Description

Explicitly closes the connection opened by `openbd5con`. Calling this is optional — the finalizer will close it on garbage collection or R exit — but explicit close is preferred to release the file handle promptly.

Usage

```
closebd5con(bd5con)
```

Arguments

`bd5con`	Object returned by `openbd5con`.

Value

`NULL` invisibly.

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 - chr:pos. 2 - chr:pos:ref:alt. 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)
```

`getbd5snp` *Read a SNP from a Format 5 binary dosage file*

Description

Seeks to the SNP's compressed block in the .bdose file, decompresses it, and returns the dosage and genotype probabilities for all samples.

Usage

```
getbd5snp(bd5info, snp)
```

Arguments

`bd5info`	Object returned by `getbdinfo`.
`snp`	The SNP to retrieve: either a 1-based integer index or a character SNP ID matching a value in `bd5info$snps$snpid`.

Value

A list with four numeric vectors, each of length n_samples:

dosage DS values in [0, 2]; NA = missing.

p0 P(g=0) values in [0, 1]; NA = missing.

p1 P(g=1) values in [0, 1]; NA = missing.

p2 P(g=2) values in [0, 1]; NA = missing.

`getbd5snp_buf` *Read a Format 5 SNP into pre-allocated vectors (buffered variant)*

Description

Like `getbd5snp` but writes results into caller-supplied vectors instead of allocating new ones. Intended for tight loops where thousands of SNPs are read sequentially; pre-allocating the output vectors once avoids repeated memory allocation.

Usage

```
getbd5snp_buf(bd5info, snp, dosage, p0, p1, p2)
```

Arguments

bd5info	Object returned by getbdinfo.
snp	1-based integer index or character SNP ID.
dosage	Pre-allocated numeric(n_samples) vector.
p0	Pre-allocated numeric(n_samples) vector.
p1	Pre-allocated numeric(n_samples) vector.
p2	Pre-allocated numeric(n_samples) vector.

Details

The four output vectors **must not** have more than one R binding at the call site (no extra variables pointing to the same object); R's copy-on-modify semantics would otherwise prevent in-place update.

Value

NULL invisibly. dosage, p0, p1, and p2 are updated in place.

---

getbd5snp_con *Read a Format 5 SNP using a persistent open connection*

Description

Like getbd5snp_buf but reuses an already-open file connection instead of opening and closing it on every call. Use openbd5con before the loop and closebd5con (or let the finalizer handle it) after.

Usage

```
getbd5snp_con(bd5info, snp, dosage, p0, p1, p2, bd5con)
```

Arguments

bd5info	Object returned by getbdinfo.
snp	1-based integer index or character SNP ID.
dosage	Pre-allocated numeric(n_samples) vector.
p0	Pre-allocated numeric(n_samples) vector.
p1	Pre-allocated numeric(n_samples) vector.
p2	Pre-allocated numeric(n_samples) vector.
bd5con	Object returned by openbd5con.

Value

NULL invisibly. dosage, p0, p1, and p2 are updated in place.

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. For Format 5 files, a single .bdose file name; the companion .bdi metadata file is read automatically from `paste0(bdfiles[1], ".bdi")`. For format 4, a single file name. For formats 1, 2, and 3, a vector of three file names: the binary dosage file, the family information file, and the SNP information file.

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 chr:pos, 2 indicates to change the snpid into chr:pos:ref:alt, 3 indicates to change the snpid into chr:pos_ref_alt, 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 - chr:pos 2 - chr:pos:ref:alt 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))
```

---

mergebd *Merge Format 5 binary dosage files*

Description

Merges two or more Format 5 binary dosage files into a single Format 5 output file. The merge type is determined automatically:

Usage

```
mergebd(bdose_files, bdose_file)
```

Arguments

bdose_files	Character vector of paths to the input .bdose files. Must contain at least two entries. The companion .bdi file for each is expected at `paste0(bdose_files[i], ".bdi")`.
bdose_file	Path for the output .bdose file. The companion .bdi metadata file is written to `paste0(bdose_file, ".bdi")`.

Details

- If subject IDs do not overlap across files, a **subject merge** is performed: the output contains all subjects from every file and the SNPs common to all files.
- If SNP IDs do not overlap across files, a **SNP merge** is performed: the output contains all SNPs from every file and the subjects common to all files.

If both subject IDs and SNP IDs overlap across files the merge cannot be performed and an error is returned.

SNPs are identified by chromosome, position, reference allele, and alternate allele.

Value

NULL (invisibly)

Examples

```
bdfile <- system.file("extdata", "vcf1a.bdose", package = "BinaryDosage")
bdinfo_src <- getbdinfo(bdfile)

# Create two format 5 files with non-overlapping subjects
bdose_a   <- tempfile(fileext = ".bdose")
bdose_b   <- tempfile(fileext = ".bdose")
bdose_tmp <- tempfile(fileext = ".bdose")
updatebd(bdfiles = bdfile, bdose_file = bdose_tmp)
subsetbd(bdfiles    = bdose_tmp,
         bdose_file = bdose_a,
         subjectids = bdinfo_src$samples$sid[1:30])
subsetbd(bdfiles    = bdose_tmp,
         bdose_file = bdose_b,
         subjectids = bdinfo_src$samples$sid[31:60])

bdose_out <- tempfile(fileext = ".bdose")
mergebd(bdose_files = c(bdose_a, bdose_b),
        bdose_file  = bdose_out)
```

---

openbd5con *Open a persistent connection to a Format 5 binary dosage file*

Description

Opens the .bdose file for reading and returns an object that holds the connection open across multiple calls to `getbd5snp_con`. The connection is closed automatically when the object is garbage-collected or when R exits; call `closebd5con` to close it explicitly.

Usage

```
openbd5con(bd5info)
```

Arguments

`bd5info`	Object returned by `getbdinfo`.

Value

An object of class `"bd5con"` to be passed to `getbd5snp_con` and `closebd5con`.

subsetbd *Subset a binary dosage file*

Description

Creates a new Format 5 binary dosage file containing a subset of the SNPs and/or subjects from an existing binary dosage file. The input file may be in any format (1-5). At least one filtering criterion must be supplied, and all supplied criteria must be met for a SNP or subject to be retained.

Usage

```
subsetbd(
  bdfiles,
  bdose_file,
  minmaf = NULL,
  locations = NULL,
  startloc = NULL,
  endloc = NULL,
  subjectids = NULL
)
```

Arguments

bdfiles	Vector of file names for the input binary dosage file. Format 4 files require one file name. Formats 1, 2, and 3 require three file names: the binary dosage file, the family file, and the map file. Format 5 files require two file names: the .bdose file and the .bdinfo file.
bdose_file	Path for the output .bdose file. The companion .bdi metadata file is written to `paste0(bdose_file, ".bdi")`.
minmaf	Minimum minor allele frequency. SNPs whose MAF (computed over the retained subjects) is below this value are excluded. Must be a single numeric value between 0 and 0.5.
locations	Integer or numeric vector of SNP base-pair locations to retain. Cannot be used together with `startloc` and `endloc`.
startloc	Start of the location range to retain (inclusive). Must be used together with `endloc`. Cannot be used together with `locations`.
endloc	End of the location range to retain (inclusive). Must be used together with `startloc`. Cannot be used together with `locations`.
subjectids	Character vector of subject IDs to retain.

Value

NULL (invisibly)

Examples

```
bdfile <- system.file("extdata", "vcf1a.bdose", package = "BinaryDosage")
bdinfo     <- getbdinfo(bdfile)
bdose_file <- tempfile(fileext = ".bdose")
subsetbd(bdfiles    = bdfile,
         bdose_file = bdose_file,
         subjectids = bdinfo$samples$sid[1:30])
```

---

updatebd *Update a binary dosage file to Format 5*

Description

Reads a binary dosage file in format 1, 2, 3, or 4, detects the format automatically, and converts it to a Format 5 file pair by calling the appropriate conversion routine. If the source file does not contain genotype probabilities, those values are stored as missing in the output.

Usage

```
updatebd(bdfiles, bdose_file)
```

Arguments

bdfiles	Vector of file names for the binary dosage file. Format 4 files require one file name. Formats 1, 2, and 3 require three file names: the binary dosage file, the family file, and the map file.
bdose_file	Path for the output .bdose file. The companion .bdi metadata file is written to `paste0(bdose_file, ".bdi")`.

Value

NULL (invisibly)

Examples

```
vcf1afile <- system.file("extdata", "set1a.vcf", package = "BinaryDosage")
bdfile <- tempfile()
suppressWarnings(
  vcftobdlegacy(vcffiles = vcf1afile,
                bdfiles = bdfile,
                format = 4L)
)
bdose_file <- tempfile(fileext = ".bdose")
updatebd(bdfiles    = bdfile,
         bdose_file = bdose_file)
```

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 Format 5 binary dosage files*

Description

Reads the DS (dosage) and GP (genotype probabilities) FORMAT fields from a bgzipped, tabix-indexed VCF file — as produced by imputation servers such as the Michigan Imputation Server — and writes a pair of Format 5 BinaryDosage files.

Usage

```
vcftobd(
  vcffile,
  bdose_file,
  region = NULL,
  snpidformat = 0L,
  bdoptions = character(0)
)
```

Arguments

vcffile	Path to the bgzipped, tabix-indexed VCF file.
bdose_file	Path for the output .bdose file. The companion .bdi metadata file is written to `paste0(bdose_file, ".bdi")`.
region	Optional genomic region string in bcftools format (e.g. `"chr21"` or `"chr21:1-5000000"`). Requires a tabix index. Default `NULL` processes the entire file.
snpidformat	Integer controlling how SNP IDs are stored. **-1** Generate IDs as `chr:pos:ref:alt`; equivalent to 2 for Format 5. **0** Use the IDs as they appear in the VCF file (default). Auto-detects format 1 or 2 if all IDs match. **1** Store IDs as `chr:pos`. An error is raised if the VCF already uses `chr:pos:ref:alt` format, as information would be lost. **2** Store IDs as `chr:pos:ref:alt`. **3** Store IDs as `chr:pos_ref_alt`.
bdoptions	Character vector specifying which per-SNP statistics to store. Any combination of `"aaf"` (alternate allele frequency), `"maf"` (minor allele frequency), and `"rsq"` (imputation r-squared). For each statistic, the corresponding VCF INFO field is used when present for the first SNP (AF, MAF, R2 respectively); otherwise the value is calculated from the dosage data. Default `character(0)` stores no statistics.

Details

The .bdose file begins with a 4-byte magic number followed by one gzip-compressed block per SNP. Each block contains the DS values for all samples followed by the GP values, encoded as unsigned 16-bit integers (round(value * 10000); 0xffff = missing).

The .bdi file is an RDS-serialised R list of class `"genetic-info"` with the following elements:

filename Path to the associated .bdose file.

usesfid Logical; always FALSE for VCF-sourced files.

samples data.frame with columns `fid` (empty) and `sid` (sample IDs).

onechr Logical; TRUE if all SNPs are on a single chromosome.

snpidformat Numeric; resolved SNP ID format (see `snpidformat` parameter).

snps data.frame with columns chromosome, location, snpid, reference, alternate.

snpinfo Named list of per-SNP annotations requested via `bdoptions`. Each element is a numeric vector of length equal to the number of SNPs. Values are read from the VCF INFO column when available for the first SNP (AF for aaf, MAF for maf, R2 for rsq); otherwise they are calculated from the dosage values.

additionalinfo List of class `"bdose-info"` with format, subformat, headersize, numgroups, and groups.

datasize Integer vector of length 0 (unused in Format 5).

indices Numeric vector of byte offsets into .bdose, one per SNP.

Value

NULL (invisibly)

`vcftobdlegacy` *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

```
vcftobdlegacy(
  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 - chr:pos 2 - chr:pos:ref:alt 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
vcftobdlegacy(vcffiles = c(vcf1afile, vcf1ainfo), bdfiles = bdfiles)
# Verify the file was written correctly
bdinfo <- getbdinfo(bdfiles)
```

Index

∗ **Iterating functions**
bdapply, 2
genapply, 5
vcfapply, 20

bdapply, 2, *6*, *20*
bdmerge, 3

closebd5con, 5

genapply, *3*, 5, *20*
gentobd, 6
getaaf, 8
getbd5snp, 9
getbd5snp_buf, 9
getbd5snp_con, 10
getbdinfo, 11
getgeninfo, *5*, 11
getmaf, 13
getrsq, 14
getsnp, 14
getvcfinfo, 15

mergebd, 16

openbd5con, 17

subsetbd, 18

updatebd, 19

vcfapply, *3*, *6*, 20
vcftobd, 21
vcftobdlegacy, 22