Documentation for Software Package MeSH

Xiaoquan Wen, Matthew Stephens

May 7, 2013

1 Introduction

Software package MeSH (**Me**ta analysis with **S**ubgroup **H**eterogeneity) implements the Bayesian statistical methods discussed in [1] using C++ programming language. It provides a flexible toolbox for Bayesian association analyses involving potentially heterogeneous subgroups. Subgroup structures are common in genetic association studies: in meta-analysis, different participating studies form subgroups; In detecting gene-environment interactions, (discretized) environmental conditions can cluster samples into different subgroups.

Briefly, we use a Bayesian hierarchical model approach to explicitly model potentially heterogeneous genetic effects presented in multiple subgroups of genetic association studies. The evidence from observed data for or against a hypothesized genetic association is summarized using Bayes factors (BF).

1.1 Method Overview

MeSH performs single SNP association test across a set of pre-defined subgroups. Within each subgroup, the association between the phenotype of a quantitative trait and the genotypes of a target SNP is modeled by a standard linear model,

$$\boldsymbol{y}_s = \mu_s \mathbf{1} + \beta_{\mathbf{s}} \boldsymbol{g}_{\mathbf{s}} + \boldsymbol{e}_{\mathbf{s}},\ \boldsymbol{e}_{\mathbf{s}} \sim \mathrm{N}(\mathbf{0}, \sigma_{\mathbf{s}}^{\mathbf{2}}\mathbf{I}), \tag{1}$$

where s indexes a particular subgroup, and $\boldsymbol{y}_s, \boldsymbol{g}_s, \boldsymbol{e}_s$ represent the vectors of sample phenotype, genotype and residual errors, respectively. Furthermore residual errors are assumed independent across subgroups.

The "global" null hypothesis of interest is that there is no genotype-phenotype association within any subgroup; that is, $\beta_s = 0$ for all s. Through prior specification, MeSH allows flexible specifications of

heterogeneous genetic effects in subgroups under the alternative models. More specifically, there four types of prior models are implemented in the current version of MeSH:

1. Exchangeable Standardize Effects (ES) Prior. A standardized effect size is defined by $b_s := \frac{\beta_s}{\sigma_s}$ and this prior model contains two parameters (ϕ, ω), and it assumes

$$b_s \sim \mathrm{N}(\bar{b}, \phi^2) \text{ and } \bar{b} \sim \mathrm{N}(0, \omega^2),$$

   where *phi* describes the degree of heterogeneity of standardized effects and ω describes the magnitude of the mean effect $(\bar{b})$.

2. Exchangeable Effects (EE) Prior. This prior model directly models regression coefficient β_s for subgroup s, it contains two parameters (ψ, w) and assumes that

$$\beta_s \sim \mathrm{N}(\bar{\beta}, \psi^2) \text{ and } \bar{\beta} \sim \mathrm{N}(0, w^2),$$

   where *psi* describes the degree of heterogeneity of genetic effects across subgroups and w describes the magnitude of the mean effect $(\bar{\beta})$.

3. Curved Exponential Family Normal Prior with Exchangeable Standardized Effects (CEF-ES). This prior model assumes the *a priori* standardized effect b_s for subgroup s is distributed as

$$b_s \sim \mathrm{N}(\bar{b}, k\bar{b}^2) \text{ and } \bar{b} \sim (0, \omega^2).$$

   Comparing with the ES model, the heterogeneity parameter k replaces ψ and it is interpreted by the following probability statement:

$$\Pr(b_s \text{ has a different sign from } \bar{b}) = \Phi(-\frac{1}{|k|}),$$

   where $\Phi(\cdot)$ is the cumulative probability function of the standard normal distribution.

4. Curved Exponential Family Normal Prior with Exchangeable Effects (CEF-EE). This prior model assumes

$$\beta_s \sim \mathrm{N}(\bar{\beta}, k\bar{\beta}^2) \text{ and } \bar{\beta} \sim (0, w^2).$$

   Similarly, the heterogeneity parameter k is interpreted as

$$\Pr(\beta_s \text{ has a different sign from } \bar{\beta}) = \Phi(-\frac{1}{|k|})$$

With a specified prior distribution, MeSH reports a Bayes factor as the support from the data for the alternative model relative to the global null model. Taking advantages of Bayesian model averaging,

MeSH allows a set of alternative models characterized by different levels of effect size heterogeneity to be considered and computes an overall Bayes factor by averaging over this set of alternative models.

1.2 Citations

Xiaoquan Wen and Matthew Stephens. Bayesian Methods for Genetic Association Analysis with Heterogeneous Subgroups: from Meta-Analyses to Gene-Environment Interactions. *arXiv pre-print: 1111.1210*.

2 Installation

We provide pre-compiled binary executables for some UNIX based operating systems. Alternatively, one can directly download the source code and compile the desired binary executable.

To do so, a GNU C++ compiler,e.g. GNU g++, and GNU Scientific Library (GSL) are required. To download and install GSL on your system, please refer to http://www.gnu.org/gsl/ for details.

3 Input Files

MeSH requires two types of input files: an input data file containing genotype, phenotype information for association testing and a grid file used for specifying heterogeneity and effect size priors.

3.1 Data File Format

Three types of format are allowed for input data files. The first two convey full information that allow all model options to be applied in the analysis. The third format is designed for "minimum" information of genetic associations, namely it only requires $\hat{\beta}, \text{se}(\hat{\beta})$ for each SNP of interest in each subgroup. It should be noted that use of minimum information format would limit model options in the analysis.

3.1.1 Complete Information Format I

The first type of format for input data file is regarded as the default by the program. This format requires a six-column data matrix with each row consisting of

`SNP_ID` `subgroup_label` `sample_size` $\hat{\beta}$ $\text{se}(\hat{\beta})$ $\text{var}(g)$

The entries are separated by spaces or tabs. Both "SNP_ID" and "subgroup_label" are strings and the rest of the row entries take numerical values. The following line is an example row entry taken from a simulated data set:

```
rs1800978  study1     100  0.308  0.111  1.01
```

In case that the sample variance, $\text{var}(g)$ is not directly available, assuming Hardy-Weinberg equilibrium, it is possible to *approximate* the quantity by

$$\text{var}(g) \approx 2f(1-f),$$

where f is the (sample) allele frequency of the target SNP.

3.1.2 Complete Information Format II

The other complete information format using a eight-column data matrix with the following row entries

`SNP_ID` `subgroup_label` `sample_size` $\bar{y}$ $\sum_i y_i^2$ $\bar{g}$ $\sum_i g_i^2$ $\sum_i g_i y_i$

In particular, $\bar{y}$ and $\bar{g}$ are the sample means of phenotype and genotypes, respectively; and $\sum_i y_i^2$ and $\sum_i g_i^2$ are the sums of squared sample phenotype and genotypes, respectively. Finally, $\sum_i g_i y_i$ is the sum of the individual product of the genotype and phenotype from all the samples in the subgroup. As an example, the SNP input shown in format I now is represented by

```
rs1800978 study1  100  -0.837  201.158  0.067  100.76  25.31
```

This data format is particularly convenient if the individual-level genotype and phenotype data are made available.

3.1.3 Minimum Information Format

Comparing with default complete information input format, the minimum information format input only requires information of $\hat{\beta}$ and se($\hat{\beta}$) from each subgroup. More specifically, the data matrix only contains 4 columns, i.e.,

SNP_ID	subgroup_label	$\hat{\beta}$	se($\hat{\beta}$)

With this input format, the prior choices are limited to EE and CEF-EE models and only approximate Bayes factors are computable. On the other hand, this format allows MeSH to work with non-quantitative phenotypes. For example, in a case-control study, $\hat{\beta}$ can be a log-odds ratio estimated by fitting a logistic regression model at subgroup level.

3.2 Grid File

The grid file contains prior specifications for various prior models. In all cases, the grid file always contains a two-column data matrix: the first column always represents the heterogeneity parameter (ϕ in ES, ψ in EE and k in CEF-ES and CEF-EE models) and the second column is used to specify the average effect size parameter (ω in ES and CEF-ES, w in EE and CEF-EE models). Each row of the grid data matrix provides a unique prior model and different rows can be used to describe different prior heterogeneity levels and/or prior average effect sizes. The program produces a final Bayes factor by averaging over all the prior models provided.

The following sample of the grid file is used by [1] to perform multiple-population eQTL analysis. The grid assumes an ES model and uses five levels of overall prior effects ($\sqrt{\omega^2+\phi^2} = 0.1, 0.2, 0.4, 0.8, 1.6$ values and seven degrees of heterogeneity ($\phi^2/\omega^2 = 0, 1/4, 1/2, 1, 2, 4, \infty$), which provides a comprehensive coverage of many possible scenarios. The order of the grid can be arbitrary.

```
0.0000  0.1000
0.1000  0.0000
0.0707  0.0707
0.0577  0.0816
0.0816  0.0577
0.0447  0.0894
0.0894  0.0447
0.0000  0.2000
0.2000  0.0000
0.1414  0.1414
0.1155  0.1633
0.1633  0.1155
0.0894  0.1789
0.1789  0.0894
0.0000  0.4000
0.4000  0.0000
0.2828  0.2828
0.2309  0.3266
0.3266  0.2309
0.1789  0.3578
0.3578  0.1789
0.0000  0.8000
0.8000  0.0000
0.5657  0.5657
0.4619  0.6532
0.6532  0.4619
0.3578  0.7155
0.7155  0.3578
0.0000  1.6000
1.6000  0.0000
1.1314  1.1314
0.9238  1.3064
1.3064  0.9238
0.7155  1.4311
1.4311  0.7155
```

4 Running Program

Once the binary executable is in place and required data and grid files are prepared, the program can be initiated by the following command from a terminal:

```
./mesh  -d data_file -g grid_file [options] [-o output_file]
```

The data and grid files, described in the previous section, are required command line arguments. Options and output file are optional, if not specified, the program assumes default model and parameter settings and output to the screen (stdout).

By default, the program assumes the data file is of complete information I format. It runs with ES model, and compute Bayes factors using numerical optimization based Laplace method.

When the input data format and/or model settings are desired to be different from the default setting, a new set of program parameters should be specified. The details are given in the following sections.

4.1 Program Parameters

The program parameters listed in this section should be provided in the command line.

4.1.1 Input/Output Options

- *-d data_file*: required, specification of the data file.
- *-g grid_file*: required, specification of the grid file.
- *-o output_file*: optional, specification of the output file. By default, the program output goes to screen.
- *-format* 1 | 2: optional, specification of the data file format, numerical values 1 and 2 refer to complete information format I and II, respectively. The default value is 1.
- *-min_info*: optional, indicating the data file is in minimum information format. With this option specified, the default prior model also changes to EE and the Bayes factors are computed based on analytic approximation.

4.1.2 Model Options

- *-es* | *-ee* | *-cef_es* | *-cef_ee* : optional, specification of the prior model. By default, ES model is assumed. Note, when *-min_info* option is set, i.e., the data file is in minimum information format,

only the EE (*-ee*) and the CEF_EE (*-cef_ee*) models are eligible.

- *-abf*: optional, specification of the algorithm to compute Bayes factors. By default, this flag is NOT set, and the numerical optimization based Laplace method is used. If the flag is set, the Bayes factors are computed using analytic approximations (also by Laplace method). When *-min_info* option is set, this option is automatically set.

- *-use_config*: optional. If this flag is set, the genetic association in each subgroup is modeled as either "active" ($\beta_s \neq 0$) or "silent" ($\beta_s = 0$), an scenario quite useful in some cases of gene-environment interactions. For s given subgroups, the program computes Bayes factors for all $2^s - 1$ possible configurations. Some useful applications of this model can be found in [1, 2].

4.1.3 Miscellaneous Options

- *-no_adjust*: optional. This option is used with *-abf* option, by default, the analytic approximation of Bayes factors adjusts for small sample sizes (recommended) and makes the results more accurate. Setting this flag avoids the adjustment, this flag is automatically set when "*min_info*" flag is automatically set. (in such case, there is no enough information to perform the adjustment).

- it -print_subgrp: optional. When this flag is set, the program outputs summary statistics for each individual subgroups, namely $\hat{\beta}$ and se($\hat{\beta}$).

- *-prep_hm*: optional. When this option is specified, the program prepares input Bayes factors suitable for the Hierarchical models implemented in software packages eqtlbma and bridge.

4.2 Output from Program

In the simplest case, the program outputs the Bayes factor for each SNP included in the data file in each line. If "*-print_subgrp*" option is specified, following the Bayes factor result, $\hat{\beta}$ and se($\hat{\beta}$) for each subgroup are also displayed.

By default, the results are displayed on screen (stdout). If "-output_file" is specified, all the results are recorded in the specified output file.

References

[1] Wen, X. and Stephens, M. "Bayesian Methods for Genetic Association Analysis with Heterogeneous Subgroups: from Meta-Analyses to Gene-Environment Interactions" *arXiv pre-print: 1111.1210*

[2] Flutre, T., Wen, X., Pritchard, JK and Stephens, M. "A Statistical Framework for Joint eQTL Analysis in Multiple Tissues" *PLoS Genetics (in press)*