Journal of Human Genetics (2009) 54, 169–173
© 2009 The Japan Society of Human Genetics All rights reserved 1434-5161/09 $32.00
www.nature.com/jhg

ORIGINAL ARTICLE

A practical case–control association test for detecting a susceptibility allele at a copy number variation locus

Jun Ohashi

The association between a copy number variant (CNV) and susceptibility to disease has drawn much attention. In this study, a case–control association test for a CNV locus with multiple alleles is proposed for detecting a single CNV allele associated with a disease. In the association test, CNV allele frequencies are estimated for cases and controls separately using an expectation-maximization (EM) algorithm, and the χ^2 values are calculated for each CNV allele to compare the estimated frequency between them. A permutation procedure is used to obtain an empirical *P*-value for each CNV allele and for controlling a global type I error rate. The statistical power of the present association test was evaluated by a computer simulation analysis with several parameter settings. The results revealed that the statistical power was markedly different among CNV alleles with different copy numbers, and a higher power could be achieved for a susceptible allele with the lowest or highest copy number in comparison with those with intermediate copy numbers.

Journal of Human Genetics (2009) **54,** 169–173; doi:10.1038/jhg.2009.8; published online 6 February 2009

Keywords: case–control association study; copy number variant; expectation-maximization algorithm; permutation test

INTRODUCTION

A copy number variant (CNV), where there is a copy number change involving a DNA fragment 1 kb or larger in size,[1] is increasingly recognized as one of the genetic factors producing human phenotypic variations, including disease susceptibility. For example, a CNV including the *FCGR3* gene has been shown to be associated with several types of autoimmune disorders, such as systemic lupus erythematosus, microscopic polyangiitis and Wegener's granulomatosis.[2] In infectious diseases, a CNV of the gene encoding CCL3L1 influences human immunodeficiency virus/acquired immunodeficiency syndrome susceptibility.[3]

Most of the reported associations between CNVs and common disorders come from an alteration of gene dosage, where a positive or negative correlation is observed between the number of CNV and the disease susceptibility.[4] A diploid copy number at a CNV locus (that is, a total copy number per diploid genome in a single individual) can be experimentally determined for an individual using various methods, including high-throughput platforms based on DNA microarray technology. Therefore, from a statistical point of view, it is easy to detect such associations resulting from gene dosage if diploid copy number at a CNV locus is treated as a quantitative variable in suitable statistical analyses (for example, a logistic regression analysis). However, there may be a case that an allele with a specific copy number shows an association with a disease directly or indirectly. For example, one of the copies on an allele may have a gain-of-function mutation increasing the risk of a disease. In addition, an allele with a specific copy number may have only functional copies or genes and the other alleles may have only disrupted or non-functional genes. Furthermore, an allele with a certain copy number may be in strong linkage disequilibrium with a nearby disease variant, and thus serve as a genetic marker. To detect such a CNV allele, the frequency of each CNV allele needs to be compared between cases and controls. In this study, a case–control association test for a CNV locus with multiple alleles was used for detecting a single CNV allele associated with a disease. The computer simulation analysis revealed that the present association test controlled a global type I error rate and the statistical power varied among CNV alleles with different copy numbers.

MATERIALS AND METHODS

EM algorithm for estimating allele frequencies

The expectation-maximization (EM) algorithm is used for estimation of allele frequencies at a CNV locus. The equation below is basically the same EM algorithm reported by Kato *et al.*,[5,6] although the description is different. It is assumed that only the diploid copy number, for an individual, can be observed as a phenotype denoted by an integer and the genotype is unknown unless the diploid copy number is 0 or 1 (the genotypes of individual with diploid copy number of 0 and 1 are 0/0 and 0/1, respectively). In a set of N individuals, the observed count of individuals with a diploid copy number of k (denoted as N_k) can be considered as the sum of multiple unobserved counts, that is the numbers of individuals with the allele combination of i copies and $k-i$ copies (denoted as $N_{i,\ k-i}$), where $0 \leqslant i \leqslant k$. At the expectation step, under the

Doctoral Program in Life System Medical Sciences, Graduate School of Comprehensive Human Sciences, University of Tsukuba, Ibaraki, Japan
Correspondence: Dr J Ohashi, Doctoral Program in Life System Medical Sciences, Graduate School of Comprehensive Human Sciences, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8575, Japan.
E-mail: juno-tky@umin.ac.jp

Received 15 October 2008; revised 8 January 2009; accepted 11 January 2009; published online 6 February 2009

assumption of the Hardy–Weinberg equilibrium, the unobserved count can be estimated by the following formula:

$$N_{i,k-i} = \begin{cases} \dfrac{2a_i a_{k-i} N_k}{\sum_{j=0}^{k} a_j a_{k-j}} & \text{for } i \neq k-i \\ \dfrac{a_i^2 N_k}{\sum_{j=0}^{k} a_j a_{k-j}} & \text{for } i = k-i \end{cases}$$

where a_j denotes the population frequency of allele with j copies. At the maximization step, the population frequencies of alleles with i copies are revised as

$$a_i = \begin{cases} \frac{N_{i,i}}{2N} + \sum_{k=i}^{H} \frac{N_{i,k-i}}{2N} & \text{for } 2i \leqslant H \\ \sum_{k=i}^{H} \frac{N_{i,k-i}}{2N} & \text{for } 2i > H \end{cases}$$

where H denotes the possible maximum number of copies at the CNV locus (that is, H is equal to the maximum value of the diploid copy number observed in a set of N individuals) and thus $\sum_{k=0}^{H} N_k = N$ and $\sum_{i=0}^{H} a_i = 1$. It is noted here that there may be an allele with H copies when individuals with a diploid copy number

Table 1 2×2 table for calculation of the χ^2 value for an allele with *i* copies

Allele	*Case*	*Control*
i	$2Nx_i$	$2My_i$
Non-*i*	$2N\,(1-x_i)$	$2M\,(1-y_i)$
Total	$2N$	$2M$

Note: The estimated frequencies of a CNV allele with *i* copies for cases and controls are denoted by x_i and y_i, respectively.

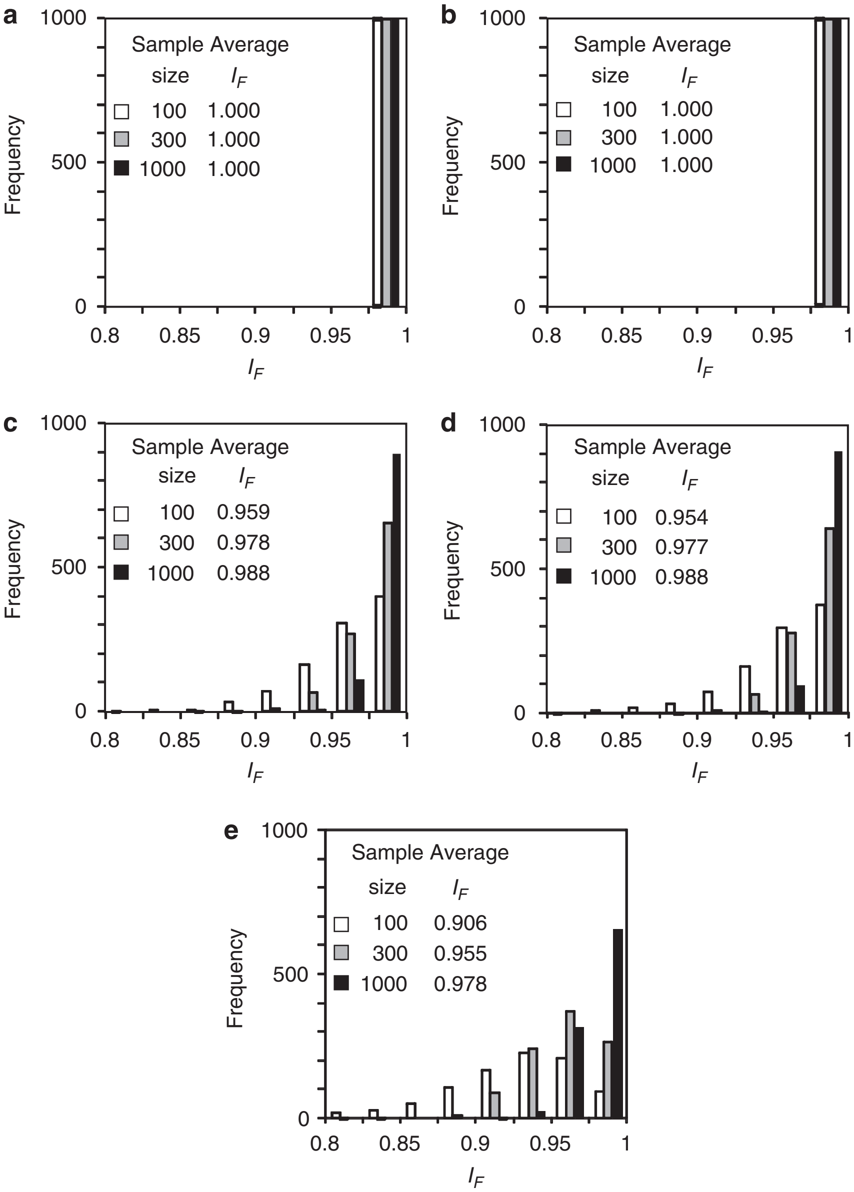

Figure 1 Frequency distribution of I_F The results were obtained from 1000 simulation runs for each CNV parameter set under the assumption of Hardy–Weinberg equilibrium. Only the frequency distribution of $I_F \geqslant 0.8$ is shown. (**a**) Two CNV alleles with copy numbers of 0 and 1; {0, 1}. (**b**) Two CNV alleles with copy numbers of 1 and 2; {1, 2}. (**c**) Three CNV alleles with copy numbers of 0, 1 and 2; {0, 1, 2}. (**d**) Three CNV alleles with copy numbers of 1, 2 and 3; {1, 2, 3}. (**e**) Four CNV alleles with copy numbers of 0, 1, 2 and 3; {0, 1, 2, 3}.

of H are observed. These revised allele frequencies are used to obtain a set of revised expected values of the unobserved counts $N_{i,j}$. The cycle of revising a_i and revising $N_{i,j}$ is repeated until the change in the log-likelihood becomes negligible. In practice, 1000 cycles are sufficient for the convergence. The log-likelihood $\ln L$ is given by the formula

$$\ln L = \sum_{k=0}^{H}\left[N_k \ln\left(\sum_{i=0}^{k} a_i a_{k-i}\right)\right]$$

As an EM algorithm can be trapped at a local maximum, a number of sets of initial allele frequencies may be examined to find a set of allele frequencies giving the maximum value of $\ln L$. In this study, 20 sets of initial CNV allele frequencies randomly selected were examined in the computer simulation.

Permutation test

A case–control association study was considered, where CNV allele frequencies in N cases and those in M controls are compared. In this test, CNV allele frequencies were estimated for cases and controls separately as described above. Next, to identify the susceptibility allele, a χ^2 value is calculated for each CNV allele based on a 2×2 contingency table, as shown in Table 1. Therefore, a total of H+1 χ^2 values are calculated. The allele frequencies estimated by the EM algorithm should not be directly used for a χ^2 test, as the test statistics include the error due to the estimation. This problem can be overcome using permutation procedures to obtain empirical P-values. The obtained χ^2 value for each allele from Table 1 is compared with the largest χ^2 value among H+1 ones calculated from a new data set provided by randomly permuting case–control labels. It should be noted here that individuals are assumed to be interchangeable under the null hypothesis of no association between CNV alleles and disease. The present permutation test can avoid an inflation of a type I error rate due to multiple testing for all CNV alleles and can control a global type I error rate.

Simulation

To evaluate the performance of the EM algorithm for estimating the CNV allele frequencies, a computer simulation was performed under the assumption of the Hardy–Weinberg equilibrium. In the simulation, the genotype of each individual was determined with random number, and the CNV allele frequencies were estimated based only on the frequencies of the diploid copy number in a set of samples (that is, the genotype data were assumed not to be available). The allele frequencies in a set of samples generated by computer simulation (referred as true allele frequencies) were compared with the estimated allele frequencies using the similarity index[7]

$$I_F = 1 - \frac{1}{2}\sum_{i=0}^{H}|\bar{a}_i - a_i|$$

where $\bar{a}_i$ is the true frequency of allele i in a set of simulated samples. The I_F value ranges from 0 to 1 and the larger I_F indicates the better estimation. In this paper, each simulation was characterized by the copy numbers of the alleles, and the alleles were assumed to have an equivalent frequency unless otherwise stated. For example, a CNV parameter set of {0, 1, 2} represents a simulation where three alleles with copy numbers of 0, 1 and 2 are assumed to have the same population frequency (that is, 1/3).

To further examine the statistical power of the present permutation test for a case–control association study, case and control samples were also generated by computer simulation. Here, a CNV allele with i copies was assumed to be associated with a disease. The penetrances for the genotypes of i/i, i/non-i and non-i/non-i were denoted by f_2, f_1 and f_0, respectively. The disease prevalence is given by $e=a_i^2f_2+2a_i\,(1-a_i)\,f_1+(1-a_i)^2\,f_0$.

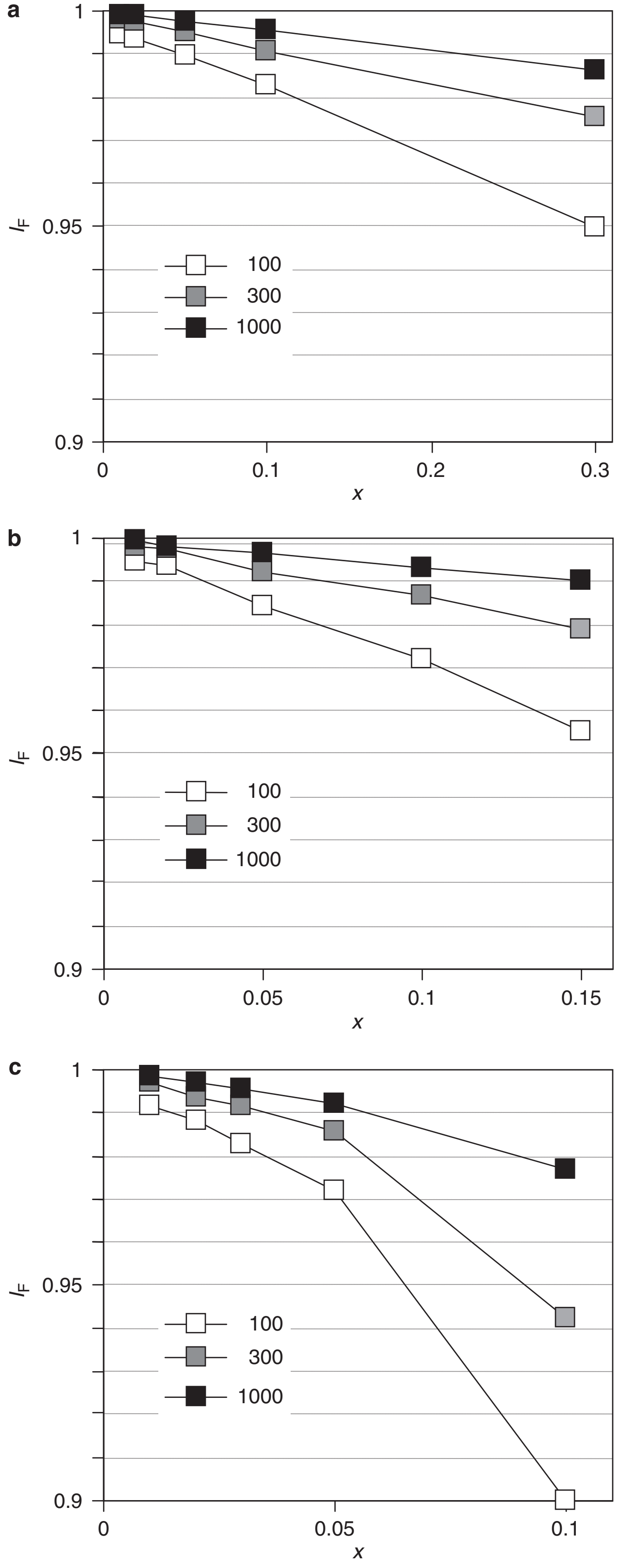

Figure 2 Average of I_F for CNV with nonequivalent allele frequencies. The results were obtained from 1000 simulation runs for each CNV parameter set under the assumption of Hardy–Weinberg equilibrium. The averages of I_F were calculated for the sample sizes of 100, 300 and 1000. (**a**) Three CNV alleles with copy numbers of 0, 1 and 2 had population frequencies of x, 1–2x and x, respectively. (**b**) Three CNV alleles with copy numbers of 1, 2 and 3 had population frequencies of 1–3x, 2x and x, respectively. (**c**) Four CNV alleles with copy numbers of 0, 1, 2 and 3 had population frequencies of 2x, 1–5x, 2x and x, respectively.

Therefore, the probabilities of a case sample being genotypes of *i*/*i*, *i*/non-*i* and non-*i*/non-*i* are $a_i^2 f_2/e$, $2a_i\ (1-a_i)\ f_1/e$ and $(1-a_i)^2\ f_0/e$. Similarly, the probabilities of a control sample being genotypes of *i*/*i*, *i*/non-*i* and non-*i*/non-*i* are $a_i^2\ (1-f_2)/(1-e)$, $2a_i\ (1-a_i)\ (1-f_1)/(1-e)$ and $(1-a_i)^2\ (1-f_0)/(1-e)$, respectively.

RESULTS

Estimation of CNV allele frequencies

To evaluate fundamental performance of the present method, the difference between the true and estimated allele frequencies was evaluated under the assumption of Hardy–Weinberg equilibrium (that is, an ideal situation for the present algorithm). The frequency distributions of I_F values for five representative CNV parameter sets {0, 1}, {1, 2}, {0, 1, 2}, {1, 2, 3} and {0, 1, 2, 3} are shown in Figure 1. For each CNV parameter set, 1000 runs were performed. As the number of samples decreased or the number of alleles increased, the average value of I_F decreased. However, the averages of I_F were more than 0.9 in these CNV parameter sets, with sample size of more than 100 (Figure 1). To further confirm the performance of a CNV locus with nonequivalent allele frequencies, the averages of I_F were calculated for other parameter sets (Figure 2). Here, a CNV allele with one copy was assumed to be the most frequent (see figure legend for details). Compared with a CNV locus with equivalent allele frequencies, the estimation was more reliable. Therefore, the present algorithm gives a reliable estimation of the CNV allele frequencies under the Hardy–Weinberg equilibrium. Although there are CNV loci with

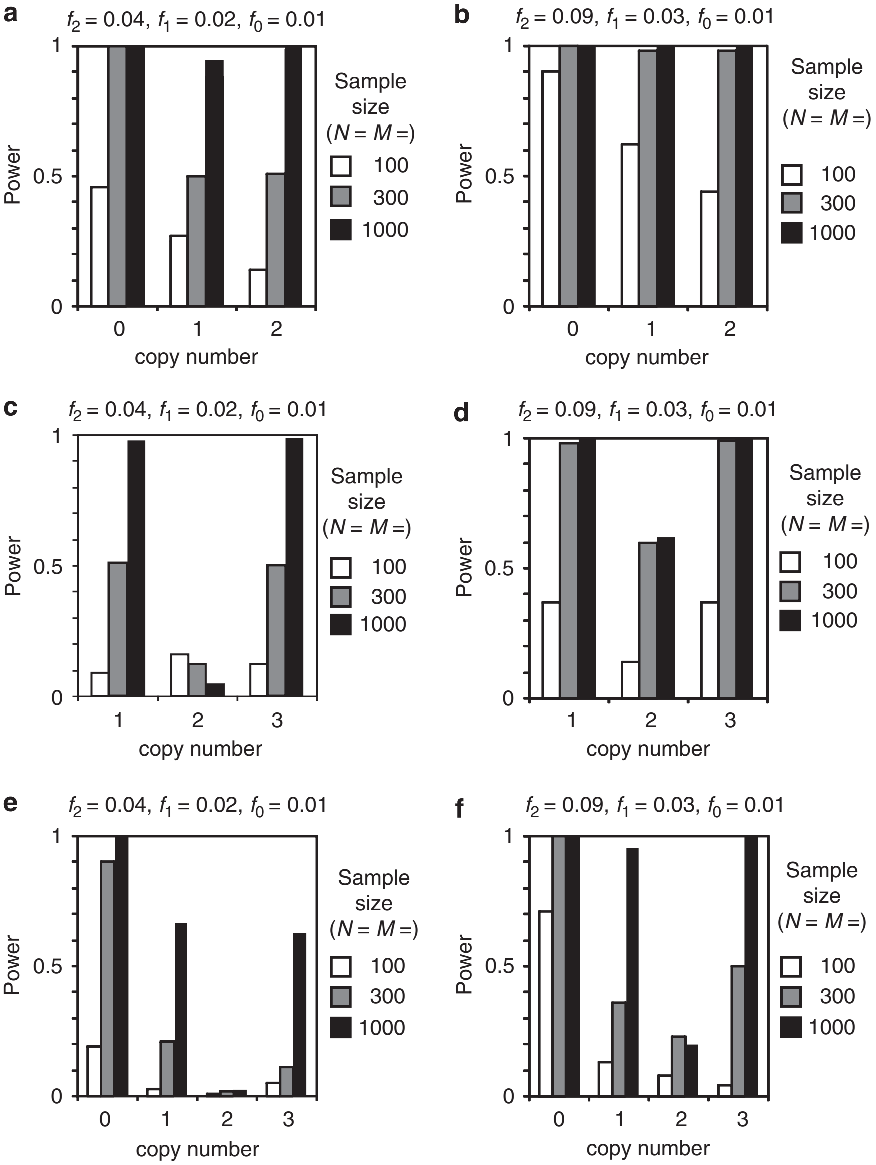

Figure 3 Power of association test for each CNV allele. The results were obtained from 100 simulation runs (sample sets) for each allele of each CNV parameter set. In the simulation, one of the CNV alleles was assumed to be associated with a disease and the horizontal axis represents the copy number. The powers for detecting it as a susceptible allele were calculated for three sets of the sample size (100: 100 cases and 100 controls; 300: 300 cases and 300 controls; and 1000: 1000 cases and 1000 controls). (**a**) The locus had three CNV alleles with copy numbers of 0, 1 and 2; {0, 1, 2} and one of CNV alleles had the penetrance f_2=0.04, f_1=0.02 and f_0=0.01. (**b**) {0, 1, 2} and the penetrance: f_2=0.09, f_1=0.03 and f_0=0.01. (**c**) {1, 2, 3} and the penetrance: f_2=0.04, f_1=0.02 and f_0=0.01. (**d**) {1, 2, 3} and the penetrance: f_2=0.09, f_1=0.03 and f_0=0.01. (**e**) {0, 1, 2, 3} and the penetrance: f_2=0.04, f_1=0.02 and f_0=0.01. (**f**) {0, 1, 2, 3} and the penetrance: f_2=0.09, f_1=0.03 and f_0=0.01.

alleles with more complicated copy numbers, only copies of 0, 1, 2 and 3 were assumed in the simulation of this study.

Power of case–control association test

To estimate the statistical power of the present association test, 100 sample sets were generated for three CNV parameter sets {0, 1, 2}, {1, 2, 3} and {0, 1, 2, 3}, and the empirical P-value for each CNV allele was calculated with 1000 permutations in each sample set. In the simulation, one of the CNV alleles was assumed to be associated with a disease and the power was calculated based on the empirical P-value of the susceptible CNV allele. Here, the proportion of the empirical P-value of <0.05 in 100 sample sets was regarded as a power for the CNV allele. The power was different among CNV alleles in the same parameter set (Figure 3). A susceptible CNV allele with the lowest copy number at the locus was easy to be detected, although the same power can be expected for each allele if the genotype of each subject is observable. In the case of {0, 1, 2}, an allele with one CNV copy (that is, allele with an intermediate copy number) was detected with a power of 0.95 in 1000 cases and 1000 controls, whereas an allele with two CNV copies was difficult to detect in the case of {1, 2, 3} (that is, the power was 0.05). Therefore, whether a susceptible CNV allele could be detected was largely dependent on the property of CNV locus.

DISCUSSION

In this study, a practical case–control association test for detecting a susceptible CNV allele was proposed and the statistical power was evaluated by a computer simulation. Although only a limited number of parameter sets were examined, the feature of the present method has been understood. The statistical power is markedly different among CNV alleles with different copy numbers, and the higher power seems to be achieved for susceptible alleles with the lowest or highest copy number in comparison with those with intermediate copy number (for example, Figures 3c–f). However, it is noted that an association between a CNV locus and a disease can be detected with high power if the frequencies of the observed diploid copy numbers are compared between cases and controls, even if a susceptible allele has an intermediate copy number (data not shown). To examine the reason why the statistical power is lower for susceptible allele with intermediate copy number than that with the lowest or highest copy number, the averages of I_F were calculated for case and control groups under various parameter sets. The results showed that the lower averages of I_F are generally observed for a case group when a susceptible allele has an intermediate copy number, whereas the averages of I_F for control group are not different regardless of copy number of a susceptible allele. For instance, in the case of Figure 3c, the averages of I_F for susceptible alleles with 1, 2, and 3 were 0.961, 0.935, and 0.961, respectively, for N=1000 (that is, 1000 case individuals), and 0.988, 0.989, and 0.988 for M=1000 (that is, 1000 control individuals). Thus, the lower power for a susceptible allele with an intermediate copy number results from larger error of allele frequency estimation in case group, although the error is not so large.

The deviation from the Hardy–Weinberg equilibrium makes it difficult to detect a susceptible CNV allele, as the CNV allele frequencies, especially in cases, cannot be estimated accurately by the EM algorithm. The penetrance of each genotype of a susceptible allele and allele frequencies in cases and controls are estimated for cohort studies using the maximum likelihood method.[8] The simultaneous estimation of penetrance will allow the accurate inference of CNV allele frequencies at a susceptible locus even in cases. Therefore, for analyzing the CNV data from cohort studies, it would be interesting to develop a new method in order to estimate both the penetrance and CNV allele frequencies to detect a susceptible CNV allele.

The present EM algorithm for estimating CNV allele frequencies requires reliable data of an integer diploid copy number. Recently, a heuristic method has been proposed to determine the integer diploid copy number of each individual based on data from a genotyping array with optimized ‘copy number probes’ (Affymetrix SNP 6.0), and the integer diploid copy numbers for 1320 copy number polymorphisms in 270 HapMap samples[9,10] have been reported with this method.[11] As the present association test requires a permutation procedure for estimating CNV allele frequencies, a lot of computational time is needed. Therefore, for analyzing genome-wide CNV data, the association between the CNV loci and the disease should be initially evaluated with a conventional χ^2 test for the integer diploid copy number. Then, to detect a single susceptible CNV allele, the present association test should be performed for each CNV locus showing P-value less than the genome-wide significance level.

ACKNOWLEDGEMENTS

This study was supported by research funds (KAKENHI) (Grant-in-Aid for Scientific Research, 20018004) on Priority Areas ‘Applied Genomics’ from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

1 Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. *Nat. Rev. Genet.* **7,** 85–97 (2006).

2 Fanciulli, M., Norsworthy, P. J., Petretto, E., Dong, R., Harper, L., Kamesh, L. *et al.* FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nat. Genet.* **39,** 721–723 (2007).

3 Gonzalez, E., Kulkarni, H., Bolivar, H., Mangano, A., Sanchez, R., Catano, G. *et al.* The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* **307,** 1434–1440 (2005).

4 Estivill, X. & Armengol, L. Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies. *PLoS Genet.* **3,** 1787–1799 (2007).

5 Kato, M., Nakamura, Y. & Tsunoda, T. An algorithm for inferring complex haplotypes in a region of copy-number variation. *Am. J. Hum. Genet.* **83,** 157–169 (2008).

6 Kato, M., Nakamura, Y. & Tsunoda, T. MOCSphaser: a haplotype inference tool from a mixture of copy number variation and single nucleotide polymorphism data. *Bioinformatics* **24,** 1645–1646 (2008).

7 Excoffier, L. & Slatkin, M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.* **12,** 921–927 (1995).

8 Ito, T., Inoue, E. & Kamatani, N. 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. *Genetics* **168,** 2339–2348 (2004).

9 The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (2005).

10 Frazer, K. A. Ballinger, D. G. Cox, D. R. Hinds, D. A. Stuve, L. L. Gibbs, R. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–861 (2007).

11 McCarroll, S. A., Kuruvilla, F. G., Korn, J. M., Cawley, S., Nemesh, J., Wysoker, A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.* **40,** 1166–1174 (2008).