SUGEN 8.1 Manual

Ran Tao (*taor@live.unc.edu*)

11/2/2016

Contents

1. GENERAL INFORMATION

SUGEN is a command-line software program written in C++ to implement the weighted and unweighted approaches described by Lin et al. (2014) for various types of association analysis under complex survey sampling. The current version of the program can accommodate continuous, binary, and right-censored time-to-event traits. It can perform single-variant and gene-based association analysis. In single-variant analysis, it can perform standard association analysis, conditional analysis, and gene-environment interaction analysis using Wald statistics. In standard association analysis, we include the SNP of interest and other covariates (if any) as predictors in the regression model. In conditional analysis, we include the SNP of interest, the SNPs that are conditioned on, and other covariates (if any) as predictors in the regression model. In gene-environment interaction analysis, we include the SNP of interest, the environment variables, the interactions between the SNP and environment variables, and other covariates (if any) as predictors in the regression model. In gene-based analysis, it generates the score statistics and covariance matrix for variants in each gene. These summary statistics can be loaded into the software program MASS (http://dlin.web.unc.edu/software/mass/) to perform all commonly used gene-based association tests.

2. DOWNLOAD AND INSTALLATION

2.1 Download

The latest version of SUGEN can be downloaded from http://dlin.web.unc.edu/software/sugen/.

2.2 Installation

```
## Step 1: Unzip the package.
$ unzip sugen.zip

## Step 2: Go to the SUGEN directory.
$ cd ./sugen8_20160929

## Step 3: Install SUGEN.
$ make

## An executable called ''SUGEN'' will be generated in the directory.
```

3. SYNOPSIS

SUGEN [**--pheno** *pheno_file*] [**--formula** *formula*] [**--id-col** *iid*] [**--family-col** *fid*] [**--weight-col** *wt*] [**--vcf** *vcf_file.gz*] [**--dosage**] [**--probmatrix** *prob_file*] [**--subset** *subset_expression*] [**--unweighted**] [**--model** *model*] [**--robust-variance**] [**--left-truncation** *left_truncation_time*] [**--cond** *cond_file*] [**--ge** *envi_covs*] [**--score**] [**--score-rescale** *rescale_rule*] [**--group** *group_file*] [**--hetero-variance** *strata*] [**--out-prefix** *out_prefix*] [**--out-zip**] [**--extract-chr** *chr*] [**--extract-range** *range*] [**--extract-file** *extract_file*] [**--group-maf** *maf_ub*] [**--group-callrate** *cr_lb*]

4. OPTIONS

4.1 Input Options

- **--pheno** {*pheno_file*}

  Specifies the phenotype file. The default name is `pheno.txt`.

- **--formula** {*formula*}

  Specifies the regression formula. In linear or logistic regression, the format of *formula* is

  $$''\texttt{trait} = \texttt{covariate}_1 + \texttt{covariate}_2 + \ldots + \texttt{covariate}_p''.$$

  The trait and covariates must appear in *pheno_file*. If there is no covariate, then we specify the formula as

  $$''\texttt{trait} =''.$$

  In Cox proportional hazards regression, the format of *formula* is

  $$''\texttt{(time, event)} = \texttt{covariate}_1 + \texttt{covariate}_2 + \ldots + \texttt{covariate}_p''.$$

  In this case, the double quotes in *formula* cannot be omitted. The time, event indicator, and covariates must appear in *pheno_file*. If there is no covariate, then we specify the formula as

  $$''\texttt{(time, event)} =''.$$

- **--id-col** {*iid*}

  Specifies the subject ID column in *pheno_file*. The default column name is `IID`.

- **--family-col** {*fid*}

Specifies the family ID column in *pheno_file*. The default column name is `FID`. If study subjects are independent, then we specify the family ID column to be the same as the subject ID column.

- **--weight-col** {*wt*}

  Specifies the weight column in *pheno_file*. The default column name is `WT`. This option is ignored if **--unweighted** is specified.

- **--vcf** {*vcf_file.gz*}

  Specifies the block compressed and indexed VCF file. The default name is `geno.vcf.gz`.

- **--dosage**

  Analyzes dosage data in the VCF file. The dosages must be stored in the "DS" field of the VCF file. An example can be found here.

- **--probmatrix** {*prob_file*}

  Specifies the file that contains the file names of the pairwise inclusion probability matrices. The default name is `probmatrix.txt`. This option is optional in weighted analysis and ignored in unweighted analysis.

- **--subset** {*subset_expression*}

  Restricts analysis to a subset of subjects in *pheno_file*. For example, if one wants to restrict the analysis to subjects whose `var_a` equals `level_1`, where `var_a` is a column in *pheno_file*, and `level_1` is one of the values of `var_a`, then we can specify *subset_expression* as "`var_a` = `level_1`".

4.2 Analysis Options

- **--unweighted**

  Uses the unweighted approach.

- **--model** {*model*}

  Specifies the regression model. There are three options: `linear` (linear regression), `logistic` (logistic regression), and `coxph` (Cox proportional hazard regression). The default value is `linear`. In linear or logistic regression, the trait is continuous or binary (0/1), respectively. In Cox proportional hazard regression, the time is positive, and the event indicator is binary (0/1).

- **--robust-variance**

  If this option is specified, then the robust variance estimator will be used. Otherwise, the model-based variance estimator will be used.

- **--left-truncation** {*left_truncation_time*}

  Specifies the left truncation time (if any) in Cox proportional hazards regression.

- **--cond** {*cond_file*}

  In single-variant analysis, performs conditional analysis conditioning on the variants included in *cond_file*. There is no default value for *cond_file*. The format of the variant IDs in *cond_file* is `chromosome:position`. This option is valid only when **--score** is not specified. In this situation, either [**--cond** *cond_file*] or [**--ge** *envi_covs*] can be specified, but not both. If neither is specified, then standard association analysis is performed.

- **--ge** {*envi_covs*}

  In single-variant analysis, performs gene-environment interaction analysis. *envi_covs* are the names of the environment variables. The format of *envi_covs* is $\texttt{covariate}_1$, $\texttt{covariate}_2, \ldots, \texttt{covariate}_k$. That is, multiple environment variables are separately by "`,`". There is no default value for *envi_covs*. This option is valid only when **--score** is not specified. In this situation, either [**--cond** *cond_file*] or [**--ge** *envi_covs*] can be specified, but not both. If neither is specified, then standard association analysis is performed.

- **--score**

  Uses score statistics.

- **--score-rescale** {*rescale_rule*}

  Specifies the method to rescale the score statistics. There are two options: `naive` and `optimal`. The default value is `naive`. This option is valid only when **--score** is specified.

- **--group** {*group_file*}

  Performs gene-based association analysis. Gene memberships of variants are defined in *group_file*. There is no default value for *group_file*. This option is valid only when **--score** is specified.

- **--hetero-variance** {*strata*}

  Allows the residual variance in linear regression to be different in different levels of *strata*.

4.3 Output Options

- **--out-prefix** {*out_prefix*}

  Specifies the prefix of the output files. The default prefix is `results`.

- **--out-zip**

  Zips the output files.

- **--extract-chr** {*chr*}

  Restricts single-variant analysis to variants in chromosome *chr*. This option is valid only when [**--group** *group_file*] is not specified.

- **--extract-range** {*range*}

  Restricts single-variant analysis to variants in chromosome *chr* and position in *range*. The format of *range* is `1000000-2000000`. This option is valid only when [**--group** *group_file*] is not specified and [**--extract-chr** *chr*] is specified.

- **--extract-file** {*extract_file*}

  Restricts single-variant analysis to variants in *extract_file*. The format of the variant IDs in *extract_file* is `chromosome:position`. This option is valid only when [**--group** *group_file*], [**--extract-chr** *chr*], and [**--extract-range** *range*] are not specified.

- **--group-maf** {*maf_ub*}

  Specifies the minor allele frequency (MAF) upper bound for gene-based association analysis. *maf_ub* is a real number between 0 and 1. Its default value is `0.05`. Variants with MAFs greater than *maf_ub* will not be included in the analysis.

- **--group-callrate** {*cr_lb*}

  Specifies the call rate lower bound for gene-based association analysis. *cr_lb* is a real number between 0 and 1. Its default value is `0`. Variants with call rates less than *cr_lb* will not be included in the analysis.

5. INPUT FILES

5.1 Phenotype File (*pheno_file*)

The phenotype file should be tab-delimited. Missing data are denoted by `NA`. The rows represent study subjects. The 1st row is the header line. This file should include the subject ID column, family ID column (unless the subjects are independent), weight column (unless the unweighted approach is used, i.e., when **--unweighted** is specified), trait column (with trait values being continuous or binary if *model* = `linear` or `logistic`, respectively), time and event indicator columns (if *model* = `coxph`), and covariates columns (unless there is no covariate in *formula*). Subjects with missing values in any of the columns specified by [**--formula** *formula*], [**--id-col** *iid*], [**--family-col** *fid*], or [**--weight-col** *wt*] are excluded from the analysis.

5.2 VCF File (*vcf_file.gz*)

The VCF file contains the genotype data. The format specifications of a VCF file can be found on this web page. The VCF file should be compressed by bgzip and indexed by tabix, using the following command:

```
## this command will produce vcf_file.gz
$ bgzip vcf_file

## this command will produce vcf_file.gz.tbi
$ tabix -p vcf -f vcf_file.gz
```

5.3 Pairwise Inclusion Probability Matrix

The files that contain the pairwise inclusion probability matrices should be tab-delimited. The 1st row is the header line containing the subject IDs. The remaining rows constitute a symmetric square matrix. That is to say, the number of rows equals the number of columns plus 1 (for the header line). The marginal inclusion probability of the ith subject is in the $(i+1)$th row and ith column. The pairwise inclusion probability of the ith and jth subjects is in the $(i+1)$th row and jth column, as well as in the $(j+1)$th row and ith column. All inclusion probabilities are strictly greater than 0 and less than or equal to 1. Missing values are not permitted. Note that there can be multiple pairwise inclusion probability matrices. Subjects in different pairwise inclusion probability matrices are assumed to be independent. Note that these pairwise inclusion probability matrices are optional in the weighted approach and not needed in the unweighted approach.

5.4 File that Contains the File Names of the Pairwise Inclusion Probability Matrices (*prob_file*)

Each row is the file name of one pairwise inclusion probability matrix. Note that this file is optional in the weighted approach and not needed in the unweighted approach.

5.5 File that Contains the Variants for Conditional Analysis (*cond_file*)

Each row is a variant ID, which should be in `chromosome:position` format. Note that this file is needed only when we perform conditional analysis (i.e., when [**--cond** *cond_file*] is specified).

5.6 File that Contains Variants' Grouping Information in Gene-based Analysis (*group_file*)

Each row is a gene, which should be in the following format:

$$\texttt{gene}_1 \quad \texttt{variant}_1, \texttt{variant}_2, \ldots, \texttt{variant}_m$$

The gene and variant IDs are separated by a tab. The variant IDs in the same gene are separately by "`,`". Variant IDs should be in `chromosome:position` format. Note that this file is needed only when we perform gene-based analysis (i.e., when [**--group** *group_file*] is specified).

5.7 File that Contains the Subset of Variants to be Analyzed in Single-Variant Analysis (*extract_file*)

Each row is a variant ID, which should be in `chromosome:position` format. Note that this file is needed only when [**--extract-file** *extract_file*] is specified.

6. OUTPUT FILES

6.1 Wald Statistics

6.1.1 Single-Variant Analysis Results (*prefix.wald.out*)

The rows represent SNPs. The first row is the header line. Missing values are denoted by `NA`. Tables 1–3 describe the columns of *prefix.wald.out* in standard association analysis, conditional analysis, and gene-environment interaction analysis, respectively.

Table 1. Column Description for *prefix.wald.out* in Standard Association Analysis

Column Name	Description
CHROM	Chromosome.
POS	Position.
VCF_ID	Varaint ID in the VCF file.
REF	Reference allele.
ALT	Alternative allele.
ALT_AF	Alternative allele frequency.
ALT_AC	Alternative allele count.
N_INFORMATIVE	Number of subjects included in the analysis.
N_REF	Number of subjects with two reference alleles.
N_HET	Number of subjects with one reference and one alternative alleles.

N_ALT	Number of subjects with two alternative alleles.
N_DOSE	Number of subjects with genotype dosages.
BETA	Effect estimate.
SE	Standard error estimate of BETA.
PVALUE	p-value.

Table 2. Column Description for *prefix.wald.out* in Conditional Analysis

Column Name	Description
CHROM	Chromosome.
POS	Position.
VCF_ID	Varaint ID in the VCF file.
REF	Reference allele.
ALT	Alternative allele.
ALT_AF	Alternative allele frequency.
ALT_AC	Alternative allele count.
N_INFORMATIVE	Number of subjects included in the analysis.
N_REF	Number of subjects with two reference alleles.
N_HET	Number of subjects with one reference and one alternative alleles.
N_ALT	Number of subjects with two alternative alleles.
N_DOSE	Number of subjects with genotype dosages.
BETA	Effect estimate.
SE	Standard error estimate of BETA.
PVALUE	p-value.
BETA_`variant`	Effect estimate of `variant` that is conditioned on.
SE_`variant`$_i$	Standard error estimate of BETA_`variant`.
PVALUE_`variant`	p-value of `variant` that is conditioned on.

Table 3. Column Description for the Gene-Environment Interaction Analysis Output File

Column Name	Description
CHROM	Chromosome.
POS	Position.
VCF_ID	Varaint ID in the VCF file.
REF	Reference allele.
ALT	Alternative allele.
ALT_AF	Alternative allele frequency.
ALT_AC	Alternative allele count.
N_INFORMATIVE	Number of subjects included in the analysis.
N_REF	Number of subjects with two reference alleles.

N_HET	Number of subjects with one reference and one alternative alleles.
N_ALT	Number of subjects with two alternative alleles.
N_DOSE	Number of subjects with genotype dosages.
PVALUE_G	p-value of the genetic variable.
PVALUE_INTER	p-value of the interaction term(s) between the genetic variable and the environment variable(s) `envi`.
PVALUE_BOTH	p-value of both the genetic variable and the interaction terms.
BETA_G	Effect estimate of the genetic variable.
BETA_`envi`	Effect estimate of the environment variable `envi`.
BETA_G:`envi`	Effect estimate of the interaction term between the genetic variable and the environment variable `envi`, denoted by G:`envi`.
COV_G_G	Variance estimate of Beta_G.
COV_`envi`_`envi`	Variance estimate of Beta_`envi`.
COV_G:`envi`_G:`envi`	Variance estimate of Beta_G:`envi`.
COV_G_`envi`	Covariance estimate between Beta_G and Beta_`envi`.
COV_G_G:`envi`	Covariance estimate between Beta_G and Beta_G:`envi`.
COV_`envi`_G:`envi`	Covariance estimate between Beta_`envi` and Beta_G:`envi`.

6.2 Score Statistics

6.2.1 Single-Variant Analysis Results (*prefix.score.out*)

The rows represent SNPs. The first row is the header line. Missing values are denoted by `NA`. Tables 4 describe the columns of *prefix.score.out* in standard association analysis.

Table 4. Column Description for *prefix.score.snp.out* in Standard Association Analysis

Column Name	Description
GENE_ID	Gene ID. In single-variant analysis (i.e., [**--group** *group_file*] is not specified), GENE_ID equals CHROM:POS.
CHROM	Chromosome.
POS	Position.
VCF_ID	Varaint ID in the VCF file.
REF	Reference allele.
ALT	Alternative allele.
ALT_AF	Alternative allele frequency.
ALT_AC	Alternative allele count.
N_INFORMATIVE	Number of subjects included in the analysis.
N_REF	Number of subjects with two reference alleles.

N_HET	Number of subjects with one reference and one alternative alleles.
N_ALT	Number of subjects with two alternative alleles.
N_DOSE	Number of subjects with genotype dosages.
U	Score statistic.
V	Variance estimate of U.
BETA	Effect estimate.
SE	Standard error estimate of BETA.
PVALUE	*p*-value.

6.2.2 Gene-Based Summary Statistics (*prefix.mass.out*)

The gene-based summary statistics are stored in MASS 7.0 format. They can be loaded into the software program MASS to perform all commonly used gene-based association tests. They can also be converted by the software program PreMeta to files that are compatible with other commonly used rare-variant meta-analysis software programs, including RAREMETAL, seqMeta, and MetaSKAT.

7. VERSION HISTORY

- 1.0 (released on May 29th, 2013)

  First version released.

- 2.0 (released on Nov 12nd, 2013)

  1. Added the capability to perform gene-environment interaction analysis.
  2. Deleted the tab delimiter at the end of each row in *outfile*.

- 3.0 (released on Dec 7th, 2013)

  Added the capability to perform logistic regression for binary (0/1) traits.

- 4.0 (released on Feb 9th, 2014)

  Added the capability to analyze data with multiple pairwise inclusion probability matrices.

- 4.1 (released on Mar 13rd, 2014)

  Added the capability to deal with imputed genotype dosages.

- 5.0 (released on May 21st, 2014)

  1. Modified the variance estimation formula. Included both the model-based and robust variance estimators.

2. Changed the format of the phenotype file.

- 5.1 (released on Aug 14th, 2014)

  Added the capability to perform conditional analysis.

- 5.2 (released on Sep 21st, 2014)

  Modified the variance estimation formula. Used a new approach to trim the pairwise inclusion probabilities.

- 6.0 (released on Oct 1st, 2014)

  Added the unweighted approach.

- 6.1 (released on Oct 6th, 2014)

  Changed some option names. Changed some column names in output files.

- 6.2 (released on Nov 18th, 2014)

  Changed the name of the software program from "SOLReg" to "SUGEN".

- 6.3 (released on Nov 13rd, 2015)

  Improved the computational efficiency of unweighted analysis.

- 7.0 (released on March 30th, 2016)

  Improved the user interface. Changed the genotype file format from plain text to VCF. Added the capability to perform gene-based association analysis.

- 7.1 (released on May 2nd, 2016)

  Added the capability to handle dosage data.

- 7.2 (released on May 5th, 2016)

  Fixed a bug in reading the phenotype file when it contains redundant columns.

- 7.3 (released on May 30th, 2016)

  1. Fixed a bug in gene-environment interaction analysis where the environment variable is the last covariate in the model.
  2. Added the **--subset** option.
  3. Added the **--hetero-variance** option.
  4. Modified the model-based variance estimator so that it is stable for rare variants.

- 8 (released on September 29, 2016)

1. Added the capability to perform Cox proportional hazard regression.
2. Modified the model-based covariance matrix estimator in gene-based tests so that it is more accurate for rare variants.
3. Fixed a bug in reading the phenotype file when the subject ID or family ID column is the last column of the phenotype file.

- 8.1 (current version, released on November 2, 2016)

1. Added p-values in the gene-environment interaction analysis output file.
2. Fixed a bug in the weighted approach.