Package 'SeqSQC'

January 21, 2021

Title A bioconductor package for sample quality check with next generation sequencing data

Version 1.12.0

Description The SeqSQC is designed to identify problematic samples in NGS data, including samples with gender mismatch, contamination, cryptic relatedness, and population outlier.

biocViews Experiment Data, Homo_sapiens_Data, Sequencing Data, Project1000genomes, Genome

Depends R (>= 3.4), ExperimentHub (>= 1.3.7), SNPRelate (>= 1.10.2)

License GPL-3

Encoding UTF-8

URL https://github.com/Liubuntu/SeqSQC

BugReports https://github.com/Liubuntu/SeqSQC/issues

LazyData true

RoxygenNote 7.0.2

VignetteBuilder knitr

Imports e1071, GenomicRanges, gdsfmt, ggplot2, GGally, IRanges, methods, rbokeh, RColorBrewer, reshape2, rmarkdown, S4Vectors, stats, utils

Suggests BiocStyle, knitr, testthat

git_url https://git.bioconductor.org/packages/SeqSQC

git_branch RELEASE_3_12

git_last_commit 0a0a0f2

git_last_commit_date 2020-10-27

Date/Publication 2021-01-20

Author Qian Liu [aut, cre]

Maintainer Qian Liu <qliu7@buffalo.edu>

R topics documented:

`SeqSQC-package` *Sample Quality Check for NGS Data using SeqSQC package*

Description

SeqSQC

Details

Sample Quality Check for NGS Data.

Author(s)

Qian Liu

See Also

`LoadVfile`
for data preparation;
`MissingRate`
`PCACheck`
`Inbreeding`
`IBDCheck`
`PCACheck`
for individual sample QC checks;
`problemList`
for the summary of problematic samples with reason and sample list to be removed;
`IBDRemove`

for the problematic sample pairs detected with cryptic relationship;
RenderReport
to generate the sample QC report;
plotQC
to generate the ggplot or interactive plots in html format for each individual QC check;
sampleQC
for wrapper of data preparation, all sample QC checks, QC result summary, and sample QC report.

CCDS.Hs37.3.reduced_chr1.bed *Example capture region file used in vignette.*

Description

This .bed file contains only CCDS capture region only in chromosome 1, which is meant to be used together with the example_sub.vcf as a runnable example in the function of LoadVfile and sampleQC in the vignette.

Author(s)

Qian Liu <qliu7@buffalo.edu>

---

example.gds *example gds file used in vignette.*

Description

This gds file contains genotype and phenotype for 92 whole-genome sequenced samples captured by CCDS region. This is a merged dataset of the 87 benchmark samples and the 5 study samples (all are assembled from the 1000 Genomes Project). The meta info for these 92 samples includes sample name, pupulation, age, relation note and group info (benchmark or study).

Author(s)

Qian Liu <qliu7@buffalo.edu>

---

example.seqfile.Rdata *Example SeqSQC file used in vignette.*

Description

The SeqSQC object is a list of two objects. The first object gdsfile is the filepath of the "example.gds" file which stores the genotype and meta info of the example data merged with the benchmark data. The second object QCresult contains the data dimensions (# of samples and variants), sample annotation, and QC results for sample missing rate, sex check, inbreeding outlier check, IBD check, and population outlier check.

Author(s)

Qian Liu <qliu7@buffalo.edu>

example_sub.vcf *Example vcf file used in vignette.*

Description

This vcf file contains only a subset (1000 lines of variants) of the original vcf file for the 5 study samples (examples assembled from the 1000 Genomes Project). This is to be used as a runnable example in the function of `LoadVfile` and `sampleQC` in the vignette.

Author(s)

Qian Liu <qliu7@buffalo.edu>

IBDCheck *Sample relationship check with SeqSQC object input file.*

Description

Function to calculate the IBD coefficients for all sample pairs and to predict related sample pairs in study cohort.

Usage

```
IBDCheck(seqfile, remove.samples = NULL, LDprune = TRUE,
  kin.filter = TRUE, missing.rate = 0.1, ss.cutoff = 300,
  maf = 0.01, hwe = 1e-06, ...)
```

Arguments

`seqfile`	SeqSQC object, which includes the merged gds file for study cohort and benchmark.
`remove.samples`	a vector of sample names for removal from IBD calculation. Could be problematic samples identified from previous QC steps, or user-defined samples.
`LDprune`	whether to use LD-pruned snp set. The default is TRUE.
`kin.filter`	whether to use "kinship coefficient >= 0.08" as the additional criteria for related samples. The default is TRUE.
`missing.rate`	to use the SNPs with "<= `missing.rate`" only; if NaN, no threshold. By default, we use `missing.rate = 0.1` to filter out variants with missing rate greater than 10%.
`ss.cutoff`	the minimum sample size (300 by default) to apply the MAF filter. This sample size is the sum of study samples and the benchmark samples of the same population as the study cohort.
`maf`	to use the SNPs with ">= `maf`" if sample size defined in above argument is greater than `ss.cutoff`; otherwise NaN is used by default for no MAF threshold.
`hwe`	to use the SNPs with Hardy-Weinberg equilibrium p >= `hwe` if sample size defined in above argument is greater than `ss.cutoff`; otherwise no hwe threshold. The default is 1e-6.
`...`	Arguments to be passed to other methods.

Details

Using LD-pruned variants (by default), we calculate the IBD coefficients for all sample pairs, and then predict related sample pairs in study cohort using the support vector machine (SVM) method with linear kernel and the known relatedness embedded in benchmark data as training set.
Sample pairs with discordant self-reported and predicted relationship are considered as problematic. All predicted related pairs are also required to have coefficient of kinship >= 0.08 by default. The sample with higher missing rate in each related pair is selected for removal from further analysis by function of `IBDRemove`.

Value

a data frame with sample names, the descent coefficients of k0, k1 and kinship, self-reported relationship and predicted relationship for each pair of samples.

Author(s)

Qian Liu <qliu7@buffalo.edu>

Examples

```
load(system.file("extdata", "example.seqfile.Rdata", package="SeqSQC"))
gfile <- system.file("extdata", "example.gds", package="SeqSQC")
seqfile <- SeqSQC(gdsfile = gfile, QCresult = QCresult(seqfile))
seqfile <- IBDCheck(seqfile, remove.samples=NULL, LDprune=TRUE, missing.rate=0.1)
res.ibd <- QCresult(seqfile)$IBD
tail(res.ibd)
```

IBDRemove *Obtain the problematic sample list from IBD relatedness.*

Description

Function to extract the related sample pairs from IBD results, and to generate the sample list for removal from the related pairs based on sample missing rate.

Usage

```
IBDRemove(seqfile, all = FALSE)
```

Arguments

`seqfile`	SeqSQC object, with IBD results.
`all`	whether to check the IBD for all sample pairs (including the benchmark samples). The default is FALSE.

Value

a list of 2 elements: `$ibd.pairs` is a data frame with 5 columns including sample names(id1, id2), IBD coefficients of k0 and k1, and kinship for samples with cryptic relatedness. `$ibd.remove` is a vector of samples to be removed, which are generated by extracting the sample with higher missing rate in each problematic sample pair.

Author(s)

Qian Liu <qliu7@buffalo.edu>

Examples

```
load(system.file("extdata", "example.seqfile.Rdata", package="SeqSQC"))
gfile <- system.file("extdata", "example.gds", package="SeqSQC")
seqfile <- SeqSQC(gdsfile = gfile, QCresult = QCresult(seqfile))
seqfile <- IBDCheck(seqfile, remove.samples=NULL, LDprune=TRUE, missing.rate=0.1)
IBDRemove(seqfile)
```

`Inbreeding` *Sample inbreeding check with SeqSQC object input file.*

Description

Function to calculate population-specific inbreeding coefficients, and to predict inbreeding outliers that are five standard deviation beyond the mean.

Usage

```
Inbreeding(seqfile, remove.samples = NULL, LDprune = TRUE,
  missing.rate = 0.1, ss.cutoff = 300, maf = 0.01, hwe = 1e-06,
  ...)
```

Arguments

`seqfile`	SeqSQC object, which includes the merged gds file for study cohort and benchmark.
`remove.samples`	a vector of sample names for removal from inbreeding coefficient calculation. Could be problematic samples identified from previous QC steps, or user-defined samples.
`LDprune`	whether to use LD-pruned snp set. The default is TRUE.
`missing.rate`	to use the SNPs with "<= `missing.rate`" only; if NaN, no threshold. By default, we use `missing.rate = 0.1` to filter out variants with missing rate greater than 10%.
`ss.cutoff`	the minimum sample size (300 by default) to apply the MAF filter. This sample size is the sum of study samples and the benchmark samples of the same population as the study cohort.
`maf`	to use the SNPs with ">= `maf`" if sample size defined in above argument is greater than `ss.cutoff`; otherwise NaN is used by default for no MAF threshold.
`hwe`	to use the SNPs with Hardy-Weinberg equilibrium p >= `hwe` if sample size defined in above argument is greater than `ss.cutoff`; otherwise no hwe threshold. The default is 1e-6.
`...`	Arguments to be passed to other methods.

Details

Using LD-pruned variants (by default), we calculate the inbreeding coefficients for each sample in the study cohort and for benchmark samples of the same population as the study cohort. Samples with inbreeding coefficients that are five standard deviations beyond the mean are considered problematic and are shown as "Yes" in the column of `outlier.5sd`. Benchmark samples in this column are set to be "NA".

Value

a data frame with sample name, inbreeding coefficient, and an indicator of whether the inbreeding coefficient is five standard deviation beyond the mean.

Author(s)

Qian Liu <qliu7@buffalo.edu>

Examples

```
load(system.file("extdata", "example.seqfile.Rdata", package="SeqSQC"))
gfile <- system.file("extdata", "example.gds", package="SeqSQC")
seqfile <- SeqSQC(gdsfile = gfile, QCresult = QCresult(seqfile))
seqfile <- Inbreeding(seqfile, remove.samples=NULL, LDprune=TRUE, missing.rate=0.1)
res.inb <- QCresult(seqfile)$Inbreeding
tail(res.inb)
```

LoadVfile *Data preprocessing for VCF or plink input from NGS or GWAS data.*

Description

Function to read VCF or plink files, merge with benchmark data, and output as SeqSQC object.

Usage

```
LoadVfile(vfile, output = "sampleqc", capture.region = NULL,
  sample.annot = NULL, LDprune = TRUE, vfile.restrict = FALSE,
  slide.max.bp = 5e+05, ld.threshold = 0.3, format.data = "NGS",
  format.file = "vcf", ...)
```

Arguments

`vfile`	vcf or PLINK input file (ped/map/bed/bim/fam with same basename). Vfile could be a vector of character strings, see details.
`output`	a character string for name of merged data of SeqSQC object. The `dirname(output)` would be used as the directory to save the QC results and plots. The default is `sampleqc` in working directory.
`capture.region`	the BED file of sequencing capture regions. The default is NULL. For exome-sequencing data, the capture region file must be provided.
`sample.annot`	sample annotation file with 3 columns (with header) in the order of sample id, sample population and sex info. The default is NULL.

LDprune	whether to use LD-pruned snp set. The default is TRUE.
vfile.restrict	whether the input vcf or plink file has already been restricted by capture region. The default is FALSE.
slide.max.bp	the window size of SNPs when calculating linkage disequilibrium. The default is 5e+05.
ld.threshold	the r^2 threshold for LD-based SNP pruning if LDprune = TRUE. The default is 0.3.
format.data	the data source. The default is NGS for sequencing data.
format.file	the data format. The default is vcf.
...	Arguments to be passed to other methods.

Details

For vfile with more than one file names, LoadVfile will merge all dataset together if they all contain the same samples. It is useful to combine genetic/genomic data together if VCF data is divided by chromosomes.
sample.annot file contains 3 columns with column names. col 1 is sample with sample ids; col 2 is population with values of "AFR/EUR/ASN/EAS/SAS"; col 3 is gender with values of "male/female".

Value

a SeqSQC object with the filepath to the gds file which stores the genotype, the summary of samples and variants, and the QCresults including the sample annotation information.

Author(s)

Qian Liu <qliu7@buffalo.edu>

Examples

```
infile <- system.file("extdata", "example_sub.vcf", package="SeqSQC")
sample.annot <- system.file("extdata", "sampleAnnotation.txt", package="SeqSQC")
cr <- system.file("extdata", "CCDS.Hs37.3.reduced_chr1.bed", package="SeqSQC")
outfile <- file.path(tempdir(), "testWrapUp")
seqfile <- LoadVfile(vfile = infile, output = outfile, capture.region = cr,
sample.annot = sample.annot)
```

MissingRate *Sample missing rate check with SeqSQC object input file.*

Description

Function to calculate sample missing rate and to identify sample outlier with high missing rate (> 0.1).

Usage

```
MissingRate(seqfile, remove.samples = NULL)
```

Arguments

seqfile	SeqSQC object, which includes the merged gds file for study cohort and benchmark.
remove.samples	a vector of sample names for removal from missing rate check. Could be problematic samples identified from other QC steps, or user-defined samples.

Details

The value of the outlier column is set to NA for benchmark samples.

Value

a data frame with sample name, sample missing rate, and an indicator of whether the sample has a missing rate greater than 0.1.

Author(s)

Qian Liu <qliu7@buffalo.edu>

Examples

```
load(system.file("extdata", "example.seqfile.Rdata", package="SeqSQC"))
gfile <- system.file("extdata", "example.gds", package="SeqSQC")
seqfile <- SeqSQC(gdsfile = gfile, QCresult = QCresult(seqfile))
seqfile <- MissingRate(seqfile, remove.samples=NULL)
res.mr <- QCresult(seqfile)$MissingRate
tail(res.mr)
```

PCACheck *Population outlier check with SeqSQC object input file.*

Description

Function to perform principle component analysis for all samples and to infer sample ancestry.

Usage

```
PCACheck(seqfile, remove.samples = NULL, npcs = 4, LDprune = TRUE,
  missing.rate = 0.1, ss.cutoff = 300, maf = 0.01, hwe = 1e-06,
  ...)
```

Arguments

seqfile	SeqSQC object, which includes the merged gds file for study cohort and benchmark.
remove.samples	a vector of sample names for removal from PCA calculation. Could be problematic samples identified from previous QC steps, or user-defined samples.
npcs	the number principle components to use for the population prediction in SVM model. The default value is 4, and it is required to be <= 10.
LDprune	whether to use LD-pruned snp set, the default is TRUE.

missing.rate	to use the SNPs with "<= missing.rate" only; if NaN, no threshold. By default, we use missing.rate = 0.1 to filter out variants with missing rate greater than 10%.
ss.cutoff	the minimum sample size (300 by default) to apply the MAF filter. This sample size is the sum of study samples and the benchmark samples of the same population as the study cohort.
maf	to use the SNPs with ">= maf" if sample size defined in above argument is greater than ss.cutoff; otherwise NaN is used by default for no MAF threshold.
hwe	to use the SNPs with Hardy-Weinberg equilibrium p >= hwe if sample size defined in above argument is greater than ss.cutoff; otherwise no hwe threshold. The default is 1e-6.
...	Arguments to be passed to other methods.

Details

Using LD-pruned autosomal variants (by default), we calculate the eigenvectors and eigenvalues for principle component analysis (PCA). We use the benchmark samples as training dataset, and predict the population group for each sample in the study cohort based on the top four eigenvectors. Samples with discordant predicted and self-reported population groups are considered problematic. The function PCACheck performs the PCA analysis and identifies population outliers in study cohort.

Value

a data frame with sample name, reported population, data resource (benchmark vs study cohort), the first four eigenvectors and the predicted population.

Author(s)

Qian Liu <qliu7@buffalo.edu>

Examples

```
load(system.file("extdata", "example.seqfile.Rdata", package="SeqSQC"))
gfile <- system.file("extdata", "example.gds", package="SeqSQC")
seqfile <- SeqSQC(gdsfile = gfile, QCresult = QCresult(seqfile))
seqfile <- PCACheck(seqfile, remove.samples=NULL, LDprune=TRUE, missing.rate=0.1)
res.pca <- QCresult(seqfile)$PCA
tail(res.pca)
```

---

plotQC *Plot the QC results for specific QC steps.*

Description

Plot QC results.

Usage

```
plotQC(seqfile, QCstep = c("MissingRate", "SexCheck", "Inbreeding",
  "IBD", "PCA"), interactive = FALSE, sdcoef = 5, pc1 = "EV1",
  pc2 = "EV2", pairedScatter = FALSE, ...)
```

Arguments

seqfile	SeqSQC object with QC results.
QCstep	which QC step the user want to do plotting. Takes values of c("MissingRate","SexCheck","Inbree
interactive	whether to generate interactive plot. Recommend to use interactive = TRUE if user perform sample QC using an rmarkdown script and output plot to html format.
sdcoef	for inbreeding outlier check, how many standard deviation we need for identification of inbreeding outliers. The default is 5.
pc1	the eigenvector on x axis for PCA result. The default is "EV1" for eigenvector 1.
pc2	the eigenvector on y axis for PCA result. The default is "EV2" for eigenvector 2.
pairedScatter	for PCA result, whether to plot the paired scatterplot for the first 4 PC axes.
...	Arguments to be passed to other methods.

Value

the ggplot or interactive plot (if output is in html format) for specific QC result. If "interactive=FALSE", it returns a ggplot and author could have the flexibility to add on any layers, scales, faceting specifications and coordinate systems.

Author(s)

Qian Liu <qliu7@buffalo.edu>

Examples

```
load(system.file("extdata", "example.seqfile.Rdata", package="SeqSQC"))
gfile <- system.file("extdata", "example.gds", package="SeqSQC")
seqfile <- SeqSQC(gdsfile = gfile, QCresult = QCresult(seqfile))
p <- plotQC(seqfile, QCstep="PCA", interactive=FALSE)
p
```

---

problemList *Generate the problematic sample list.*

Description

generate the problematic sample list from QC steps that have been done, and provide each problematic sample with a reason for removal (high missing rate, gender mismatch, inbreeding outlier, cryptic relationship or population outlier).

Usage

```
problemList(seqfile)
```

Arguments

seqfile	SeqSQC object with sample QC results.

Value

a list of 2 datasets: 1) a data frame with 2 columns: `sample` for problematic sample name, and `remove.reason` for the reason of removing the sample. 2) a data frame with 1 column `sample` for problematic samples to be removed.

Author(s)

Qian Liu <qliu7@buffalo.edu>

Examples

```
load(system.file("extdata", "example.seqfile.Rdata", package="SeqSQC"))
problemList(seqfile)
```

`RenderReport` *Render the rmarkdown file for generating a sample QC report.*

Description

Function to render the pre-compiled rmarkdown file to generate the sample QC report.

Usage

```
RenderReport(input, output, interactive = TRUE)
```

Arguments

`input`	SeqSQC object with QC results.
`output`	a character string to define the file name for the QC report.
`interactive`	whether to generate interative plots in the report. The default is TRUE.

Value

Will incure the rendering of the rmarkdown file for generating the sample QC report. The report will return to the file denoted in `output` in the function.

Author(s)

Qian Liu <qliu7@buffalo.edu>

Examples

```
load(system.file("extdata", "example.seqfile.Rdata", package="SeqSQC"))
RenderReport(seqfile, output="report.html", interactive=FALSE)
```

sampleAnnotation.txt *Sample annotation file for the example data used in vignette.*

Description

This sample annotation file is a required input from the user when using SeqSQC. It includes the sample info with sample name stored in the column of `sample`, the population info stored in the column of `population`, and the gender info stored in the column of `gender`. The `population` column must be a in the format of "AFR/EUR/ASN/EAS/SAS". The `gender` column must be in the format of "female/male". This file is meant to be used together with the `example_sub.vcf` as a runnable example in the function of `LoadVfile` and `sampleQC` in the vignette.

Author(s)

Qian Liu <qliu7@buffalo.edu>

sampleQC *The wrap-up function for sample QC of sequencing/GWAS data.*

Description

A wrap-up function for sample QC. It reads in the variant genotypes in vcf/PLINK format, merges study cohort with benchmark data, and performs sample QC for the merged dataset.

Usage

```
sampleQC(vfile = NULL, output = "sampleqc", capture.region = NULL,
  sample.annot = NULL, LDprune = TRUE, vfile.restrict = FALSE,
  slide.max.bp = 5e+05, ld.threshold = 0.3, format.data = "NGS",
  format.file = "vcf", QCreport = TRUE, out.report = "report.html",
  interactive = TRUE, results = TRUE, plotting = TRUE, ...)
```

Arguments

`vfile`	vcf or PLINK input file (ped/map/bed/bim/fam with same basename). The default is NULL. Vfile could be a vector of character strings, see details. Could also take file in `SeqSQC` object generated from `LoadVfile`.
`output`	a character string for name of merged data of SeqSQC object. The `dirname(output)` would be used as the directory to save the QC result and plots. The default is `sampleqc` in the working directory.
`capture.region`	the BED file of sequencing capture regions. The default is NULL. For exome-sequencing data, the capture region file must be provided.
`sample.annot`	sample annotation file with 3 columns (with header) in the order of sample id, sample population and sex info. The default is NULL.
`LDprune`	whether to use LD-pruned snp set. The default is TRUE.
`vfile.restrict`	whether the input vcf or plink file has already been restricted by capture region. The default is FALSE.

slide.max.bp	the window size of SNPs when calculating linkage disequilibrium. The default is 5e+05.
ld.threshold	the r^2 threshold for LD-based SNP pruning if LDprune = TRUE. The default is 0.3.
format.data	the data source. The default is NGS for sequencing data.
format.file	the data format. The default is vcf.
QCreport	Whether to generate the sample QC report in html format.
out.report	the file name for the sample QC report. The default is report.html.
interactive	whether to generate interactive plots in the sample QC report if QCreport = TRUE.
results	whether to write out the results for each QC steps in .txt files. The default is TRUE.
plotting	whether to output the plots for each QC steps in .pdf files. the default is TRUE.
...	Arguments to be passed to other methods.

Details

For vfile with more than one file names, sampleQC will merge all dataset together if they all contain the same samples. It is useful to combine genetic/genomic data together if VCF data is divided by chromosomes.
There are 3 columns in sample.annot file. col 1 is sample with sample ids, col 2 is population with values of "AFR/EUR/ASN/EAS/SAS", col 3 is gender with values of "male/female".

Value

a SeqSQC object with the filepath to the gds file which stores the genotype, the summary of samples and variants, and the QCresults including the sample annotation information and all QC results.

Author(s)

Qian Liu <qliu7@buffalo.edu>

Examples

```
## Not run:
infile <- system.file("extdata", "example_sub.vcf", package="SeqSQC")
sample.annot <- system.file("extdata", "sampleAnnotation.txt", package="SeqSQC")
cr <- system.file("extdata", "CCDS.Hs37.3.reduced_chr1.bed", package="SeqSQC")
outfile <- file.path(tempdir(), "testWrapUp")
seqfile <- sampleQC(vfile = infile, output = outfile, capture.region = cr,
sample.annot = sample.annot, format.data = "NGS", format.file = "vcf",
QCreport = TRUE, out.report="report.html", interactive = TRUE)
## save(seqfile, file="seqfile.RData")

load(system.file("extdata", "example.seqfile.Rdata", package="SeqSQC"))
gfile <- system.file("extdata", "example.gds", package="SeqSQC")
seqfile <- SeqSQC(gdsfile = gfile, QCresult = QCresult(seqfile))
seqfile <- sampleQC(sfile = seqfile, output = outfile, QCreport = FALSE,
out.report="report.html", interactive = TRUE)

## End(Not run)
```

SeqOpen *Open the gds file in SeqSQC objects.*

Description

Function to open the gds file inside the SeqSQC object.

Usage

```
SeqOpen(seqfile, readonly = TRUE, allow.duplicate = FALSE)
```

Arguments

seqfile	SeqSQC object, which has been merged with benchmark data.
readonly	whether to open the gds file in read-only mode. If "FALSE", it is allowed to write data to the file. The default is TRUE.
allow.duplicate	whether to allow to open a GDS file with read-only mode when it has been opened in the same R session. The default is FALSE.

Value

a gds file with the filepath in the input SeqSQC object.

Author(s)

Qian Liu <qliu7@buffalo.edu>

Examples

```
library(gdsfmt)
load(system.file("extdata", "example.seqfile.Rdata", package="SeqSQC"))
gfile <- system.file("extdata", "example.gds", package="SeqSQC")
seqfile <- SeqSQC(gdsfile = gfile, QCresult = QCresult(seqfile))
dat <- SeqOpen(seqfile)
dat
closefn.gds(dat)
```

SeqSQC-class *A data format to store genotype phenotype and sample QC results from SeqSQC.*

Description

A SeqSQC object is a list of two objects. The first object gdsfile is the filepath of the GDS (discussed in section below) file which stores the genotype information from the original VCF file. The second object QCresult is a list of sample information and QC results, which include the dimension (# of samples and variants), sample annotation, and QC results for sample missing rate, sex check, inbreeding outlier check, IBD check, and population outlier check.

Usage

```
SeqSQC(gdsfile, QCresult = List())

gdsfile(x)

gdsfile(x) <- value

QCresult(x)

QCresult(x) <- value

## S4 method for signature 'SeqSQC'
gdsfile(x)

## S4 replacement method for signature 'SeqSQC'
gdsfile(x) <- value

## S4 method for signature 'SeqSQC'
QCresult(x)

## S4 replacement method for signature 'SeqSQC'
QCresult(x) <- value
```

Arguments

gdsfile	A character string for the filepath of the GDS file.
QCresult	A list with sample information and sample QC results.
x	an SeqSQCClass object.
value	the new value for the SeqSQC object slots.

Value

The filepath to the gds file.

Slots

gdsfile A character string for the filepath of the GDS file.

QCresult A list with sample information and sample QC results.

Examples

```
load(system.file("extdata", "example.seqfile.Rdata", package="SeqSQC"))
gdsfile(seqfile)
QCresult(seqfile)
```

SexCheck *Sample gender check with SeqSQC object input file.*

Description

Function to calculate the X chromosome inbreeding coefficient and to predict sample gender.

Usage

```
SexCheck(seqfile, remove.samples = NULL, missing.rate = 0.1,
  ss.cutoff = 300, maf = 0.01, ...)
```

Arguments

seqfile	SeqSQC object, which includes the merged gds file for study cohort and benchmark.
remove.samples	a vector of sample names for removal from sex check. Could be problematic samples identified from previous QC steps, or user-defined samples.
missing.rate	to use the SNPs with "<= missing.rate" only; if NaN, no threshold. By default, we use missing.rate = 0.1 to filter out variants with missing rate greater than 10%.
ss.cutoff	the minimum sample size (300 by default) to apply the MAF filter. This sample size is the sum of study samples and the benchmark samples of the same population as the study cohort.
maf	to use the SNPs with ">= maf" if sample size defined in above argument is greater than ss.cutoff; otherwise NaN is used by default for no MAF threshold.
...	Arguments to be passed to other methods.

Details

Samples are predicted to be female or male if the inbreeding coefficient is below 0.2, or greater than 0.8, respectively. The samples with discordant reported gender and predicted gender are considered as problematic. When the inbreeding coefficient is within the range of [0.2, 0.8], "0" is shown in the column of pred.sex to indicate ambiguous gender, which is not considered as problematic.

Value

a data frame with sample name, reported gender, x chromosome inbreeding coefficient, and predicted gender.

Author(s)

Qian Liu <qliu7@buffalo.edu>

Examples

```
load(system.file("extdata", "example.seqfile.Rdata", package="SeqSQC"))
gfile <- system.file("extdata", "example.gds", package="SeqSQC")
seqfile <- SeqSQC(gdsfile = gfile, QCresult = QCresult(seqfile))
seqfile <- SexCheck(seqfile, remove.samples=NULL, missing.rate=0.1)
res.sexc <- QCresult(seqfile)$SexCheck
tail(res.sexc)
```

Index