Copyright © 2004 by the Genetics Society of America
DOI: 10.1534/genetics.103.024653

Association Test Algorithm Between a Qualitative Phenotype and a Haplotype or Haplotype Set Using Simultaneous Estimation of Haplotype Frequencies, Diplotype Configurations and Diplotype-Based Penetrances

Toshikazu Ito, Eisuke Inoue and Naoyuki Kamatani[1]

Division of Genomic Medicine, Department of Applied Biomedical Engineering and Science and Institute of Rheumatology, Tokyo Women's Medical University, Tokyo 162-0054, Japan and Algorithm Team, Japan Biological Information Research Center (JBIRC), Japan Biological Informatics Consortium (JBIC), Tokyo 135-0064, Japan

Manuscript received November 17, 2003
Accepted for publication August 30, 2004

ABSTRACT

Analysis of the association between haplotypes and phenotypes is becoming increasingly important. We have devised an expectation-maximization (EM)-based algorithm to test the association between a phenotype and a haplotype or a haplotype set and to estimate diplotype-based penetrance using individual genotype and phenotype data from cohort studies and clinical trials. The algorithm estimates, in addition to haplotype frequencies, penetrances for subjects with a given haplotype and those without it (dominant mode). Relative risk can thus also be estimated. In the dominant mode, the maximum likelihood under the assumption of no association between the phenotype and presence of the haplotype ($L_{0\max}$) and the maximum likelihood under the assumption of association ($L_{\max}$) were calculated. The statistic $-2\log(L_{0\max}/L_{\max})$ was used to test the association. The present algorithm along with the analyses in recessive and genotype modes was implemented in the computer program PENHAPLO. Results of analysis of simulated data indicated that the test had considerable power under certain conditions. Analyses of two real data sets from cohort studies, one concerning the *MTHFR* gene and the other the *NAT2* gene, revealed significant associations between the presence of haplotypes and occurrence of side effects. Our algorithm may be especially useful for analyzing data concerning the association between genetic information and individual responses to drugs.

ANALYSIS of polymorphism data based on linkage disequilibrium (LD) and haplotype structure is becoming increasingly important. Haplotype is also useful for associating the genetic information of subjects with various phenotypes. A phenotype is often associated not only with each SNP but also with haplotypes. If the disease association of a specific allele is dependent on *cis*-acting interactions with other loci, the disease association may not be detected unless the functional haplotypic unit itself is analyzed. Thus, many studies of various diseases and conditions have reported the importance of haplotype information in addition to single locus information (Puffenberger *et al.* 1994; Drysdale *et al.* 2000; Mummidi *et al.* 2000; Hugot *et al.* 2001; Joosten *et al.* 2001; Rioux *et al.* 2001).

We can interpret haplotypes as complete data and genotypes (for example, SNP genotypes at multiple linked loci) as incomplete data from the statistical viewpoint, since we can extract all genotype data from haplotype data, while the reverse is not the case. Therefore, it is generally more useful to consider the association between polymorphism and phenotype based on haplotypes or diplotype configurations (haplotype combinations) rather than on alleles or genotypes. Recent studies have suggested that, in some cases, phenotypes such as diabetes (Horikawa *et al.* 2000) and reaction to drugs are associated with haplotypes or diplotype configurations rather than (single-nucleotide polymorphism) genotypes (Judson *et al.* 2000; Bader 2001; Tanaka *et al.* 2002; Urano *et al.* 2002).

The phenotypes at the level of individuals are based on diplotype configurations rather than haplotypes. This is equivalent to the relationship between the genotypes and alleles at a locus. Note that, according to Mendel's law, phenotypes are associated with genotypes but not alleles. Therefore, the association between a phenotype and genetic information is in some cases detected more efficiently by comparing the proportions of affected individuals between subjects with different diplotype configurations than by comparing the haplotype frequencies between subjects with different phenotypes.

However, the haplotype or diplotype configuration of a subject cannot easily be observed, although molecular methods for the determination of haplotypes have been reported (Michalatos-Beloin *et al.* 1996). To compare the proportions of affected individuals between subjects

[1]*Corresponding author:* Division of Genomic Medicine, Department of Applied Biomedical Engineering and Science and Institute of Rheumatology, Tokyo Women's Medical University, 10-22 Kawada-cho, Shinjuku-ku, Shinjuku, Tokyo 162-0054, Japan.
E-mail: kamatani@ior.twmu.ac.jp

with different diplotype configurations, the diplotype configuration for each individual should be calculated as posterior distribution based on population haplotype frequencies, which are inferred using one of the haplotype inference algorithms, such as Clark's algorithm (Clark 1990), the expectation-maximization (EM) algorithm (Excoffier and Slatkin 1995; Hawley and Kidd 1995; Long *et al.* 1995; Schneider *et al.* 2000; Kitamura *et al.* 2002), the PHASE algorithm (Stephens *et al.* 2001), the partition-ligation (PL) algorithm (Niu *et al.* 2002), and the PL-EM algorithm (Qin *et al.* 2002). Once diplotype configurations are inferred, all the individuals are classified according to the presence or absence of a haplotype or a diplotype configuration. After classifying the affected and nonaffected individuals into categories, the test of independence is performed.

When one of these algorithms is used to calculate the posterior distribution of the diplotype configuration, at least in some individuals, diplotype configurations are not unequivocally determined. It is not clear how much type I error can occur by classifying subjects according to inferred haplotypes rather than according to the real haplotype information.

To overcome this problem, we developed an algorithm to infer diplotype-based penetrance in addition to haplotype frequencies in populations and diplotype configurations based on observed single-nucleotide polymorphism (SNP) genotypes at multiple linked loci and phenotype data. This algorithm does not require that the diplotype configuration of each individual be unequivocally determined. Rather, on the basis of the EM algorithm, it calculates the maximum-likelihood estimates of population haplotype frequencies, posterior distribution of the diplotype configuration for each individual, and the diplotype-based penetrances. Using the algorithm, the association between the presence of haplotypes and a phenotype (in the dominant mode) can be tested at the individual level using the genotype and phenotype data from cohort studies or clinical trials. We examined the usefulness of this algorithm using both simulated and real data sets and found that it was very useful for analyzing genotype and phenotype data from cohort studies and randomized clinical trials.

METHODS

Algorithm: *Sample space of the EM-based algorithm for haplotype inference:* In the EM-based algorithm for haplotype inference using genotype data, the sample space is defined as a set of outcomes from the following experiment. First, haplotype frequencies are provided for a collection of infinite haplotype copies. (Throughout this article, "haplotype" and "haplotype copy" have different concepts. If a subject has the two same haplotype copies, the number of haplotypes is still one.) According to the haplotype frequencies, each of the N subjects is given two ordered haplotype copies after randomly drawing them from the collection of haplotype copies. The observed data are the genotypes at multiple linked loci involved in the haplotypes of the group of all subjects. In this EM-based algorithm for both haplotype and penetrance inference, the experiment used to define the sample space is a little different. After two ordered haplotype copies are assigned to each subject, he or she develops or does not develop a phenotype as a stochastic process. Thus, the difference between the EM-based algorithm for haplotype inference and the new algorithm presented here is that, in the new algorithm, the process of development of phenotype is included in the experiment for construction of the sample space.

New sample space: Let us assume that there are l linked SNP loci. The number of all possible haplotypes will be $L = 2^l$. We set up a collection of an infinite number of haplotype copies. The haplotype frequencies in the collection are $\Theta = (\theta_1, \ldots, \theta_j, \ldots, \theta_L)$, where θ_j is the frequency of the jth haplotype, and $\theta_j \geq 0$, $\Sigma_{j=1}^{L}\theta_j = 1$. To each of N individuals, ordered two haplotype copies are assigned by randomly drawing them from the collection of haplotype copies. A diplotype configuration is defined as an ordered combination of two haplotype copies. (Throughout the algorithm section, "diplotype configuration" means an ordered set of two haplotype copies for a subject, while in the other parts of this article, it means an unordered two haplotype copies for a subject). Let $a_1, a_2, \ldots, a_{L^2}$ be possible diplotype configurations. The probability that the ith subject has the diplotype configuration a_k is $P(d_i = a_k \mid \Theta) = \theta_l\theta_m$, where d_i is a diplotype configuration for the ith subject, and l and m are the orders of the haplotypes that constitute a_k. This means that Hardy-Weinberg equilibrium is assumed at the haplotype level. The ith subject develops the phenotype ψ_+ at the probability determined as a function of d_i. Theoretically, the penetrances can be defined for all the diplotype configurations. However, it is not realistic to assign different penetrances to all the different diplotype configurations. We therefore defined only two or three penetrances depending on the mode of inheritance in this study.

Thus, in the dominant mode, the subjects with a haplotype and those without it were given two different penetrances, while in the recessive mode, the subjects who were homozygous for a haplotype and the others were given two different penetrances. In the genotype mode, the subjects with zero, one, and two copies of a haplotype were given three different penetrances.

Let H_{all} denote the set of all the haplotypes, and let H_+ denote the subset of H_{all} containing the haplotype or haplotypes the presence of which has a different effect than the others. H_+ typically contains only one haplotype, but may contain multiple haplotypes. If H_+ is set up so that it contains all the haplotypes with an allele at a locus, then it is equivalent to the situation of testing the association of an allele (rather than a haplotype) with the phenotype. Two penetrances were

set when the analysis was performed either in the dominant or in the recessive mode. Thus, in the dominant mode, let D_+ denote a set of diplotype configurations that contains a member or members of H_+. In the recessive mode, let D_+ denote a set of diplotype configurations whose two haplotypes are the members of H_+. Then, let q_+ denote the probability that the ith individual develops ψ_+ when $d_i \in D_+$, and let q_- denote the probability that the ith individual develops ψ_+ when $d_i \notin D_+$.

Thus, if ψ_i denotes the phenotype of ith subject,

$$P(\psi_i = \psi_+ \mid d_i \in D_+) = q_+$$

and

$$P(\psi_i = \psi_+ \mid d_i \notin D_+) = q_-.$$

Note that Θ and q_+, q_- are independent. Since dominant or recessive mode of inheritance is assumed in the above model, only two penetrances are defined.

If genotype-dependent mode is assumed, three penetrances should be defined depending on the genotype. Let D_0, D_1, and D_2 denote the sets of diplotype configurations that contain zero, one, and two copies of the members of H_+, respectively. Let q_0, q_1, and q_2 denote the probabilities that the ith individual develops ψ_+ when $d_i \in D_0$, $d_i \in D_1$, and $d_i \in D_3$, respectively. Thus,

$$P(\psi_i = \psi_+ \mid d_i \in D_0) = q_0,$$

$$P(\psi_i = \psi_+ \mid d_i \in D_1) = q_1,$$

and

$$P(\psi_i = \psi_+ \mid d_i \notin D_2) = q_2.$$

Note that Θ and q_0, q_1, and q_2 are independent.

The new experiment is different from the old experiment in that the process of developing a phenotype is included in the former. The parameters q_+ and q_- or q_0, q_1, and q_2 are, in addition to Θ, included in the new probability space. Note that ψ_i is independent of Θ conditional on d_i.

Likelihood function: The observed data are the genotypes and phenotypes of the subjects. Let $G_{\text{obs}} = (g_1, g_2, \ldots, g_N)$ and $\Psi_{\text{obs}} = (w_1, w_2, \ldots, w_N)$ denote the vectors of the observed genotypes and phenotypes, respectively, where g_i and w_i denote the observed genotype and phenotype of the ith subject.

The likelihood function under the alternative hypothesis is written as follows. Thus the likelihood function in the dominant or recessive mode is

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

where A_i denotes the set of a_k for ith subject that is consistent with g_i.

Since d_i is independent of q_+, q_- and ψ_i is independent of Θ conditional on d_i,

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

For any i and k,

$$P(\psi_i = w_i \mid d_i = a_k, q_+, q_-) = \begin{cases} q_+ & \text{if } w_i = \psi_+ \text{ and } a_k \in D_+ \\ 1 - q_+ & \text{if } w_i \neq \psi_+ \text{ and } a_k \in D_+ \\ q_- & \text{if } w_i = \psi_+ \text{ and } a_k \notin D_+ \\ 1 - q_- & \text{if } w_i \neq \psi_+ \text{ and } a_k \notin D_+. \end{cases}$$

In the genotype mode, the likelihood function is

$$L(\Theta, q_0, q_1, q_2) \propto \prod_{i=1}^{N} \sum_{a_k \in A_i} P(d_i = a_k \mid \Theta) P(\psi_i = w_i \mid d_i = a_k, q_0, q_1, q_2) \tag{2}$$

and

$$P(\psi_i = w_i \mid d_i = a_k, q_0, q_1, q_2) = \begin{cases} q_0 & \text{if } w_i = \psi_+ \text{ and } a_k \in D_0 \\ 1 - q_0 & \text{if } w_i \neq \psi_+ \text{ and } a_k \in D_0 \\ q_1 & \text{if } w_i = \psi_+ \text{ and } a_k \in D_1 \\ 1 - q_1 & \text{if } w_i \neq \psi_+ \text{ and } a_k \in D_1 \\ q_2 & \text{if } w_i = \psi_+ \text{ and } a_k \in D_2 \\ 1 - q_2 & \text{if } w_i \neq \psi_+ \text{ and } a_k \in D_2. \end{cases}$$

Under the null hypothesis that the phenotype is independent of the diplotype configuration concerning the loci examined, the likelihood function is

$$L(\Theta, q_c) \propto \prod_{i=1}^{N} \sum_{a_k \in A_i} P(\psi_i = w_i \mid d_i = a_k, q_c) P(d_i = a_k \mid \Theta), \tag{3}$$

where q_c denotes the penetrance for all the diplotype configurations.

For any i and k,

$$P(\psi_i = w_i \mid d_i = a_k, q_c) = \begin{cases} q_c & \text{if } w_i = \psi_+ \\ 1 - q_c & \text{if } w_i \neq \psi_+. \end{cases}$$

To obtain the maximum likelihood under the alternative hypothesis, Equation 1 or 2 is maximized over (Θ, q_+, and q_-) or (Θ, q_0, q_1, and q_2), and the maximum likelihood thus obtained is denoted as $L_{\max}$. Then Equation 3 is maximized over Θ and q_c, and the maximum likelihood under the null hypothesis thus obtained is denoted as $L_{0\max}$. The likelihood ratio $L_{0\max}/L_{\max}$ is used to test the association between the haplotype and the phenotype.

In the maximization of $L_{\max}$, the parameters to be estimated are $\Theta = (\theta_1, \theta_2, \ldots, \theta_L)$, q_+, and q_- in the dominant or recessive mode and Θ, q_0, q_1, and q_2 in the genotype mode, while, in maximization of $L_{0\max}$, the parameters to be estimated are $\Theta = (\theta_1, \theta_2, \ldots, \theta_L)$ and q_c. The space spanned by the latter maximization is a subspace of that spanned by the former. Under the null hypothesis, the statistic $-2 \log(L_{0\max}/L_{\max})$ is expected to follow the χ^2 distribution with 1 or 2 d.f. asymptotically depending on the number of penetrances.

The EM algorithm: In this section, the algorithm is described mainly for the dominant mode. If the com-

plete data of $d_1, d_2, \ldots, d_N$ and $\psi_1, \psi_2, \ldots, \psi_N$ were available, the maximum-likelihood estimates of $\theta_1, \theta_2, \ldots, \theta_L$ and q_+, q_- would be easily obtained as $\hat{\theta}_j = n_j/(2N)$ for $j = 1, 2, \ldots, L$ and $\hat{q}_+ = N_{+\psi+}/N_+$, $\hat{q}_- = N_{-\psi+}/N_-$, where n_j is the number of the copies of jth haplotype in the N subjects, $N_+ = \#\{i;\ d_i \in D_+\}$, $N_- = \#\{i;\ d_i \notin D_+\}$, $N_{+\psi+} = \#\{i;\ d_i \in D_+, \psi_i = \psi_+\}$, and $N_{-\psi+} = \#\{i;\ d_i \notin D_+, \psi_i = \psi_+\}$. Here, $\#\{i;\ ,\ \}$ denotes the number of subjects that meet the conditions after ;.

However, complete data are not available, and we observe only the genotypes and phenotypes of the subjects. Therefore, we substitute the expected values of $n_j/(2N)$, $N_{+\psi+}/N_+$, and $N_{-\psi+}/N_-$ for the real values in the EM algorithm to maximize the likelihood function defined by Equation 1. If we supply the condition $q_c = q_+ = q_-$, then we obtain the maximum likelihood $L_{0\max}$ for the null hypothesis defined by Equation 3. Note that q_+/q_- is usually called "relative risk," and $\hat{q}_+/\hat{q}_-$ is the maximum-likelihood estimate of relative risk. In the genotype mode, q_0, q_1, and q_1 are substituted by the expected values just as q_+ and q_- were substituted in case of the dominant mode as described above.

By use of the EM algorithm, missing data in the observed data of genotypes and phenotypes could be handled. If data were missing from the observed genotype for the ith subject, g_i was interpreted as the set of possible genotypes for the subject not inconsistent with g_i. For the missing data in phenotypes, the probability that a subject develops the unknown phenotype was interpreted to be 1.

We can test the association between a phenotype and a set of haplotypes rather than a single haplotype. Thus, a set of haplotypes H_+ typically has only a single haplotype as a member; however, multiple haplotypes can be members of H_+. We have recently defined a concept of "incomplete haplotype" (Kamatani *et al.* 2004).

Thus, let H_{all} denote the set of all the haplotypes concerning all linked loci within a region. A complete haplotype is defined as a list of alleles at all linked loci in the region. Here, incomplete haplotype H_I is defined as a subset of H_{all} whose members have certain alleles at some loci within the region (Kamatani *et al.* 2004). Therefore, an incomplete haplotype is defined by a list of alleles at a limited number of the loci. In other words, an incomplete haplotype is defined by a list of alleles at all the loci (one allele at a locus), some of which are masked. For example, AC* defines an incomplete haplotype whose members are the complete haplotypes ACT and ACC (when the alleles at the third locus are T and C). Note that the set of all incomplete haplotypes is not the same as the set of all the set of complete haplotypes. The former is included in the latter. Also note that an allele of a SNP locus can also be defined as an incomplete haplotype because the allele T at the third position in the above haplotype is defined as **T.

Any incomplete haplotype as described above can be used as a target haplotype whose association with a phenotype is tested. In PENHAPLO, we implemented the algorithm to test the association between all possible incomplete haplotypes in dominant, recessive, or genotype mode. However, when the number of loci in the region is high, the number of all the incomplete haplotypes should be very large. The problem of multiple comparison will indeed emerge and this problem is discussed in the DISCUSSION.

TABLE 1

Haplotype frequencies for *SAA* gene (Moriguchi *et al.* 2001)

Haplotype	Frequency
ACTGCC	0.394
ACCGTC[a]	0.214
AGCGCT	0.210
GCCGTC	0.036
GCTGCT	0.035
GGCACT	0.023
AGTGCT	0.023
AGCACT	0.018
GGCGCT	0.017
ACTGTC	0.013
ACCGCC	0.006
ACCATC	0.006
AGCGCC	0.003

[a] The haplotype that was considered the phenotype-associated haplotype when the simulation was performed under the alternative hypothesis.

Simulation: *Empirical distribution of the statistic* $-2\log(L_{0\max}/L_{\max})$ *under the null hypothesis:* We first examined the empirical distribution of the statistic $-2\log(L_{0\max}/L_{\max})$ under the null hypothesis by simulation. To do this, we used the frequency distribution of haplotypes Θ obtained from the real data rather than simulating it. Thus, we used Θ obtained from our previous study on *SAA* (serum amyloid A) genes (Moriguchi *et al.* 2001), which include six SNPs. Table 1 shows the haplotypes and their frequencies. For q_c, we tested various values between 0 and 1. Note that, under the null hypothesis, the penetrance is the same for all diplotype configurations.

We began the simulation by assigning ordered sets of two haplotype copies to each of the N subjects by drawing the haplotype copies using Θ. Then, the phenotype of each subject was determined according to q_c. Thus, q_c was used as the probability at which any subject develops the phenotype ψ_+. Then, after removing the phase information, the algorithm as defined above was applied to the simulated data for determination of the statistic $-2\log(L_{0\max}/L_{\max})$. The simulation was repeated 10,000 times, and the distribution of the statistic was examined.

Simulation under the alternative hypothesis: Next, simulation under the alternative hypothesis was performed in the dominant mode. Thus, one of the haplotypes was assumed to be associated with the phenotype ψ_+, and

this haplotype was denoted the "phenotype-associated haplotype." For this simulation, D_+ was defined as the set of diplotype configurations that contained at least one phenotype-associated haplotype. Various real values between 0 and 1 were given to q_- and q_+ before the simulation.

The simulation was begun by assigning each of the subjects an ordered set of two haplotype copies by drawing them using Θ. Then the phenotype of each person was determined using q_+ or q_- as the probability of developing the phenotype ψ_+. q_+ was used when the diplotype configuration of that subject contained the phenotype-associated haplotype, while q_- was used otherwise.

After removing the phase information, the SNP genotypes at multiple loci and the phenotype data were subjected to the above-defined algorithm. Simulation was repeated many times, and the results obtained were analyzed. Thus, the power was estimated at various values of q_+ and q_- by assuming that, under the null hypothesis, the statistic $-2\log(L_{0\max}/L_{\max})$ follows the χ^2 distribution with 1 d.f.

In addition, simulation data were applied to two different association tests to evaluate the performance of the present algorithm. First, using the true diplotype configurations of the subjects that could not be observed easily in real data, two-by-two contingency tables were prepared for χ^2 test. The rows of the contingency tables show whether a subject had at least one copy of phenotype-associated haplotype, and the column shows the phenotype. Second, after posterior distribution of the diplotype configuration for each subject was estimated by LDSUPPORT, which is the haplotype-inference program based on the EM algorithm (Kitamura *et al.* 2002), using the phase-removed genotype data, two-by-two contingency tables were prepared assuming that the diplotype configuration of the maximum probability was true. Then the probablity of the tables were calculated by χ^2 distribution with 1 d.f. In this simulation, another Θ from an artificial gene, which was made for six SNPs under a weak linkage disequilibrium condition, was used as the frequency distribution of haplotypes in addition to Θ from the *SAA* gene.

Analysis of real data: We analyzed two sets of data previously published. One of them was derived from a cohort study in which the association between the haplotypes at *MTHFR* and the occurrence of side effects was tested. The other was derived from another cohort study in which the association between the haplotypes at *NAT2* (*N*-acetyltransferase 2) gene and the occurrence of side effects was tested, as published previously (Tanaka *et al.* 2002). In both of the studies, haplotypes were significantly associated with the occurrence of side effects.

To evaluate the reliability of estimated parameters $\hat{q}_+$ and $\hat{q}_-$, the nonparametric bootstrap method is used to calculate means and standard errors. A bootstrap sample was constructed by drawing a new set of N subjects from the original N subjects, by permitting duplicate sampling, and was applied to the present algorithm. Bootstrap sampling was repeated 10,000 times (B = 10,000) to calculate means and standard errors of the frequencies of the haplotypes, $\hat{q}_+$, $\hat{q}_-$, and the statistic $-2\log(L_{0\max}/L_{\max})$.

RESULTS

Empirical distribution of the statistic $-2\log(L_{0\max}/L_{\max})$ under the null hypothesis: Although we implemented dominant, recessive, and genotype modes in PENHAPLO, the distribution of the statistic was examined mainly in the dominant and the genotype modes.

Figure 1, A–D, shows the histograms of the statistic $-2\log(L_{0\max}/L_{\max})$ at various values of q_c and N in the dominant mode. The histograms were compatible with the expected χ^2 distribution with 1 d.f. under all the conditions tested. The histogram in the genotype mode was compatible with the expected χ^2 distribution with 2 d.f. for $q_c = 0.1$ and $N = 1000$ (Figure 1E), while the histogram is shifted to the positive direction as compared with the curve for $q_c = 0.2$ and $N = 200$ (Figure 1F). The genotype model seems to require larger sample sizes than the dominant model for the test statistic to follow the expected distribution.

Table 2 shows the estimated $\hat{q}_+$ and $\hat{q}_-$ and empirical type I error rates ($\alpha = 0.05$) for various simulation parameters of q_c and N in the dominant mode. These results suggest that, under the null hypothesis, this statistic follows the χ^2 distribution with 1 d.f. asymptotically.

Simulation under the alternative hypothesis: Simulation under the alternative hypothesis was performed in the dominant mode. Thus, it was repeated 10,000 times for various values of q_+, q_-, and N, and the proportion of the trials that yielded values of the statistic over 3.841 (the value that yielded the cumulative density function of 0.95 for the χ^2 distribution with 1 d.f.) was determined (empirical power). Figure 2 shows that the power increased with increasing value of q_+/q_- ($q_+/q_- \geq 1$) and with increase in sample size N. We then tested the effects of frequency of the phenotype-associated haplotype on the statistic. Figure 3 shows that the empirical power peaked at the intermediate frequency of phenotype-associated haplotype.

Table 3 indicates that the present algorithm has higher power than the association test using the contingency tables made by posterior distribution of the diplotype configuration after haplotype inference, especially under weak linkage disequilibrium conditions. The association test with the contingency tables made by the true diplotype configurations has the highest power because the complete data are known, but the present algorithm can be applied to SNP genotypes at multiple loci without complete data that could not be observed easily in real data.

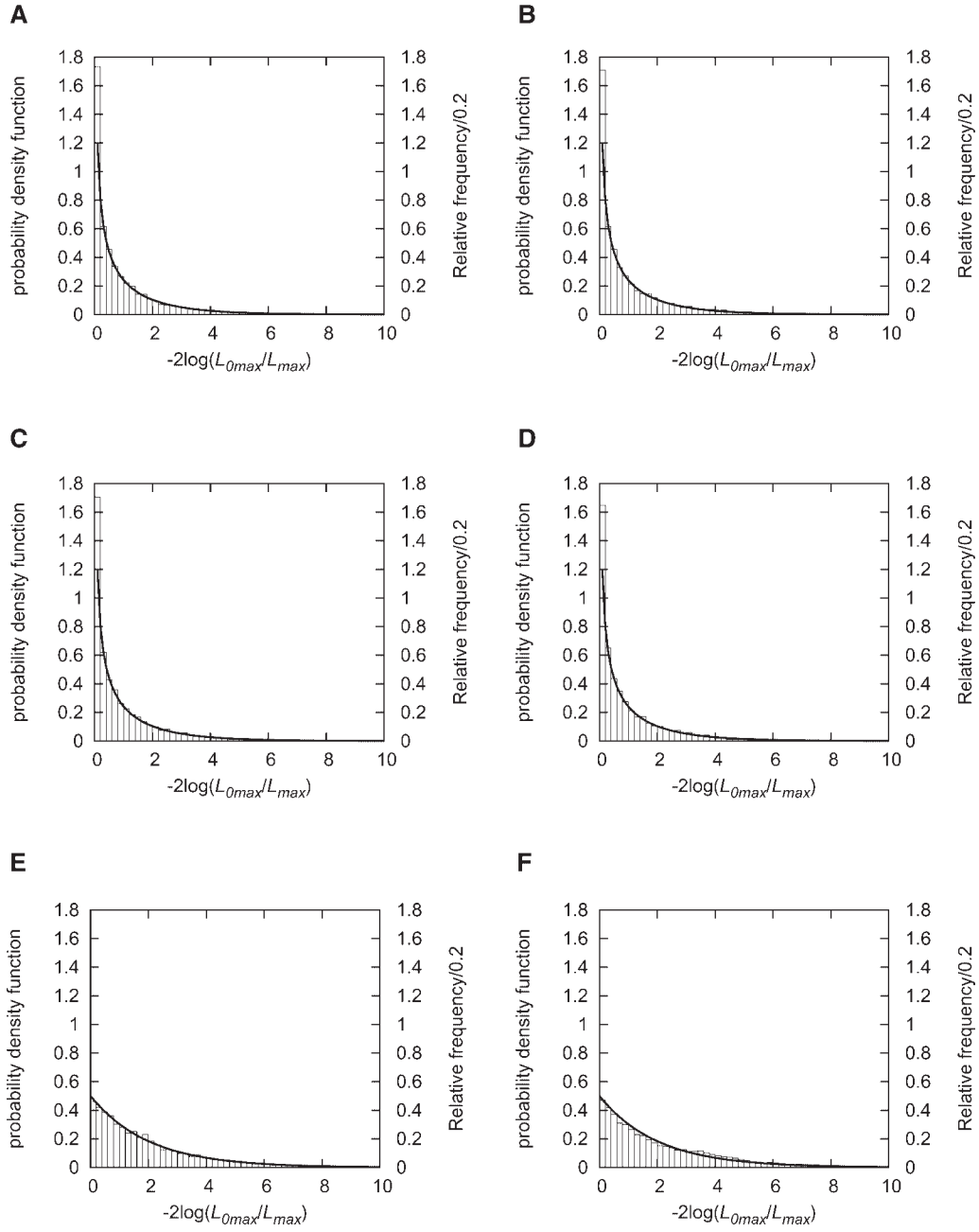

FIGURE 1.—Histograms of the statistic $-2\log(L_{0\max}/L_{\max})$ for the data produced under the null hypothesis. Simulation was performed under the null hypothesis as described in SIMULATION (under METHODS) with parameters (A) $q_c = 0.2$, $N = 1000$; (B) $q_c = 0.1$, $N = 1000$; (C) $q_c = 0.2$, $N = 200$; and (D) $q_c = 0.2$, $N = 100$ in the dominant mode or (E) $q_c = 0.1$, $N = 1000$ and (F) $q_c = 0.2$, $N = 200$ in the genotype mode. After the phase information was removed, the data were applied to PENHAPLO to calculate the statistic. This simulation was repeated 10,000 times for each parameter set. The histograms of the statistic are shown with bars. Each bar indicates the relative frequency in a 0.2 interval. The probability density function of the χ^2 distribution with 1 d.f. is shown with curves in A, B, C, and D, while the probability density function of the χ^2 distribution with 2 d.f. is shown with curves in E and F.

Distribution of estimated penetrances $\hat{q}_+$ and $\hat{q}_-$: We then examined the distribution of estimated $\hat{q}_+$ and $\hat{q}_-$ under the alternative hypothesis. Table 4 indicates that the estimation of q_+ and q_- is quite accurate and that variations are small. Accordingly, the relative risk $\hat{q}_+/\hat{q}_-$ was also reliably estimated. These results indicate that the estimated penetrances are accurate with minor variations.

Analysis of real data: We then applied the present algorithm to real data for *MTHFR* (URANO *et al.* 2002) and *NAT2* (TANAKA *et al.* 2002). Both sets of data were derived from cohort studies.

The data set for *MTHFR* was derived from a cohort study of rheumatoid arthritis patients. The 104 patients who received methotrexate were examined for both the occurrence of side effects and the genotypes at two SNP loci in the *MTHFR* gene. The precise clinical and genotype data have been published elsewhere (URANO *et al.* 2002). One of the haplotypes was assumed to be the phenotype-associated haplotype, according to a previous article. The statistic $-2\log(L_{0\max}/L_{\max})$ calculated for the data was 6.8074, which was significant ($P < 0.01$). Type II error estimated by the nonparametric bootstrap method was 0.22, which should decrease as the sample size increases. The maximum-likelihood estimates $\hat{q}_+$ and $\hat{q}_-$ were 0.2571 ± 0.0523 and 0.0588 ± 0.0405, respectively, and the maximum-likelihood estimate relative risk was 4.37. The standard errors of $\hat{q}_+$ and $\hat{q}_-$ were

TABLE 2

Estimated penetrances and empirical type I error rates in the dominant mode

q_c	Sample size N	$\hat{q}_+$[a]	$\hat{q}_-$[a]	Type I error rate[b]
0.2	1000	0.20025 ± 0.02085	0.20000 ± 0.01625	0.05060
0.1	1000	0.09996 ± 0.01570	0.10008 ± 0.01241	0.05580
0.2	200	0.19929 ± 0.04719	0.19973 ± 0.03654	0.05040
0.2	100	0.19987 ± 0.06644	0.20118 ± 0.05238	0.05640
0.4	100	0.39986 ± 0.08245	0.40040 ± 0.06400	0.05980
0.5	100	0.50117 ± 0.08471	0.49938 ± 0.06418	0.05280

Each simulation was performed with a given penetrance (q_c) and a sample size (N) under the null hypothesis in the dominant mode as stated in SIMULATION (under METHODS). After removing the phase information, PENHAPLO was used to estimate $\hat{q}_+$ and $\hat{q}_-$ under the alternative hypothesis. The same software also calculates $L_{\max}$ by analysis under the alternative hypothesis and $L_{0\max}$ by analysis under the null hypothesis. The statistic $-2\log(L_{0\max}/L_{\max})$ was then calculated. This simulation was repeated 10,000 times for each parameter set.

[a] Mean ± SD of the estimates for two penetrances obtained by analysis under the alternative hypothesis.

[b] The proportion of attempts that yielded values of the statistic over 3.841 (the value that yields the cumulative density function of 0.95 for the χ^2 distribution with 1 d.f.).

estimated by the nonparametric bootstrap method. Thus, the presence of the phenotype-associated haplotype was associated with susceptibility to development of the phenotype. The diplotype distribution of each subject was concentrated on a single event. For all the subjects, the diplotype configurations estimated under the alternative hypothesis were the same as those estimated by the null hypothesis or by LDSUPPORT, which is the haplotype inference program based on the EM algorithm (KITAMURA *et al.* 2002). Thus, the diplotype configurations estimated were in this case the same regardless of the presence of phenotypes. The maximum-likelihood estimates $\hat{\Theta}$ were not different between the results obtained with and without incorporation of phenotype data.

We next analyzed the data set for *NAT2*. This data set was also derived from a cohort study of rheumatoid arthritis patients. The 144 patients who received sulfasalazine were examined for both the occurrence of side effects and the genotypes at seven SNP loci in the *NAT2* gene. One of the haplotypes, known to be the wild-type haplotype, was assumed to be the phenotype-associated haplotype, according to a previous article. The statistic $-2\log(L_{0\max}/L_{\max})$ calculated for the data was 13.4629, which was significant ($P < 0.001$), showing that this haplotype was significantly associated with side effects. Type II error estimated by the nonparametric bootstrap method is 0.088. The maximum-likelihood estimates $\hat{q}_+$ and $\hat{q}_-$ were 0.0809 ± 0.0233 and 0.6248 ± 0.1839, respectively, and the maximum-likelihood estimate relative risk was 0.129. Thus, more precisely, the presence of the phenotype-associated haplotype was associated with a reduced probability of side effects. The diplotype

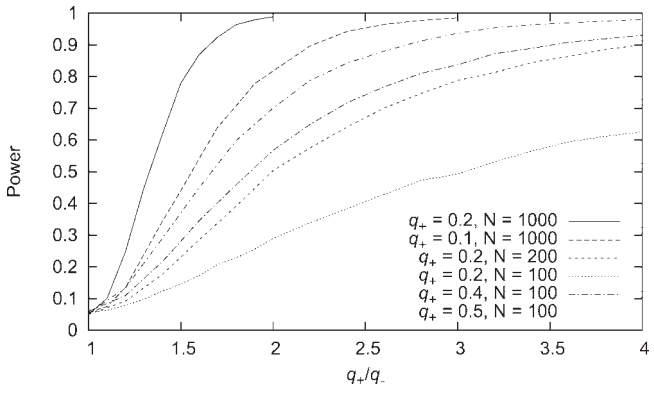

FIGURE 2.—Increase in power with increase in q_+/q_- (relative risk) and sample size N. Simulation under the alternative hypothesis in the dominant mode was performed at fixed values of q_+ and N and varying values of q_-. After removing the phase information, the data were analyzed by PENHAPLO for calculation of the statistic $-2\log(L_{0\max}/L_{\max})$. The proportion of attempts with significant statistic values (3.841) was plotted.

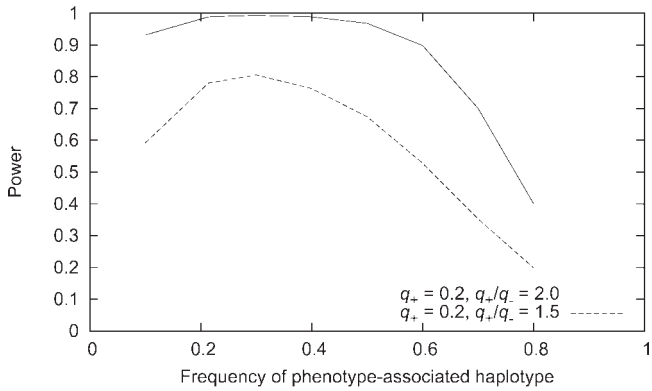

FIGURE 3.—Relationship between the frequency of phenotype-associated haplotype and power. The frequency of the phenotype-associated haplotype in Table 1 was changed, and the effect of this change was absorbed by the other haplotypes. Then simulation under the alternative hypothesis in the dominant mode was performed. Power was calculated as in Figure 2.

TABLE 3

Empirical power of three different association tests in the dominant mode

Θ[a]	q_+	q_+/q_-	Sample size N	Significance level	Diplotype counting[b]	Haplotype inference[c]	Present algorithm
SAA	0.500	4.0	200	0.050	1.000	1.000	1.000
				0.010	1.000	0.999	0.999
				0.001	0.996	0.992	0.993
SAA	0.500	3.0	200	0.050	0.999	0.999	0.999
				0.010	0.994	0.989	0.991
				0.001	0.958	0.943	0.946
SAA	0.500	2.0	200	0.050	0.955	0.941	0.943
				0.010	0.855	0.823	0.831
				0.001	0.633	0.587	0.600
SAA	0.500	1.0	200	0.050	0.050	0.051	0.051
				0.010	0.010	0.009	0.011
				0.001	0.001	0.001	0.001
ART	0.500	4.0	200	0.050	1.000	0.983	0.995
				0.010	0.998	0.938	0.979
				0.001	0.985	0.808	0.903
ART	0.500	3.0	200	0.050	0.995	0.931	0.973
				0.010	0.976	0.830	0.909
				0.001	0.910	0.609	0.741
ART	0.500	2.0	200	0.050	0.914	0.716	0.810
				0.010	0.782	0.496	0.610
				0.001	0.531	0.246	0.338
ART	0.500	1.0	200	0.050	0.053	0.047	0.059
				0.010	0.009	0.010	0.014
				0.001	0.000	0.001	0.001
ART	0.200	2.0	1000	0.050	0.976	0.870	0.923
				0.010	0.920	0.712	0.798
				0.001	0.786	0.466	0.563
ART	0.200	1.5	1000	0.050	0.745	0.546	0.616
				0.010	0.533	0.317	0.382
				0.001	0.277	0.132	0.152
ART	0.200	1.0	1000	0.050	0.052	0.050	0.054
				0.010	0.010	0.010	0.011
				0.001	0.001	0.001	0.001

Each simulation was performed as described in SIMULATION (under METHODS) under the alternative hypothesis in the dominant mode with a given penetrance of q_+ and varying the other penetrance of q_- to change the relative risk (q_+/q_-). This simulation was repeated 10,000 times for each parameter set.

[a] Θ from *SAA* or an artificial gene was used. To construct the collection of haplotype copies for the artificial gene, the data were made for six SNPs, each pair of which had weak linkage disequilibrium.

[b] Two-by-two contingency tables were prepared for the χ^2 test using the true diplotype configurations of the subjects from simulation data, and the probablity of the tables were calculated by χ^2 distribution with 1 d.f.

[c] Two-by-two contingency tables were prepared for χ^2 test using the posterior distribution of diplotype configurations estimated by LDSUPPORT, and the probablity of the tables was calculated by χ^2 distribution with 1 d.f.

distribution of each subject was concentrated on a single event in all cases but one. For all the subjects, the diplotype configurations estimated under the alternative hypothesis were the same as those estimated under the null hypothesis or by LDSUPPORT (results not shown).

DISCUSSION

In this article, we describe an algorithm that tests the association between a phenotype and a haplotype or a haplotype set at the level of individuals using observed data for genotypes and phenotypes. This algorithm can be applied to data from either clinical trials or cohort studies. The algorithm can also calculate maximum-likelihood estimates of penetrances in dominant, recessive, or genotype modes.

This algorithm implemented in PENHAPLO was found to be both reliable and powerful. As shown by the simulation studies in the dominant mode, the test statistic $-2\log(L_{0\max}/L_{\max})$ followed the χ^2 distribution with 1 d.f. asymptotically when the data were derived under the assumption of no association between haplotype and phenotype. It was also shown that this method had considerable power under certain conditions (when

TABLE 4

Estimated penetrances and empirical power in the dominant mode

q_+/q_-	$\hat{q}_+$ (mean ± SD) [a]	$\hat{q}_-$ (mean ± SD) [a]	Empirical power [b]
1.0	0.20025 ± 0.02085	0.20000 ± 0.01625	0.05060
1.1	0.19958 ± 0.02056	0.18202 ± 0.01568	0.10070
1.2	0.19999 ± 0.02076	0.16722 ± 0.01518	0.24860
1.3	0.20020 ± 0.02078	0.15390 ± 0.01447	0.45060
1.4	0.19968 ± 0.02098	0.14314 ± 0.01432	0.61810
1.5	0.20008 ± 0.02074	0.13298 ± 0.01377	0.77940
1.6	0.20003 ± 0.02077	0.12499 ± 0.01340	0.86990
1.7	0.19998 ± 0.02066	0.11807 ± 0.01315	0.92470
1.8	0.20021 ± 0.02060	0.11083 ± 0.01272	0.96390
1.9	0.20007 ± 0.02046	0.10493 ± 0.01265	0.97860
2.0	0.20003 ± 0.02059	0.09990 ± 0.01236	0.98910

Each simulation was performed as described in SIMULATION (under METHODS) under the alternative hypothesis with a given penetrance of $q_+ = 0.2$ and varying the other penetrance of q_- in the dominant mode. The relative risk (q_+/q_-) was changed from 1.0 to 2.0. The sample size N was fixed at 1000. PENHAPLO was used to estimate $\hat{q}_+$ and $\hat{q}_-$ under the alternative hypothesis and to calculate the statistic $-2\log(L_{0\max}/L_{\max})$. This simulation was repeated 10,000 times for each parameter set.

[a] Mean ± SD of the estimates for two penetrances obtained by analysis under the alternative hypothesis.

[b] The proportion of attempts that yielded values of the statistic over 3.841 (the value that yields the cumulative density function of 0.95 for the χ^2 distribution with 1 d.f.).

q_+/q_- was not close to 1, N was sufficiently large, and the frequency of the phenotype-associated haplotype was close to neither 0 nor 1). Furthermore, compared with the association test using contingency tables made by the posterior distribution of diplotype configuration after haplotype inference, this algorithm has greater power. When this method was applied to real data published previously, it could effectively detect the difference in penetrance between subjects with a given haplotype and those without it.

Although the association between phenotype and haplotypes is tested by comparing the frequencies of a haplotype between two groups with different phenotypes, comparison of penetrances between two groups with different diplotype configurations is even more important in some cases. This method provided the basis for such studies.

In this algorithm, the diplotype configuration for each subject should not necessarily be unequivocally determined. Therefore, this method is expected to be robust and superior to the methods in which ambiguous diplotype configurations are not allowed. When the haplotypes in question involve only polymorphic loci within a haplotype block, the diplotype distribution for each subject is often concentrated on a single event and the diplotype configuration thus estimated is reliable. However, if the haplotypes involve loci located beyond the boundaries of a haplotype block, the diplotype distribution for each subject is often dispersed, and even if the most likely diplotype configuration is selected, estimation is not reliable. Therefore, the test presented here is likely to be more practical for real data.

This method considers the association between haplotypes and phenotypes by focusing on penetrances, while other methods view the same association on the basis of the difference in the frequencies of the haplotypes between groups with different phenotypes. Recent studies on drug-related genes have clarified that some genetic differences are significantly related to the occurrence of side effects and to efficacy. Some studies have shown that haplotypes are associated with reactions to drugs (DRYSDALE *et al.* 2000; JUDSON *et al.* 2000). If one intends to apply the results of such studies to a single individual, it is important to evaluate the probability that he or she develops the phenotype (for example, the occurrence of side effects). In such a case, the probability of developing the phenotype may be estimated on the basis of the observed data for the genotypes before the administration of drugs. This article provides one such method for estimation of probabilities on which decisions regarding treatment may depend.

The problem in association studies based on haplotypes is that the candidate phenotype-associated haplotype should be known in advance. In some cases, it is known from previous independent studies or the data from the functional studies. However, when it is not, one must test all the haplotypes, each of which may be associated with the phenotype. In PENHAPLO, the function to test all incomplete haplotypes for the association with a phenotype has been implemented. When each incomplete haplotype in a region is tested, the problem of the multiple comparison will emerge because the number of all the incomplete haplotypes will be huge. The correction for multiple comparisons, such as Bonferroni's correction, is definitely too conservative since such multiple tests are tightly dependent on each

other. One of the solutions is to perform the association study using haplotypes constructed with haplotype-tagging SNPs (htSNPs) extracted from a block. The extraction of htSNPs is one of the methods with which to remove the redundancy from the haplotype-based association analysis. Then we use only htSNPs within a block as unmasked loci in the incomplete haplotypes. When the above procedures are employed, the number of incomplete haplotypes within a block with considerable frequencies is not so large (KAMATANI *et al.* 2004). We admit, however, that the problem of multiple comparison is the greatest problem to be considered in the future studies in the haplotype-based association studies. The computer program PENHAPLO will be sent to researchers on request under certain conditions.

In summary, we have devised an algorithm with which to test the association between a haplotype or a haplotype set and a phenotype in various modes and to estimate diplotype-based penetrances using data from cohort studies and clinical trials. We have implemented the algorithm in the computer program PENHAPLO. Both simulated data and real data were analyzed by our algorithm. The results suggested that our method is especially useful for analyzing data concerning the association between genetic information and individual responses to drugs.

This study was supported by a grant for Research for the Future Program from the Japan Society for the Promotion of Science (JSPS).

LITERATURE CITED

BADER, J. S., 2001 The relative power of SNPs and haplotype as genetic markers for association tests. Pharmacogenomics **2:** 11–24.

CLARK, A. G., 1990 Inference of haplotypes from PCR-amplified samples of diploid populations. Mol. Biol. Evol. **7:** 111–122.

DRYSDALE, C. M., D. W. MCGRAW, C. B. STACK, J. C. STEPHENS, R. S. JUDSON *et al.*, 2000 Complex promoter and coding region beta 2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. Proc. Natl. Acad. Sci. USA **97:** 10483–10488.

EXCOFFIER, L., and M. SLATKIN, 1995 Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol. Biol. Evol. **12:** 921–927.

HAWLEY, M. E., and K. K. KIDD, 1995 HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. J. Hered. **86:** 409–411.

HORIKAWA, Y., N. ODA, N. J. COX, X. LI, M. ORHO-MELANDER *et al.*, 2000 Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. Nat. Genet. **26:** 163–175.

HUGOT, J. P., M. CHAMAILLARD, H. ZOUALI, S. LESAGE, J. P. CEZARD *et al.*, 2001 Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature **411:** 599–603.

JOOSTEN, P. H., M. TOEPOEL, E. C. MARIMAN and E. J. VAN ZOELEN, 2001 Promoter haplotype combinations of the platelet-derived growth factor alpha-receptor gene predispose to human neural tube defects. Nat. Genet. **27:** 215–217.

JUDSON, R., J. C. STEPHENS and A. WINDEMUTH, 2000 The predictive power of haplotypes in clinical response. Pharmacogenomics **1:** 5–16.

KAMATANI, N., A. SEKINE, T. KITAMOTO, A. IIDA, A. SAITO *et al.*, 2004 Large-scale single-nucleotide polymorphism (SNP) and haplotype analyses, using dense SNP maps, of 199 drug-related genes in 752 subjects: the analysis of the association between uncommon SNPs within haplotype blocks and the haplotypes constructed with haplotype-tagging SNPs. Am. J. Hum. Genet. **75:** 190–203.

KITAMURA, Y., M. MORIGUCHI, H. KANEKO, H. MORISAKI, T. MORISAKI *et al.*, 2002 Determination of probability distribution of diplotype configuration (diplotype distribution) for each subject from genotypic data using the EM algorithm. Ann. Hum. Genet. **66:** 183–193.

LONG, J. C., R. C. WILLIAMS and M. URBANEK, 1995 An E-M algorithm and testing strategy for multiple locus haplotypes. Am. J. Hum. Genet. **56:** 799–810

MICHALATOS-BELOIN, S., S. A. TISHKOFF, K. L. BENTLEY, K. K. KIDD and G. RUANO, 1996 Molecular haplotyping of genetic markers 10 kb apart by allele-specific long-range PCR. Nucleic Acids Res. **24:** 4841–4843.

MORIGUCHI, M., C. TERAI, H. KANEKO, Y. KOSEKI, H. KAJIYAMA *et al.*, 2001 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:** 1266–1272.

MUMMIDI, S., M. BAMSHAD, S. S. AHUJA, E. GONZALEZ, P. M. FEUILLET *et al.*, 2000 Evolution of human and non-human primate CC chemokine receptor 5 gene and mRNA. Potential roles for haplotype and mRNA diversity, differential haplotype-specific transcriptional activity, and altered transcription factor binding to polymorphic nucleotides in the pathogenesis of HIV-1 and simian immunodeficiency virus. J. Biol. Chem. **275:** 18946–18961.

NIU, T., Z. S. QIN, X. XU and J. S. LIU, 2002 Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. Am. J. Hum. Genet. **70:** 157–169.

PUFFENBERGER, E. G., E. R. KAUFFMAN, S. BOLK, T. C. MATISE, S. S. WASHINGTON *et al.*, 1994 Identity-by-descent and association mapping of a recessive gene for Hirschsprung disease on human chromosome 13q22. Hum. Mol. Genet. **3:** 1217–1225.

QIN, Z. S., T. NIU and A. S. LIU, 2002 Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. Am. J. Hum. Genet. **71:** 1242–1247.

RIOUX, J. D., M. J. DALY, M. S. SILVERBERG, K. LINDBLAD, H. STEINHART *et al.*, 2001 Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. Nat. Genet. **29:** 223–228.

SCHNEIDER, S., D. ROESSLI and L. EXCOFFIER, 2000 *Arlequin: A Software for Population Genetics Data Analysis*, Version 2.000. Genetics and Biometry Laboratory, Department of Anthropology, University of Geneva, Geneva

STEPHENS, M., N. J. SMITH and P. DONNELLY, 2001 A new statistical method for haplotype reconstruction from population data. Am. J. Hum. Genet. **68:** 978–989.

TANAKA, E., A. TANIGUCHI, W. URANO, H. NAKAJIMA, Y. MATSUDA *et al.*, 2002 Adverse effects of sulfasalazine in patients with rheumatoid arthritis are associated with diplotype configuration at the N-acetyltransferase 2 gene. J. Rheumatol. **29:** 2492–2499.

URANO, W., A. TANIGUCHI, H. YAMANAKA, E. TANAKA, H. NAKAJIMA *et al.*, 2002 Polymorphisms in the methylenetetrahydrofolate reductase gene were associated with both the efficacy and the toxicity of methotrexate used for the treatment of rheumatoid arthritis, as evidenced by single locus and haplotype analyses. Pharmacogenetics **12:** 183–190.

Communicating editor: Y.-X. FU