

Association tests and estimation of haplotype frequencies and of penetrance-related parameters in a case-control study

Shiori Furihata

Japan Biological Information Research Center, Japan Biological Informatics Consortium
Time24 Bldg. 10F, 2-45 Aomi, Koto-ku, Tokyo 135-0064, Japan
sfurihat@jbirc.aist.go.jp

Toshikazu Ito

Mitsubishi Research Institute, INC.
Otemachi 2-3-6, Chiyoda-ku, Tokyo 100-8141, JAPAN

Naoyuki Kamatani

Division of Genomic Medicine, Department of Advanced Biomedical Engineering and Science,
and Institute of Rheumatology, Tokyo Women's Medical University,
10-22 Kawada-cho, Shinjuku-ku, Shinjuku, Tokyo 162-0054, Japan

Abstract

This study shows the applicability of our algorithm which tests haplotype and qualitative phenotype association, in a case-control study. Although our algorithm named PENHAPLO is originally developed for analysis of a cohort study, it is shown that PENHAPLO is a useful tool for analysis of case-control studies as well. Simulations confirm that type I errors under null hypothesis are accurately estimated. The statistical power for alternative hypothesis calculated using the algorithm is comparable to those analyzed using known-phase data.

1. Introduction

With the increasing number of single-nucleotide polymorphism (SNP) markers, analysis of polymorphism data based on linkage disequilibrium and haplotypes structure becomes important. Some studies suggest that phenotypes of individuals are associated with a diplotype configuration (a combination of haplotypes) rather than a haplotype. Since neither haplotypes nor diplotype configurations of a subject are usually observed, the haplotype frequencies of cases and controls are inferred using an algorithm. For analysis of either cohort studies or clinical trials, we developed an algorithm for association tests between an individual qualitative phenotype and specific haplotypes[1]. The algorithm estimates diplotype-based penetrances as well as haplo-

type frequencies in a population. Based on an Expectation-maximization (EM) method, the algorithm is implemented in the computer program PENHAPLO. We investigate the applicability of PENHAPLO to a case-control study.

2. Methods

Suppose that a subject develops a qualitative phenotype ψ , such as a disease, with probability q_+ when possessing a haplotype H_+ . Those subjects who do not have haplotype H_+ develop the phenotype with probability q_- . If haplotypes are in Hardy-Weinberg equilibrium in the underlying population with frequencies $\Theta = (\theta_1, \dots, \theta_L)$, the likelihood function is expressed as:

$$L(\Theta, q_+, q_-) \propto \prod_{i=1}^N \sum_{a_k \in A_i} P(d_i = a_k | \Theta) P(\psi_i = w_i | d_i = a_k, q_+, q_-). \quad (1)$$

Here, A_i denotes the set of diplotype configuration a_k for i th subject that are consistent with the observed genotype g_i , w_i is the observed phenotype, and N is the number of subjects in the observed data. The maximum likelihood under the null hypothesis, i.e., $q_0 = q_+ = q_-$, and under the alternative hypothesis, i.e., $q_+ \neq q_-$, is calculated. The likelihood ratio is used to test the association between the phenotype and the haplotype. The parameters Θ , q_+ , q_- and q_0 are estimated using the EM method.

In this study, we apply the algorithm to a case-control study. Since the haplotype frequencies in a case-control

study are biased, we assume that the biases are expressed as r_+ and r_- . When PENHAPLO is used to analyze a case-control study, r_+ and r_- are estimated instead of q_+ and q_- . If prevalence of disease λ is known for a population, penetrances q_+ for specific haplotypes can be obtained using r_+ and r_- .

$$q_+ = \frac{1}{\frac{s}{(1-s)} \frac{(1-r_+)(1-\lambda)}{r_+} + 1}, \quad (2)$$

where s is a ratio of the number of cases to total number of subjects in a case-control study. Therefore, parameters r_+ and r_- are related to the penetrances. Odds ratio is calculated as $(\frac{r_+}{1-r_+})(\frac{1-r_-}{r_-})$.

To investigate type I errors under the null hypothesis, we perform simulations. The flow of the simulation is shown in Fig. 1. Based on the haplotype frequencies obtained for SAA genes[2], the test statistics are calculated using PENHAPLO for each 20,000 simulated data. At the same time, two by two contingency tables based on known-phase data are tested using the Pearson's χ^2 statistic as well.

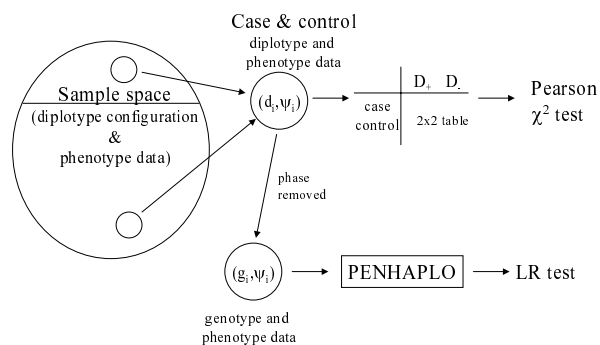


Figure 1. Schematic representation of the simulation (where $q_+ = q_- = 0.2$ is assumed for the null hypothesis).

3. Results

Figure 2 shows the type I errors under the null hypothesis for the results analyzed using PENHAPLO and those using 2x2 tables based on known-phase data. The range of the statistical error bars for both PENHAPLO and χ^2 test cross the 0.05 level of significance.

We also perform simulations under the alternative hypothesis to estimate powers. The results are shown in Fig. 3. The powers estimated using PENHAPLO are comparable to those analyzed using the Pearson's χ^2 test based on known-phase data.

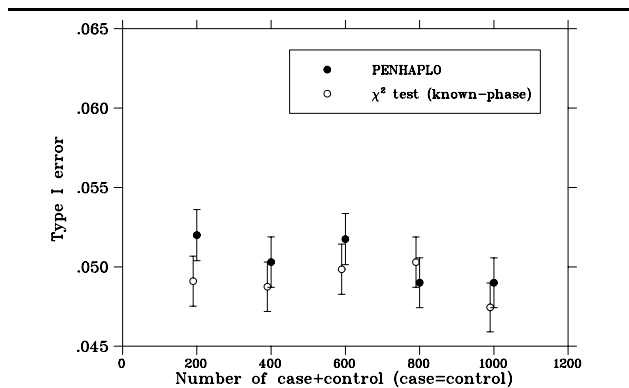


Figure 2. Type I errors

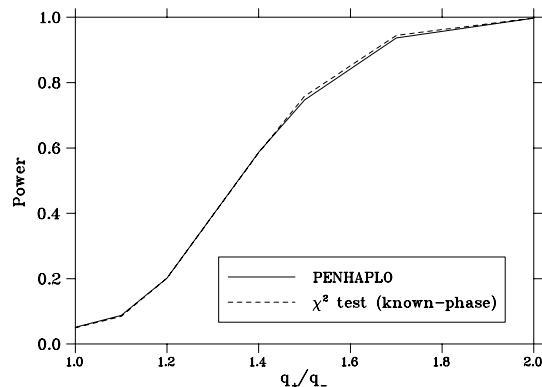


Figure 3. Powers

4. Discussion

Simulations confirm that type I errors under null hypothesis are accurately estimated. Under the alternative hypothesis, the statistical power calculated using PENHAPLO is comparable to that calculated using known-phase data. It is shown in this work that PENHAPLO is a useful tool for analysis of case-control studies concerning haplotype and qualitative phenotype association.

References

- [1] T. Ito, E. Inoue, and N. Kamatani. Association test algorithm between individual phenotype and a haplotype using simultaneous estimation of haplotype frequencies, diplotype configurations, and diplotype-based penetrances (submitted for publication).
- [2] M. Moriguchi, C. Terai, H. Kaneko, Y. Koseki, H. Kajiyama, M. Uesato, S. Inada, and N. Kamatani. A novel single-nucleotide polymorphism at the 5'-flanking region of *SAA1* associated with risk of type aa amyloidosis secondary to rheumatoid arthritis. *Arthritis Rheum*, 44:1273-1280, 2001.