SUGEN 8.6 Overview

Misa Graff, July 2017

General Information

- By Ran Tao, https://sites.google.com/site/dragontaoran/home
- Website: http://dlin.web.unc.edu/software/sugen/
- Standalone command-line software program in C++
- GEE to account for relatedness
- Capable of performing a pooled-analysis of 50,000+ subjects
- Model-based variance estimator accurate for low-frequency variants
  - *equivalent to standard regression for independent samples*
- Continuous and binary traits
- Single-variant analysis using Wald statistics
  - *standard, conditional, gene-environment interaction analysis*
- Gene-based rare-variant analysis based on score statistics

Synopsis

- $ SUGEN [--pheno pheno_file] [--formula formula] [--id-col iid] [--family-col fid] [--weight-col wt] [--vcf vcf_file.gz] [--probmatrix prob_file] [--unweighted] [--model model] [--robust-variance] [--cond cond_file] [--ge envi_covs] [--score] [--score-rescale rescale_rule] [--group group_file] [--group-maf maf_ub] [--group-callrate cr_lb] [--out-prex out prex] [--out-zip] [--extract-chr chr] [--extract-range range] [--extract-file extract_file]

Input options

- **--pheno** {pheno_file}: specifies the phenotype file
  - *Missing = NA*
- **--formula** {formula}: specifies the regression formula
  - *“trait = age + gender + pc1 + pc2”*
  - *“trait ="*
  - *“(time, event)=age + gender + pc1 + pc2”*
  - *“(time, event=”*
- **--id-col** {iid}: specifies the subject ID column in the pheno file
- **--family-col** {fid}: specifies the family ID column in the pheno file
  - *when subjects are independent, set fid = iid*

Input options

- **--weight-col** {WT}: for weighted analyses, specifies the weight column in pheno_file. The default column name is WT.
- **--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.
- **--unweighted**: specifies unweighted analyses
- **--vcf** {vcf_file.gz}: specifies the VCF file
- **--dosage**: Analyzes dosage data in the VCF file. The dosages must be stored in the *DS* field of the VCF file.

Input options

- **--model** {model}: Specifies the regression model. The default value is linear.
  - *linear (linear regression: trait is continuous),*
  - *logistic (logistic regression: trait is binary 0/1),*
  - *coxph (Cox proportional hazards regression: event time is positive, and the event indicator is binary)*
- **--left-truncation** {left_truncation_time}: Specifies the left truncation time (if any) in Cox proportional hazards regression
- **--robust-variance**: the robust variance estimator will be used, otherwise the model-based variance estimator will be used.
- **--cond** {cond-file}: performs conditional analysis conditioning on the variants included in cond_file.

Input options

- **--ge** {envi_covs}: In single-variant analysis, performs gene-environment interaction analysis.
  - *envi_covs are the names of the environment variables.*
  - *format of envi_covs is covariate_1,covariate_2,...,covariate_k.*
    - multiple environment variables are separately by commas.
  - *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. Option for weighted analyses.
- **--group** {group_file}: Performs gene-based association analysis. Gene memberships of variants are defined in 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. Can help reduce inflation.

Output options

- **--out-prefix** {out_prefix}: specifies the prefix of the output files
- **--out-zip**: compresses the output files
- **--extract-chr** {chr}: restricts single-variant analysis to variants in chromosome chr
- **--extract-range** {range}: restricts single-variant analysis to variants
- in a specic range in chromosome chr
  - *range format: 1000000-2000000*
- **--extract-file** {extract_file}: restricts single-variant analysis to variants in extract file
  - variant IDs in *chromosome:position format*

Phenotype File

fid	scanID	EV1	EV2	sex	sex_num	age	trait	disease	group	group_unc	group_uw
1	p36	2.074	0.4816	M	1	23	89.075	0	uw	0	1
1	p207	0.9547	-0.0854	F	2	69	273.06	0	unc	1	0
2	p370	0.7555	-0.5965	F	2	59	234.74	1	unc	1	0
2	p202	0.9292	-0.8418	F	2	19	87.782	0	unc	1	0
3	p290	-0.71	1.2707	M	1	39	140.03	0	unc	1	0
3	p421	-0.2321	-0.9616	M	1	56	239.42	0	unc	1	0

- Tab delimited text file.
- Missing data denoted by \NA"
- Required columns: subject ID (must match VCF file), family ID, trait
- binary trait coded as 0/1
- Optional columns: covariates
- Trait and covariates must be numeric

VCF File

- Compressed by bgzip and indexed by tabix, using the following commands:
- To produce the vcf_file.gz

  $ bgzip vcf_file
- To produce vcf_file.gz.tbi

  $ tabix -p vcf -f vcf_file.gz

Other input files

- cond_file specified in option **--cond** cond_file
  - *each row is a variant*
  - *variant ID in chromosome:position format*
- group_file specified in option **--group** group_file

  gene1 1:1000,1:1003

  gene2 3:50000,3:50354,3:54352
  - *group and variant IDs are separated by a tab*
- extract_file specified in option **--extract-file** extract_file
  - *same format as cond_file*

Output files

- Single-variant analysis
  - *chromosome, position, reference/alternative alleles, allele counts, allele frequency, effect estimates, standard error estimates, p-values*
- Gene-based analysis
  - *single-variant results*
  - *score statistics and their covariance matrices in MASS format*

Gene-based Rare-variant Tests Using MASS

- MASS website: http://dlin.web.unc.edu/software/mass/
- Performs burden, CMC, variable threshold, SKAT, SKAT-O tests
- Calculates asymptotic and resampling p-values
- Performs fixed effects and trans-ethnic meta-analysis
- Performs conditional analysis
- Summary statistics generated by SUGEN can be directly loaded into MASS gene association software.