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 |