Package 'Gimpute'

November 28, 2018

Type Package

Title Gimpute: An efficient genetic data processing and imputation pipeline

Version 1.0

Date 2018-11-28

Author Junfang Chen, Dietmar Lippold, Emanuel Schwarz

Maintainer Junfang Chen <junfang.chen33@gmail.com>

Description Genotype imputation is essential for genome-wide association studies (GWAS) to retrieve information of untyped variants and facilitate comparability across studies. Based on widely used and freely available tools, we have developed Gimpute, an automated processing and imputation pipeline for genome-wide association data. Gimpute includes processing steps for genotype liftOver, quality control, population outlier detection, haplotype pre-phasing, imputation, post imputation, data management and the extension to other existing pipeline.

License GPL-3

Imports doParallel, lattice, grDevices, utils

Depends R (>= 3.5.0)

Suggests BiocStyle, knitr

VignetteBuilder knitr

biocViews Genetics, GenomicVariation, SNP, GenomeWideAssociation, QualityControl, VariantDetection, Alignment

Encoding UTF-8

LazyData true

RoxygenNote 6.1.1

R topics documented:

`.convertImpute2ByGtool` *Convert IMPUTE2 format files into PLINK format*

Description

Convert all chunks of IMPUTE2 format files into binary PLINK format using GTOOL.

Usage

```
.convertImpute2ByGtool(gtool, chrs, prefixChunk, phaseDIR, imputedDIR,
  prefix4eachChr, suffix4imputed, postImputeDIR, threshold, nCore)
```

Arguments

`gtool`	an executable program in either the current working directory or somewhere in the command path.
`chrs`	specifiy the chromosome codes for conversion.
`prefixChunk`	the prefix of the chunk files for each chromosome, along with the location directory.
`phaseDIR`	the directory where pre-phased files are located.
`imputedDIR`	the directory where the imputated files are located.
`prefix4eachChr`	the prefix of the input IMPUTE2 files and also the output PLINK binary files for each chunk.
`suffix4imputed`	the suffix of the IMPUTE2 format file that stores the imputed value. Both ".impute2" and ".gen" are accepted.
`postImputeDIR`	the directory where converted PLINK binary files will be located.
`threshold`	threshold for merging genotypes from GEN probability. Default 0.9.
`nCore`	the number of cores used for computation.

Value

The converted binary PLINK format files for each chunk from IMPUTE2 results.

Author(s)

Junfang Chen

`.filterImputeData` *Filter genetic variants*

Description

Filter out genetic variants accoring to the imputation quality score.

Usage

```
.filterImputeData(plink, suffix4impute2info, outputInfoFile,
  infoScore = 0.6, inputPrefix, outputPrefix)
```

Arguments

`plink`	an executable program in either the current working directory or somewhere in the command path.
`suffix4impute2info`	the suffix of input IMPUTE2 generated files that store the imputation quality score for each variant from .impute2_info files.
`outputInfoFile`	the output file of impute2 info scores consisting of two columns: all imputed SNPs and their info scores.
`infoScore`	the cutoff of filtering imputation quality score for each variant. The default value is 0.6.
`inputPrefix`	the prefix of the input imputed PLINK binary files.
`outputPrefix`	the prefix of the output filtered PLINK binary files.

Details

Filter genetic variants accoring to the imputation quality score with the help of .impute2_info files generated by IMPUTE2. Often, we keep variants with imputation info score of greater than 0.6. Note that imputed SNPs with more than two alleles are not considered.

Value

A pure text file contains the info scores of all imputed SNPs with two columns: SNP names and the corresponding info scores. A pure text file with all excluded SNPs having bad info scores. The filtered PLINK binary imputed files.

Author(s)

Junfang Chen

`.filterImputeData2` *Filter genetic variants*

Description

Filter out genetic variants accoring to the info score.

Usage

```
.filterImputeData2(plink, outputInfoFile, infoScore = 0.6, inputPrefix,
  outputPrefix)
```

Arguments

`plink`	an executable program in either the current working directory or somewhere in the command path.
`outputInfoFile`	the output file of info scores consisting of two columns: SNP names and their info scores. The headers are c("rsid", "info").
`infoScore`	the cutoff of filtering imputation quality score for each variant. The default value is 0.6.
`inputPrefix`	the prefix of the input imputed PLINK binary files.
`outputPrefix`	the prefix of the output filtered PLINK binary files.

Details

Filter genetic variants accoring to the imputation quality score with the help of .impute2_info files generated by IMPUTE2. Often, we keep variants with imputation info score of greater than 0.6. Note that imputed SNPs with more than two alleles are not considered.

Value

A pure text file contains the info scores of all imputed SNPs with two columns: SNP names and the corresponding info scores. A pure text file with all excluded SNPs having bad info scores. The filtered PLINK binary imputed files,

Author(s)

Junfang Chen

`.getInfoScoreImpute2` *Extract info score*

Description

Extract info score file summarized from the output of IMPUTE2.

Usage

```
.getInfoScoreImpute2(suffix4impute2info, outputInfoFile)
```

Arguments

`suffix4impute2info`	the suffix of input IMPUTE2 generated files that store the imputation quality score for each variant from .impute2_info files.
`outputInfoFile`	the output file of impute2 info scores consisting of two columns: all imputed SNPs and their info scores.

Value

A pure text file contains the info scores of all imputed SNPs with two columns: SNP names and the corresponding info scores. The header names are "rsid" and "info".

Author(s)

Junfang Chen

`.imputedByImpute2` *Impute genotypes using IMPUTE2*

Description

Perform imputation by IMPUTE2 for the autosomal and sex chromosome prephased known haplotypes with a reference panel.

Usage

```
.imputedByImpute2(impute2, chrs, prefixChunk, phaseDIR, referencePanel,
  impRefDIR, imputedDIR, prefix4eachChr, nCore, effectiveSize = 20000)
```

Arguments

impute2	an executable program in either the current working directory or somewhere in the command path.
chrs	specifiy the chromosome codes for imputation.
prefixChunk	the prefix of the chunk files for each chromosome, along with the proper location directory.
phaseDIR	the directory where prephased haplotypes are located.
referencePanel	a string indicating the type of imputation reference panels is used: c("1000Gphase1v3_macGT1", "1000Gphase3").
impRefDIR	the directory where the imputation reference files are located.
imputedDIR	the directory where imputed files will be located.
prefix4eachChr	the prefix of IMPUTE2 files for each chunk.
nCore	the number of cores used for computation.
effectiveSize	this parameter controls the effective population size. Commonly denoted as Ne. A universal -Ne value of 20000 is suggested.

Details

If ChrX is available then it is done automatically by passing the flag –chrX to IMPUTE2.

Value

The imputed files for all chunks from the given chromosomes.

Author(s)

Junfang Chen

See Also

phaseImpute2.

.imputedByImpute4 *Impute genotypes using IMPUTE4*

Description

Perform imputation by IMPUTE4 for the autosomal prephased known haplotypes with a reference panel.

Usage

```
.imputedByImpute4(impute4, chrs, prefixChunk, phaseDIR, referencePanel,
  impRefDIR, imputedDIR, prefix4eachChr, nCore, effectiveSize = 20000)
```

Arguments

impute4	an executable program in either the current working directory or somewhere in the command path.
chrs	specifiy the chromosome codes for imputation.
prefixChunk	the prefix of the chunk files for each chromosome, along with the proper location directory.
phaseDIR	the directory where prephased haplotypes are located.
referencePanel	a string indicating the type of imputation reference panels is used: c("1000Gphase1v3_macGT1", "1000Gphase3").
impRefDIR	the directory where the imputation reference files are located.
imputedDIR	the directory where imputed files will be located.
prefix4eachChr	the prefix of IMPUTE2 files for each chunk.
nCore	the number of cores used for computation.
effectiveSize	this parameter controls the effective population size. Commonly denoted as Ne. A universal -Ne value of 20000 is suggested.

Value

The imputed files for all chunks from given chromosomes, except sex chromosomes.

Author(s)

Junfang Chen

See Also

phaseImpute4.

.mergePlinkData *Merge chunk-wise PLINK files*

Description

Merge all chunk-wise PLINK binary files into chromosome-wise PLINK binary files then assemble into a genome-wide PLINK binary file set.

Usage

```
.mergePlinkData(plink, chrs, prefix4eachChr, prefix4mergedPlink, nCore)
```

Arguments

plink	an executable program in either the current working directory or somewhere in the command path.
chrs	specifiy the chromosome codes to be merged.
prefix4eachChr	the prefix of the input chunk-wise PLINK files.
prefix4mergedPlink	the prefix of the final output PLINK binary files.
nCore	the number of cores used for computation.

Details

Create a file containing a list chunk-wise PLINK PED and MAP file names. The prefix of these files must already indicate in which chromosome they belong to and files from the same chromosome will be combined. Then all chromosomal PLINK files are assembled together into one whole genome PLINK binary file set.

Value

The merged genome-wide PLINK binary files.

Author(s)

Junfang Chen

.prepareLegend2bim *Prepare a bim-like reference file*

Description

Prepare a bim-like file from the imputation reference legend file (e.g. 1000 genome project).

Usage

```
.prepareLegend2bim(inputFile, referencePanel, outputFile, ncore = 20)
```

Arguments

inputFile	a set of legend file from the imputation reference panel. For example, 1000 genome projects.
referencePanel	a string indicating the type of imputation reference panels is used: c("1000Gphase1v3_macGT1", "1000Gphase3").
outputFile	the pure text file that stores the prepared PLINK BIM file alike format data.
ncore	the number of cores used for computation. The default is 20.

Details

To prepare a bim-like reference file from legend files. One should first extract the specific content from these legend files after downloading. Note that extract only biallelic SNPs (only 1 allele in column3 and 4, and start with 'rs") and remove duplicated snp IDs. Column names are added in the end.

Value

Prepare a bim-like reference file. Note that the column names are already defined, i.e. "chr", "rsID", "pos", "a0", "a1."

Author(s)

Junfang Chen

.prePhasingByShapeit

Prephasing genotypes using SHAPEIT

Description

Perform prephasing for study genotypes by SHAPEIT for the autosomal and sex chromosome haplotypes using a reference panel (pre-set).

Usage

```
.prePhasingByShapeit(shapeit, chrs, dataDIR, prefix4eachChr,
  referencePanel, impRefDIR, phaseDIR, nThread = 40,
  effectiveSize = 20000, nCore = 1)
```

Arguments

`shapeit`	an executable program in either the current working directory or somewhere in the command path.
`chrs`	specifiy the chromosome codes for phasing.
`dataDIR`	the directory where genotype PLINK binary files are located.
`prefix4eachChr`	the prefix of PLINK binary files for each chromosome.
`referencePanel`	a string indicating the type of imputation reference panels is used: c("1000Gphase1v3_macGT1", "1000Gphase3").
`impRefDIR`	the directory where the imputation reference files are located.
`phaseDIR`	the directory where resulting pre-phased files will be located.
`nThread`	the number of threads used for computation. The default value is 40.
`effectiveSize`	this parameter controls the effective population size. The default value is 20000.
`nCore`	the number of cores used for computation. This can be tuned along with nThread. The default value is 1.

Details

If ChrX is available then it is done automatically by passing the flag –chrX to SHAPEIT.

Value

The pre-phased haplotypes for given chromosomes.

Author(s)

Junfang Chen

See Also

`phaseImpute2`.

.snpSharedPos *Find shared genomic position between two files.*

Description

Find shared genomic position between two files and return the snp names of the second input file.

Usage

```
.snpSharedPos(inputFile1, inputFile2, outputFile, nCore = 20)
```

Arguments

`inputFile1`	the pure text file that has at least three columns: chromosomal location, snp name and base-pair position, with the name c("chr", "rsID", "pos")
`inputFile2`	the pure text file that has at least three columns: chromosomal location, snp name and base-pair position, with the name c("chr", "rsID", "pos")
`outputFile`	the pure text file return the snp name of the second input file.
`nCore`	the number of cores used for computation. The default is 20.

Value

The snp name of the second input file which shares the same genomic position with that of the first input file.

`checkAlign2ref` *Check the alignment with the imputation reference panel*

Description

Perform the alignment against a reference panel by considering the following parameters: variant name, genomic position and the allele profile. Output files are generated sequentially depending on their previous PLINK data.

Usage

```
checkAlign2ref(plink, inputPrefix, referencePanel, bimReferenceFile, out2,
  out2.snp, out3, out3.snp, out4, out4.snp, out4.snpRetained, nCore = 25)
```

Arguments

`plink`	an executable program in either the current working directory or somewhere in the command path.
`inputPrefix`	the prefix of the input PLINK files.
`referencePanel`	a string indicating the type of imputation reference panels is used: c("1000Gphase1v3_macGT1", "1000Gphase3").
`bimReferenceFile`	the reference file used for the alignment, which is a PLINK BIM alike format file.
`out2`	the prefix of the output PLINK binary files after removing SNPs whose genomic positions are not in the imputation reference, taking SNP names into account.
`out2.snp`	the output plain text file that stores the removed SNPs whose genomic positions are not in the imputation reference, taking SNP names into account.
`out3`	the prefix of the output PLINK binary files after removing SNPs whose genomic positions are not in the imputation reference, ingoring SNP names.
`out3.snp`	the output plain text file that stores the removed SNPs whose genomic positions are not in the imputation reference, ingoring SNP names.
`out4`	the prefix of the output PLINK binary files after removing SNPs whose alleles are not in the imputation reference, taking their genomic positions into account.
`out4.snp`	the output plain text file that stores the removed SNPs whose alleles are not in the imputation reference, taking their genomic positions into account.
`out4.snpRetained`	the output plain text file that stores the removed SNPs whose alleles are in the imputation reference, taking their genomic positions into account.
`nCore`	the number of cores used for computation. This can be tuned along with nThread.

Details

The output files are genrated in order. Genomic position includes chromosomal location and base-pair position of the individual variant. All monomorphic SNPs are retained for further processing.

Value

The set of aligned PLINK files from your own study compared with the imputation reference.

Author(s)

Junfang Chen

`chrWiseSplit` *Split genome-wide genotyping data into chromosome-wide PLINK binary files.*

Description

Split the whole genome genotyping data chromosome-wise; allow parallel computating for all chromosomes.

Usage

```
chrWiseSplit(plink, inputPrefix, chrXPAR1suffix, chrXPAR2suffix,
  nCore = 25)
```

Arguments

`plink`	an executable program in either the current working directory or somewhere in the command path.
`inputPrefix`	the prefix of the input PLINK binary files before splitting.
`chrXPAR1suffix`	if chromosome 25 is available and with PAR1, then generate the suffix with X_PAR1 for chrX_PAR1.
`chrXPAR2suffix`	if chromosome 25 is available and with PAR2, then generate the suffix with X_PAR2 for chrX_PAR2.
`nCore`	the number of cores used for parallel computation. The default value is 25.

Details

If chromosome 25 is also available, namely the pseudo-autosomal region of chromosome X, then further split chr25 (PAR or Chr_XY) into PAR1 and PAR2 according to the genomic coordination GRCh37 from https://en.wikipedia.org/wiki/Pseudoautosomal_region. The locations of the PARs within GRCh37 are: PAR1 X 60001 2699520; PAR2 X 154931044 155260560.

Value

The output PLINK binary files for each chromosome with the same prefix as the inputPrefix but appended with the chromosome codes, and possibly also the logical value for the pseudo-autosomal region (PAR) indicating if PAR exists in the input genotyping data or not.

Author(s)

Junfang Chen

Examples

```
## In the current working directory
bedFile <- system.file("extdata", "alignedData.bed", package="Gimpute")
bimFile <- system.file("extdata", "alignedData.bim", package="Gimpute")
famFile <- system.file("extdata", "alignedData.fam", package="Gimpute")
system(paste0("scp ", bedFile, " ."))
system(paste0("scp ", bimFile, " ."))
system(paste0("scp ", famFile, " ."))
inputPrefix <- "alignedData"
chrXPAR1suffix <- "X_PAR1"
chrXPAR2suffix <- "X_PAR2"
## Not run: Requires an executable program PLINK, e.g.
## plink <- "/home/tools/plink"
## chrWiseSplit(plink, inputPrefix, chrXPAR1suffix, chrXPAR2suffix, nCore)
```

---

chunk4eachChr *Chunk each chromosome into multiple segments*

Description

Chunk each chromosome genotyping data into multiple segments by a predefined window size.

Usage

```
chunk4eachChr(inputPrefix, outputPrefix, chrs, windowSize = 3e+06)
```

Arguments

`inputPrefix`	the prefix of the input PLINK .bim file for each chromosome, without the chromosome codes.
`outputPrefix`	the prefix of the output pure text files that keep all the chunks for each chromosome separately.
`chrs`	specifiy the chromosome codes for chunking.
`windowSize`	the window size of each chunk. The default value is 3000000.

Value

The output pure text files include all the chunks for each chromosome separately.

Author(s)

Junfang Chen

Examples

```
## In the current working directory
bimFile <- system.file("extdata", "gwas_data_chr23.bim", package="Gimpute")
system(paste0("scp ", bimFile, " ."))
inputPrefix <- "gwas_data_chr"
outputPrefix <- "chunks_chr"
```

```
chrs <- 23
print(chrs)
chunk4eachChr(inputPrefix, outputPrefix, chrs, windowSize=3000000)
```

`computeInfoByQctool`
Calculate the info score by QCTOOL

Description

Calculate the info score by QCTOOL for a set of SNPs from .GEN files.

Usage

```
computeInfoByQctool(qctool, inputSuffix, outputInfoFile)
```

Arguments

`qctool`	an executable program in either the current working directory or somewhere in the command path.
`inputSuffix`	the suffix of input .GEN files within current directory.
`outputInfoFile`	the output info scores file consisting of two columns: all SNPs from .GEN files and their info scores.

Details

These .GEN files may come from the output of impute4 results. The intermediate generated files are retained in the directory.

Value

A pure text file contains the info scores of all SNPs from .GEN files with two columns: SNP names and the corresponding info scores.

Author(s)

Junfang Chen

Examples

```
## In the current working directory
inputSuffix <- "noINDEL.gen"
outputInfoFile <- "infoScore.txt"
## computeInfoByQctool(qctool, inputSuffix, outputInfoFile)
```

`extractByGenipe` *Extract imputed markers using Genipe*

Description

Extract imputed markers located in a specific genomic region using Genipe. Note that, 1.) 'bed' PLINK binary format is specifically used for the output format. 2.) The markers of the whole chromosome are extracted together. For the filtering of maf and info will be done during post imputation.

Usage

```
extractByGenipe(inputImpute2, inputMAP, outputPrefix, format, prob)
```

Arguments

`inputImpute2`	the output from IMPUTE2.
`inputMAP`	the output PLINK MAP file from Genipe, which will be used for generating markers in a text file (only one column without column name).
`outputPrefix`	the prefix of the output files. [impute2_extractor]
`format`	the output format. Can specify either "impute2" for probabilities (same as impute2 format, i.e. 3 values per sample), "dosage" for dosage values (one value between 0 and 2 by sample), "calls" for hard calls, or "bed" for Plink binary format (with hard calls). [impute2]
`prob`	the probability threshold used when creating a file in the dosage or call format. [0.9]

Value

The extracted imputed files using genipe.

Author(s)

Junfang Chen

References

Lemieux Perreault, L. P., et al. (2016). genipe: an automated genome-wide imputation pipeline with automatic reporting and statistical tools. Bioinformatics, 32(23), 3661-3663.

genoQC *Quality control for genotype data*

Description

Perform quality control on the genotype data.

Usage

```
genoQC(plink, inputPrefix, snpMissCutOffpre = 0.05,
  sampleMissCutOff = 0.02, Fhet = 0.2, cutoffSubject, cutoffSNP,
  snpMissCutOffpost = 0.02, snpMissDifCutOff = 0.02,
  femaleChrXmissCutoff = 0.05, pval4autoCtl = 1e-06,
  pval4femaleXctl = 1e-06, outputPrefix, keepInterFile = TRUE)
```

Arguments

plink	an executable program in either the current working directory or somewhere in the command path.
inputPrefix	the prefix of the input PLINK binary files.
snpMissCutOffpre	the cutoff of the missingness for removing SNPs before subject removal. The default is 0.05.
sampleMissCutOff	the cutoff of the missingness for removing subjects/instances. The default is 0.02.
Fhet	the cutoff of the autosomal heterozygosity deviation. The default is 0.2.
cutoffSubject	the cutoff determines that families (subjects) with more than the predefined cutoff of Mendel errors by considering all SNPs will be removed. The default is 0.05.
cutoffSNP	the cutoff indicates that SNPs with more than the predefined cutoff of Mendel error rate will be excluded (i.e. based on the number of trios/duos). The default is 0.1.
snpMissCutOffpost	the cutoff of the missingness for removing SNPs after subject removal. The default is 0.02.
snpMissDifCutOff	the cutoff of the difference in missingness between cases and controls. The default is 0.02.
femaleChrXmissCutoff	the cutoff of the missingness in female chromosome X SNPs. The default is 0.05.
pval4autoCtl	the p-value cutoff for controlling HWE test in either control or case subjects. Only autosomal SNPs are considered. The default is 0.000001
pval4femaleXctl	the p-value cutoff for controlling HWE test in female control subjects. Only chromosome X SNPs are considered. The default is 0.000001
outputPrefix	the prefix of the output PLINK binary files after QC.

keepInterFile
: a logical value indicating if the intermediate processed files should be kept or not. The default is TRUE.

Details

The original PLINK files are implicitly processed by the following default steps: 1.) Set all heterozygous alleles of SNPs on male chrX as missing; 2.) SNP missingness < 0.05 (before sample removal); 3.) Subject missingness < 0.02; 4.) Remove subjects with |Fhet| >= 0.2; 5.) Reset paternal and maternal codes; 6.) SNP missingness < 0.02 (after sample removal); 7.) Remove SNPs with difference >= 0.02 of SNP missingness between cases and controls; 8.) Remove subjects or SNPs with Mendel errors for family based data. 9.) Remove chrX SNPs with missingness >= 0.05 in females. (Optional, if no chrX data); 10.) Remove autosomal SNPs with HWE p < 10-6 in controls; 11.) Remove chrX SNPs with HWE p < 10-6 in female controls. (Optional, if no chrX data).

Value

The output PLINK binary files after QC.

Author(s)

Junfang Chen

References

Schizophrenia Working Group of the Psychiatric Genomics, C. (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature 511(7510): 421-427.

Examples

```
## In the current working directory
bedFile <- system.file("extdata", "genoUpdatedData.bed", package="Gimpute")
bimFile <- system.file("extdata", "genoUpdatedData.bim", package="Gimpute")
famFile <- system.file("extdata", "genoUpdatedData.fam", package="Gimpute")
system(paste0("scp ", bedFile, bimFile, famFile, " ."))
inputPrefix <- "genoUpdatedData"
outputPrefix <- "2_13_removedSnpHweFemaleX"
## Not run: Requires an executable program PLINK, e.g.
## plink <- "/home/tools/plink"
## genoQC(plink, inputPrefix,
##        snpMissCutOffpre=0.05,
##        sampleMissCutOff=0.02,
##        Fhet=0.2, cutoffSubject, cutoffSNP,
##        snpMissCutOffpost=0.02,
##        snpMissDifCutOff=0.02,
##        femaleChrXmissCutoff=0.05,
##        pval4autoCtl=0.000001,
##        pval4femaleXctl=0.000001,
##        outputPrefix, keepInterFile=TRUE)
```

getGroupLabel *Get the outcome label of the genotype data*

Description

Get the group label from the PLINK FAM file.

Usage

```
getGroupLabel(inputFAMfile)
```

Arguments

`inputFAMfile` the PLINK FAM file.

Details

If the input FAM file also has missing outcomes, which are shown in the sixth column of FAM file as "0", then an error is given.

Value

The group label of the genotype data: "control" or "case" or "caseControl" indicating both groups exist.

Author(s)

Junfang Chen

Examples

```
famFile <- system.file("extdata", "genoUpdatedData.fam", package="Gimpute")
getGroupLabel(inputFAMfile=famFile)
```

imputedByGenipe *Impute genotypes using Genipe*

Description

Perform imputation by Genipe for the autosomal and sex chromosome prephased known haplotypes with a reference panel. Note that pre-phasing using SHAPEIT is done without the refernce haplotypes.

Usage

```
imputedByGenipe(chrs, impRefDir, inputPrefix, shapeit, impute2, plink,
  fastaFile, segmentSize, thread4impute2, thread4shapeit)
```

Arguments

chrs	specifiy the chromosomes for imputation. There are four different options ("autosomes", "1, or 2, or 3...or 22", "23", "25"). 1.) 'autosomes': will impute chromosome 1 to 22 together; 2.) 'chrs' belongs to one of (1, 2,...22) then the imputation is done just for one autosomal chromosome. 3.) '23' will do the imputation for the non-pseudoautosomal region of chromosome 23. 4.) '25' imputes for the pseudoautosomal regions of chromosome 23.
impRefDir	the directory where the imputation reference files are located.
inputPrefix	the prefix of the input PLINK binary files.
shapeit	an executable SHAPEIT binary program in either the current working directory or somewhere in the command path.
impute2	an executable IMPUTE2 binary program in either the current working directory or somewhere in the command path.
plink	an executable program in either the current working directory or somewhere in the command path.
fastaFile	the human reference files for the initial strand check.
segmentSize	the length of a single segment for imputation.
thread4impute2	the number of threads for the imputation.
thread4shapeit	the number of threads for phasing.

Value

The imputed files using genipe.

Author(s)

Junfang Chen

References

Lemieux Perreault, L. P., et al. (2016). genipe: an automated genome-wide imputation pipeline with automatic reporting and statistical tools. Bioinformatics, 32(23), 3661-3663.

---

mergeByGenipe *Merge imputed files using Genipe*

Description

Concatenate IMPUTE2 output files and retrieve some statistics. This is automatically called by the main genipe pipeline to merge IMPUTE2 files generated for all the genomic segments. For details, see Genipe IMPUTE2 merger options.

Usage

```
mergeByGenipe(inputImpute2, chr, probability, completionRate, info,
  outputPrefix)
```

Arguments

inputImpute2	the output from IMPUTE2.
chr	specifiy the chromosome segment to be merged, on which the imputation was made.
probability	the probability threshold for no calls. [<0.9]
completionRate	the completion rate threshold for site exclusion. [<0.98]
info	the measure of the observed statistical information associated with the allele frequency estimate threshold for site exclusion. (<0.00)
outputPrefix	the prefix for the imputed output files.

Value

The merged imputed files using genipe.

Author(s)

Junfang Chen

References

Lemieux Perreault, L. P., et al. (2016). genipe: an automated genome-wide imputation pipeline with automatic reporting and statistical tools. Bioinformatics, 32(23), 3661-3663.

---

phaseImpute *Phasing and imputation*

Description

Perform phasing, imputation and conversion from IMPUTE2 or GEN format into PLINK binary files.

Usage

```
phaseImpute(inputPrefix, outputPrefix, plink, shapeit, imputeTool, impute,
  qctool, gtool, windowSize = 3e+06, effectiveSize = 20000,
  nCore = 40, threshold = 0.9, outputInfoFile, referencePanel,
  impRefDIR, tmpImputeDir, keepTmpDir = TRUE)
```

Arguments

inputPrefix	the prefix of the input PLINK binary files for the imputation.
outputPrefix	the prefix of the output PLINK binary files after imputation.
plink	an executable program in either the current working directory or somewhere in the command path.
shapeit	an executable program in either the current working directory or somewhere in the command path.
imputeTool	a string indicating the type of imputation tool is used: "impute2" or "impute4".

`impute` an executable program in either the current working directory or somewhere in the command path. It can be either "impute2" or "impute4".

`qctool` an executable program in either the current working directory or somewhere in the command path. This is only used if imputeTool is "impute4".

`gtool` an executable program in either the current working directory or somewhere in the command path.

`windowSize` the window size of each chunk. The default value is 3000000.

`effectiveSize`
this parameter controls the effective population size. Commonly denoted as Ne. A universal -Ne value of 20000 is suggested.

`nCore` the number of cores used for splitting chromosome by PLINK, phasing, imputation, genotype format modification, genotype conversion, and merging genotypes. The default value is 40.

`threshold` threshold for merging genotypes from GEN probability. Default 0.9.

`outputInfoFile`
the output file of impute2 info scores consisting of two columns: all imputed SNPs and their info scores.

`referencePanel`
a string indicating the type of imputation reference panels is used: "1000Gphase1v3_macGT1" or "1000Gphase3".

`impRefDIR` the directory where the imputation reference files are located.

`tmpImputeDir` the name of the temporary directory used for storing phasing and imputation results.

`keepTmpDir` a logical value indicating if the directory 'tmpImputeDir' should be kept or not. The default is TRUE.

Details

The whole imputation process mainly consists of the following steps: 1.) Phasing the input PLINK data using an existing imputation reference; 2.) Imputing the input PLINK data using phased results and an existing reference data; 3.) Converting IMPUTE2 or GEN format data into PLINK format. 4.) Combining all imputed data into whole-genome PLINK binary files. 5.) Filtering out imputed variants with bad imputation quality. Parallel computing in R is supported.

Value

Note that chromosome X is not supported for the impute4. 1.) The filtered imputed PLINK binary files; 2.) The final PLINK binary files including bad imputed variants; 3.) A pure text file contains the info scores of all imputed SNPs with two columns: SNP names and the corresponding info scores.

Author(s)

Junfang Chen

References

1. Howie, B., et al. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet 44(8): 955-959.
2. Howie, B. N., et al. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 5(6): e1000529.

3. Bycroft, C., et al. Genome-wide genetic data on~ 500,000 UK Biobank participants. BioRxiv (2017): 166298.

Examples

```
## In the current working directory
bedFile <- system.file("extdata", "alignedData.bed", package="Gimpute")
bimFile <- system.file("extdata", "alignedData.bim", package="Gimpute")
famFile <- system.file("extdata", "alignedData.fam", package="Gimpute")
system(paste0("scp ", bedFile, " ."))
system(paste0("scp ", bimFile, " ."))
system(paste0("scp ", famFile, " ."))
inputPrefix <- "alignedData"
outputPrefix <- "gwasImputed"
outputInfoFile <- "infoScore.txt"
tmpImputeDir <- "tmpImpute"
## Not run: Requires an executable program PLINK, e.g.
## plink <- "/home/tools/plink"
## phaseImpute(inputPrefix, outputPrefix,
##             plink, shapeit, imputeTool, impute, qctool, gtool,
##             windowSize=3000000, effectiveSize=20000,
##             nCore=40, threshold=0.9, outputInfoFile,
##             referencePanel, impRefDIR, tmpImputeDir, keepTmpDir=TRUE)
```

---

plotPCA4plink *Population outlier detection*

Description

Principle component analysis (PCA) on the genotype data is performed to detect population outliers, and the first two PCs are plotted for the visualization.

Usage

```
plotPCA4plink(gcta, inputPrefix, nThread = 20, outputPC4subjFile,
  outputPCplotFile)
```

Arguments

`gcta`	an executable program in either the current working directory or somewhere in the command path.
`inputPrefix`	the prefix of the input PLINK binary files.
`nThread`	the number of threads used for computation. The default is 20.
`outputPC4subjFile`	the pure text file that stores all the subject IDs and their corresponding eigenvalues of the first two principle components.
`outputPCplotFile`	the plot file for visualizing the first two principle components of all investigated subjects.

Details

Before population outlier detection, it's better to perform QC on the genotype data. Only autosomal genotypes are used for principle component analysis.

Value

The output pure text file and plot file for storing first two principle components of study subjects.

Author(s)

Junfang Chen

Examples

```
## In the current working directory
bedFile <- system.file("extdata", "QCdata.bed", package="Gimpute")
bimFile <- system.file("extdata", "QCdata.bim", package="Gimpute")
famFile <- system.file("extdata", "QCdata.fam", package="Gimpute")
system(paste0("scp ", bedFile, bimFile, famFile, " ."))
inputPrefix <- "QCdata"
outputPC4subjFile <- "2_13_eigenvalAfterQC.txt"
outputPCplotFile <- "2_13_eigenvalAfterQC.png" ## png format
## Not run: Requires an executable program GCTA, e.g.
## gcta <- "/home/tools/gcta64"
## plotPCA4plink(gcta, inputPrefix, nThread=20,
##               outputPC4subjFile, outputPrefix)
```

---

postImpQC *Post imputation quality control*

Description

Perform quality control and data management after imputation.

Usage

```
postImpQC(plink, inputPrefix, out1, out2, out3, out4, outputInfoFile,
  infoScore = 0.6, outputMonoSNPfile, prefixAlign2ref, missCutoff = 20,
  outRemovedSNPfile, outRetainSNPfile, referencePanel)
```

Arguments

plink	an executable program in either the current working directory or somewhere in the command path.
inputPrefix	the prefix of the final imputed PLINK files.
out1	the prefix of well imputed PLINK files with the index.
out2	the prefix of well imputed PLINK files after removing any SNPs with the same positions (if any), the index is also appended.
out3	the prefix of well imputed PLINK files with the index after adding previously identified monomorphic SNPs if any.

`out4` the prefix of final well imputed PLINK files with the index.

`outputInfoFile`
the output file of impute2 info scores consisting of two columns: all imputed SNPs and their info scores.

`infoScore` the cutoff of filtering imputation quality score for each variant. The default value is 0.6.

`outputMonoSNPfile`
the output pure text file that stores the removed monomorphic SNPs, one per line, if any.

`prefixAlign2ref`
the prefix of the output PLINK binary files after removing SNPs whose alleles are not in the imputation reference, taking their genomic positions into account.

`missCutoff` the cutoff of the least number of instances for a SNP that is not missing. The default is 20.

`outRemovedSNPfile`
the output file of SNPs with pre-defined missing values that are removed.

`outRetainSNPfile`
the output file of SNPs that are retained.

`referencePanel`
a string indicating the type of imputation reference panels is used: c("1000Gphase1v3_macGT1", "1000Gphase3").

Value

All imputed genotype data in PLINK format, well imputed data and its variants, as well as a set of pure text files containing removed or retained SNPs.

Author(s)

Junfang Chen

`prepareAnnoFile4affy`
prepare Affymetrix chip annotation file

Description

Prepare Affymetrix chip annotation file into the format of interest.

Usage

```
prepareAnnoFile4affy(inputFile, outputFile, chipType)
```

Arguments

`inputFile` an input pure text file that contains the chip annotation information.

`outputFile` an output pure text file that stores the chip annotation information in a user-defined format.

`chipType` a string name defines the type of the chip annotation file: 'SNPIDstudy', and 'rsIDstudy'.

Details

If the chip annotation file is not available for your study, it can be downloaded from http://www.well.ox.ac.uk/~wrayner/st
The chip annotation file is organized into two different types:

1. If the snp name of your study genotype data starts with "SNP_", then the chip type "SNPID-study" is used; Usually, Affymetrix chip data belongs to this category. The prepared output annotation file must at least consist of the following column names: SNPIDstudy, rs, chr, pos, strand.
2. If the snp name of your study genotype data starts with "rs", then the chip type "rsIDstudy" is used; The prepared output annotation file must at least consist of the following column names: SNPIDstudy, rs, chr, pos, strand. Illumina chip is often specified in this format.

The column "strand" must only have two kinds of values "-" and "+". Variants with all other values should be excluded.

Value

a pure text file that stores the prepared chip annotation information in a user-defined format.

Author(s)

Junfang Chen

---

reductExpand *Post imputation data extraction and expansion*

Description

Reduce well imputed dataset to have SNPs before imputation and then added genotype data that are different from the imputation reference panel.

Usage

```
reductExpand(plink, referencePanel, inputPrefix, inputQCprefix,
  snpRefAlleleFile, snpDiffAlleleFile, snpMissPosFile,
  snpSameNameDifPosFile, reducedToSpecificfn, specificDiffAllelefn,
  specificMissPosfn, specificDiffPosfn)
```

Arguments

`plink`	an executable program in either the current working directory or somewhere in the command path.
`referencePanel`	a string indicating the type of imputation reference panels is used: c("1000Gphase1v3_macGT1", "1000Gphase3").
`inputPrefix`	the prefix of final well imputed PLINK files.
`inputQCprefix`	the prefix of QCed PLINK files.
`snpRefAlleleFile`	the output plain text file that stores the removed SNPs whose alleles are in the imputation reference, taking their genomic positions into account.

`snpDiffAlleleFile`
: the output plain text file that stores the removed SNPs whose alleles are not in the imputation reference, taking their genomic positions into account.

`snpMissPosFile`
: the output plain text file that stores the removed SNPs whose genomic positions are not in the imputation reference, ingoring SNP names.

`snpSameNameDifPosFile`
: the output plain text file that stores the removed SNPs whose genomic positions are not in the imputation reference, taking SNP names into account.

`reducedToSpecificfn`
: the prefix of PLINK files which are reduced in the number of SNPs and contain only SNPs prior to imputation.

`specificDiffAllelefn`
: the prefix of PLINK files which added genotypes with different alleles other than the imputation to the file with the prefix: reducedToSpecificfn.

`specificMissPosfn`
: the prefix of PLINK files which added genotypes with missing positions other than the imputation to the file with the prefix: specificDiffAllelefn.

`specificDiffPosfn`
: the prefix of PLINK files which added genotypes with different positions other than the imputation to the file with the prefix: specificMissPosfn .

Value

Extracted genotypes from the imputed data which contain SNPs before imputation. On the basis of extracted genotypes, genotypes different from the imputation reference panel are appended, including SNPs with different alleles, missing positions, and different positions.

Author(s)

Junfang Chen

`removedDoubleProbes` *Remove duplicated SNPs*

Description

Remove duplicated SNPs that have same rs-names or duplicated genomic position.

Usage

```
removedDoubleProbes(plink, inputPrefix, chipAnnoFile, chipType,
  outputSNPdupFile, outputPrefix)
```

Arguments

plink	an executable program in either the current working directory or somewhere in the command path.
inputPrefix	the prefix of the input PLINK binary files.
chipAnnoFile	an input chip annotation file. If the chip annotation file is not available for your study, it can be downloaded from http://www.well.ox.ac.uk/~wrayner/strand/.
chipType	a string name defines the type of the chip annotation file: 'SNPIDstudy', and 'rsIDstudy'. The detail is described in prepareAnnoFile4affy.
outputSNPdupFile	a pure text file that stores the duplicated SNP IDs, which are detected by the use of the chip annotation file.
outputPrefix	the prefix of the output PLINK binary files.

Details

Duplicated SNPs have two levels of meaning: 1.) SNPs have same rs-names but different versions of SNP ID in chip annotation file. e.g. SNP-A IDs for Affymetrix chip. 2.) SNPs with duplicated genomic position: the combination of base pair position and chromosomal location.

Value

The output PLINK binary files after removing duplicated SNP IDs.

Author(s)

Junfang Chen

See Also

prepareAnnoFile4affy

Examples

```
## In the current working directory
bedFile <- system.file("extdata", "controlData.bed", package="Gimpute")
bimFile <- system.file("extdata", "controlData.bim", package="Gimpute")
famFile <- system.file("extdata", "controlData.fam", package="Gimpute")
chipAnnoFile <- system.file("extdata", "chipAnno.txt", package="Gimpute")
system(paste0("scp ", bedFile, bimFile, famFile, " ."))
inputPrefix <- "controlData"
chipType <- "rsIDstudy"
outputSNPdupFile <- "snpDup.txt"
outputPrefix <- "removedDoubleProbes"
## Not run: Requires an executable program PLINK, e.g.
## plink <- "/home/tools/plink"
## removedDoubleProbes(plink, inputPrefix, chipAnnoFile,
##                     chipType, outputSNPdupFile, outputPrefix)
```

`removedExclProbe` *Remove improper SNPs*

Description

Remove SNPs that may be duplicated, or with unexpected SNP names.

Usage

```
removedExclProbe(plink, inputPrefix, excludedProbeIdsFile, outputPrefix)
```

Arguments

`plink`	an executable program in either the current working directory or somewhere in the command path.
`inputPrefix`	the prefix of the input PLINK binary files.
`excludedProbeIdsFile`	a pure text file that stores the SNP IDs, one per line, which need to be removed. If it is null, then no SNPs are removed.
`outputPrefix`	the prefix of the output PLINK binary files.

Details

excludedProbeIdsFile should be defined in a plain text file in advance. Improper SNPs such as AFFX, cnvi etc.. with unexpect format must be excluded.

Value

The output PLINK binary files after removing unwanted SNP IDs.

Author(s)

Junfang Chen

Examples

```
## In the current working directory
bedFile <- system.file("extdata", "controlData.bed", package="Gimpute")
bimFile <- system.file("extdata", "controlData.bim", package="Gimpute")
famFile <- system.file("extdata", "controlData.fam", package="Gimpute")
excludedProbeIdsFile <- system.file("extdata", "excludedProbeIDs.txt",
                                    package="Gimpute")
system(paste0("scp ", bedFile, bimFile, famFile, " ."))
inputPrefix <- "controlData" ## Specify the input PLINK file prefix
outputPrefix <- "1_06_removedExclProbe"
## Not run: Requires an executable program PLINK, e.g.
## plink <- "/home/tools/plink"
## removedExclProbe(plink, inputPrefix, excludedProbeIdsFile, outputPrefix)
```

`removedInstFhet` *Remove subjects with abnormal autosomal heterozygosity deviation*

Description

Remove subjects with great autosomal heterozygosity deviation.

Usage

```
removedInstFhet(plink, Fhet, inputPrefix, outputPrefix)
```

Arguments

`plink`	an executable program in either the current working directory or somewhere in the command path.
`Fhet`	the cutoff of the autosomal heterozygosity deviation.
`inputPrefix`	the prefix of the input PLINK binary files.
`outputPrefix`	the prefix of the output PLINK binary files.

Details

If the cutoff of the autosomal heterozygosity deviation is set to be greater than 0.2, i.e. |Fhet| >= 0.2, then this analysis will automatically skip haploid markers (male X and Y chromosome markers).

Value

The output PLINK binary files after removing subjects with great autosomal heterozygosity deviation.

Author(s)

Junfang Chen

Examples

```
## In the current working directory
bedFile <- system.file("extdata", "genoUpdatedData.bed", package="Gimpute")
bimFile <- system.file("extdata", "genoUpdatedData.bim", package="Gimpute")
famFile <- system.file("extdata", "genoUpdatedData.fam", package="Gimpute")
system(paste0("scp ", bedFile, bimFile, famFile, " ."))
Fhet <- 0.2
inputPrefix <- "genoUpdatedData"
outputPrefix <- "2_06_removedInstFhet"
## Not run: Requires an executable program PLINK, e.g.
## plink <- "/home/tools/plink"
## removedInstFhet(plink, Fhet, inputPrefix, outputPrefix)
```

`removedInstMiss` *Remove subjects with missing values*

Description

Remove Subjects or instances with missingness of greater than a certain threshold.

Usage

```
removedInstMiss(plink, sampleMissCutOff, inputPrefix, outputPrefix)
```

Arguments

`plink`	an executable program in either the current working directory or somewhere in the command path.
`sampleMissCutOff`	the cutoff of the missingness for removing subjects/instances.
`inputPrefix`	the prefix of the input PLINK binary files.
`outputPrefix`	the prefix of the output PLINK binary files.

Value

The output PLINK binary files after removing subjects with a pre-defined removal cutoff.

Author(s)

Junfang Chen

Examples

```
## In the current working directory
bedFile <- system.file("extdata", "genoUpdatedData.bed", package="Gimpute")
bimFile <- system.file("extdata", "genoUpdatedData.bim", package="Gimpute")
famFile <- system.file("extdata", "genoUpdatedData.fam", package="Gimpute")
system(paste0("scp ", bedFile, bimFile, famFile, " ."))
sampleMissCutOff <- 0.02
inputPrefix <- "genoUpdatedData"
outputPrefix <- "2_05_removedInstMiss"
## Not run: Requires an executable program PLINK, e.g.
## plink <- "/home/tools/plink"
## removedInstMiss(plink, sampleMissCutOff, inputPrefix, outputPrefix)
```

`removedMaleHetX` *Remove male subjects with haploid heterozygous SNPs*

Description

Determine the frequency of male subjects that have heterozygous SNPs on chromosome X and a reasonable cutoff to remove those affect males, if chromosome X data exists.

Usage

```
removedMaleHetX(plink, inputPrefix, hhSubjCutOff = 15, outputPrefix,
  outputSubjHetFile, outputRetainSubjectFile, outputHetSNPfile)
```

Arguments

`plink`	an executable program in either the current working directory or somewhere in the command path.
`inputPrefix`	the prefix of the input PLINK binary files.
`hhSubjCutOff`	the cutoff for removing male subjects with haploid heterozygous SNPs on the chromosome X. The default is 15.
`outputPrefix`	the prefix of the output PLINK binary files.
`outputSubjHetFile`	the output pure text file that stores male subjects that have heterozygous SNPs with their frequency (if any), i.e. the number of .hh SNPs in this male. Lines are sorted by descending number.
`outputRetainSubjectFile`	the output pure text file that stores male subjects that have heterozygous SNPs with their frequency after subject removal (if any). Lines are sorted by descending number.
`outputHetSNPfile`	the output pure text file that stores all heterozygous SNPs with their frequency (the number of males for this SNP) , if any. Lines are sorted by descending number.

Details

A haploid heterozygous is a male genotype that is heterozygous, which could be an error given the haploid nature of the male X chromosome. In principle, one has to remove all males that have heterozygous SNPs on the chromosome X. However, too many males might be removed in some data sets. Therefore a small percentage of such males in the data set is allowed.

Value

1.) The output PLINK binary files. 2.) A pure text file with two columns: heterozygous male subjects and their corresponding heterozygous SNPs. 3.) After subject removal, a pure text file consisting of two columns: heterozygous male subjects and their corresponding heterozygous SNPs. A pure text file with two columns: all heterozygous SNPs and their frequency.

Author(s)

Junfang Chen

Examples

```
## In the current working directory
bedFile <- system.file("extdata", "genoUpdatedData.bed", package="Gimpute")
bimFile <- system.file("extdata", "genoUpdatedData.bim", package="Gimpute")
famFile <- system.file("extdata", "genoUpdatedData.fam", package="Gimpute")
system(paste0("scp ", bedFile, bimFile, famFile, " ."))
inputPrefix <- "genoUpdatedData"
hhSubjCutOff <- 15 ##  can be tuned
outputPrefix <- "2_02_removedInstHetX"
outputSubjHetFile <- "2_02_instHetXfreqAll.txt"
outputRetainSubjectFile <- "2_02_instHetXfreqRetained.txt"
outputHetSNPfile <- "2_02_snpHHfreqAll.txt"
## Not run: Requires an executable program PLINK, e.g.
## plink <- "/home/tools/plink"
## removedMaleHetX(plink, inputPrefix, hhSubjCutOff,
##                 outputPrefix, outputSubjHetFile,
##                 outputRetainSubjectFile, outputHetSNPfile)
```

---

removedMendelErr *Check Mendel errors for family-based data*

Description

Exclude subjects and/or genetic variants (SNPs) based on Mendel errors in the family data (trio/duo).

Usage

```
removedMendelErr(plink, inputPrefix, cutoffSubject = 0.05,
  cutoffSNP = 0.1, outputPrefix)
```

Arguments

`plink`	an executable program in either the current working directory or somewhere in the command path.
`inputPrefix`	the prefix of the input PLINK binary files.
`cutoffSubject`	the cutoff determines that families (subjects) with more than the predefined cutoff of Mendel errors by considering all SNPs will be removed. The default is 0.05.
`cutoffSNP`	the cutoff indicates that SNPs with more than the predefined cutoff of Mendel error rate will be excluded (i.e. based on the number of trios/duos). The default is 0.1.
`outputPrefix`	the prefix of the output PLINK binary files.

Details

Do make sure that all your family relationships are correct in your input data before applying this function. The input PLINK data should have complete sex and group/outcome information. By default, trios and duos are both considered. If no family information is given at all (only founders), then this function will not remove any subjects or variants but give the warning showing that no duos or trios are present.

Value

The output PLINK binary files.

Author(s)

Junfang Chen

Examples

```
## In the current working directory
bedFile <- system.file("extdata", "dataWithFamChr21.bed", package="Gimpute")
bimFile <- system.file("extdata", "dataWithFamChr21.bim", package="Gimpute")
famFile <- system.file("extdata", "dataWithFamChr21.fam", package="Gimpute")
system(paste0("scp ", bedFile, bimFile, famFile, " ."))
inputPrefix <- "dataWithFamChr21"
outputPrefix <- "removedMendelErr"
## Not run: Requires an executable program PLINK, e.g.
## plink <- "/home/tools/plink"
## removedMendelErr(plink, inputPrefix,
##                  cutoffSubject, cutoffSNP, outputPrefix)
```

---

removedMonoSnp *Exclude monomorphic SNPs*

Description

Detect monomorphic SNPs from PLINK BIM file and exclude them if any.

Usage

```
removedMonoSnp(plink, inputPrefix, outputPrefix, outputSNPfile)
```

Arguments

`plink`	an executable program in either the current working directory or somewhere in the command path.
`inputPrefix`	the prefix of the input PLINK binary files.
`outputPrefix`	the prefix of the output PLINK binary files after removing monomorphic SNPs.
`outputSNPfile`	the output pure text file that stores the removed monomorphic SNPs, one per line, if any.

Value

The output PLINK binary files after removing monomorphic SNPs and a pure text file with removed monomorphic SNPs.

Author(s)

Junfang Chen

Examples

```
## In the current working directory
bedFile <- system.file("extdata", "alignedData.bed", package="Gimpute")
bimFile <- system.file("extdata", "alignedData.bim", package="Gimpute")
famFile <- system.file("extdata", "alignedData.fam", package="Gimpute")
system(paste0("scp ", bedFile, " ."))
system(paste0("scp ", bimFile, " ."))
system(paste0("scp ", famFile, " ."))
inputPrefix <- "alignedData"
outputPrefix <- "removedMonoSnp"
outputSNPfile <- "monoSNP.txt"
## Not run: Requires an executable program PLINK, e.g.
## plink <- "/home/tools/plink"
## removedMonoSnp(plink, inputPrefix, outputPrefix, outputSNPfile)
```

`removedParentIdsMiss`

Reset paternal and maternal codes

Description

Reset paternal and maternal codes of non-founders if parents not present. Replace the paternal ID and maternal ID of subjects (childs) by the value zero if the paternal ID and the maternal ID do not belong to any subject (parent) with the same family ID as the child.

Usage

```
removedParentIdsMiss(plink, inputPrefix, outputPrefix)
```

Arguments

`plink`	an executable program in either the current working directory or somewhere in the command path.
`inputPrefix`	the prefix of the input PLINK binary files.
`outputPrefix`	the prefix of the output PLINK binary files.

Details

Do make sure that all your family relationships are correct in your input data before applying this function. By default, if parental IDs are provided for a sample, they are not treated as a founder even if neither parent is in the dataset. With no modifiers, –make-founders clears both parental IDs whenever at least one parent is not in the dataset, and the affected samples are now considered as founders.

Value

The output PLINK binary files.

Author(s)

Junfang Chen

Examples

```
## In the current working directory
bedFile <- system.file("extdata", "genoUpdatedData.bed", package="Gimpute")
bimFile <- system.file("extdata", "genoUpdatedData.bim", package="Gimpute")
famFile <- system.file("extdata", "genoUpdatedData.fam", package="Gimpute")
system(paste0("scp ", bedFile, bimFile, famFile, " ."))
inputPrefix <- "genoUpdatedData"
outputPrefix <- "2_07_removedParentIdsMiss"
## Not run: Requires an executable program PLINK, e.g.
## plink <- "/home/tools/plink"
## removedParentIdsMiss(plink, inputPrefix, outputPrefix)
```

---

removedSnpFemaleChrXhweControl

Hardy Weinberg Equilibrium test for chromosome X SNPs in female controls.

Description

Hardy Weinberg Equilibrium test for SNPs on the chromosome X in female controls.

Usage

```
removedSnpFemaleChrXhweControl(plink, inputPrefix, pval = 1e-06,
  outputPvalFile, outputSNPfile, outputPrefix)
```

Arguments

`plink`	an executable program in either the current working directory or somewhere in the command path.
`inputPrefix`	the prefix of the input PLINK binary files.
`pval`	the p-value cutoff for controlling HWE test in female control subjects. Only chromosome X SNPs are considered. The default value is 0.000001.
`outputPvalFile`	the output pure text file that stores chromosome X SNPs and their sorted HWE p-values.
`outputSNPfile`	the output pure text file that stores the removed SNPs, one per line.
`outputPrefix`	the prefix of the output PLINK binary files.

Value

The output PLINK binary files after HWE test on chromosomal X in female controls.

Author(s)

Junfang Chen

See Also

`removedSnpHWEauto`

Examples

```
## In the current working directory
bedFile <- system.file("extdata", "genoUpdatedData.bed", package="Gimpute")
bimFile <- system.file("extdata", "genoUpdatedData.bim", package="Gimpute")
famFile <- system.file("extdata", "genoUpdatedData.fam", package="Gimpute")
system(paste0("scp ", bedFile, bimFile, famFile, " ."))
inputPrefix <- "genoUpdatedData"
outputPvalFile <- "2_12_snpHwePvalfemaleXct.txt"
outputSNPfile <- "2_12_snpRemovedHweFemaleXct.txt"
outputPrefix <- "2_12_removedSnpHweFemaleX"
## Not run: Requires an executable program PLINK, e.g.
## plink <- "/home/tools/plink"
## removedSnpFemaleChrXhweControl(plink, inputPrefix, pval=0.000001,
##                                outputPvalFile, outputSNPfile,
##                                outputPrefix)
```

`removedSnpFemaleChrXmiss` *remove chromosome X SNPs in females*

Description

Remove SNPs on the chromosome X with a pre-defined cutoff for missingness in females.

Usage

```
removedSnpFemaleChrXmiss(plink, femaleChrXmissCutoff, inputPrefix,
  outputPrefix)
```

Arguments

`plink`	an executable program in either the current working directory or somewhere in the command path.
`femaleChrXmissCutoff`	the cutoff of the missingness in female chromosome X SNPs.
`inputPrefix`	the prefix of the input PLINK binary files.
`outputPrefix`	the prefix of the output PLINK binary files.

Value

The output PLINK binary files.

Author(s)

Junfang Chen

Examples

```
## In the current working directory
bedFile <- system.file("extdata", "genoUpdatedData.bed", package="Gimpute")
bimFile <- system.file("extdata", "genoUpdatedData.bim", package="Gimpute")
famFile <- system.file("extdata", "genoUpdatedData.fam", package="Gimpute")
system(paste0("scp ", bedFile, bimFile, famFile, " ."))
femaleChrXmissCutoff <- 0.05
inputPrefix <- "genoUpdatedData"
outputPrefix <- "2_10_removedSnpFemaleChrXmiss"
## Not run: Requires an executable program PLINK, e.g.
## plink <- "/home/tools/plink"
## removedSnpFemaleChrXmiss(plink, femaleChrXmissCutoff,
##                          inputPrefix, outputPrefix)
```

---

removedSnpHetX *Remove heterozygous SNPs in male chromosome X*

Description

Remove heterozygous SNPs in haploid male chromosome X only if chromosome X data exists.

Usage

```
removedSnpHetX(plink, inputPrefix, hhCutOff, outputPrefix,
  outputHetSNPfile, outputRetainSNPfile)
```

Arguments

`plink`	an executable program in either the current working directory or somewhere in the command path.
`inputPrefix`	the prefix of the input PLINK binary files.
`hhCutOff`	the cutoff for removing male haploid heterozygous SNPs on the chromosome X. The default is 0.005.
`outputPrefix`	the prefix of the output PLINK binary files.
`outputHetSNPfile`	the output pure text file that stores all heterozygous SNPs with their frequency (the number of males for the corresponding SNP), if any. Lines are sorted by descending number.
`outputRetainSNPfile`	the output pure text file that stores retained heterozygous SNPs with their frequency (the number of males for the corresponding SNP), if any. Lines are sorted by descending number.

Details

A haploid heterozygous is a male genotype that is heterozygous, which could be an error given the haploid nature of the male X chromosome. In principle, one has to remove all heterozygous SNPs of chromosome X in males. However, too many SNPs might be removed in some data sets. Therefore a small percentage of such SNPs in the data set is allowed.

Value

1.) The output PLINK binary files. 2.) A pure text files (if any) with two columns: SNPs and their corresponding frequency. 3.) After SNP removal, a pure text files (if any) with two columns: SNPs and their corresponding frequency.

Author(s)

Junfang Chen

Examples

```
## In the current working directory
bedFile <- system.file("extdata", "genoUpdatedData.bed", package="Gimpute")
bimFile <- system.file("extdata", "genoUpdatedData.bim", package="Gimpute")
famFile <- system.file("extdata", "genoUpdatedData.fam", package="Gimpute")
system(paste0("scp ", bedFile, bimFile, famFile, " ."))
inputPrefix <- "genoUpdatedData"
hhCutOff <- 0.005 ##  can be tuned
outputPrefix <- "2_01_removedSnpHetX"
outputHetSNPfile <- "2_01_snpHHfreqAll.txt"
outputRetainSNPfile <- "2_01_snpHHfreqRetained.txt"
## Not run: Requires an executable program PLINK, e.g.
## plink <- "/home/tools/plink"
## removedSnpHetX(plink, inputPrefix, hhCutOff, outputPrefix,
##                outputHetSNPfile, outputRetainSNPfile)
```

---

removedSnpHWEauto *Hardy Weinberg Equilibrium test for autosomal SNPs*

Description

Remove autosomal SNPs deviating from Hardy Weinberg Equilibrium (HWE).

Usage

```
removedSnpHWEauto(groupLabel, plink, inputPrefix, pval = 1e-06,
  outputPvalFile, outputSNPfile, outputPrefix)
```

Arguments

`groupLabel`	a string value indicating the outcome label: "control", or, "case" or "caseControl" for both existing groups. For more details, see `getGroupLabel`.
`plink`	an executable program in either the current working directory or somewhere in the command path.
`inputPrefix`	the prefix of the input PLINK binary files.
`pval`	the p-value cutoff for controlling HWE test in either control or case subjects. Only autosomal SNPs are considered. The default value is 0.000001.
`outputPvalFile`	the output pure text file that stores autosomal SNPs and their sorted HWE p-values.
`outputSNPfile`	the output pure text file that stores the removed SNPs, one per line.
`outputPrefix`	the prefix of the output PLINK binary files.

Value

The output PLINK binary files after HWE test on the autosome.

Author(s)

Junfang Chen

See Also

removedSnpFemaleChrXhweControl, getGroupLabel.

Examples

```
## In the current working directory
bedFile <- system.file("extdata", "genoUpdatedData.bed", package="Gimpute")
bimFile <- system.file("extdata", "genoUpdatedData.bim", package="Gimpute")
famFile <- system.file("extdata", "genoUpdatedData.fam", package="Gimpute")
system(paste0("scp ", bedFile, bimFile, famFile, " ."))
groupLabel <- "control"
inputPrefix <- "genoUpdatedData" ## Specify the input PLINK file prefix
outputPvalFile <- "2_11_snpHwePvalAuto.txt"
outputSNPfile <- "2_11_snpRemovedHweAuto.txt"
outputPrefix <- "2_11_removedSnpHweAuto"
## Not run: Requires an executable program PLINK, e.g.
## plink <- "/home/tools/plink"
## removedSnpHWEauto(groupLabel, plink, inputPrefix, pval=0.000001,
##                   outputPvalFile, outputSNPfile, outputPrefix)
```

---

removedSnpMiss *Remove SNPs with missing values*

Description

Remove SNPs with missingness of greater than a certain threshold.

Usage

```
removedSnpMiss(plink, snpMissCutOff, inputPrefix, outputPrefix)
```

Arguments

plink	an executable program in either the current working directory or somewhere in the command path.
snpMissCutOff	the cutoff of the missingness for removing SNPs.
inputPrefix	the prefix of the input PLINK binary files.
outputPrefix	the prefix of the output PLINK binary files.

Value

The output PLINK binary files after removing SNPs with pre-defined missing values.

Author(s)

Junfang Chen

Examples

```
## In the current working directory
bedFile <- system.file("extdata", "genoUpdatedData.bed", package="Gimpute")
bimFile <- system.file("extdata", "genoUpdatedData.bim", package="Gimpute")
famFile <- system.file("extdata", "genoUpdatedData.fam", package="Gimpute")
system(paste0("scp ", bedFile, bimFile, famFile, " ."))
snpMissCutOff <- 0.05
inputPrefix <- "genoUpdatedData"
outputPrefix <- "2_04_removedSnpMissPre"
## Not run: Requires an executable program PLINK, e.g.
## plink <- "/home/tools/plink"
## removedSnpMiss(plink, snpMissCutOff, inputPrefix, outputPrefix)
```

---

removedSnpMissDiff *Remove SNPs with difference in SNP missingness between cases and controls.*

Description

Remove SNPs with difference in SNP missingness between cases and controls. To test for differential call rates between cases and controls for each SNP

Usage

```
removedSnpMissDiff(plink, inputPrefix, snpMissDifCutOff, outputPrefix,
  groupLabel)
```

Arguments

`plink`	an executable program in either the current working directory or somewhere in the command path.
`inputPrefix`	the prefix of the input PLINK binary files.
`snpMissDifCutOff`	the cutoff of the difference in missingness between cases and controls.
`outputPrefix`	the prefix of the output PLINK binary files.
`groupLabel`	a string value indicating the outcome label: "control", or, "case" or "caseControl" for both existing groups. For more details, see `getGroupLabel`.

Details

Only if both case-control groups exist in the input genotype data, differential SNPs are removed.

Value

The output PLINK binary files.

Author(s)

Junfang Chen

See Also

getGroupLabel.

Examples

```
## In the current working directory
bedFile <- system.file("extdata", "genoUpdatedData.bed", package="Gimpute")
bimFile <- system.file("extdata", "genoUpdatedData.bim", package="Gimpute")
famFile <- system.file("extdata", "genoUpdatedData.fam", package="Gimpute")
system(paste0("scp ", bedFile, bimFile, famFile, " ."))
inputPrefix <- "genoUpdatedData"
snpMissDifCutOff <- 0.02
outputPrefix <- "2_09_removedSnpMissDiff"
groupLabel <- "control"
## Not run: Requires an executable program PLINK, e.g.
## plink <- "/home/tools/plink"
## removedSnpMissDiff(plink, inputPrefix, snpMissDifCutOff,
##                    outputPrefix, groupLabel)
```

---

removedSnpMissPostImp

Remove SNPs after post imputation

Description

Remove SNPs which have a non missing value for less than a predefined number of instances.

Usage

```
removedSnpMissPostImp(plink, inputPrefix, missCutoff, outRemovedSNPfile,
  outRetainSNPfile, outputPrefix)
```

Arguments

plink	an executable program in either the current working directory or somewhere in the command path.
inputPrefix	the prefix of the input PLINK binary files.
missCutoff	the cutoff of the least number of instances for a SNP that is not missing. The default is 20.
outRemovedSNPfile	the output file of SNPs with pre-defined missing values that are removed.
outRetainSNPfile	the output file of SNPs that are retained.
outputPrefix	the prefix of the PLINK binary files.

Value

The PLINK binary files after post imputation quality control and a pure text file contains SNPs with pre-defined missing values.

Author(s)

Junfang Chen

Examples

```
## In the current working directory
bedFile <- system.file("extdata", "alignedData.bed", package="Gimpute")
bimFile <- system.file("extdata", "alignedData.bim", package="Gimpute")
famFile <- system.file("extdata", "alignedData.fam", package="Gimpute")
system(paste0("scp ", bedFile, " ."))
system(paste0("scp ", bimFile, " ."))
system(paste0("scp ", famFile, " ."))
inputPrefix <- "alignedData"
outputPrefix <- "removedSnpMissPostImp"
## Not run: Requires an executable program PLINK, e.g.
## plink <- "/home/tools/plink"
## removedSnpMissPostImp(plink, inputPrefix, missCutoff=20,
##                       outRemovedSNPfile, outRetainSNPfile, outputPrefix)
##
```

---

removedUnmapProbes *Remove SNPs not in the chip annotation file*

Description

Check if all SNPs are included in the chip annotation file. If some of the input SNPs are not included, then remove them.

Usage

```
removedUnmapProbes(plink, inputPrefix, chipAnnoFile, chipType,
  outputPrefix, outputSNPfile)
```

Arguments

plink	an executable program in either the current working directory or somewhere in the command path.
inputPrefix	the prefix of the input PLINK binary files.
chipAnnoFile	a pure text file that stores the chip annotation information.
chipType	a string name defines the type of the chip annotation file: 'SNPIDstudy', and 'rsIDstudy'. The detail is described in prepareAnnoFile4affy.
outputPrefix	the prefix of the output PLINK binary files.
outputSNPfile	a pure text file that stores the SNP IDs, one per line, which are not mapped to the chip annotation file.

Details

If the chip annotation file is not available for your study, you can download it from http://www.well.ox.ac.uk/~wrayner/str

Value

The output text file contains the removed SNP IDs, one per line. The PLINK binary files after removing unmapped SNP IDs.

Author(s)

Junfang Chen

See Also

prepareAnnoFile4affy

Examples

```
## In the current working directory
bedFile <- system.file("extdata", "controlData.bed", package="Gimpute")
bimFile <- system.file("extdata", "controlData.bim", package="Gimpute")
famFile <- system.file("extdata", "controlData.fam", package="Gimpute")
chipAnnoFile <- system.file("extdata", "chipAnno.txt", package="Gimpute")
system(paste0("scp ", bedFile, bimFile, famFile, " ."))
inputPrefix <- "controlData" ## Specify the input PLINK file prefix
outputSNPfile <- "1_07_removedUnmapProbes"
## Not run: Requires an executable program PLINK, e.g.
## plink <- "/home/tools/plink"
## removedUnmapProbes(plink, inputPrefix, chipAnnoFile, chipType,
##                    outputPrefix, outputSNPfile)
```

---

removedWrongAnceInst

Remove samples with incorrect ancestry

Description

Remove samples with the incorrect ancestry or keep samples at your own chioce.

Usage

```
removedWrongAnceInst(plink, inputPrefix, metaDataFile, ancestrySymbol,
  outputPrefix)
```

Arguments

plink	an executable program in either the current working directory or somewhere in the command path.
inputPrefix	the prefix of the input PLINK binary files.

`metaDataFile` a pure text file that stores the meta information of the samples. This file must contain at least the following content (column names are in parentheses): family ID in the PLINK files (FID), individual ID in the PLINK files (IID), ID in the description files (descID), self identified ancestry (ance; e.g. AFR: African, AMR: Ad Mixed American, EAS: East Asian, EUR: European, SAS: South Asian), sex (sex; 1 = male, 2 = female, 0 = missing), age (age), group (group; 0 = control/unaffected, 1 = case/affected). All unknown and missing values are represented by the value NA. Lines with a missing value for FID or IID are not contained.

`ancestrySymbol` an indicator that shows the symbol of genetic ancestry. If it is null, then all samples are selected.

`outputPrefix` the prefix of the output PLINK binary files.

Details

ancestrySymbol, such as 'EUR' stands for the European, 'EAS' for East Asian. See the metaDataFile for more details.

Value

The output PLINK binary files after checking the ancestry information.

Author(s)

Junfang Chen

Examples

```
## In the current working directory
bedFile <- system.file("extdata", "controlData.bed", package="Gimpute")
bimFile <- system.file("extdata", "controlData.bim", package="Gimpute")
famFile <- system.file("extdata", "controlData.fam", package="Gimpute")
metaDataFile <- system.file("extdata", "1_01_metaData.txt",
                            package="Gimpute")
system(paste0("scp ", bedFile, bimFile, famFile, " ."))
inputPrefix <- "controlData" ## Specify the input PLINK file prefix
ancestrySymbol <- "EAS"
outputPrefix <- "1_05_removedWrongAnceInst"
## Not run: Requires an executable program PLINK, e.g.
## plink <- "/home/tools/plink"
## removedWrongAnceInst(plink, inputPrefix, metaDataFile,
##                      ancestrySymbol, outputPrefix)
```

---

removedYMtSnp *Remove SNPs on the chromosome Y and mitochondrial DNA*

Description

Remove SNPs on the chromosome Y and mitochondrial DNA.

Usage

```
removedYMtSnp(plink, inputPrefix, outputPrefix)
```

Arguments

plink	an executable program in either the current working directory or somewhere in the command path.
inputPrefix	the prefix of the input PLINK binary files.
outputPrefix	the prefix of the output PLINK binary files.

Details

Note that if chromosome Y and mitochondrial DNA are available, they must be coded as 24 and 26, respectively.

Value

The output PLINK binary files after removing SNPs on the chromosome Y and mitochondrial DNA.

Author(s)

Junfang Chen

Examples

```
## In the current working directory
bedFile <- system.file("extdata", "controlData.bed", package="Gimpute")
bimFile <- system.file("extdata", "controlData.bim", package="Gimpute")
famFile <- system.file("extdata", "controlData.fam", package="Gimpute")
system(paste0("scp ", bedFile, bimFile, famFile, " ."))
inputPrefix <- "controlData" ## Specify the input PLINK file prefix
outputPrefix <- "1_11_removedYMtSnp"
## Not run: Requires an executable program PLINK, e.g.
## plink <- "/home/tools/plink"
## removedYMtSnp(plink, inputPrefix, outputPrefix)
```

---

removeNoGroupId *Remove samples without group information*

Description

Remove samples without group/outcome/phenotype information, which is coded as -9 in the PLINK .FAM file.

Usage

```
removeNoGroupId(plink, inputPrefix, outputPrefix)
```

Arguments

plink	an executable program in either the current working directory or somewhere in the command path.
inputPrefix	the prefix of the input PLINK binary files.
outputPrefix	the prefix of the output PLINK binary files.

Value

The output PLINK binary files after removing samples without group information.

Author(s)

Junfang Chen

Examples

```
## In the current working directory
bedFile <- system.file("extdata", "controlData.bed", package="Gimpute")
bimFile <- system.file("extdata", "controlData.bim", package="Gimpute")
famFile <- system.file("extdata", "controlData.fam", package="Gimpute")
system(paste0("scp ", bedFile, bimFile, famFile, " ."))
inputPrefix <- "controlData" ## Specify the input PLINK file prefix
outputPrefix <- "1_04_removedNoGroupId"
## Not run: Requires an executable program PLINK, e.g.
## plink <- "/home/tools/plink"
## removeNoGroupId(plink, inputPrefix, outputPrefix)
```

---

removeOutlierByPCs *Remove population outliers*

Description

Remove population outliers by using principle component analysis.

Usage

```
removeOutlierByPCs(plink, gcta, inputPrefix, nThread = 20,
  cutoff = NULL, cutoffSign, inputPC4subjFile, outputPC4outlierFile,
  outputPCplotFile, outputPrefix)
```

Arguments

plink	an executable program in either the current working directory or somewhere in the command path.
gcta	an executable program in either the current working directory or somewhere in the command path.
inputPrefix	the prefix of the input PLINK binary files.
nThread	the number of threads used for computation. The default is 20.
cutoff	the cutoff that distinguishes the outliers from ordinary population using PCA. If it is null, then there are no outliers or outliers are not required to be removed. The default is NULL.

`cutoffSign` the cutoff sign: 'greater' or 'smaller' that determines if the outliers should be greater or smaller than the cutoff value.

`inputPC4subjFile`
the pure text file that stores all the subject IDs and their corresponding eigenvalues of the first two principle components.

`outputPC4outlierFile`
the pure text file that stores the outlier IDs and their corresponding eigenvalues of the first two principle components.

`outputPCplotFile`
the plot file for visualizing the first two principle components of all subjects without population outliers.

`outputPrefix` the prefix of the output PLINK binary files.

Details

This function is used for removing population outliers. If the outliers are necessary to be removed, then one uses the eigenvalues from the first principle component as a criterion to find out the outliers by assigning an appropriate cutoff.

Value

1.) The output PLINK binary files after outlier removal. 2.) The output pure text file (if any) for storing removed outlier IDs and their corresponding PCs. 3.) The plot file (if any) for visualizing the first two principle components after outlier removal.

Author(s)

Junfang Chen

Examples

```
## In the current working directory
bedFile <- system.file("extdata", "QCdata.bed", package="Gimpute")
bimFile <- system.file("extdata", "QCdata.bim", package="Gimpute")
famFile <- system.file("extdata", "QCdata.fam", package="Gimpute")
system(paste0("scp ", bedFile, bimFile, famFile, " ."))
inputPrefix <- "QCdata"
cutoff <- NULL ## no outlier to be removed
cutoffSign <- "greater" ## not used if cutoff == NULL
inputPC4subjFile <- "2_13_eigenvalAfterQC.txt"
outputPC4outlierFile <- "2_13_eigenval4outliers.txt"
outputPCplotFile <- "2_13_removedOutliers.png"
outputPrefix <- "2_13_removedOutliers"
## Not run: Requires an executable program PLINK and GCTA, e.g.
## plink <- "/home/tools/plink"
## gcta <- "/home/tools/gcta64"
## removeOutlierByPCs(plink, gcta, inputPrefix, nThread=20,
##                    cutoff, cutoffSign, inputPC4subjFile,
##                    outputPC4outlierFile, outputPCplotFile, outputPrefix)
```

`removeSampID` *Remove samples in PLINK files*

Description

Remove sample IDs that are useless such as duplicated or related IDs from PLINK binary files. These IDs are defined in a plain text file.

Usage

```
removeSampID(plink, removedSampIDFile, inputPrefix, outputPrefix)
```

Arguments

`plink`	an executable program in either the current working directory or somewhere in the command path.
`removedSampIDFile`	a pure text file that stores the useless sample IDs, each ID per line. If it is null, then duplicate the input PLINK files as the output files.
`inputPrefix`	the prefix of the input PLINK binary files.
`outputPrefix`	the prefix of the output PLINK binary files.

Value

The output PLINK binary files after removing certain sample IDs.

Author(s)

Junfang Chen

Examples

```
## In the current working directory
bedFile <- system.file("extdata", "controlData.bed", package="Gimpute")
bimFile <- system.file("extdata", "controlData.bim", package="Gimpute")
famFile <- system.file("extdata", "controlData.fam", package="Gimpute")
system(paste0("scp ", bedFile, bimFile, famFile, " ."))
inputPrefix <- "controlData" ## Specify the input PLINK file prefix
removedSampIDFile <- system.file("extdata", "excludedSampIDsV1.txt",
                                 package="Gimpute")
outputPrefix <- "1_02_removedExclInst"
## Not run: Requires an executable program PLINK, e.g.
## plink <- "/home/tools/plink"
## removeSampID(plink, removedSampIDFile, inputPrefix, outputPrefix)
```

`renamePlinkBFile` *Rename PLINK binary files*

Description

Rename a set of PLINK binary files (.BED, .BIM and .FAM).

Usage

```
renamePlinkBFile(inputPrefix, outputPrefix, action)
```

Arguments

`inputPrefix`	the prefix of the input PLINK binary files.
`outputPrefix`	the prefix of the output PLINK binary files.
`action`	a string indicating if the action is "copy" or "move".

Details

The original input files can be retained using the action "copy" or removed by using "move".

Value

Renamed PLINK binary files.

Author(s)

Junfang Chen

Examples

```
## In the current working directory
bedFile <- system.file("extdata", "controlData.bed", package="Gimpute")
bimFile <- system.file("extdata", "controlData.bim", package="Gimpute")
famFile <- system.file("extdata", "controlData.fam", package="Gimpute")
system(paste0("scp ", bedFile, bimFile, famFile, " ."))
inputPrefix <- "controlData"
outputPrefix <- "dataCtl"
outputPrefix <- "1_02_removedExclInst"
## renamePlinkBFile(inputPrefix, outputPrefix, action="move")
```

setHeteroHaploMissing *Set haploid heterozygous SNPs as missing*

Description

Set all heterozygous alleles of chromosome X SNPs in male as missing.

Usage

```
setHeteroHaploMissing(plink, inputPrefix, outputPrefix)
```

Arguments

plink	an executable program in either the current working directory or somewhere in the command path.
inputPrefix	the prefix of the input PLINK binary files.
outputPrefix	the prefix of the output PLINK binary files.

Value

The output PLINK binary files after setting haploid heterozygous SNPs as missing.

Author(s)

Junfang Chen

Examples

```
## In the current working directory
bedFile <- system.file("extdata", "genoUpdatedData.bed", package="Gimpute")
bimFile <- system.file("extdata", "genoUpdatedData.bim", package="Gimpute")
famFile <- system.file("extdata", "genoUpdatedData.fam", package="Gimpute")
system(paste0("scp ", bedFile, bimFile, famFile, " ."))
inputPrefix <- "genoUpdatedData"
outputPrefix <- "2_03_setHeteroHaploMissing"
## Not run: Requires an executable program PLINK, e.g.
## plink <- "/home/tools/plink"
## setHeteroHaploMissing(plink, inputPrefix, outputPrefix)
```

splitXchr *Split chromosome X into pseudoautosomal region and non-pseudoautosomal region.*

Description

Split chromosome X into pseudoautosomal region and non-pseudoautosomal region, if chromosome X data is available.

Usage

```
splitXchr(plink, inputPrefix, outputPrefix)
```

Arguments

plink	an executable program in either the current working directory or somewhere in the command path.
inputPrefix	the prefix of the input PLINK binary files.
outputPrefix	the prefix of the output PLINK binary files.

Details

Genomic coordinate system is on genome build hg19.

Value

The output PLINK binary files after splitting chromosome X into pseudoautosomal region and non-pseudoautosomal region.

Author(s)

Junfang Chen

Examples

```
## In the current working directory
bedFile <- system.file("extdata", "controlData.bed", package="Gimpute")
bimFile <- system.file("extdata", "controlData.bim", package="Gimpute")
famFile <- system.file("extdata", "controlData.fam", package="Gimpute")
system(paste0("scp ", bedFile, bimFile, famFile, " ."))
inputPrefix <- "controlData" ## Specify the input PLINK file prefix
outputPrefix <- "1_10_splitXchr"
## Not run: Requires an executable program PLINK, e.g.
## plink <- "/home/tools/plink"
## removeSampID(plink, inputPrefix, outputPrefix)
```

---

updatedSnpInfo *Update the SNP information*

Description

Update SNP information including SNP name, base-pair position, chromosomal location and the strand information.

Usage

```
updatedSnpInfo(plink, inputPrefix, chipAnnoFile, chipType, outputPrefix)
```

Arguments

plink	an executable program in either the current working directory or somewhere in the command path.
inputPrefix	the prefix of the input PLINK binary files.
chipAnnoFile	a pure text file that stores the chip annotation information. If the chip annotation file is not available for your study, it can be downloaded from http://www.well.ox.ac.uk/~wrayner/stra
chipType	a string name defines the type of the chip annotation file: 'SNPIDstudy', and 'rsIDstudy'. The detail is described in prepareAnnoFile4affy.
outputPrefix	the prefix of the output PLINK binary files.

Details

The SNP information in the chip annotation file is used as the reference.

Value

The output PLINK binary files after updating SNP information.

Author(s)

Junfang Chen

See Also

prepareAnnoFile4affy

Examples

```
## In the current working directory
bedFile <- system.file("extdata", "controlData.bed", package="Gimpute")
bimFile <- system.file("extdata", "controlData.bim", package="Gimpute")
famFile <- system.file("extdata", "controlData.fam", package="Gimpute")
chipAnnoFile <- system.file("extdata", "chipAnno.txt", package="Gimpute")
system(paste0("scp ", bedFile, bimFile, famFile, " ."))
inputPrefix <- "controlData"
chipType <- "rsIDstudy"
outputPrefix <- "updatedSnpInfo"
## Not run: Requires an executable program PLINK, e.g.
## plink <- "/home/tools/plink"
## updatedSnpInfo(plink, inputPrefix, chipAnnoFile,
##                chipType, outputPrefix, outputSNPfile)
```

---

updateGenoInfo *Update genotype information*

Description

Update genotype information of the original PLINK binary files involving subject metadata information remapping and SNP information rearrangement and conversion according to the annotation file.

Usage

```
updateGenoInfo(plink, inputPrefix, metaDataFile, removedSampIDFile,
  ancestrySymbol, excludedProbeIdsFile, chipAnnoFile, chipType,
  outputPrefix, keepInterFile = TRUE)
```

Arguments

plink	an executable program in either the current working directory or somewhere in the command path.
inputPrefix	the prefix of the input PLINK binary files.
metaDataFile	a pure text file that stores the meta information of the samples. This file must contain at least the following content (column names are in parentheses): family ID in the PLINK files (FID), individual ID in the PLINK files (IID), ID in the description files (descID), self identified ancestry (ance; e.g. AFR: African, AMR: Ad Mixed American, EAS: East Asian, EUR: European, SAS: South Asian), sex (sex; 1 = male, 2 = female), age (age), group (group; 0 = control/unaffected, 1 = case/affected). All unknown and missing values are represented by the value NA. Lines with a missing value for FID or IID are not contained.
removedSampIDFile	a pure text file that stores the useless sample IDs, each ID per line. If it is null, then duplicate the input PLINK files from the last step as the output files.
ancestrySymbol	an indicator that shows the symbol of genetic ancestry. If it is null, then all samples are selected.
excludedProbeIdsFile	a pure text file that stores the SNP IDs, one per line, which need to be removed. If it is null, no SNPs are removed.
chipAnnoFile	a pure text file that stores the chip annotation information.
chipType	a string name defines the type of the chip annotation file: 'SNPIDstudy', and 'rsIDstudy'. The detail is described in prepareAnnoFile4affy.
outputPrefix	the prefix of the output PLINK binary files.
keepInterFile	a logical value indicating if the intermediate processed files should be kept or not. The default is TRUE.

Details

The original PLINK files are implicitly processed by the following steps: 1.) remove duplicated subjects; 2.) update group ID and sex information; 3.) remove not labelled subjects; 4.) remove subjects with wrong ancestry; 5.) remove incorrectly annotated SNPs; 6.) remove SNPs that are not in the annotation file; 7.) remove duplicated SNPs; 8.) update SNP genomic position and strand information; 9.) split chromosome X into pseudoautosomal region (PAR) and non-PAR; 10.) remove SNPs on the chromosome Y and mitochondrial DNA. The metadata information file and the chip annotation file are used as the reference for the update. If the chip annotation file is not available for your study, it can be downloaded from http://www.well.ox.ac.uk/~wrayner/strand/.

Value

The output PLINK binary files after genotype information remapping.

Author(s)

Junfang Chen

References

Purcell, Shaun, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics 81.3 (2007): 559-575.

Examples

```
## In the current working directory
bedFile <- system.file("extdata", "controlData.bed", package="Gimpute")
bimFile <- system.file("extdata", "controlData.bim", package="Gimpute")
famFile <- system.file("extdata", "controlData.fam", package="Gimpute")
system(paste0("scp ", bedFile, bimFile, famFile, " ."))
inputPrefix <- "controlData"
metaDataFile <- system.file("extdata", "1_01_metaData.txt",
                            package="Gimpute")
excludedProbeIdsFile <- system.file("extdata", "excludedProbeIDs.txt",
                                    package="Gimpute")
removedSampIDFile <- system.file("extdata", "excludedSampIDsV1.txt",
                                 package="Gimpute")
chipAnnoFile <- system.file("extdata", "chipAnno.txt", package="Gimpute")
ancestrySymbol <- "EUR"
outputPrefix <- "1_11_removedYMtSnp"
metaDataFile <- "1_01_metaData.txt"
chipType <- "rsIDstudy"
## Not run: Requires an executable program PLINK, e.g.
## plink <- "/home/tools/plink"
## updateGenoInfo(plink, inputPrefix, metaDataFile, removedSampIDFile,
##                ancestrySymbol, excludedProbeIdsFile, chipAnnoFile,
##                chipType, outputPrefix, keepInterFile=TRUE)
```

---

updateGroupIdAndSex

Update group and geneder information

Description

Replace group and gender information in the PLINK binary files by using the information from the metadata file.

Usage

```
updateGroupIdAndSex(plink, inputPrefix, metaDataFile, outputPrefix)
```

Arguments

plink	an executable program in either the current working directory or somewhere in the command path.
inputPrefix	the prefix of the input PLINK binary files.

`metaDataFile` a pure text file that stores the meta information of the samples. This file must contain at least the following content (column names are in parentheses): family ID in the PLINK files (FID), individual ID in the PLINK files (IID), ID in the description files (descID), self identified ancestry (ance; e.g. AFR: African, AMR: Ad Mixed American, EAS: East Asian, EUR: European, SAS: South Asian), sex (sex; 1 = male, 2 = female), age (age), group (group; 0 = control/unaffected, 1 = case/affected). All unknown and missing values are represented by the value NA. Lines with a missing value for FID or IID are not contained.

`outputPrefix` the prefix of the output PLINK binary files.

Details

Find the shared sample IDs between PLINK input files and metadata file. Use the information from the metadata file as the reference and update the group information/the outcome in the PLINK file. Group label should be 1 and 2. (1=unaff, 2=aff, 0=miss); missing phenotype will be indicated as -9.

Value

The output PLINK binary files after updating the gender and grouping information.

Author(s)

Junfang Chen

Examples

```
## In the current working directory
bedFile <- system.file("extdata", "controlData.bed", package="Gimpute")
bimFile <- system.file("extdata", "controlData.bim", package="Gimpute")
famFile <- system.file("extdata", "controlData.fam", package="Gimpute")
metaDataFile <- system.file("extdata", "1_01_metaData.txt",
                            package="Gimpute")
system(paste0("scp ", bedFile, bimFile, famFile, " ."))
inputPrefix <- "controlData" ## Specify the input PLINK file prefix
outputPrefix <- "1_03_replacedGroupAndSex"
## Not run: Requires an executable program PLINK, e.g.
## plink <- "/home/tools/plink"
## updateGroupIdAndSex(plink, inputPrefix, metaDataFile, outputPrefix)
```

Index