J Hum Genet (2008) 53:789–801
DOI 10.1007/s10038-008-0312-0

ORIGINAL ARTICLE

# New correction algorithms for multiple comparisons in case–control multilocus association studies based on haplotypes and diplotype configurations

**Kazuharu Misawa · Shoogo Fujii · Toshimasa Yamazaki · Atsushi Takahashi · Junichi Takasaki · Masao Yanagisawa · Yozo Ohnishi · Yusuke Nakamura · Naoyuki Kamatani**

Received: 21 March 2008 / Accepted: 3 June 2008 / Published online: 24 July 2008


**Abstract** The multiple comparison problem arises in population-based studies when the association between phenotypes and multilocus genotypes is examined. Although Bonferroni's correction is often used to cope with such a problem, it may yield too conservative conclusions because all of the tests are assumed to be independent. We have developed new correction algorithms for the test of independence between phenotypes and multilocus genotypes at loci in linkage disequilibrium. In one of the algorithms, the exact type I error rate is calculated for the independency test. We found that such exact probabilities can be calculated using a 128 CPU PC cluster if the numbers of cases and controls are not more than 50. As an alternative method, we developed algorithms to calculate asymptotically the type I error rates using a Markov-chain Monte Carlo sampler that provided a good approximation to values calculated by the exact method. When the new algorithms were applied to both simulation and real data, the real overall type I error rates for the loci in linkage disequilibrium were from one-third to half as high as those obtained by Bonferroni's correction. These algorithms are likely to be useful for multilocus association studies for data obtained by case–control and cohort studies.

**Keywords** Linkage disequilibrium · Type I error · Single nucleotide polymorphism · Markov chain Monte Carlo · Haplotype

K. Misawa (✉)
Research Program for Computational Science, Research and Development Group for Next-Generation Integrated Living Matter Simulation, Fusion of Data and Analysis Research and Development Team, RIKEN, 4-6-1 Shirokane-dai, Minato-ku, Tokyo 108-8639, Japan
e-mail: kazumisawa@riken.jp

S. Fujii · T. Yamazaki · A. Takahashi · J. Takasaki · N. Kamatani
Laboratory for Statistical Analysis, RIKEN Center for Genomic Medicine, Tokyo, Japan

S. Fujii · M. Yanagisawa
Department of Computer Science, Waseda University, Tokyo, Japan

Y. Ohnishi
Laboratory for SNP Analysis, RIKEN Center for Genomic Medicine, Tokyo, Japan

Y. Nakamura
Laboratory for Pharmacogenetics, RIKEN Center for Genomic Medicine, Tokyo, Japan

N. Kamatani
Division of Genomic Medicine, Department of Advanced Biomedical Engineering and Science, and Institute of Rheumatology, Tokyo Women's Medical University, Tokyo, Japan

## Introduction

Various genotyping technologies targeted at single-nucleotide polymorphism (SNP) have been introduced. Thus, the Taqman technique (Rickert et al. 2004), the Invader method (Ohnishi et al. 2001), matrix-assisted laser desorption/ionization (MALDI) time-of-flight (TOF) mass spectrometry (MS) (Jurinke et al. 2004), BeadArray technology (Oliphant et al. 2002), and oligonucleotide arrays (Cutler et al. 2001) have been very efficient for high-throughput genotyping of a large number of single-nucleotide polymorphism (SNP) loci at reasonable costs. Such high throughput, low-cost, and accurate technologies have made genome-wide association studies feasible. For example, over 50,000 SNPs have been selected from the

entire human genome and used for the identification of the genes related to diseases such as cardiac infarction (Ozaki et al. 2002, 2004), rheumatoid arthritis (Suzuki et al. 2003; Tokuhiro et al. 2003), osteoarthritis (Kizawa et al. 2005), and diabetes mellitus (Kanazawa et al. 2004).

The analyses of these data involve statistical tests on hundreds of thousands of SNP loci in the genome. One of the crucial problems in such tests is correction for multiple comparisons (Cardon and Bell 2001; Lander and Kruglyak 1995; Seaman and Muller-Myhsok 2005; Thakkinstian et al. 2004). Thus, even though the type I error rate may be 0.05 for a SNP, the global type I error rate for all SNPs is much higher. Bonferroni's correction is commonly used for the multiple-comparison problem. However, this method is known to be too conservative and may drop truly significant SNPs (type II error). In Bonferroni's correction, all of the multiple tests are assumed to be independent. In the case of SNP-based association studies, however, this assumption does not hold. Thus, several SNP loci are likely to be in linkage disequilibrium, and association tests using SNPs that are in linkage disequilibrium with each other are not independent. Thus, if one of the multiple SNPs that are in linkage disequilibrium with each other is judged to be significantly associated with a disease by a statistical test, then the other SNPs are more likely to be judged to be significant by the same test, compared to independent SNPs, regardless of whether the association is true or merely sample-dependent. This problem arises not only in genome-wide association studies, but also in various association tests in which multiple linked loci are tested.

Several approaches have been proposed for the multiple comparison problem in association studies based on multilocus genotypes. Sabatti et al. (2003) applied the false discovery rate (FDR) proposed by Benjamini and Hochberg (1995), and succeeded in increasing the power. Cheverud (2001) introduced a method for the correction of multiple comparisons in genome scans through the use of the variance of the eigenvalues of the observed marker correlation matrix. Subsequently, Nyholt (2004) proposed a simple correction for multiple testing on the basis of the spectral decomposition of matrices of pairwise linkage disequilibrium between SNPs. However, no standard correction method has been established for the multiple comparisons for association studies in such a context.

We have developed new algorithms to correct for the multiple comparisons at multiple SNP loci in linkage disequilibrium. In the present study, we calculated the exact probability of the type I error under the condition that the haplotype frequencies in the population are known. We assumed Hardy–Weinberg's equilibrium at the haplotype level, and either the number of alleles or the number of genotypes was assumed to follow a multinomial distribution. We found that the exact probability of the type I error can be calculated for a case–control study in which the test of independence between a phenotype and the genetic information is performed for all linked SNP loci within a chromosomal region. Since the new method incorporates the inheritance model, the inheritance mode could be incorporated into the calculation.

However, the calculation of the exact probability was possible only when the numbers of the cases and controls were small ($n < 50$), even for the case in which a high-speed machine was used. We devised a Markov-chain Monte Carlo (MCMC) algorithm to calculate asymptotically the probability of the type I error for the above test. We also devised an algorithm to calculate the exact probability of the type I error under the assumption of a hypergeometric distribution for either the number of alleles or the number of individuals. Although the exact probabilities of the type I error under the assumption of the hypergeometric distribution were difficult to calculate, a MCMC algorithm was developed to calculate such values asymptotically.

## Methods

### Notation

In the present study, a haplotype denotes a list of alleles (one allele per locus) at multiple linked polymorphic loci (only biallelic loci are considered in the present manuscript), and a haplotype copy denotes a list of alleles possessed in a gamete. Therefore, if a subject is homozygous for a haplotype, then the subject possesses a haplotype, but two haplotype copies. A combinational diplotype configuration is defined as a combination of two (unordered) haplotype copies possessed by an individual, and an ordered diplotype configuration denotes an ordered list of two haplotype copies arranged according to the derivation (father and mother) (Shibata et al. 2004).

### Calculation of type I error rate in allele frequency mode when haplotype frequencies are known

#### *Calculation of exact probability under the assumption of a multinomial distribution*

Let $l$ denote the number of linked SNP loci and $L$ the number of possible haplotypes in the population. The number of possible haplotypes will be $L = 2^l$. Imagine that the population frequencies of all $L$ haplotypes are given, and the frequency of the $i$th haplotype is $h_i$ ($i = 1, 2,\ldots, L$), where $\sum_{i=1}^{L} h_i = 1$. Let $n_1$ and $n_2$ denote the sizes of the case and the control groups, respectively, that were independently and randomly drawn from the population. We

test whether the allele frequency in each SNP locus differs between the case and the control groups. Let us next consider an experiment in which $2n_1$ and $2n_2$ haplotype copies are drawn as the case and control groups, respectively, from the population. For the test of independence between the phenotype and the allele frequency, the null hypothesis ($H_0$) is defined such that there is no allele frequency difference between the case and control groups. Suppose that the numbers of copies of the $i$th haplotype from the first (case) and second (control) groups are random variables $X_{1i}$ and $X_{2i}$, respectively, where $\sum_{i=1}^{L} X_{1i} = 2n_1$ and $\sum_{i=1}^{L} X_{2i} = 2n_2$. Then, because the two samplings are independent, both $X_{1i}$ and $X_{2i}$ are binomially distributed, and the joint distributions of $X_{11}$, $X_{12}$,..., $X_{1L}$ and $X_{21}$, $X_{22}$,..., $X_{2L}$ are multinomial. Therefore, the probabilities can be given by

$$P(X_{11} = x_{11}, X_{12} = x_{12}, \ldots, X_{1L} = x_{1L}) = \frac{(2n_1)!}{x_{11}!x_{12}!\ldots x_{1L}!} h_1^{x_{11}} h_2^{x_{12}} \ldots h_L^{x_{1L}}, \quad (1)$$

and

$$P(X_{21} = x_{21}, X_{22} = x_{22}, \ldots, X_{2L} = x_{2L}) = \frac{(2n_2)!}{x_{21}!x_{22}!\ldots x_{2L}!} h_1^{x_{21}} h_2^{x_{22}} \ldots h_L^{x_{2L}}. \quad (2)$$

Moreover, the joint distribution of $X_{11}$, $X_{12}$,..., $X_{1L}$ and $X_{21}$, $X_{22}$,..., $X_{2L}$ is described as a product of Eqs. (1) and (2):

$$P(X_{11}, X_{12}, \ldots, X_{1L}, X_{21}, X_{22}, \ldots, X_{2L}) = \frac{(2n_1)!(2n_2)!}{\prod_{i=1}^{L}\prod_{j=1}^{2} x_{ji}!} \prod_{i=1}^{L}\prod_{j=1}^{2} h_i^{x_{ji}}. \quad (3)$$

A matrix, the elements of which are given by $\alpha_{ik}$ ($i$ = 1, 2,...,$L$; $k$ = 1, 2,..., $l$), is defined as $\alpha_{ik} = 1$ if the $k$th locus of the $i$th haplotype has a minor allele and as $\alpha_{ik} = 0$ if the $k$th locus of the $i$th haplotype has a major allele. Then, a minor allele frequency (MAF) $p_k$ of the $k$th locus is given by a function of the haplotype frequencies, as follows:

$$p_k = \sum_{i=1}^{L} \alpha_{ik} h_i$$

The numbers of minor alleles at the $k$th locus for the first and second groups, $Y_{1k}$ and $Y_{2k}$, are random variables, and are described as

$$Y_{jk} = \sum_{i=1}^{L} \alpha_{ik} X_{ji}. \quad (4)$$

One result of this experiment is represented, with respect to the $k$th locus, by a 2 × 2 contingency table (Table 1), and a test of the independence in the allele frequency mode is performed using this table. A variable $Z$ is defined such that, after the tests for $l$ loci ($k$ = 1, 2,..., $l$), $Z$ = 1 if the test is significant at any locus, otherwise $Z$ = 0. Then, $Z$ is a function of $Y_{jk}$, which is a function of $X_{ji}$, as shown by Eq. (4), and is therefore a random variable. Therefore, $Z$ can be described as

$$Z = f\left(X_{11}, X_{12}, \ldots, X_{1L-1}, X_{21}, X_{22}, \ldots, X_{2,L-1}\right), \text{ and } P\left[f\left(X_{11}, X_{12}, \ldots, X_{1L-1}, X_{21}, X_{22}, \ldots, X_{2,L-1}\right) = 1\right] \quad (5)$$

defines the probability of block type I error. Here, the block type I error is defined as an event in which a type I error occurs in at least one of the tests. Since the distribution of $\{X_{ji}\}$ is given by Eq. (3), Eq. (5) can be obtained by enumerating all possible values of $\{X_{ji}\}$ and calculating $f(X_{11}, X_{12},\ldots, X_{1L-1}, X_{21}, X_{22},\ldots, X_{2,L-1})$, as follows:

$$P\left[f\left(X_{11}, X_{12}, \ldots, X_{1L-1}, X_{21}, X_{22}, \ldots, X_{2,L-1}\right) = 1\right] = \sum_{x_{11}=0}^{2n_1} \sum_{x_{12}=0}^{2n_1-x_{11}} \sum_{x_{13}=0}^{2n_1-x_{11}-x_{12}} \cdots \sum_{x_{1,L-1}=0}^{2n_1-x_{11}-x_{12}-\cdots-x_{1,L-2}} \sum_{x_{21}=0}^{2n_2} \sum_{x_{22}=0}^{2n_2-x_{21}} \sum_{x_{23}=0}^{2n_2-x_{21}-x_{22}} \cdots \sum_{x_{2,L-1}=0}^{2n_2-x_{21}-x_{22}-\cdots-x_{2,L-2}} f\left(x_{11}, x_{12}, \ldots, x_{1L-1}, x_{21}, x_{22}, \ldots, x_{2,L-1}\right) \frac{(2n_1)!(2n_2)!}{\prod_{i=1}^{L}\prod_{j=1}^{2} x_{ji}!} \prod_{i=1}^{L}\prod_{j=1}^{2} h_i^{x_{ji}}, \quad (6)$$

where $x_{1L} = 2n_1 - \sum_{i=1}^{L-1} x_{1i}$ and $x_{2L} = 2n_2 - \sum_{i=1}^{L-1} x_{2i}$.

**Table 1** Contingency table I for the allele frequency mode

| Group | Number of minor allele copies | Number of major allele copies |
|---|---|---|
| 1 | $y_{1k}$ | $2n_1 - y_{1k}$ |
| 2 | $y_{2k}$ | $2n_2 - y_{2k}$ |
| Total | $y_k$ | $2n_1 + 2n_2 - y_k$ |

$n_1$ and $n_2$ denote the number of subjects in groups 1 and 2, respectively. $y_{jk}$ denotes the observed value of a random variable $Y_{jk}$, denoting the number of minor allele copies at the $k$th locus in group $j$. $k$ denotes the order of the locus at which the test of independence is performed. The number in each cell of the 2 × 2 contingency table indicates the observed number of allele copies

*Calculation of type I error rate under the assumption of a multinomial distribution by MCMC*

The time required to calculate formula (6) is enormous, even with the assistance of computers. The possible number of different $\{x_{ji}\}$ in the right-hand side of Eq. (6) is given as

$$\frac{1}{(L-1)!}(2n_1+1)(2n_1+2)\cdots(2n_1+L-1) \times \frac{1}{(L-1)!}(2n_2+1)(2n_2+2)\cdots(2n_2+L-1) \quad (7)$$

and so increases very rapidly with respect to $n_1$, $n_2$, and $L$. When $L = 10$, $n_1 = 100$ and $n_2 = 100$, Eq. (7) exceeds $10^{30}$. We then attempted to calculate the approximate values of Eq. (6) with given $n_1$, $n_2$, and $L$ using the MCMC method. We also developed an approximate algorithm to calculate formula (6) (see Appendix).

First, we consider a case in which the population haplotype frequencies are given. Thus, we used the MCMC method based on the Metropolis–Hastings algorithm (Sorensen and Gianola 2002) to generate $\{x_{ji}\}$ assuming that each of $\{X_{1i}\}$ and $\{X_{2i}\}$ follows the same multinomial distribution with the frequency parameters of the given population haplotype frequencies. A sample from the MCMC contains the values of $x_{1i}$, $i = 1, 2,\ldots, L$ and $x_{2i}$, $i = 1, 2,\ldots, L$, representing the numbers of haplotypes in the case and control groups. The number of minor alleles in each group at each locus is given by Eq. (4). Using the function $f$, we can test whether the independence test at least at one of the $l$ loci is significant. If the samples of $\{X_{1i}\}$ and $\{X_{2i}\}$ are generated appropriately, then the proportion of the samples that are shown to be significant by the above test can be calculated.

Thus, the following MCMC sampler was produced:

1. The state space is defined by a set of all the different $\{x_{ji}\}$, $x_{ji} \geq 0$, $j = 1, 2$; $i = 1, 2,\ldots, L$.
2. As an initial state, arbitrary integer values of $x_{ji}$ ($j = 1, 2$; $i = 1, 2,\ldots,L$) are given so that $\sum_{i=1}^{L} x_{ji} = 2n_j, x_{ij} \geq 0$.
3. $j = 1$ or $j = 2$ is selected in equal probability.
4. An integer value $u$ is selected in equal probability from the integers from 1 to $L$.
5. If $x_{ju} = 0$, then the state is kept invariant, the step is advanced, a test of significance is performed as in (10) and the process returns to (3).
6. If $x_{ju} > 0$, an integer value $v$ other than $u$ ($1 \leq v \leq L$) is selected, and new candidates $x^*_{ju} = x_{ju} - 1$ and $x^*_{jv} = x_{jv} + 1$ are calculated.
7. Then, the following value is calculated,
$$c = \frac{h_v x_{ju}! x_{jv}!}{h_u x^*_{ju}! x^*_{jv}!} = \frac{h_v x_{ju}}{h_u (x_{jv} + 1)}$$
8. If $c \geq 1$, then $\{x_{ji}\}$ is updated by substituting $x^*_{ju}$ for $x_{ju}$, and substituting $x^*_{jv}$ for $x_{jv}$, the step is advanced, a significance test is performed as described in (10) and the process then returns to (3).
9. If $c < 1$, then $\{x_{ji}\}$ is updated by substituting $x^*_{ju}$ for $x_{ju}$, and substituting $x^*_{jv}$ for $x_{jv}$ with probability $c$, else the state is kept invariant (probability $1 - c$), the step is advanced, a significance test is performed as described in (10) and the process then returns to (3).
10. A test of independence between the phenotype and alleles at each of the $l$ loci is performed using a contingency table (Table 1) obtained from the values of $x_{ji}$ ($j = 1, 2$; $i = 1, 2,\ldots,L$). Thus, to construct a contingency table for the $k$th locus ($k = 1, 2,\ldots, l$), $y_{jk} = \sum_{i=1}^{L} \alpha_{ik} x_{ji}$ is calculated for the groups $j = 1$ and $j = 2$ according to Eq. (4). If a significant result is obtained by the test for at least one of the $l$ loci, then the block test is considered to be significant.
11. After the MCMC has been run in a sufficient number of steps, the proportion of the steps judged to be significant is considered to be the empirical type I error rate for the block test.

## Calculation of type I error rate for the test in dominant, recessive, or genotype mode

### *Calculation of exact probability for the test in dominant, recessive, or genotype mode under the assumption of a multinomial distribution*

In order to calculate the type I error rate for the test in dominant or recessive mode, the concept of diplotype configuration is necessary. In the present study, the test in the dominant mode is defined as the test of the difference in the proportion of the individuals with the minor allele between the two groups. On the other hand, the test in the recessive model refers to the test of the difference in the proportion of the individuals with the major allele. Note that the latter test is equivalent to the test of the difference in the proportion of the individuals with the minor allele as homozygotes.

In the present study, the ordered diplotype configuration denotes the ordered combination of two haplotype copies in a subject. Let $d_{ij}$ denote an ordered diplotype configuration consisting of the $i$th and $j$th (in this order) haplotypes. Since the total number of haplotypes is $L$, the total number of ordered diplotype configurations is $L^2$. If the Hardy–Weinberg's equilibrium holds at the haplotype level, then the frequency of $d_{ij}$ in the population is equal to $h_i h_j$. For the test of the difference in the dominant or recessive mode, subjects rather than alleles should be considered. Thus, the frequencies of the subjects with a certain category of genotypes are compared between the cases and controls. The sample space in this case, therefore, is slightly different. To construct the sample space, an experiment is defined as drawing at random $n_1$ and $n_2$ ordered diplotype configurations for the first (case) and the second (control) groups, respectively. Each ordered diplotype configuration corresponds to a subject. In this sample space, the total number of individuals having $d_{ij}$ in the $k$th ($k = 1, 2$) group is denoted by a random variable $X_{kij}$ that follows a binomial distribution $B(n_k, h_i h_j)$. For a specific $k$, the joint distribution of all $X_{kij}$ ($i = 1, 2,\ldots,L$; $j = 1,$

2,…,*L*), that is, the distribution of each of the matrices $\{X_{1ij}\}$ and $\{X_{2ij}\}$ is multinomial. The probability for the *k*th group is given by

$$P\left(\{X_{kij}\} = \{x_{kij}\}\right) = \frac{n_k!}{\prod_{i=1}^{L}\prod_{j=1}^{L} x_{kij}!}\prod_{i=1}^{L}\prod_{j=1}^{L}\left(h_i h_j\right)^{x_{kij}}$$

The joint distribution of $X_{kij}$ for all *k*, *i* and *j* is given by

$$P\left(\{X_{1ij}\} = \{x_{1ij}\}, \{X_{2ij}\} = \{x_{2ij}\}\right) = \frac{n_k!}{\prod_{k=1}^{2}\prod_{i=1}^{L}\prod_{j=1}^{L} x_{kji}!}\prod_{k=1}^{2}\prod_{i=1}^{L}\prod_{j=1}^{L}\left(h_i h_j\right)^{x_{kji}}.$$

For each locus *q*, a contingency table is constructed using $\alpha_{iq}$ and $\{X_{kij}\}$ for the test of independence in the recessive mode as shown in Table 2. Here, $\alpha_{iq}$ ($i = 1, 2,\ldots,L$; $q = 1, 2,\ldots, l$), is defined so that $\alpha_{iq} = 1$ if the *q*th locus of the *i*th haplotype has a minor allele and $\alpha_{iq} = 0$ otherwise. Note that $\alpha_{iq}\alpha_{jq}$ in Table 2 is equal to 1 only when both of the alleles at locus *q* in the subject with the ordered diplotype configuration $d_{ij}$ are minor alleles. In addition, note that the mode (recessive or dominant) depends on which allele (minor or major allele) we consider, and, in the present study, we consider the minor allele in all cases. Next, the function $f(\{X_{kij}\}$, $k = 1,2$; $i = 1, 2, \ldots, L$; $j = 1, 2, \ldots, L$) is defined such that, based on an independence test at each *q*th locus ($q = 1, 2,\ldots, l$) using the above contingency table, $f = 1$ if significance is found at any locus, otherwise $f = 0$.

**Table 2** Contingency table for the recessive mode

| Group | Number of subjects without major allele | Number of subjects with major allele |
|---|---|---|
| 1 | $\sum_{i=1}^{L}\sum_{j=1}^{L}\alpha_{iq}\alpha_{jq}x_{1ij}$ | $n_1 - \sum_{i=1}^{L}\sum_{j=1}^{L}\alpha_{iq}\alpha_{jq}x_{1ij}$ |
| 2 | $\sum_{i=1}^{L}\sum_{j=1}^{L}\alpha_{iq}\alpha_{jq}x_{2ij}$ | $n_2 - \sum_{i=1}^{L}\sum_{j=1}^{L}\alpha_{iq}\alpha_{jq}x_{2ij}$ |

$n_1$ and $n_2$ denote the number of subjects in groups 1 and 2, respectively. *L* and *l* denote the total possible numbers of haplotypes and loci, respectively. $\alpha_{rs}$ ($r = 1, 2, \ldots, L$; $s = 1, 2, \ldots, l$), is defined such that $\alpha_{rs} = 1$ if the *s*th locus of the *r*th haplotype has the minor allele and $\alpha_{rs} = 0$ if the *s*th locus of the *r*th haplotype has the major allele. $x_{kij}$ denotes the observed number of subjects with the ordered diplotype configuration with the *i*th and *j*th (in this order) haplotypes in group *k*. *q* denotes the order of the locus at which the test of independence is performed

Then, $Z = f(\{X_{kij}\}$, $k = 1,2$; $i = 1, 2, \ldots, L$; $j = 1, 2, \ldots, L$) is a random variable, and the type I error rate $P(Z = 1)$ is obtained by

$$\begin{aligned} P(Z=1) = & \sum_{x_{111}=0}^{s_{111}}\sum_{x_{112}=0}^{s_{112}}\sum_{x_{113}=0}^{s_{113}}\cdots\sum_{x_{11L}=0}^{s_{11L}} \\ & \sum_{x_{121}=0}^{s_{121}}\sum_{x_{122}=0}^{s_{122}}\sum_{x_{123}=0}^{s_{123}}\cdots\sum_{x_{12L}=0}^{s_{12L}} \\ & \cdots \\ & \sum_{x_{1L1}=0}^{s_{1L1}}\sum_{x_{1L2}=0}^{s_{1L2}}\sum_{x_{1L3}=0}^{s_{1L3}}\cdots\sum_{x_{1L,L-1}=0}^{s_{1L,L-1}} \\ & \sum_{x_{211}=0}^{s_{211}}\sum_{x_{212}=0}^{s_{212}}\sum_{x_{213}=0}^{s_{213}}\cdots\sum_{x_{21L}=0}^{s_{21L}} \\ & \sum_{x_{221}=0}^{s_{221}}\sum_{x_{222}=0}^{s_{222}}\sum_{x_{223}=0}^{s_{223}}\cdots\sum_{x_{22L}=0}^{s_{22L}} \\ & \cdots \\ & \sum_{x_{2L1}=0}^{s_{2L1}}\sum_{x_{2L2}=0}^{s_{2L2}}\sum_{x_{2L3}=0}^{s_{2L3}}\cdots\sum_{x_{2L,L-1}=0}^{s_{2L,L-1}} \\ & \frac{\prod_{k=1}^{2} f\left(\{x_{kij}\}\right) n_k!}{\prod_{k=1}^{2}\prod_{i=1}^{L}\prod_{j=1}^{L} x_{kji}!}\prod_{k=1}^{2}\prod_{i=1}^{L}\prod_{j=1}^{L}\left(h_i h_j\right)^{x_{kji}}, \end{aligned} \tag{8}$$

where $k = 1, 2$; $i = 1, 2,\ldots,L$; $j = 1, 2,\ldots,L$, $x_{KLL} = n_k - \sum_{i=1}^{L}\sum_{j=1}^{L-1} x_{kij} - \sum_{j=1}^{L-1} x_{kiL}$, and $s_{kij}$ are defined as $s_{k11} = n_k, s_{k12} = s_{k11} - x_{k11}, \ldots, s_{k1L} = s_{k1,L-1} - x_{k1,L-1}$, $s_{k21} = s_{k1L} - x_{k1L}, s_{k22} = s_{k21} - x_{k21}, \ldots, s_{k2L} = s_{k2,L-1} - x_{k2,L-1}, \ldots, s_{kL1} = s_{k,L-1,L} - x_{k,L-1,L}, s_{kL2} = s_{kL1} - x_{kL1}$, $s_{k,L,L-1} = s_{k,L,L-2} - x_{kL,L-2}$.

On the other hand, when testing in the dominant mode, a contingency table is given by Table 3, and the independence is tested at each $q = 1, 2,\ldots, l$ locus, and $f(\{x_{kij}\}) = 1$ if significance is found at any locus, otherwise $f(\{x_{kij}\}) = 0$. Note that, in this case, $\alpha_{iq} + \alpha_{jq} - \alpha_{iq}\alpha_{jq}$ is equal to 1 if either of the two alleles at locus *q* for ordered diplotype configuration $d_{ij}$ has the minor allele.

The contingency table for the test of independence in the allele frequency mode can be constructed using $\{x_{kij}\}$ and $\{\alpha_{iq}\}$. Thus, Table 4 gives the contingency table for the test of independence in the allele frequency mode. Note that

**Table 3** Contingency table for the dominant mode

| Group | Number of subjects with minor allele | Number of subjects without minor allele |
|---|---|---|
| 1 | $\sum_{i=1}^{L}\sum_{j=1}^{L}(\alpha_{iq} + \alpha_{jq} - \alpha_{iq}\alpha_{jq})x_{1ij}$ | $n_1 - \sum_{i=1}^{L}\sum_{j=1}^{L}(\alpha_{iq} + \alpha_{jq} - \alpha_{iq}\alpha_{jq})x_{1ij}$ |
| 2 | $\sum_{i=1}^{L}\sum_{j=1}^{L}(\alpha_{iq} + \alpha_{jq} - \alpha_{iq}\alpha_{jq})x_{2ij}$ | $n_2 - \sum_{i=1}^{L}\sum_{j=1}^{L}(\alpha_{iq} + \alpha_{jq} - \alpha_{iq}\alpha_{jq})x_{2ij}$ |

The notations in this table are identical to those described for Table 2

$\alpha_{iq} + \alpha_{jq}$ is equal to 0, 1 and 2 when the $q$th locus in the ordered diplotype configuration $d_{ij}$ contains 0, 1 and 2, respectively, copies of the minor allele.

Due to its generality, the present sample space is more useful than the sample space defined for the test of independence only in the allele frequency mode. The multiple comparison problem occurs not only due to the test at multiple loci, but also due to the test by multiple modes. Thus, researchers often test the independence in the allele frequency, dominant and recessive modes at each locus, and the test is considered to be significant if any of the tests in any of the modes is found to be significant. Since the results of the tests in different modes are not independent of each other, the application of Bonferroni's correction leads to conclusions that are too conservative. Using $\{x_{kij}\}$ and $\{\alpha_{iq}\}$, we can calculate the exact type I error rate when the test of independence is performed at multiple linked loci in three different modes. Thus, using $\{x_{kij}\}$ and$\{\alpha_{iq}\}$, the test of independence is performed for each $q$th locus using Tables 2, 3 and 4, and the function $f$ is defined such that $f(\{x_{kij}\})$ is equal to 1 if the test at any locus in any mode is significant, otherwise $f(\{x_{kij}\}) = 0$. We can also perform the test of independence in the genotype mode. Thus, Table 5 is used as a contingency table for genotype mode for each locus $q$. If the test of independence is significant at any locus, $f(\{x_{kij}\}) = 1$, otherwise $f(\{x_{kij}\}) = 0$. The exact calculation of the probability of the block type I error is performed using Eq. (8).

*Calculation of type I error rate for the test in the dominant, recessive, or genotype mode by MCMC under the assumption of a multinomial distribution*

The algorithm of the MCMC method for calculating the approximate probability of the block type I error in dominant, recessive, and genotype modes consists of generating a sample of $\{X_{kij}\}$ ($k$ denotes a group, and $i$ and $j$ denote the orders of the haplotypes) so that each $\{X_{1ij}\}$ and $\{X_{2ij}\}$ follows a multinomial distribution using the Metropolis–Hastings method. In each step, a test of independence was performed for the sample at each locus in any of the three modes. The empirical type I error rate was defined as the proportion of steps in which the test of independence at any of the loci exhibited significance among selected steps. The processes of the procedure are as follows:

1. The state space of this Markov-chain is the set of all different tensors $\{x_{kij}\}$, $k = 1, 2$; $i = 1, 2,\ldots,L$; $j = 1, 2,\ldots,L$, and $x_{kij} \geq 0$ and $\sum_{i=1}^{L}\sum_{j=1}^{L} x_{kij} = n_k$ for all $k$.
2. As an initial state, arbitrary integer values of $x_{kij}$ ($k = 1, 2$; $i = 1, 2,\ldots,L$; $j = 1, 2,\ldots,L$) are given such that $\sum_{i=1}^{L}\sum_{j=1}^{L} x_{kij} = n_k$,and $x_{kij} \geq 0$ 0 for all $k$, $i$, $j$.
3. $k = 1$ or $k = 2$ is selected at random.
4. Two ordered integers $(u, v)$ $(1 \leq u \leq L, 1 \leq v \leq L)$ are selected at random.
5. If $x_{kuv} = 0$, then the state is kept invariant, the step is advanced, the test of independence is performed as described in (10), and the process returns to (3).
6. If $x_{kuv} > 0$, additional two ordered integers $(w, s)$, which are different from $(u, v)$ $(1 \leq w \leq L, 1 \leq s \leq L)$, are selected.
7. New candidates $x^*_{kuv} = x_{kuv} - 1$ and $x^*_{kuv} = x_{kuv} + 1$ are calculated followed by the calculation of $c = \frac{h_w h_s x_{kuv}! x_{kws}!}{h_u h_v x^*_{kuv}! x^*_{kws}!} = \frac{h_w h_s x_{kuv}}{h_u h_v (x_{kws}+1)}$
8. If $c \geq 1$, then $x^*_{kuv}$ and $x^*_{kws}$ are substituted for $x_{kuv}$ and $x_{kws}$, respectively, the step is advanced, the test of independence is performed as in described in (10) and the process returns to (3).
9. If $c < 1$, then $x^*_{kuv}$ and $x^*_{kws}$ are substituted for $x_{kuv}$ and $x_{kws}$, respectively, with probability $c$, else (probability $1 - c$) the state is kept invariant. The step is advanced, the test of independence is performed as described in (10) and the process returns to (3).
10. A test of independence between the phenotype and genotypes at each of the $L$ loci was performed using any one of four contingency tables (Tables 2, 3, 4, 5) obtained from the values of $x_{kij}$ ($k = 1, 2$; $i = 1,$

**Table 4** Contingency table II for the allele frequency mode

| Group | Number of minor alleles | Number of major alleles |
|---|---|---|
| 1 | $\sum_{i=1}^{L}\sum_{j=1}^{L}(\alpha_{iq}+\alpha_{jq})x_{1ij}$ | $2n_1 - \sum_{i=1}^{L}\sum_{j=1}^{L}(\alpha_{iq}+\alpha_{jq})x_{1ij}$ |
| 2 | $\sum_{i=1}^{L}\sum_{j=1}^{L}(\alpha_{iq}+\alpha_{jq})x_{2ij}$ | $2n_2 - \sum_{i=1}^{L}\sum_{j=1}^{L}(\alpha_{iq}+\alpha_{jq})x_{2ij}$ |

The notations in this table are identical to those described for Table 2

**Table 5** Contingency table for the genotype mode

| Genotype[a] | Group 1 | Group 2 |
|---|---|---|
| *m/m* | $\sum_{i=1}^{L}\sum_{j=1}^{L}\alpha_{iq}\alpha_{jq}x_{1ij}$ | $\sum_{i=1}^{L}\sum_{j=1}^{L}\alpha_{iq}\alpha_{jq}x_{2ij}$ |
| *m/M* | $A$[b] | $B$[b] |
| *M/M* | $\sum_{i=1}^{L}\sum_{j=1}^{L}(1-\alpha_{iq})(1-\alpha_{jq})x_{1ij}$ | $\sum_{i=1}^{L}\sum_{j=1}^{L}(1-\alpha_{iq})(1-\alpha_{jq})x_{2ij}$ |

[a] $m$ minor allele, $M$ major allele [b] $A = n_1 - \sum_{i=1}^{L}\sum_{j=1}^{L}\alpha_{iq}\alpha_{jq}x_{1ij} - \sum_{i=1}^{L}\sum_{j=1}^{L}(1-\alpha_{iq})(1-\alpha_{jq})x_{1ij}, B = n_2 - \sum_{i=1}^{L}\sum_{j=1}^{L}\alpha_{iq}\alpha_{jq}x_{2ij} - \sum_{i=1}^{L}\sum_{j=1}^{L}(1-\alpha_{iq})(1-\alpha_{jq})x_{2ij}$ Other notations in this table are identical to those described for Table 2

2,…,$L$; $j = 1, 2,\ldots,L$). If a significant result is obtained by the test for at least one locus, then the block test in that step is considered to be significant.

11. After the MCMC has been run in a sufficient number of steps, the proportion of the steps judged to be significant is considered to be the empirical type I error rate for the block test.
12. Although the above procedure describes the test of independence in any of the three modes (dominant, recessive, and genotype modes), the method can easily be extended to the test of independence in all three modes. Thus, as described in (10), the test of independence is performed using three different contingency tables at each locus, and the block test is considered to be significant if the test at any locus in any mode is found to be significant.

## Calculation of type I error rate in the allele frequency mode when haplotype frequencies are unknown

In general, true haplotype frequencies are unknown. In this case, the above methods cannot be used because the haplotype frequencies $\{h_i\}$ are included in the probability functions for the multinomial distributions. If the haplotype frequencies are unknown, we may consider the probabilities of block type I errors conditional on the observed data. Thus, we may consider the probabilities of block type I errors conditional on the observed marginal numbers of either haplotypes or combinational diplotype configurations. The difference between the combinational and ordered diplotype configurations is that, in the combinational diplotype configuration $d_{ij} = d_{ji}$, whereas in the ordered diplotype configuration $d_{ij} \neq d_{ji}$. Note that observing the ordered diplotype configurations is usually not possible unless family data are available.

Suppose that the numbers of the $i$th haplotype from the first (case) and second (control) groups are random variables $X_{1i}$ and $X_{2i}$, respectively, where $\sum_{i=1}^{L} X_{1i} = 2n_1$, $\sum_{i=1}^{L} X_{2i} = 2n_2$ and $x_{1i}$ and $x_{2i}$ are observed values. We consider the distribution of $X_{1i}$ and $X_{2i}$, conditional on and $X_{1i} + X_{2i} = x_{1i} + x_{2i}$. Then, the joint distribution of the random variables $X_{1i}$ and $X_{2i}$ is known to be multivariate hypergeometric with the following probability function:

$$P\left(\{X_{1i}\} = \{x_{1i}\}, \{X_{2i}\} = \{x_{2i}\} \middle| \left\{\sum\nolimits_{j=1}^{2} X_{ji}\right\} = \{x_i\}\right) = \frac{(2n_1)!(2n_2)!\prod_{i=1}^{L} x_i!}{(2n_1+2n_2)!\prod_{j=1}^{2}\prod_{i=1}^{L} x_{ji}!},$$

where $x_i = x_{1i} + x_{2i}$.

The number of the minor alleles at the $k$th locus for the $j$th group, $Y_{jk}$, can be calculated using $X_{ji}$ and $\alpha_{ik}$, as defined previously in Eq. (4). Using Table 1 as the contingency table, we can perform the test of independence between alleles and the phenotype at each of the $L$ loci, and, if a significant result is obtained from at least one of the loci, then the block test is judged to be significant.

## Calculation of type I error rate in the dominant, recessive, or genotype mode by MCMC when haplotype frequencies are unknown

For the calculation of the type I error rate in dominant, recessive, or genotype mode, when haplotype frequencies are not available, new random variables $Y_{kij}$, ($k = 1, 2$; $i = 1, 2,\ldots,L$; $j \leq i$) are defined as follows:

$$Y_{kij} = \begin{cases} X_{kij} + X_{kji} & \text{if } i \neq j, \\ X_{kij} & \text{if } i = j, \end{cases}$$

Here, $Y_{kij}$ denotes the number of the subjects with the combinational diplotype configuration with the $i$th and $j$th haplotypes in the $k$th group, whereas $X_{kij}$ denotes the number of subjects with the ordered diplotype configuration. We then consider the joint distribution of $\{Y_{kij}\}$ under the constraint of $\sum_{k=1}^{2} Y_{kij} = y_{ij}$ ($i \geq j$), where $Y_{ij} = \sum_{k=1}^{2} Y_{kij}$ denotes the observed value of $Y_{1ij} + Y_{2ij}$.

Under the above condition, the random variable tensor $\{Y_{kij}\}$, $k = 1, 2$; $i = 1, 2,\ldots,L$; $1 \leq j \leq i$ follows the multivariate hypergeometric distribution with the probability function

$$P\left(\{Y_{kij}\} = \{y_{kij}\} \middle| \left\{\sum_{k=1}^{2} Y_{kij}\right\}\{y_{ij}\}\right) = \frac{n_1!n_2!\prod_{i=1}^{L}\prod_{j=1}^{i} y_{ij}!}{(n_1+n_2)!\prod_{k=1}^{2}\prod_{i=1}^{L}\prod_{j=1}^{i} y_{kij}!},$$

where $n_k = \sum_{i=1}^{L}\sum_{j=1}^{i} Y_{kij}$ denotes the number of subjects in the $k$th group. The method used to calculate the exact type I error rates under the assumption of a multivariate hypergeometric distribution is equivalent to the method described under the assumption of a multinomial distribution, as stated previously herein. Next, we describe the method used to calculate the asymptotic type I error rates by the MCMC method, rather than by the exact method, in order to avoid redundancy. The processes of the procedure are as follows:

1. The state space of this Markov-chain is the set of all different $\{Y_{kij}\}$, $k = 1, 2$; $i = 1, 2,\ldots,L$; $1 \leq j \leq i$ under the constraints of $\sum_{k=1}^{2} y_{kij} = y_{ij}$ for any $i$ and $j$, and $\sum_{i=1}^{L}\sum_{j=1}^{i} y_{kij} = n_k$ for any $k$, where $y_{ij}$, $i = 1, 2,\ldots,L$; $1 \leq j \leq i$ are the observed (given) fixed nonnegative integer values.
2. As an initial state, arbitrary integer values of $y_{kij}$ ($k = 1, 2$; $i = 1, 2,\ldots,L$; $1 \leq j \leq i$) are given under

the constraints of $\sum_{k=1}^{2} y_{kij} = y_{ij}$ for any $i$ and $j$, and $\sum_{i=1}^{L}\sum_{j=1}^{i} y_{kij} = n_k$ for any $k$.

3. $k = 1$ or $k = 2$ is selected at random. Let $k' = 1$ if $k = 2$, and $k' = 2$ if $k = 1$.
4. Two ordered integers $(u, v)$ $(1 \le u \le L, 1 \le v \le u)$ are selected at random.
5. If $y_{kuv} = 0$, then the state is kept invariant, the step is advanced, the test of independence is performed as described in (10), and the process returns to (3).
6. If $y_{kuv} > 0$, then two additional ordered integers $(w, s)$ that are different from $(u, v)$ $(1 \le w \le L, 1 \le s \le w)$ are selected. If $y_{k'ws} = 0$, then the state is kept invariant, the step is advanced, the test of independence is performed as described in (10), and the process returns to (3).
7. New candidates $y^{*}_{kuv} = y_{kuv} - 1$, $y^{*}_{kws} = y_{kws} + 1$, $y^{*}_{k'uv} = y_{k'uv} + 1$, and $y^{*}_{k'ws} = y_{k'ws} - 1$ are calculated, and the following calculation is performed: $c = \frac{y_{kuv}!y_{kws}!y_{k'uv}!y_{k'ws}!}{y^{*}_{kuv}!y^{*}_{kws}!y^{*}_{k'uv}!y^{*}_{k'ws}!} = \frac{y_{kuv}y_{k'ws}}{(y_{k'uv}+1)(y_{kws}+1)}$.
8. If $c \ge 1$, then $y^{*}_{kuv}$, $y^{*}_{kws}$, $y^{*}_{k'uv}$, and $y^{*}_{k'ws}$ are substituted for $y_{kuv}$, $y_{kws}$, $y_{k'uv}$, and $y_{k'ws}$, respectively, the step is advanced, the test of independence is performed as described in (10) and the process returns to (3).
9. If $c < 1$, then $y^{*}_{kuv}$, $y^{*}_{kws}$, $y^{*}_{k'uv}$, and $y^{*}_{k'ws}$ are substituted for $y_{kuv}$, $y_{kws}$, $y_{k'uv}$, and $y_{k'ws}$, respectively, with probability $c$, else (probability $1 - c$) the state is kept invariant, the step is advanced, the test of independence is performed as described in (10), and the process returns to (3).
10. A test of independence between the phenotype and genotypes at each of the $L$ loci was performed using any of the three contingency tables, depending on the mode of test. Thus, for the test in the recessive mode, Table 2 is used, whereas for the dominant mode, Table 3 is used, after substituting $y_{1ij}$ and $y_{2ij}$ for $x_{1ij}$ and $x_{2ij}$, respectively. For the tests in the allele frequency mode and the genotype mode, Tables 4 and 5 are used, respectively, after substituting $y_{1ij}$ and $y_{2ij}$ for $x_{1ij}$ and $x_{2ij}$, respectively. If a significant result is obtained by the test for at least one locus, then the block test in that step is considered to be significant.
11. After the MCMC has been run in a sufficient number of steps, the proportion of the steps judged to be significant is considered to be the empirical type I error rate for the block test.
12. Although the above procedure describes the test of independence in any of the three modes (dominant, recessive, and genotype modes), the method can easily be extended to the test in all three modes. Thus, as described in (10), the test of independence is performed using three different contingency tables at each locus, and the block test is considered to be significant if the test at any locus in any mode is found to be significant.

### Method for calculating the appropriate significance level at each locus in order to achieve a desirable block significance level

In the above-mentioned methods, the probability of global type I error given the significance level (type I error rate) at each SNP locus is calculated. However, in several cases, the appropriate significance level must be calculated at each locus in order to achieve a desirable global significance level (e.g., $P = 0.05$). This is achieved by the following procedure:

1. The significance level for the test at each SNP locus is increased, for example, from $P_{\text{locus}} = 0.01$ to $P_{\text{locus}} = 0.05$ by an increment of 0.005, and the global type I error rate $P_{\text{block}}$ is calculated at each $P_{\text{locus}}$ value using either the exact method or the MCMC method.
2. Since SNPs of different haplotype blocks are at linkage equilibrium, the global type I error $P_{\text{global}}$ is obtained from $P_{\text{block}}$ of all haplotype blocks by using Bonferroni's correction. Then, the appropriate $P_{\text{locus}}$ value is calculated in order to achieve the desired $P_{\text{global}}$ value.

## Results

### Calculation of corrected block type I error rates using simulated data

Calculation by the exact method is expected to take an enormous amount of time, even when the numbers of subjects in the groups are rather small and the test is performed in the allele frequency mode. For example, when the numbers of subjects in each of the two groups are $n_1 = n_2 = 15$, the number of different values of $\{x_{ji}\}$ in Eq. (6) increases with increasing numbers of haplotypes $L$. Thus, when $L = 2$, 3, 4, and 5, the numbers of different $\{x_{ji}\}$ in Eq. (6) will be 961, 246,016, 29,767,936, and 2,150,733,376, respectively.

First, we used simulated data for the calculation of block type I error rates. The data for nine linked SNP loci were considered. Thus, the possible number of haplotypes is $L = 2^9 = 512$. However, we assumed that only five haplotypes, ATAATTTAC, ACGGCCGGT, GTAATTTAT, ATAATTTAT, and GTAATTTAC, are present in the population in the frequencies of 0.5252, 0.1902, 0.1776, 0.0420, and 0.0650, respectively. The frequencies of the other 507 haplotypes were assumed to be zero. This is often the case for the real data. Thus, in this case, the real $L$ is

512, but the effective $L$ is 5. Even if $n_1 = n_2 = 50$ and $L = 5$, approximately 4 days are required to calculate the exact probability with the allele frequency model by a parallel computation with 128 CPU PC cluster (Dell PowerEdge 1750, CPU Intel Xeon 3.06 GHz) (Fig. 1). Therefore, the calculation of the exact probabilities for actual study designs is not realistic. The time required for the MCMC method is much shorter than that for the exact method. However, whether the MCMC method gives acceptable approximations under real conditions has yet to be clarified. We therefore compared the block type I error rates calculated by the exact method and the MCMC method.

Figure 2 shows an example of such a comparison, in which the test was in the allele frequency mode. The significance level for each locus was $P = 0.01$, and the numbers of subjects in the two groups were varied from 1 to 50. Figure 2 clearly demonstrates that the type I error rates calculated by the MCMC method are almost identical to those calculated by the exact method as long as the number of subjects in each group is not larger than 50. When $n_1 = n_2 > 16$, the block type I error rates corrected by both the exact method and the MCMC method were found to be between 0.025 and 0.03 (Fig. 2). Because the number of SNPs in this case is nine, Bonferroni's correction yields a block type I error rate of $P = 0.09$ at any number of subjects (Fig. 2). Thus, in this particular situation, the proposed method was approximately one-third as conservative as Bonferroni correction.

Figure 2 also shows the block type I error rates calculated assuming a hypergeometric distribution for the test in the allele frequency mode. The calculation by this method appears to be unstable for small sample sizes. This is probably because the number of haplotypes in each group does not accurately reflect the population frequency. Since the samples were made by drawing haplotypes from the population, the proportions of the haplotypes should be unstable when the sample size is small.

We then examined whether the MCMC method gives results in accord with those generated by the exact method when the test was performed in dominant and recessive modes. In this experiment, the number of the subjects in each of the two groups was 17.

Figure 3 indicates that the MCMC accords with the exact calculation for dominant and recessive modes (at least when $n_1 = n_2 \leq 17$). Although the block type I error rates obtained by the correction using the exact method and the MCMC method fluctuate with the numbers of subjects, both methods concurred very well for any number of subjects for the test in any mode (Fig. 3). The approximate method also gave almost the same results even when $n_1$ and $n_2$ are large (data not shown). Although the corrected block type I error rates for the test in the dominant mode were higher than those for the test in the recessive mode, this difference disappeared when the number of subjects was higher (e.g., $n_1 = n_2 = 100$) when the calculation was

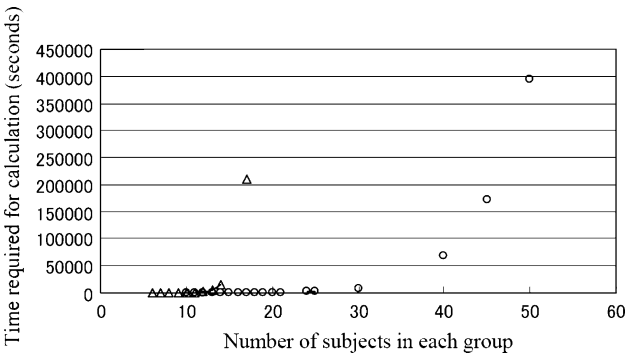


**Fig. 1** Time in seconds required for calculation of the type I error rate by the exact method for the tests in allele frequency (*open circle*) and recessive (*open triangle*) modes. The number of subjects in each group ($n_1 = n_2$) was varied up to 50. The calculation was performed by a parallel computation using a 128-CPU Linux Cluster System. The results in Figs. 2, 3, 4, and 5 were obtained using the same system

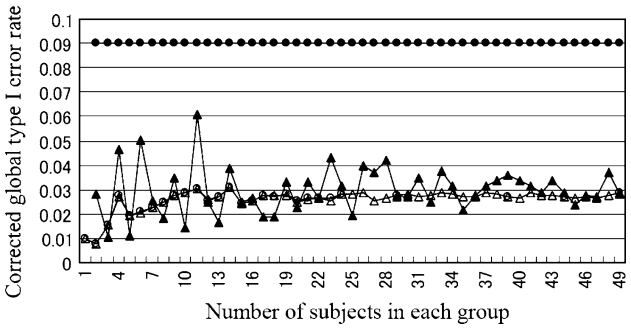


**Fig. 2** Comparison of the block type I error rates corrected by the exact method, MCMC methods, and Bonferroni's method. The data of the population haplotype frequencies were simulated as described in "Methods." The simulated data for this experiment were composed of nine linked SNP loci. The haplotype frequencies were ATAATTTAC, ACGGCCGGT, GTAATTTAT, ATAATTTAT, and GTAATTTAC at the population frequencies of 0.5252, 0.1902, 0.1776, 0.0420, and 0.0650, respectively. Based on the population haplotype frequencies, the block type I error rate was calculated for the test in the allele frequency mode by the exact method (*open circle*) and the MCMC method (*open triangle*), as described in "Methods." The significance level for each locus ($P_{\text{locus}}$) was $P = 0.01$, and the numbers of subjects in the two groups were varied from 1 to 50. For the calculation of the type I error rates assuming a hypergeometric distribution (haplotype frequencies unknown), the haplotypes in both groups (cases and controls) were drawn according to the above haplotype frequencies. Based on the combinational diplotype configurations in the subjects of the two groups, the type I error rate (*closed triangle*) was calculated according to the MCMC method described in "Methods." Since the number of loci was nine, the block type I error rate corrected by Bonferroni's method (*closed circle*) is always 0.09. The computer system used to generate the data in the following figures was the same as that used to generate the data in Fig. 1

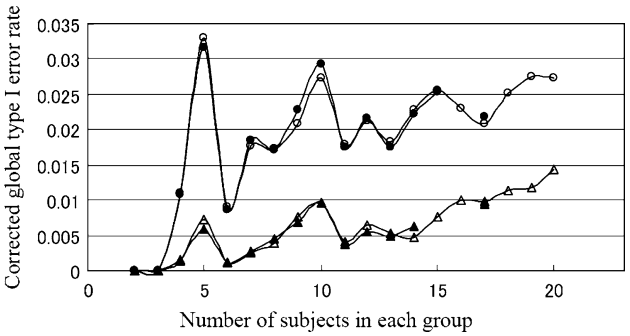


**Fig. 3** Comparison of the block type I error rates corrected by the exact method and the MCMC method for the test in the dominant and the recessive modes. The haplotype frequencies used in the experiment were identical to those in Fig. 2. The calculation of the block type I error rates for the test in the dominant and recessive modes was performed as described in "Methods". The number of subjects in each of the two groups ($n_1 = n_2$) was varied from 2 to 20. The lines in the figure show the results for (**a**) the dominant mode and the MCMC method (*open circles*), (**b**) the dominant mode and the exact method (*closed circles*), (**c**) the recessive mode and the MCMC method (*open triangles*), and (**d**) the recessive mode and the exact method (*closed triangles*)

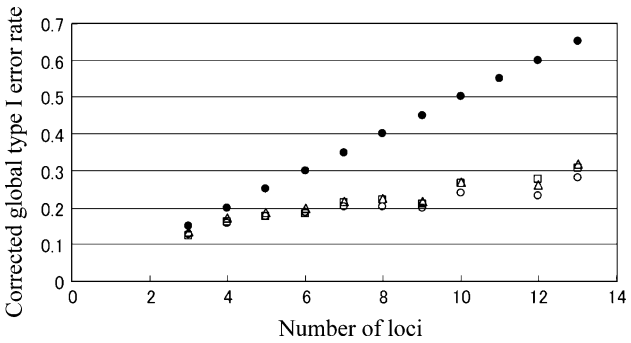


**Fig. 4** Comparison of the block type I error rates corrected by the MCMC method for the tests in the allele frequency, dominant, and recessive modes using the real data. The data from 1,145 SNPs in 249 haplotype blocks on chromosome 8 were used. The genotype data from each block from 752 control subjects were used for the inference of population haplotype frequencies. Using the haplotype frequencies for each block thus obtained, we calculated the corrected block type I error rate for the test in the allele frequency, dominant or recessive mode by the MCMC method, as described in "Methods," with a fixed type I error rate for each SNP ($P_{\text{locus}}$) of 0.01. *Each plot* indicates the average of the corrected type I error rate for the test in the allele frequency (*open circles*), dominant (*open triangles*), or recessive (*open squares*) mode. Note that the corrected block type I error rate is always 0.01 multiplied by the number of loci when the correction is made using Bonferroni's method (*closed circles*)

performed by the MCMC method. Note that the exact method cannot handle a large number of subjects.

## Calculation of the corrected block type I error rates using real data

Next, the proposed methods were applied to real data. The data were obtained by an SNP genotyping study targeted at >200 drug-related genes. DNA from 752 Japanese volunteers was used for SNP genotyping. The methods for genotyping, data processing, haplotype-block construction, and haplotype inference are described in our previous paper (Kamatani et al. 2004). For the present experiment, only the data for 1,145 SNPs in 249 haplotype blocks on chromosome 8 were used. There were 3–13 SNPs in each block, and the block type I error rates were calculated separately for each block. Figure 4 shows the corrected block type I error rates for the tests in the allele frequency, dominant, and recessive modes plotted against the number of SNP loci, compared to Bonferroni's correction. This figure demonstrates that the difference between the block type I error rates obtained by the correction using the MCMC method and Bonferroni's method increases with the number of SNP loci (Fig. 4). Thus, using the Bonferroni's correction, the corrected block type I error rate increases linearly. The rate of increase by the proposed method is much lower. In addition, the averaged block type I error rates for the tests in the allele frequency, dominant and recessive modes did not differ greatly for the same number of loci.

The correction for the multiple comparison is likely to be more efficient when the linkage disequilibrium between the loci involved is stronger. As a measure of the strength of the linkage disequilibrium for multiple loci, we used haplotype heterozygosity. Thus, we examined whether there was any positive correlation between the haplotype heterozygosity and the corrected block type I error rate. We collected the blocks having the same numbers of SNPs and tested the correlation between the haplotype heterozygosity and the corrected block type I error rate for each SNP number. The correlation coefficients ($r$) between the haplotype heterozygosity and the corrected block type I error when $P_{\text{locus}}$ was set to 0.01 were 0.3383181 ($P = 0.1058716$), 0.6603017 ($P = 8.637 \times 10^{-11}$), 0.4102507 ($P = 0.0012505$), 0.6779765 ($P = 0.0010194$), 0.69114404 ($P = 0.00185095$), 0.6291491 ($P = 0.069478$), 0.9846587 ($P = 0.0022758$), and 0.0334933 ($P = 0.9573629$) for the blocks with 3, 4, 5, 6, 7, 8, 9, and 10 SNPs, respectively. The $P$ values in parentheses indicate those for the test of the null hypothesis $r = 0$. Thus, in many cases, there was a positive correlation between the haplotype heterozygosity and the corrected block type I error rate. Figure 5 shows the relationship between the haplotype heterozygosity and the corrected block type I error rate when the number of SNPs was four. These data indicate that the corrected block type I error rate is higher when the linkage disequilibrium between the SNP loci is low (haplotype heterozygosity is high). This means that the problem of the correction for multiple comparisons

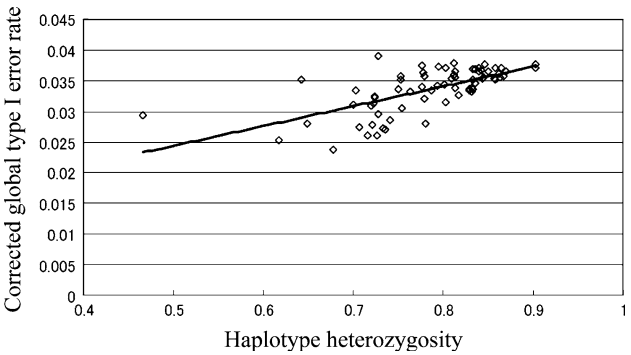


**Fig. 5** Relationship between the haplotype heterozygosity and the corrected block type I error rate when the number of SNPs is four. The corrected block type I error rates for the test were calculated in the allele frequency mode by the MCMC method for blocks with four SNPs. The regression line is shown, and the Pearson's correlation coefficient $r$ was 0.66 ($P = 8.637 \times 10^{-11}$)

using Bonferroni's method becomes more problematic as the linkage disequilibrium between separate loci becomes stronger.

## Discussion

A multiple comparison problem emerges when the association between phenotypes and multilocus genotypes is examined. Since Bonferroni's correction gives conclusions that are too conservative, several researchers have proposed better corrections for the tests on the multilocus data.

Becker and Knapp (2004) proposed a strategy to account for multiple testing in the context of haplotype analysis. They proposed a method in which the smallest raw $P$ value was used. On the other hand, Nyholt (2004) described a simple correction method for multiple testing for SNPs in linkage disequilibrium with each other. In Nyhot's method, correction for multiple testing is performed on the basis of the spectral decomposition of matrices (Cheverud 2001) of pairwise LD between SNPs. His method provides a simple and useful alternative to more computationally intensive permutation tests. In addition, Meng et al. (2003) have proposed a method based on the spectral decomposition of matrices of pairwise LD between SNPs. Recently, Moskvina and Schmidt (2008) proposed a correction method for multiple testing of genome-wide association studies on the basis of the spectral decomposition of matrices. The spectral decomposition method utilizes the fact that the individual marker association tests are not statistically independent, but dependent to a degree which can be measured in terms of the pairwise haplotypic correlation between markers observed in the empirical data. The permutation method is also used for correction for multiple testing of the genome-wide association studies (Kimmel and Shamir 2006; Sladek et al. 2007). In addition, Sabatti et al. (2003) proposed the application of the concept of the false-discovery rate (FDR) (Benjamini and Hochberg 1995) to the multilocus association studies for the purpose of the correction of multiple comparisons. The method proposed in this study differs from all of these previous methods. The permutation method and the spectral decomposition method are known to provide much less conservative results than Bonferroni correction. However, it is still unclear how conservative these methods are. Further studies are necessary to compare the probability of type I errors and type II errors of these methods to that of the exact calculation shown here.

We have shown that the exact type I error rates in the allele frequency, dominant and recessive modes can be calculated assuming multinomial distributions when the population haplotype frequencies are known. However, calculation of such exact rates requires a long computational time, and using this method for real data is impractical. For example, it would require approximately 70 years to calculate the exact type I error rate in the allele frequency mode for $n_1 = n_2 = 1{,}000$ and $L = 19$ even using one of the highest speed computers in the world (12 TFLOPS). To overcome this problem, the approximate calculation method was constructed using a MCMC sampler of a multinomial distribution. Thus, we presented methods for calculating the asymptotic type I error rates in the allele frequency, dominant, recessive, and genotype modes when the population haplotype frequencies are given.

However, the population haplotype frequencies are not always known. We have proposed methods to calculate the type I error rates in such cases. Thus, hypergeometric distributions were assumed given fixed marginal frequencies of haplotypes and diplotype configurations. The usefulness of these methods was not fully examined in the present study and so should be confirmed in future studies.

It is generally difficult to assess how many steps are necessary for convergence of MCMC chains. In this study, convergence was monitored by comparing several independent runs. Namely, several MCMC runs were conducted with the starting values of the gene frequencies being randomly set. Proportions of the cases type I errors of all runs were calculated for every million steps. When proportions of the cases type I errors of all runs became similar to each other, the MCMC chains were considered to be converged. In the case of Figs. 1 and 2, 100 million steps are enough to reach the steady states of the MCMC chains. Thus, we set the length of the burn-in as 100 million steps so that first 100 million steps were thrown away. Then we used 100 million steps after burn-in for calculating the proportions of the cases type I errors. It took a few minutes for the calculation of 200 million steps of the MCMC method.

The methods proposed herein were applied to both simulated and real data. Using the simulated data, we found that the exact block type I errors can be calculated at least in the allele frequency mode when the number of subjects in each group does not exceed 50. On the other hand, the proposed MCMC methods could be used to calculate the block type I error rates in any mode for any number of subjects in any group. The block type I error rates calculated by the exact methods and the equivalent MCMC methods were in good agreement, indicating that the MCMC methods give good approximations. One can conduct computer simulations by generating random numbers that follow the multinomial or the hypergeometric distributions to obtain the type I errors of tests. The algorithm is simpler than the algorithms generating random numbers that follow the multinomial or the hypergeometric distributions, especially when the number of terms is large (Press et al. 1999).

Using real data, in which 1,145 SNPs on chromosome 8 were partitioned into 249 haplotype blocks, we found that the block type I error rates calculated by the proposed methods increased with the number of SNP loci. However, the rate of increase was not as high as when the correction was performed by Bonferroni's method. The block type I error rates obtained by the proposed methods were half as high as those obtained by the correction by Bonferroni's method, which suggests that the application of Bonferroni's method yields corrections that are too conservative when the association is tested in case–control studies. The problem of corrections that are too conservative becomes more remarkable as the linkage disequilibrium between the SNP loci becomes stronger.

Since haplotype frequencies are expected to become available for each ethnic group, the use of the proposed correction methods is likely to be practical. The proposed methods should be tested for validity, along with other correction procedures, using large amounts of data.

**Acknowledgments** We thank two anonymous reviewers who gave us helpful comments and suggestions. The present study was supported in part by grants from the Research Project for Personalized Medicine (MEXT). This work was supported by the National Project on 'Next-generation Integrated Living Matter Simulation' of the Ministry of Education, Culture, Sports, Science and Technology (MEXT). All programs used in this study are available from the authors on request.

## Appendix

We developed a fast approximation algorithm to calculate the type I error when the haplotype frequencies are known. The notations in appendix are identical to those described in the text. The basic idea of this algorithm is to skip $x_{ji}$ ($j = 1,2$; $i = 1,\ldots, L$), whose probability of occurrence is very small. In this algorithm, we obtain $Q = 1$—(the probability of type I error). We explain this algorithm by using the genotype mode. $x_{ji}$ follows the binomial distribution; the mean of $x_{ji}$, $E[x_{ji}]$, is $2n_jh_i$, and the variance of $x_{ji}$, $V[x_{ji}]$, is $2n_jh_i(1 - h_i)$. We define $M_{ji}(t)$ and $m_{ji}(t)$ as $M_{ji}(t) = E[x_{ij}] + t\sqrt{V[x_{ji}]} = 2n_jh_i + t\sqrt{2n_jh_i(1-h_i)}$ and $m_{ji}(t) = E[x_{ji}] - t\sqrt{V[x_{ji}]} = 2n_ih_j - t\sqrt{2n_jh_i(1-h_i)}$, where $t > 0$. When $t$ is large, the binomial distribution is approximately equal to the normal distribution, whose mean and variance are $E[x_{ji}]$ and $V[x_{ji}]$, respectively, so that the probability of $x_{ji} > M_{ji}(t)$ or $x_{ji} < m_{ji}(t)$ is as large as 1-erf$(\frac{t}{\sqrt{2}})$, where erf is the error function. When $t = 4$, the probability that $x_{ji} > M_{ji}(t)$ or $x_{ji} > M_{ji}(t)$ is smaller than $10^{-7}$. Thus, the terms with $x_{ij}$ can be ignored when $t$ is large enough and $x_{ji} > M_{ji}(t)$ or $x_{ji} < m_{ji}(t)$.

For the sake of convenience, we define define $N_{ji}$ and $n_{ji}$ by because $0 \le x_{ji} \le 2n_j$ as

$$N_{ji} = \begin{cases} M_{ji}(t), & \text{when } M_{ji}(t) < 2n_j, \\ 2n_j, & \text{otherwize} \end{cases}$$

$$n_{ji} = \begin{cases} m_{ji}(t), & \text{when } m_{ji}(t) < 2n_i, \\ 0, & \text{otherwise} \end{cases}$$

Since $\sum_{i=1}^{L} x_{1i}$ must be $2n_1$ and $\sum_{i=1}^{L} x_{2i}$ must be $2n_2$, when we calculate $Q$, a new function $g$ is defined as

$$g(x_{11}, x_{12}, \ldots, x_{1L}, x_{11}, x_{12}, \ldots, x_{1L}) = \begin{cases} 1 & \text{when } \sum_{i=1}^{L} x_{1i} = 2n_1 \text{ and } \sum_{i=1}^{L} x_{2i} = 2n_2, \\ 0, & \text{otherwise.} \end{cases}$$

Then, $Q$ is obtained as

$$Q \approx \sum_{x_{11}=n_{11}}^{N_{11}} \sum_{x_{12}=n_{12}}^{N_{12}} \sum_{x_{13}=n_{13}}^{N_{13}} \cdots \sum_{x_{1L}=n_{1L}}^{N_{1L}} \sum_{x_{21}=n_{21}}^{N_{21}} \sum_{x_{22}=n_{22}}^{N_{22}} \sum_{x_{23}=n_{23}}^{N_{23}} \cdots \sum_{x_{2L}=n_{2L}}^{N_{2L}} g(x_{11}, x_{12}, \ldots, x_{1L-1}, x_{21}, x_{22}, \ldots, x_{2,L-1}) \times \bar{f}(x_{11}, x_{12}, \ldots, x_{1L-1}, x_{21}, x_{22}, \ldots, x_{2,L-1}) \frac{(2n_1)!(2n_2)!}{\prod_{i=1}^{L}\prod_{j=1}^{2} x_{ji}!} \prod_{i=1}^{L}\prod_{j=1}^{2} h_i^{x_{ji}},$$

where $\bar{f}(x_{11}, x_{12},\ldots, x_{1L-1}, x_{21}, x_{22},\ldots, x_{2,L-1}) = 1$ when $f(x_{11}, x_{12},\ldots, x_{1L-1}, x_{21}, x_{22},\ldots, x_{2,L-1}) = 0$ and $\bar{f}(x_{11}, x_{12},\ldots, x_{1L-1}, x_{21}, x_{22},\ldots, x_{2,L-1}) = 0$ when $f(x_{11}, x_{12},\ldots, x_{1L-1}, x_{21}, x_{22},\ldots, x_{2,L-1}) = 1$. In the same way, we can calculate the probability of the dominant and recessive modes.

In this algorithm, the possible number of different $\{x_{ji}\}$ on the right-hand side of Eq. (6) is reduced to

$$(2t)^L (2n_1)^{\frac{L}{2}} (2n_2)^{\frac{L}{2}} \prod_{j=1}^{L} \sqrt{h_j(1-h_j)}.$$

Since $h_i(1 - h_i) \le 0.5$ for $i = 1, \ldots, L$, the computational complexity of this algorithm is as large as $O\left[t^L (n_1 n_2)^{\frac{L}{2}}\right]$. Our numerical experiments showed that the

approximate values obtained by this algorithm are good enough when $t = 4$. Under the condition where $\sum_{i=1}^{L} x_{1i} = 2n_1$ and $\sum_{i=1}^{L} x_{2i} = 2n_2$, the actual number of different $\{x_{ji}\}$ is smaller than the value given by the above formula.

## References


Becker T, Knapp M (2004) A powerful strategy to account for multiple testing in the context of haplotype analysis. Am J Hum Genet 75:561–570

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Statist Soc 57:289–300

Cardon LR, Bell JI (2001) Association study designs for complex diseases. Nat Rev Genet 2:91–99

Cheverud JM (2001) A simple correction for multiple comparisons in interval mapping genome scans. Heredity 87:52–58

Cutler DJ, Zwick ME, Carrasquillo MM, Yohn CT, Tobin KP, Kashuk C, Mathews DJ, Shah NA, Eichler EE, Warrington JA, Chakravarti A (2001) High-throughput variation detection and genotyping using microarrays. Genome Res 11:1913–1925

Jurinke C, Oeth P, van den Boom D (2004) MALDI-TOF mass spectrometry: a versatile tool for high-performance DNA analysis. Mol Biotechnol 26:147–164

Kamatani N, Sekine A, Kitamoto T, Iida A, Saito S, Kogame A, Inoue E, Kawamoto M, Harigai M, Nakamura Y (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

Kanazawa A, Tsukada S, Sekine A, Tsunoda T, Takahashi A, Kashiwagi A, Tanaka Y, Babazono T, Matsuda M, Kaku K, Iwamoto Y, Kawamori R, Kikkawa R, Nakamura Y, Maeda S (2004) Association of the gene encoding wingless-type mammary tumor virus integration-site family member 5B (WNT5B) with type 2 diabetes. Am J Hum Genet 75:832–843

Kimmel G, Shamir R (2006) A fast method for computing high-significance disease association in large population-based studies. Am J Hum Genet 79:481–492

Kizawa H, Kou I, Iida A, Sudo A, Miyamoto Y, Fukuda A, Mabuchi A, Kotani A, Kawakami A, Yamamoto S, Uchida A, Nakamura K, Notoya K, Nakamura Y, Ikegawa S (2005) An aspartic acid repeat polymorphism in asporin inhibits chondrogenesis and increases susceptibility to osteoarthritis. Nat Genet 37:138–144

Lander E, Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11:241–247

Meng Z, Zaykin DV, Xu CF, Wagner M, Ehm MG (2003) Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes. Am J Hum Genet 73:115–130

Moskvina V, Schmidt KM (2008) On multiple-testing correction in genome-wide association studies. Genet Epidemiol (in press)

Nyholt DR (2004) A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. Am J Hum Genet 74:765–769

Ohnishi Y, Tanaka T, Ozaki K, Yamada R, Suzuki H, Nakamura Y (2001) A high-throughput SNP typing system for genome-wide association studies. J Hum Genet 46:471–477

Oliphant A, Barker DL, Stuelpnagel JR, Chee MS (2002) BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping. Biotechniques Suppl:56–8, 60–61

Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, Tsunoda T, Sato H, Hori M, Nakamura Y, Tanaka T (2002) Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. Nat Genet 32:650–654

Ozaki K, Inoue K, Sato H, Iida A, Ohnishi Y, Sekine A, Odashiro K, Nobuyoshi M, Hori M, Nakamura Y, Tanaka T (2004) Functional variation in LGALS2 confers risk of myocardial infarction and regulates lymphotoxin-alpha secretion in vitro. Nature 429:72–75

Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1999) Numerical recipes in C, The art of scientific computing, 2nd edn. Cambridge University Press, Cambridge

Rickert AM, Borodina TA, Kuhn EJ, Lehrach H, Sperling S (2004) Refinement of single-nucleotide polymorphism genotyping methods on human genomic DNA: amplifluor allele-specific polymerase chain reaction versus ligation detection reaction-TaqMan. Anal Biochem 330:288–297

Sabatti C, Service S, Freimer N (2003) False discovery rate in linkage and association genome screens for complex disorders. Genetics 164:829–833

Seaman SR, Muller-Myhsok B (2005) Rapid simulation of *P* values for product methods and multiple-testing adjustment in association studies. Am J Hum Genet 76:399–408

Shibata K, Ito T, Kitamura Y, Iwasaki N, Tanaka H, Kamatani N (2004) Simultaneous estimation of haplotype frequencies and quantitative trait parameters: applications to the test of association between phenotype and diplotype configuration. Genetics 168:525–539

Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S, Balkau B, Heude B, Charpentier G, Hudson TJ, Montpetit A, Pshezhetsky AV, Prentki M, Posner BI, Balding DJ, Meyre D, Polychronakos C, Froguel P (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 445:881–885

Sorensen D, Gianola D (2002) Likelihood, Bayesian, and MCMC methods in quantitative genetics. Springer, New York

Suzuki A, Yamada R, Chang X, Tokuhiro S, Sawada T, Suzuki M, Nagasaki M, Nakayama-Hamada M, Kawaida R, Ono M, Ohtsuki M, Furukawa H, Yoshino S, Yukioka M, Tohma S, Matsubara T, Wakitani S, Teshima R, Nishioka Y, Sekine A, Iida A, Takahashi A, Tsunoda T, Nakamura Y, Yamamoto K (2003) Functional haplotypes of PADI4, encoding citrullinating enzyme peptidylarginine deiminase 4, are associated with rheumatoid arthritis. Nat Genet 34:395–402

Thakkinstian A, D'Este C, Eisman J, Nguyen T, Attia J (2004) Meta-analysis of molecular association studies: vitamin D receptor gene polymorphisms and BMD as a case study. J Bone Miner Res 19:419–428

Tokuhiro S, Yamada R, Chang X, Suzuki A, Kochi Y, Sawada T, Suzuki M, Nagasaki M, Ohtsuki M, Ono M, Furukawa H, Nagashima M, Yoshino S, Mabuchi A, Sekine A, Saito S, Takahashi A, Tsunoda T, Nakamura Y, Yamamoto K (2003) An intronic SNP in a RUNX1 binding site of SLC22A4, encoding an organic cation transporter, is associated with rheumatoid arthritis. Nat Genet 35:341–348