BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 18 2005, pages 3672–3673
doi:10.1093/bioinformatics/bti536

Genome analysis

HEGESMA: genome search meta-analysis and heterogeneity testing

Elias Zintzaras[1,*] and John P. A. Ioannidis[2,3]

[1]Department of Biomathematics, University of Thessaly School of Medicine, Papakyriazi 22 street, Larissa 41222, Greece, [2]Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece and [3]Department of Medicine, Tufts University School of Medicine, Boston, MA, USA

Received on May 26, 2005; accepted on June 8, 2005
Advance Access publication June 14, 2005

ABSTRACT

Summary: Heterogeneity and genome search meta-analysis (HEGESMA) is a comprehensive software for performing genome scan meta-analysis, a quantitative method to identify genetic regions (bins) with consistently increased linkage score across multiple genome scans, and for testing the heterogeneity of the results of each bin across scans. The program provides as an output the average of ranks and three heterogeneity statistics, as well as corresponding significance levels. Statistical inferences are based on Monte Carlo permutation tests. The program allows both unweighted and weighted analysis, with the weights for each study as specified by the user. Furthermore, the program performs heterogeneity analyses restricted to the bins with similar average ranks.

Availability: http://biomath.med.uth.gr

Contact: zintza@med.uth.gr

Genome scans (genome searches) provide useful information on chromosomal loci that are linked to specific complex diseases. It is increasingly common for many independent teams to perform genome scans for the same disease with the important loci often becoming difficult to discern owing to low linkage signals (Altmuller *et al*., 2001) and/or apparent diversity in the results of different searches. It is however, important to compile the results of different genome scans on the same disease. Genome Scan Meta-analysis (GSMA) is an established method for summation of the data from diverse genome scans through meta-analysis (Wise *et al*., 1999) and for testing formally whether there is small or large heterogeneity for specific chromosomal loci across genome scans (Zintzaras and Ioannidis, 2005).

In this paper, we present HEGESMA (heterogeneity and genome search meta-analysis), a comprehensive software program for GSMA that includes heterogeneity testing. The program produces summary estimates for the ranks for each bin, and three different heterogeneity statistics, with their corresponding statistical significance levels.

A typical GSMA starts by splitting the chromosomes into 'bins' (loci) of similar length. Usually each bin has a width of ∼30cM giving 120 bins in total for the whole genome). For each genome scan, the most significant result of the test statistic (LOD, MLS, NLP, z-statistics, P-values) obtained within the bin is recorded (Terwilliger and Ott, 1994). Then, for each scan the bins are ranked according to their significance of results; and, the ranks for each bin are summed and averaged across scans. The extent of heterogeneity between studies for each bin can be estimated by the statistics Q, Ha and B which have been developed by the authors (Zintzaras and Ioannidis, 2005).

The Q-statistic is defined as the sum of the squared deviations of bin rank from the mean of the ranks of each study. The Ha-statistic is defined as the sum of absolute differences of each rank from the mean of the ranks. The B-statistic is defined as the sum of the distinct absolute differences in ranks between studies.

Weighted analyses (ranks of the bins in each study can be weighted by factors reflecting study size, such as the number of pedigrees) provide more weight to larger studies. Weighted analyses may be more appropriate in the evaluation of the overall significance of a specific bin, but probably not for heterogeneity testing (Zintzaras and Ioannidis, 2005).

The statistical significance of the average rank of each bin and the heterogeneity metrics are assessed by using a Monte Carlo method (Zintzaras and Ioannidis, 2005). In a run using this method, we randomly permute the ranks of each study and the simulated average rank and heterogeneity metrics are calculated; we then repeat the procedure for the number of runs specified by the user, and a null distribution of the metric is constructed. The significance levels are determined based on this distribution. The method tests for both high and low heterogeneity between studies. In addition, the probability of observing a given average rank for a bin by chance in bins with the same 'place' in the ascending order of average ranks in the runs (ordered ranks) is calculated (Levinson *et al*., 2003).

The program is written in Compaq Visual Fortran Professional Edition 6.6.0 and may be used in a PC with DOS. In its development the IMSL library was used. An executable file can be downloaded from http://biomath.med.uth.gr, and the Fortran code is available upon request (zintza@med.uth.gr). The code is suitable for other operating systems (e.g. Unix), but the IMSL library must be accessible in order to compile it.

The program is designed for biostatisticians and other researchers who need to perform a GSMA, and for testing heterogeneity across scans. The data entry is from an ASCII file, previously created by the user, containing the ranks of the bins of each study. The study weights are entered from a separate ASCII file. The data and weights files are stored in the same directory as the HEGESMA executable file, and they are identified as the xxx.dat and weights.dat, respectively.

In executing the program there is a series of questions for entering the parameters required: (1) the number of studies included in the

*To whom correspondence should be addressed.

Table 1. Output with the results of the unweighted analysis

observed mean	observed Q	observed Ha	observed B	right-sided P-value for mean	right-sided P order.-value for mean	left-sided P-value for Q	left-sided P-value for Ha	left-sided P-value for B
88.88	3538.19	97.75	248.50	0.0514	0.0022	0.5268	0.4724	0.5193
97.88	2207.19	79.75	185.50	0.0133	0.0701	0.3139	0.3025	0.2702
94.00	1554.50	77.00	171.00	0.0253	0.0237	0.2023	0.2759	0.2179
...								

The first three lines correspond to the first three bins: 1.1 (1st bin), 1.2 (2nd bin), 1.3 (3rd bin), etc. The "monte_weight" file contains the results of the weighted analysis in the same format as in the unweighted analysis.

GSMA (in the available version, the maximum number is limited to 20 studies); (2) the number of bins used (the maximum number is limited to 130); (3) the number of Monte Carlo permutations to determine the significance levels of the average ranks and the heterogeneity statistics (the maximum number is limited to 50 000); and (4) the type of analysis to be performed (main analysis: weighted or unweighted). For a specific bin, the heterogeneity analysis can be restricted to bins with average ranks in the neighbourhood (± 2) of the average rank of the specific bin, since heterogeneity may depend on the average rank (Zintzaras and Ioannidis, 2005).

The output consists of the following material: (1) average rank, right-sided P-value for each bin and P ordered-value for each bin; (2) Q-statistic and left-sided P-value for each bin; (3) Ha-statistic and left-sided P-value for each bin; and (4) B-statistic and left-sided P-value for each bin. When the heterogeneity analysis is restricted to the neighbourhood of a specific bin, the output consists of the Q-, Ha- and B-statistics with their corresponding left-sided P-values for the specific bin.

The following output files are produced according to the type of analysis: the file 'monte_unweighted' when an unweighted analysis is performed; the file 'monte_weighted' for a weighted analysis; the file 'monte_link_unweighted' for an unweighted analysis restricted to a specific bin; and the 'monte_link_weighted' for a weighted analysis restricted to a specific bin. The output files are in the ASCII format, and they are located in the same directory.

Here, we use a published GSMA (Fisher *et al*., 2003) to illustrate the features of HEGESMA. The GSMA evaluated data from four independent genome searches on rheumatoid arthritis and each genome search was divided into 120 bins. Then, the ranks of the bins for each study were calculated. The weights were defined as the scaled numbers of affected sibling pairs.

The input consists of the ASCII file with the ranks of the bins for each study, and the ASCII file with the four weights. The input file ("xxx.dat") with the ranks of the bins from the four studies in rheumatoid arthritis GSMA is:

```
104   40   119     92.5
104   58   112.5   117
...   ...  ...     ...
 19   49    90.5    65
```

Each column corresponds to an individual study. 1st line corresponds to bin 1.1, 2nd line corresponds to bin 1.2, 120th (last) line to bin 22.2 bin (in total 120 lines corresponding to 120 bins).

The input file ("weights.dat") with the weights given to the four studies is:

```
0.33  0.45  0.07  0.15
```

The output with the average ranks for each bin, the values of the heterogeneity statistics, the right-sided P-values for summed ranks and the left-sided P-values for the heterogeneity statistics are as shown in Table 1 (named: "monte_unweight") (metrics are shown in adjacent columns for illustration).

The output, when the analysis (unweighted and weighted) was restricted to a specific bin (bin 16.3), is as follows (named: "monte_link_unweighted"):

```
Bin of interest
  99
observed mean value for bin of interest
  93.625000
left-sided P-value for Q
  0.033708
left-sided P-value for Ha
  0.030163
left-sided P-value for B
  0.032040
```

Bin 99 is the 99th bin, corresponding to the third bin on chromosome 16 (16.3). The "monte_link_weighted" output file contains the results of the corresponding weighted analysis, and has the same format as above.

Currently, another software for available GSMA is provided with free access at http://mmg.umds.ac.uk. However, HEGESMA is the only software available for performing both GSMA and heterogeneity testing across genome searches. Besides its published applications in schizophrenia and rheumatoid arthritis, HEGESMA is currently being used in a number of ongoing meta-analyses of genome scans, including osteoporosis (bone mineral density and fracture risk), Parkinson's disease and autism. The availability of software to test for heterogeneity should enable this to be more widely implemented.

REFERENCES

Altmuller,J. *et al*. (2001) Genomewide scans of complex human diseases: true linkage is hard to find. *Am. J. Hum. Genet.*, **69**, 936–950.

Fisher,S.A. *et al*. (2003) Meta-analysis of four rheumatoid arthritis genome-wide linkage studies: confirmation of a susceptibility locus on chromosome 16. *Arthritis Rheum.*, **48**, 1200–1206.

Levinson,D.F. *et al*. (2003) Genome scan meta-analysis of schizophrenia and bipolar disorder, part I: methods and power analysis. *Am. J. Hum. Genet.*, **73**, 17–33.

Terwilliger,J.D. and Ott,J. (1994) *Handbook of Human Genetic Linkage.* Johns Hopkins University Press, Baltimore, MD.

Wise,L.H. *et al*. (1999) Meta-analysis of genome searches. *Ann. Hum. Genet.*, **63**, 263–272.

Zintzaras,E. and Ioannidis,J.P. (2005) Heterogeneity testing in Meta-Analysis of Genome Searches. *Genet. Epidemiol.*, **28**, 123–137.