MEGHA – Massively Expedited Genome-wide Heritability Analysis

Accompanying publication

Tian Ge, Thomas E. Nichols, Phil H. Lee, Avram J. Holmes, Joshua L. Roffman, Randy L. Buckner, Mert R. Sabuncu, and Jordan W. Smoller.
Massively Expedited Genome-wide Heritability Analysis (MEGHA).
Proceedings of the National Academy of Sciences, 12(8), 2479-2484, 2015.

Documentation

http://scholar.harvard.edu/tge/software/megha

Contact

Please direct all bug reports and questions to Tian Ge at
tge1@mgh.harvard.edu

System and software requirements

- MATLAB
- FreeSurfer (https://surfer.nmr.mgh.harvard.edu/fswiki) MATLAB tools (freesurfer/matlab) and SurfStat (http://www.math.mcgill.ca/keith/surfstat) for surface-based analysis and clustering
- May require huge RAM, depending on the sample size and the number of phenotypes

MATLAB scripts included in the downloaded

- Example.m (examples)
- MEGHA.m (MATLAB function for MEGHA that can take the same phenotypic, covariate and GRM files as GCTA)
- MEGHAmat.m (a simplified version of MEGHA.m when phenotypic data, covariates and GRM have been prepared as MATLAB .mat format)
- MEGHASurf.m (MATLAB function for applying MEGHA to surface data)
- MEGHASurfmat.m (a simplified version of MEGHAsurf.m when surface data, covariates and GRM have been prepared as MATLAB .mat format)
- ParseFile.m
- ParseGRM.m
- ReadFile.m
- ExtractSurf.m
- ReadFileSurf.m
- WriteStats.m
- WriteMap.m

Installation

No installation needed. But for surface-based analysis and clustering, add FreeSurfer MATLAB tools and SurfStat package to the Path.

MEGHA.m

Input arguments:

- PhenoFile: a plain text file containing phenotypic data (Col 1: family ID; Col 2: subject ID; From Col 3: numerical values)
- header: 1 – PhenoFile contains a headerline; 0 – PhenoFile does not contain a headerline
- CovFile: a plain text file containing covariates (intercept NOT included. Col 1: family ID; Col 2: subject ID; From Col 3: numerical values. If no covariate needs to be included in the model, set CovFile = '')
- delimiter: delimiter used in the phenotypic file and covariates file
- GRMFile: a plain text file containing the lower trianglar elements of the GRM (Col 1 & 2: indices of pairs of individuals; Col 3: number of non-missing SNPs; Col 4: the estimate of genetic relatedness)
- GRMid: a plain text file of subject IDs corresponding to GRMFile (Col 1: family ID; Col 2: subject ID)

[Note: Subjects in PhenoFile, CovFile and GRMFile do not need to be exactly the same. This function will find the subjects in common in these files and sort the order of the subjects.]

- Nperm: number of permutations; set Nperm = 0 if permutation inference is not needed; default Nperm = 0
- WriteStat: 1 – write point estimates and significant measures of heritability to the output directory "OutDir"; 0 – do not write statistics; default WriteStat = 0; default output directory is the working directory

Output arguments:

- Pval: MEGHA p-values
- h2: MAGHA estimates of heritability magnitude
- SE: estimated standard error of heritability magnitude
- PermPval: MEGHA permutation p-values
- PermFWEcPval: MEGHA family-wise error corrected permutation p-values
- Nsubj: total number of subjects in common
- Npheno: number of phenotypes
- Ncov: number of covariates (including intercept)

Output files (if WriteStat=1):

- MEGHAstat.txt: point estimates and significant measures of heritability for each phenotype written to the output directory "OutDir"

MEGHAmat.m

Input arguments:

- Pheno: an Nsubj x Npheno phenotypic data matrix
- Cov: an Nsubj x Ncov covariates matrix (intercept included)
- K: an Nsubj x Nsubj GRM

[Note: Subjects in Pheno, Cov and K must be exactly the same and arranged in the same order.]

For other input and output arguments see MEGHA.m.

MEGHASurf.m

Input arguments:

- SurfDir: directory of the surface data (SUBJECTS_DIR)
- ImgSubj: a plain text file containing a list of subject IDs with imaging data to be included in the analysis
- ImgFileLh/ImgFileRh: name of the file containing surface data for the left/right hemisphere
- FSDir: directory of FreeSurfer where the folder "subjects" can be found (FREESURFER_DIR)

[Note: Image files need to be organized in the FreeSurfer convention, i.e., SurfDir/ImgSubj{i}/surf/ImageFile?h.
FSDir/subjects/fsaverage/label/?h.cortex.label and
FSDir/subjects/fsaverage/surf/?h.sphere.reg must be existing.
Please make sure that each path directory specified ends with a '/']

- CovFile: a plain text file containing covariates (intercept NOT included. Col 1: family ID; Col 2: subject ID; From Col 3: numerical values. If no covariates need to be included in the model, set CovFile = '')
- delimiter: delimiter used in the covariates file
- GRMFile: a plain text file containing the lower trianglar elements of the GRM (Col 1 & 2: indices of pairs of individuals; Col 3: number of non-missing SNPs; Col 4: the estimate of genetic relatedness)
- GRMid: a plain text file of subject IDs corresponding to GRMFile (Col 1: family ID; Col 2: subject ID)

[Note: Subjects in ImgSubj, CovFile and GRMFile do not need to be exactly the same. This function will find the subjects in common in these files and sort the order of the subjects.]

- WriteImg: 1 - write spatial maps for heritability estimates and -log10(p-values) to the output directory "OutDir"; 0 - do not write spatial maps; default WriteImg = 0; default output directory is the working directory
- Nperm: number of permutations for cluster inference; set Nperm = 0 if permutation inference is not needed; default Nperm = 0
- Pthre: p-value threshold for permutation based cluster inference

Output arguments:

- PvalLh/PvalRh: MEGHA p-values for in-mask vertices on the left/right hemisphere
- h2Lh/h2Rh: MEGHA heritability estimates for in-mask vertices on the left/right hemisphere
- SE: estimated standard error of heritability magnitude
- ClusPLh/ClusPRh: family-wise error corrected p-values for clusters above the threshold on the left/right hemisphere obtained by permutation based cluster inference
- PeakLh.t/PeakRh.t: a vector of peaks (local maxima) above the threshold on the left/right hemisphere
- PeakLh.vertid/PeakRh.vertid: vertex IDs (1-based) for the peaks above the threshold on the left/right hemisphere
- PeakLh.clusid/PeakRh.clusid: cluster IDs that contain the peak on

the left/right hemisphere

- ClusLh.clusid/ClusRh.clusid: cluster IDs on the left/right hemisphere
- ClusLh.nverts/ClusRh.nverts: number of vertices in each cluster on the left/right hemisphere
- ClusLh.resels/ClusRh.resels: resels in each cluster on the left/right hemisphere
- ClusidLh/ClusidRh: cluster IDs for each vertex on the left/right hemisphere
- Nsubj: total number of subjects in common
- NvetLh/NvetRh: number of in-mask vertices on the left/right hemisphere
- Ncov: number of covariates (including intercept)

Output files:

- LogPvalLh.mgh/LogPvalRh.mgh: vertex-wise surface maps of heritability significance ($-\log_{10}$(P-val))
- h2Lh.mgh/h2Rh.mgh: vertex-wise surface maps of heritability estimates

MEGHAsurfmat.m

Input arguments:

- SurfLh/SurfRh: an Nsubj x NvetLh/NvetRh data matrix for the left/right hemisphere
- Cov: an Nsubj x Ncov covariates matrix (intercept included)
- K: an Nsubj x Nsubj GRM

[Note: Subjects in SurfLh/SurfRh, Cov and K must be exactly the same and arranged in the same order.]

For other input and output arguments see MEGHAsurf.m.

Example Data

One could try the sample data posted on the GCTA website (http://www.complextraitgenomics.com/software/gcta/index.html). GRM can be obtained by running:

```
gcta64 --bfile test --make-grm-gz --out test
```

and then extracting test.grm from the zip file test.grm.gz

See Example.m for some simple examples.

R Implementation

We include an R implementation of MEGHA for reference. The R functions have not been wrapped up as an R package, and the computation is less efficient than the MATLAB implementation. But it might be helpful to researchers that are familiar with R programming. We will improve this implementation in the near future.