Introduction to THRESHOLD

THRESHOLD is a command-line package to implement a bootstrap approach to determining genome-wide significance threshold for association tests in sequencing studies. It currently covers single-variant tests, Burden tests, and SKAT tests.

There are three downloads for THRESHOLD: 1) this documentation file 2) The THRESHOLD source files and compiled Linux executables and 3) examples\_THRESHOLD containing a set of data and scripts for a complete GWAS example using the genomic data from the 1000 Genomes project.

Compilation of the THRESHOLD programs

The executables for threshold analysis are compiled by issuing the command “bash ./makescript.sh” within the THRESHOLD base directory. This produces the following executables in the base directory: single\_var\_thresh for single-variant simulations, thresh\_region for region-based simulations, and thresh\_results for the calculation of the threshold value from the results produced by either single\_var\_thresh or thresh\_region. In addition to the executables in this directory, the makescript.sh utility also produces output files in the single\_var\_thresh\_code and thresh\_region\_code subdirectories. These output files and the executables in the base directory may be removed with the command “bash ./makescript.sh clean”.

Basic Usage

The single\_var\_thresh program takes one command line argument, the name of a text file containing simulation parameters. The THRESHOLD root directory contains a sample configuration file config\_file\_no\_pheno.txt. A single variant simulation may be run by entering the following at the command line prompt:

./single\_var\_thresh config\_file\_no\_pheno.txt

This will produce as output config\_out\_no\_pheno.txt. The name of the output file is determined by the line containing “out\_filename” in the config\_file\_no\_pheno.txt file. The config\_out\_no\_pheno.txt file contains ascii format intermediate results that can be analyzed by the thresh\_results program:

./ thresh\_results list\_file.txt > thresh\_results\_config\_out\_no\_pheno.txt

The list\_file.txt file in the THRESHOLD directory is a text file containing the string “config\_out\_no\_pheno.txt”.

The output from the threshold calculations (thresh\_results\_config\_out\_no\_pheno.txt) contains the threshold calculated for the single-variant analysis that was performed.

Region-based simulations may be performed with the thresh\_region executable, in a manner analogous to the use of single\_var\_thresh for single-variant simulations. The thresh\_results executable may be used to analyze the results of region-based simulations just as it is used for single-variant analysis.

Details of program options are given in the sections below.

Bootstrap with single\_var\_thresh

The single\_var\_thresh program takes one command line argument.

Usage:./single\_var\_thresh config\_filename

The config\_filename argument should be the name of a text file containing program settings. Each line of the configuration file should have the form

keyword value

where “keyword” is one of several strings indicating a program parameter, and value is a setting for that parameter. The keyword and value may be separated by any combination of tabs or spaces at least one character long.

The keywords and values that may be given in the configuration file are the following:

**geno\_filename**

The required geno\_filename value should be the name of a text file containing genomic data for all subjects. Each line of the file should contain the name of one SNP followed by a string of whitespace delimited values of 0, 1, 2, or the string “NA”. The string “NA” represents missing data. Missing data for a SNP will be imputed as the mean value of all the non-missing subjects for that SNP. The geno.txt file included in the single\_var\_code directory is a sample file in the correct format containing artificial data for five SNPs and 100 subjects.

**out\_filename**

The required out\_filename argument will be used as the name of a file containing bootstrap intermediate results that will be used as input to the thresh\_results program. If a file with this name already exists it will be overwritten. Each line in the output file corresponds to one bootstrap step.

**start\_rep**

The required start\_rep parameter should be a non-negative integer value and is used to determine the sequence of pseudo-random numbers generated by the simulation. If a simulation is rerun with identical input including this value, then the results will be identical to the previous run. This aspect of the program allows for a large number of replicates to be simulated on different CPUs independently. The output files may then be concatenated.

**num\_reps**

The num\_reps argument should be a positive integer indicating the number of bootstrap samples to be run. This value is required.

**pheno\_filename**

This value must be specified for simulations using phenotype and covariate.

In order to conduct simulations with phenotype and covariates, the pheno\_filename argument must be provided. This should be the name of a file containing phenotype and any covariate information. The first (leftmost) column of the file contains phenotype information and any subsequent column contains covariate data. There should be one row for each subject that is present in the genotype file, in the same relative order. The phenotype column is the only column that is required. No constant covariate should be included in the phenotype file. If the phenotype is missing for a subject it should be indicated with the string “NA”, in which case that subject will be excluded from the simulation. An example phenotype file is included in the distribution (single\_var\_thresh\_code/pheno.txt), which contains three columns: the first column gives phenotype data, and the next two columns contain covariate information for a continuous and a binary covariate respectively.

**min\_call\_rate**

This is an optional parameter with a default value of 0.9. If given, min\_call\_rate must be set to a positive floating point value less than or equal to 1.0. The default value of 0.9 indicates that for a given SNP to be considered in the calculations, at least 90% of the subjects must have a non-missing value for that SNP.

**MAF\_cutoff**

The value of MAF\_cutoff whould be a floating point value between 0 and 1. The simulation will remove any SNP with a minor allele frequency that is less than this value. The MAF\_cutoff must be specified for all simulations.

**out\_freq**

If specified, out\_freq must be a non-negative integer. This parameter controls the number of steps between successive writes of informational data to standard output during the bootstrap simulation. The default value is 1000.

**Sample Files**

The single\_var\_thresh\_code/ config\_file\_pheno.txt and single\_var\_thresh\_code/ config\_file\_no\_pheno.txt files are sample configuration files for simulations with and without phenotype and covariate data. The output files from these simulations, which contain the data that is to be used as input to the thresh\_results executable, are config\_out\_pheno.txt and config\_out\_no\_pheno.txt respectively.

Bootstrap with thresh\_region

The thresh\_region program takes one command line arguments.

Usage:./thresh\_region config\_filename

The config\_filename argument should be the name of a text file containing program settings. Each line of the configuration file should have the form

keyword value

where “keyword” is one of several strings indicating a program parameter, and value is a setting for that parameter. The keyword and value may be separated by any combination of tabs or spaces at least one character long.

The keywords and values that may be given in the configuration file are the following:

**geno\_filename**

The required geno\_filename value should be the name of a text file containing genomic data for all subjects. Each line of the file should contain the name of one SNP followed by a string of whitespace delimited values of 0, 1, or 2. The string “NA” represents missing data. Missing data for a SNP will be imputed as the mean value of all the non-missing subjects for that SNP. The geno.txt file included in the single\_var\_code directory is a sample file in the correct format containing artificial data for five SNPs and 100 subjects.

**out\_filename**

The required out\_filename argument will be used as a prefix for the name of a file containing bootstrap intermediate results that will be used as input to the thresh\_results program. The file that is written to disk will have the suffix “..BURDEN.txt” or “.SKAT.txt” (both fileswill be produced if both Burden and SKAT simulations are performed). If a file with this name already exists it will be overwritten. Each line in the output file corresponds to one bootstrap step.

**start\_rep**

The required start\_rep parameter should be a non-negative integer value and is used to determine the sequence of pseudo-random numbers generated by the simulation. If a simulation is rerun with identical input including this value, then the results will be identical to the previous run. This aspect of the program allows for a large number of replicates to be simulated on different CPUs independently. The output files may then be concatenated.

**num\_reps**

The num\_reps argument should be a positive integer indicating the number of bootstrap samples to be run. This value is required.

**pheno\_filename**

This value must be specified for simulations using phenotype and covariate.

In order to conduct simulations with phenotype and covariates, the pheno\_filename argument must be provided. This should be the name of a file containing phenotype and any covariate information. The first (leftmost) column of the file contains phenotype information and any subsequent column contains covariate data. There should be one row for each subject that is present in the genotype file, in the same relative order. The phenotype column is the only column that is required. No constant covariate should be included in the phenotype file. If the phenotype is missing for a subject it should be indicated with the string “NA”, in which case that subject will be excluded from the simulation. An example phenotype file is included in the distribution (single\_var\_thresh\_code/pheno.txt), which contains three columns: the first column gives phenotype data, and the next two columns contain covariate information for a continuous and a binary covariate respectively.

**MAF\_cutoff**

The value of MAF\_cutoff whould be a floating point value between 0 and 1. The simulation will remove any SNP with a minor allele frequency (MAF) that is greater than this value. The MAF\_cutoff must be specified for all simulations.

**min\_call\_rate**

This is an optional parameter with a default value of 0.9. If given, min\_call\_rate must be set to a positive floating point value less than or equal to 1.0. The default value of 0.9 indicates that for a given SNP to be considered in the calculations, at least 90% of the subjects must have a non-missing value for that SNP.

**min\_MAC**

The value of min\_MAC should be a non-negative integer. Any SNP with a minor allele count less than this value will be removed from the simulation. This value may be omitted - the default value is 0.

**out\_freq**

If specified, out\_freq must be a non-negative integer. This parameter controls the number of steps between successive writes of informational data to standard output during the bootstrap simulation. The default value is 1000.

**sim\_type**

The sim\_type keyword is required and it must be set to BURDEN, SKAT, or BOTH. The last setting will results in both Burden and SKAT simulations being performed.

**map\_filename**

The map\_filename is required for gene-based simulations. This parameter whould be set to the filename of a text file that maps SNPs to genes. The map file should contain lines of the form “GENE SNP” where GENE is the name of a specific gene with SNP as one of its components. The file thresh\_region\_code/map.txt is an example of a map file.

The map\_filename parameter is incompatible with the window\_size and window\_sgtep parameters.

**window\_size**

The length in SNPs of a region for windows-based analysis. This parameter should be set to a positive integer, and it is incompatible with the map\_filename parameter. If window\_size is specified, then window\_step must also be specified.

**window\_step**

The number of SNPs to advance between successive regions for region-based analysis.. If window\_snp is set equal to sindow\_size then non-overlapping regions will be used. If these values are not equal, there will be overlap between successive regions. For example, if the SNPs to be divided into regions are “ABCDE” a window\_size of three with a window\_step of two would result in the following regions: “ABC”, and “CDE”. If window\_step is specified then window\_size must also be specified. These parameters are incompatible with the map\_file parameter.

Note that the window\_step and window\_size options result in the creation of a mapping file with a name created from the geno\_filename and the window options. The name of this file will be written to standard output during the simulation. For subsequent simulations using the same regions, the produced mapping file may be read instead of using the windows options again.

**Sample Files**

The thresh\_region\_code/config\_pheno\_windows\_Burden.txt and thresh\_region\_code/ config\_no\_pheno\_map\_SKAT.txt files are sample configuration files for a simulations with phenotype and covariates using the Burden score and a simulation without phenotype and covariate data using SKAT respectively. The output files from these simulations, which contain the data that is to be used as input to the thresh\_results executable, are config\_pheno\_windows\_out.BURDEN.txt and config\_no\_pheno\_map\_out.SKAT.txt respectively.

Threshold calculation with thresh\_results

The thresh\_results program takes one command line argument.

Usage: ../thresh\_results listfilename

The listfilename parameter should be the name of a filename containing the names of output files from either single\_var\_thresh or thresh\_region, along with an optional single character. The character should be either b or B for Burden or single-variant simulations, or s or S for SKAT. If a single character value is not given then the file will be assumed to be from a single-variant or Burden simulation. The file thresh\_region\_code/listfile.txt is an example of this format.

Blank lines in the listfile are ignored by the thresh\_results program.

Note that for the threshold calculations to be valid, all threshold results files listed in the listfile must come from simulations with the same start\_rep and num\_reps values.

If there is more than one column in the output file to be analyzed then the minimum p-value for each step will be found prior to the calculation of the threshold. An example of when this may be useful is if the simulations were conducted on two different processors after division of the input data into two groups. In that case, the output files from the two simulations could be combined (using the Unix paste command) to determine a threshold.

The thresh\_region\_code/thresh\_results.txt file was generated with the following command:

thresh\_results listfile.txt >& thresh\_results.txt

Examples Using Genomic Data from the 1000 Genomes Project

The archive examplesTHRESHOLD.tar.gz, available for download, contains input files and results for calculations using the genomic data from the 1000 genomes project. For these calculations, 1000 bootstrap samples were run. The genomic data for each chromosome is provided in a subdirectory of the base examples directory e.g. CHR01, CHR02, etc. By splitting the genomic and map information for each chromosome into many different files, the simulations may be parallelized.

The 1000 genome data does not involve covariates or phenotype, so synthetic covariate and phenotype data was prepared for demonstration purposes. The example/pheno\_random.txt file contains simulated phenotype data (sampled from a standard normal distribution), along with two covariates: one continuous covariate from a standard normal distribution and one binary covariate sampled from a Bernoulli distribution with p=0.5.

There are simulations using window-derived map files, as well as those using map files determined from functional annotation. The map files in the chromosome subdirectories were prepared using non-overlapping windows of length 50 or 5000. The 1000gen\_allchr\_LOF.map.txt map file in examples was prepared using functional annotations (note that this file is to be used with the examples/ 1000gen\_no\_multiallelic\_allchr\_nonsyn.geno\_1.txt genome file).

Running Simulations using single\_var\_thresh and thresh\_region

There are Bash shell scripts in the examples that prepare and submit simulations using either the single\_var\_thresh or the thresh\_region executable. These scripts prepare the necessary configuration files for each simulation in an automated fashion. The Bash scripts were written for systems using the Slurm job scheduler, and must be modified for other environments. The location of the single\_var\_thresh or thresh\_region executable may be modified by changing the value of the “PART1” variable in these files. Note that many of the data files used in the calculation of the thresholds were compressed and may be uncompressed with the Bash script gunzip\_loop.sh in the examples directory.

Simulations may be run using the functional annotation mapping files with the scripts 1000gen\_LOF\_Burden.sh and 1000gen\_LOF\_Burden\_pheno.sh, for Burden simulations without and with phenotype respectively. Note that all LOF example files are in the base examples/ directory. The submit\_thresh\_region.LOF.SKAT.sh script is for SKAT simulations without phenotype, and submit\_thresh\_region.LOF.pheno.SKAT.sh is for SKAT simulations with phenotype. For the SKAT scripts, each simulation of 1000 bootsstrap samples was broken up into 4 groups of 250 steps each. The output data from these simulations can be combined with the combine\_LOF\_SKAT.sh and combine\_LOF\_SKAT\_pheno.sh scripts.

There are several Bash scripts in the examples/ directory that were used for the remaining simulations. Each of the following scripts uses the genomic and mapping file data in the chromosome subdirectories. These scripts are:

submit\_single\_var\_thresh\_pheno.sh : Single-variant with phenotype.

submit\_single\_var\_thresh.sh : Single-variant without phenotype.

submit\_thresh\_region.5000.Burden.sh : Burden, no phenotype, window 5000

submit\_thresh\_region.50.Burden.pheno.sh : Burden, with phenotype, window 50

submit\_thresh\_region.50.Burden.sh : Burden, no phenotype, window 50

submit\_thresh\_region.50.SKAT.pheno.sh : SKAT, with phenotype, window 50

submit\_thresh\_region.50.SKAT.sh : SKAT, no phenotype, window 50

Scripts to Combine Output Files from Simulations

Many of the Bash scripts that were used for simulation produce many different output files for each chromosome. In order to combine all the individual results for each chromosome, a number of separate Bash scripts were used for post-simulation processing. These scripts, all in the examples directory are:

thresh\_combine\_window\_5000\_no\_pheno\_BURDEN.sh

thresh\_combine\_window\_50\_no\_pheno\_BURDEN.sh

thresh\_combine\_window\_50\_no\_pheno\_SKAT.sh

thresh\_combine\_window\_50\_pheno\_BURDEN.sh

thresh\_combine\_window\_50\_pheno\_SKAT.sh

thresh\_combine\_window\_sv\_no\_pheno.sh

thresh\_combine\_window\_sv\_pheno.sh

The output files from these utility scripts are the following text files, which are located in the examples directory, with copies in the examples/results subdirectory.

all\_temp.thresh\_combine.5000.no\_pheno\_BURDEN.txt

all\_temp.thresh\_combine.50.no\_pheno\_BURDEN.txt

all\_temp.thresh\_combine.50.no\_pheno\_SKAT.txt

all\_temp.thresh\_combine.50.pheno\_BURDEN.txt

all\_temp.thresh\_combine.50.pheno\_SKAT.txt

all\_temp.thresh\_combine.sv.no\_pheno.txt

all\_temp.thresh\_combine.sv.pheno.txt

These text files are used as input data for the thresh\_results program. Each of these files contains 1000 lines, one for every bootstrap resampling step. Each chromosome corresponds to a number of different columns. For example the calculations for chromosome 22 with the Burden score for windows of length 5000 used 106 different jobs, so there are 106 columns corresponding to this chromosome in the file all\_temp.thresh\_combine.5000.no\_pheno\_BURDEN.txt, and similarly for the other chromosomes. This file format allows for a simple parallelization of the simulations. If the data for all chromosomes had been performed sequentially on one CPU then the final output would contain one column and 1000 rows. The final threshold value would be the same regardless of this detail.

Threshold Calculation with thresh\_results

The examples/results directory contains input data and utilities to calculate the thresholds from the simulation data. To calculate the thresholds given these files, the thresh\_results executable should be in the user’s path.

The results of the threshold calculations (significance level = 0.05) are given in the table below.

|  |  |
| --- | --- |
| Simulation | Threshold |
| Window: 5000, Burden | 6.78931e-06 |
| Window: 50, Burden | 2.78633e-08 |
| Window: 50, SKAT | 5.84275e-08 |
| Window: 50, Burden, phenotype | 2.85728e-08 |
| Window: 50, SKAT, phenotype | 5.21845e-08 |
| Single-variant | 1.01998e-08 |
| Single-variant, phenotype | 1.22293e-08 |
| LOF, Burden | 2.25087e-05 |
| LOF, Burden, phenotype | 2.21253e-05 |
| LOF, SKAT | 2.18543e-05 |
| LOF, SKAT, phenotype | 2.32581e-05 |
| All (Combined) | 6.0944e-09 |