Application Notes Genomics & Informatics Vol. 2(2) 107-109, June 2004

HapAnalyzer: Minimum Haplotype Analysis System for Association Studies

Ho-Youl Jung†, Jung-Sun Park†, Yun-Ju Park, Young-Jin Kim, Kuchan Kimm and InSong Koh*

Division of Epidemiology and Bioinformatics, National Genome Research Institute, National Institute of Health, 5 Nokbun- dong, Eunpyeng-gu, Seoul, 122-701, Korea

Abstract

Summary: HapAnalyzer is an analysis system that provides minimum analysis methods for the SNP-based association studies. It consists of Hardy-Weinberg equilibrium (HWE) test, linkage disequilibrium (LD) computation, haplotype reconstruction, and SNP (or haplotype)-phenotype association assessment. It is well suited to a case-control association study for the unrelated population.
Availability: The *HapAnalyzer* is freely available from *http://www.ngri.re.kr/HapAnalyzer/*
Contact: insong@nih.go.kr

Keywords: SNP, haplotype, haplotype block, disease association

In recent years, the SNP-based association studies give the clues which may allow us to unravel complex genetic traits. Because of the high-throughput genotyping technologies, it is strongly required to devise a software system to effectively analyze the association between genotypes and phenotypes. The desired system should include the following minimum analysis components- the Hardy-Weinberg equilibrium (HWE) test, linkage disequilibrium (LD) test, *in-silico* haplotype reconstruction module, and SNP (or haplotype)-phenotype association module. At present, however, there is no such integrated system which provides all these minimum analysis methods. There is only a system which provides partial analysis methods HWE test, LD computation, and SNP-phenotype association (Barrett, 2003).

*Corresponding author: E-mail insong@nih.go.kr,
Tel +82-2-380-1416, Fax +82-2-354-1068
Accepted 10 May 2004
†These authors contributed equally to this work.

When we do the analysis of the SNP-based association studies, we generally use the several software systems separately for each specific purpose, e.g. the SAS or other statistical packages for HWE test and analysis of the phenotype association, the GOLD (Abecasis and Cookson, 2000) for LD visualization, and the PL-EM (Qin *et al.*, 2002) or other software systems for haplotype reconstruction.

In this manner, it is very tedious to use various software packages for each step of analysis. Additionally, we have to convert the input file format in order to use any specific software or analysis tools, and/or have to run the whole software packages although we only need to use the small fraction of their functions. *HapAnalyzer* is an integrated system for the SNP-based association studies, and it effectively includes all the above mentioned components.

Fig. 1 shows the system flow of *HapAnalyzer*. First, a genotype file with the tab-delimitated format (you can easily find a sample file in our system) is loaded. If the analysis of phenotype-association is needed, a file which defines the sample's phenotype (case or control) must be loaded with the appropriate genotype file. We support importing the linkage format file. If the genotypes contain any missing data, the SNP loci or sample's genotype must be filtered out or predicted by the missing imputation module for further analysis. Second, users can take the test of the Hardy-Weinberg equilibrium. Third, in order to identify LD blocks *HapAnalyzer* provides the algorithm which was proposed by Gabriel *et al* (Gabriel *et al.*, 2002). The confidential interval and block threshold are easily adjusted by the user interface. Fourth, users can make the individual's haplotype from their genotypes using the PL-EM (it must be installed in the *HapAnalyzer* working directory). Furthermore we provide the tagging SNP information after the haplotype reconstruction. Finally, users can make the assessment of the SNP (or haplotype)-phenotype association by logistic regression.

HapAnalyzer consists of five subparts - HWE test, LD block computation, haplotype reconstruction, SNP-phenotype association, and haplotype-phenotype association.

Hardy-Weinberg Equilibrium test

HWE test is to tell whether the population is in equilibrium according to the Hardy-Weinberg law. If the population is in equilibrium, the genotype frequency is the same in parents and progeny. Our system provides the observed

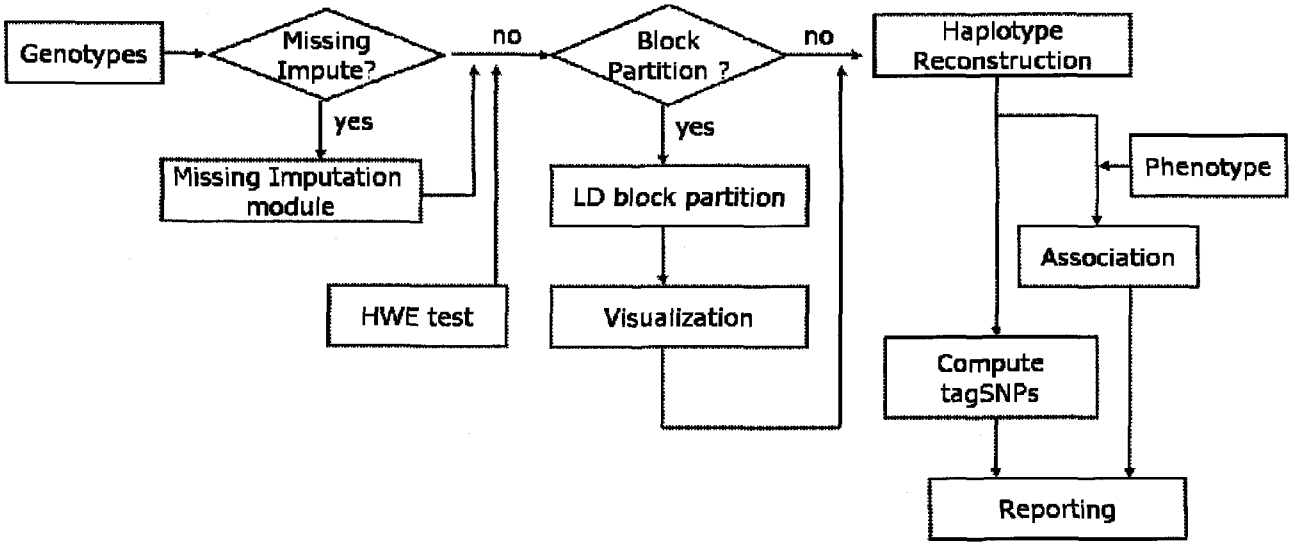

Fig. 1. The system flow of *HapAnalyzer.*

genotypic frequencies, the allelic frequencies, the expected genotypic frequencies, the Chi-square test value, and its *p*-value for each SNP locus. Users can define the cut-threshold for HWE, which is generally set to 0.01 or 0.05.

Linkage Disequilibrium & block computation

For two loci A and B, LD is said to exist when alleles at A and B tend to co-occur on haplotypes in proportions different than would be expected under statistical independence. There are many LD measures, e.g. *D*, *D'*, $|D'|$, r^2, and so on (Devlin and Risch, 1995). Our system provides two measures, $|D'|$ and r^2 with confidential interval and *p*-value, and shows the LD blocks by computing the confidential interval between every set of pairwise SNP loci (Gabriel *et al.,* 2002). We applied a dynamic programming algorithm to computing candidate LD blocks as follows:

$$C_i = \min\{C_{i-1} + \text{block}(r_i, r_j), \text{ if } 1 \le i \le j\},$$

where block(r_i, r_j) is 1 if the region between SNP site r_i and r_j over which more than 95% of SNP pairs show high levels of LD else it is a positive infinite value. High level of LD is determined by user defined values, however conventionally LD of which lower bound of confidential interval is over 0.7 and upper bound is over 0.98 is regarded as a high level of LD (Gabriel *et al.,* 2002).

In our system, we additionally provide another dynamic programming algorithm of computing haplotype blocks based-on haplotype diversity (Zhang *et al.,* 2002). Linkage disequilibrium values between every set of pairwise SNP loci are also visualized by its equivalent image format as GOLD, and provided by a Microsoft Excel spread sheet and a text file format.

Haplotype reconstruction

Haplotype reconstruction module determines the individual's haplotypes given the genotypes by using an *in-silico* haplotyping method. There are so many publicly available methods. Our system includes the PL-EM and also provides our novel method, *iHaplor* (Jung *et al.,* 2003), modified Clark's method (Clark, 1990). Users can alternatively select the reconstruction method. The system also gives the information of the tagging SNP after the haplotype reconstruction using the *BEST* algorithm (Sebastiani *et al.,* 2003).

SNP-phenotype association

By assessing the SNP-phenotype association, we can identify the candidate disease-causing SNP locus. We provide the major (A) vs. minor (a) allelic difference between case and control population. We also give the genotypic difference between these populations according to three genetic models - dominant (AA+Aa vs. aa), recessive (AA vs. Aa+aa), and co-dominant model (AA vs. Aa vs. aa). In each analysis, we compute the odds ratio, Chi-square test value, and its p-value by logistic regression.

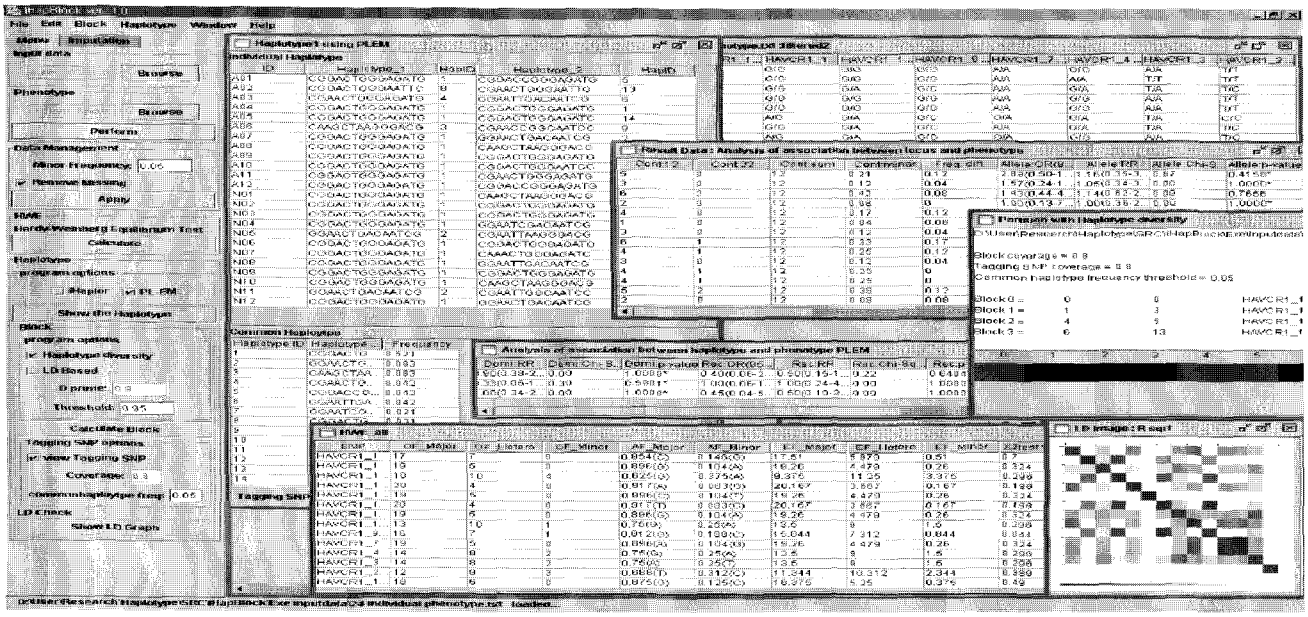

Fig. 2. A snapshot of *HapAnalyzer*: it is run on all platforms on which the Java Running Environment (JRE 1.4) is installed.

Haplotype-phenotype association

After the haplotype reconstruction, we can assess the association of the haplotypes with phenotype between case and control population. Our system only gives the result of assessment of the association of common haplotypes whose frequencies are larger than 0.05. We consider a haplotype (H) as if it is an allele. Our system provides the allelic difference between case and control (H vs. -, '-' means that the sample who does not contain the haplotype H). Our system also gives the genotypic difference between these population according to three genetic models - dominant (HH+H- vs. --), recessive (HH vs. H-+--), and co-dominant model (HH vs. H- vs. --).

Fig. 2 shows a snapshot of our system. *HapAnalyzer* is implemented using the Java 1.4, and it is run on all platforms on which the Java Running Environment (JRE 1.4) is installed. We have tested our system on both WindowsXP and Linux machines with the 512MB main memory and Pentium-4 2.0 GHz CPU.

Acknowledgements

This study was supported by the intramural fund of the National Institute of Health, Korea.

References

Abecasis, G. R. and Cookson, W. O. C. (2000). GOLD-Graphical Overview of Linkage Disequilibrium. *Bioinformatics* 16, 182-183.

Clark, A. G. (1990). Inference of haplotypes from PCR-amplified samples of diploid populations. *Molecular Biology* and *Evolution* 7, 111-122.

Devlin, B. and Risch, N. (1995). A Comparison of Linkage Disequilibrium Measures for Fine-Scale Mapping. *Genomics* 29, 311-322.

Gabriel, S. B. *et al.* (2002). The Structure of Haplotype Blocks in the Human Genome. *Science* 296, 2225-2229.

Jung, H-.Y. *et al.* (2003). iHaplor: A Hybrid Method for Haplotype Reconstruction, In *Proceedings of the Second Annual Conference of the Korean Society for Bioinformatics (KSBI2003)*, 2, 221-228.

Barrett, J. (2003). HaploView. *http://www.broad.mit.edu/personal/jcbarret/haplo/ index.php.*

Qin, Z. S., Niu, T., and Liu, J. S. (2002). Partition-Ligation-Expectation-Maximization Algorithm for Haplotype Inference with Single- Nucleotide Polymorphisms. *American Journal of Human Genetics* 71, 1242-1247.

Sebastiani, P., Lazarus, R., Weiss, S. T., Kunkel, L. M., Kohane, I. S., and Ramoni, M. F.. (2003). Minimal haplotype tagging. *Proceedings of the National Academy of Science*, 100, 9900-9905.

Zhang, K., Deng, M., Chen, T., Waterman, M. S., and Sun, F. (2002). A dynamic programming algorithm for haplotype block partitioning. *Proc. Natl. Acad. Sci. USA* 99, 7335-7339.