

# Package ‘SAIGEgds’

October 17, 2020

**Type** Package

**Title** Scalable Implementation of Generalized mixed models using GDS files in Phenome-Wide Association Studies

**Version** 1.2.3

**Date** 2020-09-03

**Depends** R (>= 3.5.0), gdsfmt (>= 1.20.0), SeqArray (>= 1.24.1), Rcpp

**LinkingTo** Rcpp, RcppArmadillo, RcppParallel

**Imports** methods, stats, utils, RcppParallel, SPAtest (>= 3.0.0)

**Suggests** parallel, crayon, RUnit, knitr, BiocGenerics, SNPRelate

**Description** Scalable implementation of generalized mixed models with highly optimized C++ implementation and integration with Genomic Data Structure (GDS) files. It is designed for single variant tests in large-scale phenome-wide association studies (PheWAS) with millions of variants and samples, controlling for sample structure and case-control imbalance. The implementation is based on the original SAIGE R package (v0.29.4.4). SAIGEgds also implements some of the SPAtest functions in C to speed up the calculation of Saddlepoint approximation. Benchmarks show that SAIGEgds is 5 to 6 times faster than the original SAIGE R package.

**License** GPL-3

**SystemRequirements** C++11, GNU make

**VignetteBuilder** knitr

**ByteCompile** TRUE

**URL** <https://github.com/AbbVie-ComputationalGenomics/SAIGEgds>

**biocViews** Software, Genetics, StatisticalMethod

**git\_url** <https://git.bioconductor.org/packages/SAIGEgds>

**git\_branch** RELEASE\_3\_11

**git\_last\_commit** 7663323

**git\_last\_commit\_date** 2020-09-18

**Date/Publication** 2020-10-16

**Author** Xiuwen Zheng [aut, cre] (<<https://orcid.org/0000-0002-1390-0708>>),  
Wei Zhou [ctb] (the original author of the SAIGE R package),  
J. Wade Davis [ctb]

**Maintainer** Xiuwen Zheng <[xiuwen.zheng@abbvie.com](mailto:xiuwen.zheng@abbvie.com)>

## R topics documented:

SAIGEgds-package . . . . .	2
seqAssocGLMM_SPA . . . . .	3
seqFitNullGLMM_SPA . . . . .	5
seqSAIGE_LoadPval . . . . .	8

<b>Index</b>	<b>9</b>
--------------	----------

---

SAIGEgds-package	<i>Scalable Implementation of Generalized mixed models in Phenome-Wide Association Studies using GDS files</i>
------------------	--

---

## Description

Scalable and accurate implementation of generalized mixed mode with the support of Genomic Data Structure (GDS) files and highly optimized C++ implementation. It is designed for single variant tests in large-scale phenome-wide association studies (PheWAS) with millions of variants and hundreds of thousands of samples, e.g., UK Biobank genotype data, controlling for case-control imbalance and sample structure in single variant association studies.

The implementation of SAIGEgds is based on the original SAIGE R package (v0.29.4.4) [Zhou et al. 2018] <https://github.com/weizhouUMICH/SAIGE/releases/tag/v0.29.4.4>. All of the calculation with single-precision floating-point numbers in SAIGE are replaced by the double-precision calculation in SAIGEgds. SAIGEgds also implements some of the SPAtest functions in C to speed up the calculation of Saddlepoint Approximation.

## Details

Package: SAIGEgds  
 Type: Package  
 License: GPL version 3

## Author(s)

Xiuwen Zheng <[xiuwen.zheng@abbvie.com](mailto:xiuwen.zheng@abbvie.com)>, Wei Zhou (the original author of the SAIGE R package, <https://github.com/weizhouUMICH/SAIGE>)

## References

- Zheng X, Davis J.Wade. SAIGEgds – an efficient statistical tool for large-scale PheWAS with mixed models. *\*Bioinformatics\** (2020). DOI: 10.1093/bioinformatics/btaa731.
- Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN, LeFaive J, VandeHaar P, Gagliano SA, Gifford A, Bastarache LA, Wei WQ, Denny JC, Lin M, Hveem K, Kang HM, Abecasis GR, Willer CJ, Lee S. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *\*Nat Genet\** (2018). Sep;50(9):1335-1341.
- Zheng X, Gogarten S, Lawrence M, Stilp A, Conomos M, Weir BS, Laurie C, Levine D. SeqArray – A storage-efficient high-performance data format for WGS variant calls. *\*Bioinformatics\** (2017). DOI: 10.1093/bioinformatics/btx145.

**Examples**

```

# open the GDS file
fn <- system.file("extdata", "grm1k_10k_snp.gds", package="SAIGEgds")
gdsfile <- seqOpen(fn)

# load phenotype
phenofn <- system.file("extdata", "pheno.txt.gz", package="SAIGEgds")
pheno <- read.table(phenofn, header=TRUE, as.is=TRUE)
head(pheno)

# fit the null model
glmm <- seqFitNullGLMM_SPA(y ~ x1 + x2, pheno, gdsfile, trait.type="binary")

# p-value calculation
assoc <- seqAssocGLMM_SPA(gdsfile, glmm, mac=10)

head(assoc)

# close the GDS file
seqClose(gdsfile)

```

---

seqAssocGLMM_SPA	<i>P-value calculation</i>
------------------	----------------------------

---

**Description**

P-value calculations using variance approximation and an adjustment of Saddlepoint approximation.

**Usage**

```

seqAssocGLMM_SPA(gdsfile, modobj, maf=NaN, mac=10, missing=0.1, dsnode="",
  spa.pval=0.05, var.ratio=NaN, res.savefn="", res.compress="LZMA",
  parallel=FALSE, verbose=TRUE)

```

**Arguments**

gdsfile	a SeqArray GDS filename, or a GDS object
modobj	an R object for SAIGE model parameters
maf	minor allele frequency threshold (checking $\geq$ maf), NaN for no filter
mac	minor allele count threshold (checking $\geq$ mac), NaN for no filter
missing	missing threshold for variants (checking $\leq$ missing), NaN for no filter
dsnode	"" for automatically searching the GDS nodes "genotype" and "annotation/format/DS", or use a user-defined GDS node in the file
spa.pval	the p-value threshold for SPA adjustment, 0.05 by default (since normal approximation performs well when the test statistic is close to the mean)
var.ratio	NaN for using the estimated variance ratio in the model fitting, or a user-defined variance ratio
res.savefn	an RData or GDS file name, "" for no saving

res.compress	the compression method for the output file, it should be one of LZMA, LZMA_RA, ZIP, ZIP_RA and none
parallel	FALSE (serial processing), TRUE (multicore processing), a numeric value for the number of cores, or other value; parallel is passed to the argument c1 in <a href="#">seqParallel</a> , see <a href="#">seqParallel</a> for more details
verbose	if TRUE, show information

### Details

The original SAIGE R package uses 0.05 as a threshold for unadjusted p-values (based on asymptotic normality) to further calculate adjusted p-values (Saddlepoint approximation, SPA). If `var.ratio=NaN`, the average of variance ratios (`mean(modobj$var.ratio$ratio)`) is used instead. For more details of SAIGE algorithm, please refer to the SAIGE paper [Zhou et al. 2018] (see the reference section).

### Value

Return a data.frame with the following components if not saving to a file:

id	variant ID in the GDS file;
chr	chromosome;
pos	position;
rs.id	the RS IDs if it is available in the GDS file;
ref	the reference allele;
alt	the alternative allele;
AF.alt	allele frequency for the alternative allele; the minor allele frequency is $\min(\text{AF.alt}, 1-\text{AF.alt})$ ;
mac	minor allele count; the allele count for the alternative allele is $\text{ifelse}(\text{AF.alt} \leq 0.5, \text{mac}, 2 * \text{num} - \text{mac})$ ;
num	the number of samples with non-missing genotypes;
beta	beta coefficient, odds ratio if binary outcomes (alternative allele vs. reference allele);
SE	standard error for beta coefficient;
pval	adjusted p-value with the Saddlepoint approximation method;
pval.noadj	p-values based on asymptotic normality (could be 0 if it is too small, e.g., $\text{pnorm}(-50) = 0$ in R);
converged	whether the SPA algorithm converges or not for adjusted p-values.

### Author(s)

Xiuwen Zheng

### References

Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN, LeFaive J, VandeHaar P, Gagliano SA, Gifford A, Bastarache LA, Wei WQ, Denny JC, Lin M, Hveem K, Kang HM, Abecasis GR, Willer CJ, Lee S. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet* (2018). Sep;50(9):1335-1341.

### See Also

[seqAssocGLMM\\_SPA](#), [seqSAIGE\\_LoadPval](#)

**Examples**

```
# open a GDS file
fn <- system.file("extdata", "grm1k_10k_snp.gds", package="SAIGEgds")
gdsfile <- seqOpen(fn)

# load phenotype
phenofn <- system.file("extdata", "pheno.txt.gz", package="SAIGEgds")
pheno <- read.table(phenofn, header=TRUE, as.is=TRUE)
head(pheno)

# fit the null model
glmm <- seqFitNullGLMM_SPA(y ~ x1 + x2, pheno, gdsfile, trait.type="binary")

# p-value calculation
assoc <- seqAssocGLMM_SPA(gdsfile, glmm, mac=10)

head(assoc)

# close the GDS file
seqClose(gdsfile)
```

---

```
seqFitNullGLMM_SPA      Fit the null model with GRM
```

---

**Description**

Fit the null model in the mixed model framework with genetic relationship matrix (GRM).

**Usage**

```
seqFitNullGLMM_SPA(formula, data, gdsfile, trait.type=c("binary", "quantitative"),
  sample.col="sample.id", maf=0.005, missing.rate=0.01, max.num.snp=1000000L,
  variant.id=NULL, inv.norm=TRUE, X.transform=TRUE, tol=0.02, maxiter=20L,
  nrun=30L, tolPCG=1e-5, maxiterPCG=500L, num.marker=30L, tau.init=c(0,0),
  traceCVcutoff=0.0025, ratioCVcutoff=0.001, geno.sparse=TRUE, num.thread=1L,
  model.savefn="", seed=200L, fork.loading=FALSE, verbose=TRUE)
```

**Arguments**

formula	an object of class formula (or one that can be coerced to that class), e.g., $y \sim x1 + x2$ , see <a href="#">lm</a>
data	a data frame for the formulas
gdsfile	a SeqArray GDS filename, or a GDS object
trait.type	"binary" for binary outcomes, "quantitative" for continuous outcomes
sample.col	the column name of sample IDs corresponding to the GDS file
maf	minor allele frequency for imported genotypes (checking $\geq$ maf), if <code>variant.id=NULL</code> ; NaN for no filter
missing.rate	threshold of missing rate (checking $\leq$ missing.rate), if <code>variant.id=NULL</code> ; NaN for no filter
max.num.snp	the maximum number of SNPs used, or -1 for no limit

<code>variant.id</code>	a list of variant IDs, used to construct GRM
<code>inv.norm</code>	if TRUE, perform inverse normal transformation on residuals for quantitative outcomes, see the reference [Sofer, 2019]
<code>X.transform</code>	if TRUE, perform QR decomposition on the design matrix
<code>tol</code>	overall tolerance for model fitting
<code>maxiter</code>	the maximum number of iterations for model fitting
<code>nrun</code>	the number of random vectors in the trace estimation
<code>tolPCG</code>	tolerance of PCG iterations
<code>maxiterPCG</code>	the maximum number of PCG iterations
<code>num.marker</code>	the number of SNPs used to calculate the variance ratio
<code>tau.init</code>	a 2-length numeric vector, the initial values for variance components, tau; for binary traits, the first element is always be set to 1. if <code>tau.init</code> is not specified, the second element will be 0.5 for binary traits
<code>traceCVcutoff</code>	the threshold for coefficient of variation (CV) for the trace estimator, and the number of runs for trace estimation will be increased until the CV is below the threshold
<code>ratioCVcutoff</code>	the threshold for coefficient of variation (CV) for estimating the variance ratio, and the number of randomly selected markers will be increased until the CV is below the threshold
<code>geno.sparse</code>	if TRUE, store the sparse structure for genotypes; otherwise, save genotypes in a 2-bit dense matrix; see details
<code>num.thread</code>	the number of threads
<code>model.savefn</code>	the filename of model output, R data file <code>'.rda'</code> , <code>'.RData'</code> , or <code>'.rds'</code>
<code>seed</code>	an integer as a seed for random numbers
<code>fork.loading</code>	load genotypes via forking or not; forking processes in Unix can reduce loading time of genotypes, but may double the memory usage; not applicable on Windows
<code>verbose</code>	if TRUE, show information

## Details

Utilizing the sparse structure of genotypes could significantly improve the computational efficiency of model fitting, but it also increases the memory usage. For more details of SAIGE algorithm, please refer to the SAIGE paper [Zhou et al. 2018] (see the reference section).

## Value

Returns a list with the following components:

<code>coefficients</code>	the beta coefficients for fixed effects;
<code>tau</code>	a numeric vector of variance components <code>'Sigma_E'</code> and <code>'Sigma_G'</code> ;
<code>linear.predictors</code>	the linear fit on link scale;
<code>fitted.values</code>	fitted values from objects returned by modeling functions using <code>glm.fit</code> ;
<code>residuals</code>	residuals;
<code>cov</code>	covariance matrix of beta coefficients;

<code>converged</code>	whether the model is fitted or not;
<code>obj.noK</code>	internal use, returned object from the SPAtest package;
<code>var.ratio</code>	a data.frame with columns 'id' (variant.id), 'maf' (minor allele frequency), 'mac' (minor allele count), 'var1' (the variance of score statistic), 'var2' (a variance estimate without accounting for estimated random effects) and 'ratio' (var1/var2, estimated variance ratio for variance approximation);
<code>trait.type</code>	either "binary" or "quantitative";
<code>sample.id</code>	the sample IDs used in the model fitting;
<code>variant.id</code>	the variant IDs used in the model fitting.

### Author(s)

Xiuwen Zheng

### References

Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN, LeFaive J, VandeHaar P, Gagliano SA, Gifford A, Bastarache LA, Wei WQ, Denny JC, Lin M, Hveem K, Kang HM, Abecasis GR, Willer CJ, Lee S. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet* (2018). Sep;50(9):1335-1341.

T Sofer, X Zheng, SM Gogarten, CA Laurie, etc. A fully adjusted two-stage procedure for rank-normalization in genetic association studies. 2019. *Genetic Epidemiology* 43(3), 263-275

### See Also

[seqAssocGLMM\\_SPA](#)

### Examples

```
# open a GDS file
fn <- system.file("extdata", "grm1k_10k_snp.gds", package="SAIGEgds")
gdsfile <- seqOpen(fn)

# load phenotype
phenofn <- system.file("extdata", "pheno.txt.gz", package="SAIGEgds")
pheno <- read.table(phenofn, header=TRUE, as.is=TRUE)
head(pheno)

# fit the null model
glmm <- seqFitNullGLMM_SPA(y ~ x1 + x2, pheno, gdsfile, trait.type="binary")
glmm

# close the GDS file
seqClose(gdsfile)
```

---

seqSAIGE\_LoadPval      *Load the association results*

---

### Description

Load the association results from an RData or GDS file.

### Usage

```
seqSAIGE_LoadPval(fn, varnm=NULL, index=NULL, verbose=TRUE)
```

### Arguments

fn	RData, RDS or GDS file names, merging datasets if multiple files
varnm	NULL, or a character vector to include the column names; e.g., c("chr", "position", "rs.id", "ref")
index	NULL, or a logical/numeric vector for a set of rows
verbose	if TRUE, show information

### Value

Return a data.frame including p-values.

### Author(s)

Xiuwen Zheng

### See Also

[seqFitNullGLMM\\_SPA](#), [seqAssocGLMM\\_SPA](#)

### Examples

```
(fn <- system.file("unitTests", "saige_pval.rda", package="SAIGEgds"))
pval <- seqSAIGE_LoadPval(fn)
```

```
names(pval)
# [1] "id"           "chr"          "pos"          "rs.id"        "ref"
# [6] "alt"          "AF.alt"       "AC.alt"       "num"          "beta"
# [11] "SE"           "pval"         "pval.noadj"  "converged"
```

```
head(pval)
```



# Index

## \* **GDS**

- SAIGEgds-package, [2](#)
- seqAssocGLMM\_SPA, [3](#)
- seqFitNullGLMM\_SPA, [5](#)
- seqSAIGE\_LoadPval, [8](#)

## \* **association**

- SAIGEgds-package, [2](#)
- seqAssocGLMM\_SPA, [3](#)
- seqFitNullGLMM\_SPA, [5](#)
- seqSAIGE\_LoadPval, [8](#)

## \* **genetics**

- SAIGEgds-package, [2](#)
- seqAssocGLMM\_SPA, [3](#)
- seqFitNullGLMM\_SPA, [5](#)
- seqSAIGE\_LoadPval, [8](#)

lm, [5](#)

SAIGEgds (SAIGEgds-package), [2](#)  
SAIGEgds-package, [2](#)  
seqAssocGLMM\_SPA, [3](#), [4](#), [7](#), [8](#)  
seqFitNullGLMM\_SPA, [5](#), [8](#)  
seqParallel, [4](#)  
seqSAIGE\_LoadPval, [4](#), [8](#)