

# Package ‘SCANG’

August 21, 2019

**Type** Package

**Title** Scan Procedure for Whole Genome Sequencing Study

**Version** 1.0.0

**Date** 2019-08-14

**Author** Zilin Li [aut, cre], XiHao Li [aut, cre], Han Chen [aut]

**Maintainer** Zilin Li <li@hsph.harvard.edu>, Xihao Li <xihao.li@g.harvard.edu>

**Description** R package for performing SCANG procedure in whole genome sequencing studies.

**License** GPL-3

**Copyright** See COPYRIGHTS for details.

**Imports** Rcpp, GMMAT, Matrix, GENESIS, kinship2

**Encoding** UTF-8

**LazyData** true

**Depends** R (>= 3.0.0)

**LinkingTo** Rcpp, RcppArmadillo

**RoxygenNote** 6.1.1

**Suggests** knitr, rmarkdown

**VignetteBuilder** knitr

## R topics documented:

fit_null_glm . . . . .	2
fit_null_glmkin . . . . .	2
SCANG . . . . .	4

<b>Index</b>	<b>6</b>
--------------	----------

---

fit_null_glm	<i>Fit generalized linear model under the null hypothesis for unrelated samples.</i>
--------------	--

---

### Description

The `fit_null_glm` function is a wrapper of the `glm` function from the `stats` package that fits a regression model under the null hypothesis for unrelated samples, which provides the preliminary step for subsequent variant-set tests in whole genome sequencing data analysis.

### Usage

```
fit_null_glm(fixed, data, family = binomial(link = "logit"),
            times = 2000, ...)
```

### Arguments

<code>fixed</code>	an object of class <code>formula</code> (or one that can be coerced to that class): a symbolic description of the fixed effects model to be fitted.
<code>data</code>	a data frame or list (or object coercible by <code>as.data.frame</code> to a data frame) containing the variables in the model.
<code>family</code>	a description of the error distribution and link function to be used in the model. This can be a character string naming a family function, a family function or the result of a call to a family function. (See <code>family</code> for details of family functions). Can be either "gaussian" for continuous phenotype or "binomial" for binary phenotype.
<code>times</code>	a number of pseudo-residuals (default = 2000).
<code>...</code>	additional arguments that could be passed to <code>glm</code> .

### Value

The function returns an object of the model fit from `glm` (`obj_nullmodel`), with an additional element indicating the samples are unrelated (`obj_nullmodel$relatedness = FALSE`). See `glm` for more details.

---

fit_null_glmkin	<i>Fitting generalized linear mixed model with known relationship matrices under the null hypothesis for related samples.</i>
-----------------	---

---

### Description

The `fit_null_glmkin` function is a wrapper of the `glmkin` function from the `GMMAT` package that fits a regression model under the null hypothesis for related samples, which provides the preliminary step for subsequent variant-set tests in whole genome sequencing data analysis. More details see `glmkin`.

**Usage**

```
fit_null_glmkin(fixed, data = parent.frame(), kins, use_sparse = NULL,
  kins_cutoff = 0.022, id, random.slope = NULL, groups = NULL,
  family = binomial(link = "logit"), method = "REML",
  method.optim = "AI", maxiter = 500, tol = 1e-05, taumin = 1e-05,
  taumax = 1e+05, tauregion = 10, times = 2000, verbose = FALSE,
  ...)
```

**Arguments**

fixed	an object of class <code>formula</code> (or one that can be coerced to that class): a symbolic description of the fixed effects model to be fitted.
data	a data frame or list (or object coercible by <code>as.data.frame</code> to a data frame) containing the variables in the model.
kins	a known positive semi-definite relationship matrix (e.g. kinship matrix in genetic association studies) or a list of known positive semi-definite relationship matrices. The rownames and colnames of these matrices must at least include all samples as specified in the <code>id</code> column of the data frame <code>data</code> .
use_sparse	a logical switch of whether the provided dense <code>kins</code> matrix should be transformed to a sparse matrix (default = <code>NULL</code> ).
kins_cutoff	the cutoff of setting all entries with smaller values to 0 in <code>kins</code> matrix (default = 0.022).
id	a column in the data frame <code>data</code> , indicating the id of samples. When there are duplicates in <code>id</code> , the data is assumed to be longitudinal with repeated measures.
random.slope	an optional column indicating the random slope for time effect used in a mixed effects model for longitudinal data. It must be included in the names of <code>data</code> . There must be duplicates in <code>id</code> and <code>method.optim</code> must be "AI" (default = <code>NULL</code> ).
groups	an optional categorical variable indicating the groups used in a heteroscedastic linear mixed model (allowing residual variances in different groups to be different). This variable must be included in the names of <code>data</code> , and <code>family</code> must be "gaussian" and <code>method.optim</code> must be "AI" (default = <code>NULL</code> ).
family	a description of the error distribution and link function to be used in the model. This can be a character string naming a family function, a family function or the result of a call to a family function. (See <code>family</code> for details of family functions).
method	method of fitting the generalized linear mixed model. Either "REML" or "ML" (default = "REML").
method.optim	optimization method of fitting the generalized linear mixed model. Either "AI", "Brent" or "Nelder-Mead" (default = "AI").
maxiter	a positive integer specifying the maximum number of iterations when fitting the generalized linear mixed model (default = 500).
tol	a positive number specifying tolerance, the difference threshold for parameter estimates below which iterations should be stopped (default = 1e-5).
taumin	the lower bound of search space for the variance component parameter $\tau$ (default = 1e-5), used when <code>method.optim</code> = "Brent". See Details.
taumax	the upper bound of search space for the variance component parameter $\tau$ (default = 1e5), used when <code>method.optim</code> = "Brent". See Details.

tauregion	the number of search intervals for the REML or ML estimate of the variance component parameter $\tau$ (default = 10), used when <code>method.optim = "Brent"</code> . See Details.
times	a number of pseudo-residuals (default = 2000).
verbose	a logical switch for printing detailed information (parameter estimates in each iteration) for testing and debugging purpose (default = FALSE).
...	additional arguments that could be passed to <a href="#">glm</a> .

### Value

The function returns an object of the model fit from [glmmkin](#) (`obj_nullmodel`), with additional elements indicating the samples are related (`obj_nullmodel$relatedness = TRUE`), whether the `kins` matrix is sparse when fitting the null model, and the matrix of pseudo residuals. See [glmmkin](#) for more details.

### References

- Chen, H. et al. (2016). Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *American Journal of Humann Genetics* 98(4), 653-666. ([pub](#))
- Chen, H. et al. (2019). Efficient variant set mixed model association tests for continuous and binary traits in large-scale whole genome sequencing studies. *American Journal of Humann Genetics* 104(2), 260-274. ([pub](#))
- Chen, H. & Conomos, M.P. (2019). GMMAT-package: Generalized Linear Mixed Model Association Tests. ([web](#))

---

SCANG

*SCANG procedure using omnibus test*

---

### Description

The SCANG function takes in genotype and the object from fitting the null model and detect the association between a quantitative/dichotomous phenotype and a variant-set in a sequence by using SCANG procedure, including SCANG-O, SCANG-B and SCANG-S. For each region, the scan statistic of SCANG-O is the set-based p-value of an omnibus test that aggregated SKAT(1,1), SKAT(1,25), Burden(1,1) and Burden(1,25) using Cauchy method; the scan statistic of SCANG-S is the set-based p-value of an omnibus test that aggregated SKAT(1,1) and SKAT(1,25) using Cauchy method; the scan statistic of SCANG-B is the set-based p-value of an omnibus test that aggregated Burden(1,1) and Burden(1,25) using Cauchy method.

### Usage

```
SCANG(genotype, obj_nullmodel, Lmin, Lmax, rare_maf_cutoff = 0.05,
      steplength = 5, alpha = 0.05, filter = 1e-04, f = 0.5)
```

**Arguments**

genotype	an n*p genotype matrix (dosage matrix) of the target sequence, where n is the sample size and p is the number of variants.
obj_nullmodel	an object from fitting the null model, which is the output from either <code>fit_null_glm</code> function for unrelated samples or <code>fit_null_glmkin</code> function for related samples. Note that <code>fit_null_glmkin</code> is a wrapper of <code>glmkin</code> function from the <code>GMMAT</code> package.
Lmin	minimum number of variants in searching windows.
Lmax	maximum number of variants in searching windows.
rare_maf_cutoff	the cutoff of maximum minor allele frequency in defining rare variants. (Default is 0.05).
steplength	difference of number of variants in searching windows, that is, the number of variants in searching windows are Lmin, Lmin+steplength, Lmin+steplength,...,Lmax. (Default is 5).
alpha	family-wise/genome-wide significance level. (Default is 0.05).
filter	a filtering threshold of screening method for SKAT. SKAT p-values are calculated for regions whose p-value is possibly smaller than the filtering threshold. (Default is 1e-4).
f	an overlap fraction, which controls for the overlapping proportion of detected regions. For example, when f=0, the detected regions are non-overlapped with each other, and when f=1, we keep every susceptible region as detected regions. (Default is 0.5.)

**Value**

The function returns a list with the following members:

`SCANG_0_res`: A matrix that summarized the significant region detected by SCANG-O. The first column is the  $-\log(p\text{-value})$  of the detected region. The next two columns are the location of the detected region (in sense of variants order). The last column is the family-wise/genome-wide error rate of the detected region. The result (0,0,0,1) means there is no significant region.

`SCANG_0_top1`: A vector of length 4 which summarized the top 1 region detected by SCANG-O. The first element is the  $-\log(p\text{-value})$  of the region. The next two elements are the location of the detected region (in sense of variants order). The last element is the family-wise/genome-wide p-value.

`SCANG_0_thres`: Empirical threshold of SCANG-O for controlling the family-wise type I error at alpha level.

`SCANG_0_thres_boot`: A vector of Monte Carlo simulation sample for generating the empirical threshold. The 1-alpha quantile of this vector is the empirical threshold.

`SCANG_S_res`, `SCANG_S_thres`, `SCANG_S_top1`, `SCANG_S_thres_boot`: Analysis results using SCANG-S. Details see SCANG-O.

`SCANG_B_res`, `SCANG_B_thres`, `SCANG_B_top1`, `SCANG_B_thres_boot`: Analysis results using SCANG-B. Details see SCANG-O.

# Index

`as.data.frame`, [2](#), [3](#)

`family`, [2](#), [3](#)

`fit_null_glm`, [2](#), [5](#)

`fit_null_glmkin`, [2](#), [5](#)

`formula`, [2](#), [3](#)

`glm`, [2](#), [4](#)

`glmkin`, [2](#), [4](#), [5](#)

`GMMAT`, [2](#), [5](#)

`SCANG`, [4](#)

`stats`, [2](#)