J Hum Genet (2003) 48:487–491
DOI 10.1007/s10038-003-0058-7

SHORT COMMUNICATION

Jun Ohashi · Katsushi Tokunaga

Power of genome-wide linkage disequilibrium testing by using microsatellite markers

Received: 6 May 2003 / Accepted: 2 July 2003 / Published online: 22 August 2003

Abstract Linkage disequilibrium (LD) testing is often used in the search for disease genes. In this study, we developed a method for calculating the expected power of genome-wide LD testing by using microsatellite markers under the following assumptions: (1) microsatellite markers have unequally frequent alleles, (2) markers are equally spaced through the human genome, (3) the degree of LD between the disease variant and the marker decays gradually because of recombination and mutation, (4) the population frequency of the disease variant is low (e.g., 0.05), (5) a single-marker test is performed in a case-control study, and (6) the significance level is adjusted by the number of tests to avoid inflation of the type I error. Our calculations revealed a markedly higher power for microsatellite markers than for single nucleotide polymorphism (SNP) markers, even if more SNPs are analyzed, suggesting that the use of microsatellite markers is preferable to the use of SNPs for genome-wide screening under the above assumptions. This method will be helpful to researchers who design genome-wide LD testing with microsatellite markers.

Keywords Case-control study · Low-frequency disease variant · χ^2 statistic · Correction of significance level · One-step stepwise mutation model · Recombination · Single nucleotide polymorphism

Introduction

The advent of inexpensive genotyping technologies has made it possible to use polymorphic markers in genome-wide linkage disequilibrium (LD) testing to detect genes involved in genetically complex diseases. Two types of markers are suitable for such a purpose: (1) single nucleotide polymorphism (SNP) markers, which are the most abundant genetic marker in the human genome, and (2) microsatellite markers, which may have many alleles and show high degrees of polymorphism.

Based on a simple deterministic model assuming steady decay of LD, Ohashi and Tokunaga (2002) have concluded that LD testing with SNP markers does not allow the detection of low-frequency disease variants with a modest contribution to the onset of a disease. This is mainly because there would be a difference in allele frequency between the low-frequency disease variant and the associated allele at the adjacent SNP marker, regardless of the degree of LD between them. The difference in allele frequency is known to reduce the power of LD studies markedly (Ohashi and Tokunaga 2001). If, however, the genetic markers have many alleles, the difference in allele frequency is expected to be small.

When a marker locus with equally frequent alleles is analyzed in LD testing, the statistical power increases with an increasing number of alleles (Chapman and Wijsman 1998; Xiong and Jin 1999), because of the increased probability that one of the marker alleles in strong LD with the disease variant has a similar allele frequency as the disease variant. However, the assumption of equally frequent alleles leads to an overestimation of the power of LD testing by using microsatellite markers (Chapman and Wijsman 1998). To consider unequally frequent alleles in the calculation of power, we have developed a model with unequal mutation rates among alleles. This paper aims to provide a reliable method for the calculation of the power of genome-wide LD testing by using microsatellite markers. Furthermore, we demonstrate the way in which microsatellite markers are more powerful for genome-wide LD testing compared with SNP markers.

J. Ohashi · K. Tokunaga
Department of Human Genetics,
Graduate School of Medicine, University of Tokyo,
Tokyo 113-0033, Japan

J. Ohashi (✉)
Department of Molecular Biology and Genetics,
Cornell University, Ithaca, NY 14853, USA
E-mail: juno-tky@umin.ac.jp
Tel.: +1-607-2551707
Fax: +1-607-2556249

Materials and methods

Genetic model

It is assumed that a disease locus has two alleles, a disease allele, *D*, and a normal allele, *d*. Allele frequencies of *D* and *d* are *p* and 1-*p*, respectively. At a microsatellite marker locus, there are *m* different alleles, M_i, $i=1, ..., m$, with allele frequencies q_i ($\sum_{i=1}^{m} q_i = 1$). Throughout the paper, these frequencies are assumed to be constant, with the population being in Hardy-Weinberg equilibrium. The recombination fraction between the two loci is denoted by θ. The frequency of the haplotype carrying *D* and M_i at time *t* (measured in generations) is represented by $H_{Di}(t)$. Similarly, the frequency of the haplotype carrying *d* and M_i at time *t* is represented by $H_{di}(t)$. Note that $q_i = H_{Di} + H_{di}$. Because microsatellite markers usually show a high mutation rate (e.g., $10^{-3}\sim10^{-5}$ per generation), mutations are assumed only for the marker. Here, a one-step stepwise mutation model (SMM) is assumed. In this model, a marker allele M_i mutates to the next larger allele M_{i+1}, and to the next smaller allele M_{i-1}, each with a probability of u_i. The smallest allele, M_1, can mutate only to M_2 with a probability of u_1, and the largest allele, M_m, can mutate only to M_{m-1} with a probability of u_m. Under these assumptions, the following deterministic equations hold for the microsatellite markers: $H_{D1}(t+1) = (1-\theta-u_1)H_{D1}(t) + u_2H_{D2}(t) + \theta pq_1$, $H_{Di}(t+1) = (1-\theta-2u_i)H_{Di}(t) + u_{i-1}H_{Di-1}(t) + u_{i+1}H_{Di+1}(t) + \theta pq_i$ ($i=2, ..., m-1$), $H_{Dm}(t+1) = (1-\theta-u_m)H_{Dm}(t) + u_{m-1}H_{Dm-1}(t) + \theta pq_m$, $H_{d1}(t+1) = (1-\theta-u_1)H_{d1}(t) + u_2H_{d2}(t) + \theta(1-p)q_1$, $H_{di}(t+1) = (1-\theta-2u_i)H_{di}(t) + u_{i-1}H_{di-1}(t) + u_{i+1}H_{di+1}(t) + \theta(1-p)q_i$ ($i=2, ..., m-1$), and $H_{dm}(t+1) = (1-\theta-u_m)H_{dm}(t) + u_{m-1}H_{dm-1}(t) + \theta(1-p)q_m$. When initial parameter values are given, we obtain each haplotype frequency for any *t*, by using these recurrence formulae.

Equilibrium distribution of allele-frequency at marker locus

Farrall and Weeks (1998) investigated the equilibrium distributions of microsatellite allele frequency of (CA)$_n$ microsatellite makers in the Généthon database. Although they examined only microsatellite markers with a CA repeat, we regard their results as representative of microsatellite markers. From their web page (http://www.well.ox.ac.uk/~mfarrall/microsatellite.html), we obtained the observed equilibrium distributions of (CA)$_n$ microsatellite makers with various numbers of alleles. In the SMM, which assumes a fixed number of marker alleles and an equilibrium distribution of allele frequency, proper mutation rates, u_i, should be given to achieve the equilibrium distribution of $\bar{q}_i$. When the mean mutation rate at a marker is *u*, the following equation should be satisfied:

$$\bar{q}_1u_1 + \bar{q}_mu_m + \sum_{i=2}^{m-1} 2\bar{q}_iu_i = u$$

It should be noted here that $u_i\bar{q}_i$ requires to be constant regardless of *i*, because allele frequencies of microsatellite marker are not changed though the calculation. Thus, we get $u_i = \frac{u}{2(m-1)\bar{q}_i}$ as the mutation rate for M_i.

Disease model

We will consider a case-control study of complex disease genes. For a complex disease, the case group will contain some case individuals who do not possess the susceptibility allele, and the control group will contain some control individuals who possess the susceptibility allele. Here, penetrances for genotypes *DD*, *Dd*, and *dd* are denoted by f_2, f_1, and f_0, respectively ($f_2 \geq f_1 \geq f_0$), and a multiplicative mode of inheritance with a genotype relative risk of *r* (i.e., $f_2 = r^2f_0$ and $f_1 = rf_0$) is examined in this study. Note that the present method can be applied to any other modes of inheritance if penetrances are specified. The conditional probabilities of the *DD*, *Dd*, and *dd* genotypes, given that the individual is affected (*case*), are given by $P(DD|case) = p^2f_2/e$, $P(Dd|case) = 2p(1-p)f_1/e$, and $P(dd|case) = (1-p)^2f_0/e$, respectively, where *e* represents the disease prevalence, $p^2f_2 + 2p(1-p)f_1 + (1-p)^2f_0$, in the studied population. Similarly, the conditional probabilities of each genotype, given that the individual is not affected (*control*), are given by $P(DD|control) = p^2(1-f_2)/(1-e)$, $P(Dd|control) = 2p(1-p)(1-f_1)/(1-e)$, and $P(dd|control) = (1-p)^2(1-f_0)/(1-e)$. By Bayes' theorem, the probability of an affected individual being of the M_iM_i genotype at time *t* is given as $P(M_iM_i|case,\ t) = \frac{1}{e}\left\{f_2H_{Di}(t)^2 + 2f_1H_{Di}(t)H_{di}(t) + f_0H_{di}(t)^2\right\}$, and the probability of an affected individual being of the M_iM_j ($j \neq i$) is given as $P(M_iM_j|case,\ t) = \frac{1}{e}\{2f_2H_{Di}(t)H_{Dj}(t) + 2f_1\left(H_{Di}(t)H_{dj}(t) + H_{di}(t)H_{Dj}(t)\right) + 2f_0H_{di}(t)H_{dj}(t)\}$. For control, $P(M_iM_i|control,\ t)$ and $P(M_iM_j|control,\ t)$ are represented by $P(M_iM_i|control,\ t) = \frac{1}{1-e}\{(1-f_2)H_{Di}(t)^2 + 2(1-f_1)H_{Di}(t)H_{di}(t) + (1-f_0)H_{di}(t)^2\}$ and $P(M_iM_j|control,\ t) = \frac{1}{1-e}\{2(1-f_2)H_{Di}(t)H_{Dj}(t) + 2(1-f_1).\left(H_{Di}(t)H_{dj}(t) + H_{di}(t)H_{Dj}(t)\right) + 2(1-f_0)H_{di}(t)H_{dj}(t)\}$, respectively.

Power

In genome-wide LD testing with a case-control design, marker allele frequencies are compared between case individuals and control individuals. For a comparison of allele frequencies, the data for a microsatellite marker with *m* alleles are summarized in an *m*×2 contingency table (Table 1). When the number of cases, *N*, is equal to that of controls, a χ^2 statistic for an *m*×2 contingency table, $X^2 = 2N\sum_{i=1}^{m}\frac{(\tilde{x}_i-\tilde{y}_i)^2}{\tilde{x}_i+\tilde{y}_i}$, can be used to test the null hypothesis of no difference in frequencies of *m* alleles between cases and controls. Here, $\tilde{x}_i$ and $\tilde{y}_i$ indicate the observed M_i frequencies in cases and controls, respectively (Chapman and Wijsman 1998). That is, $\tilde{x}_i$ and $\tilde{y}_i$ represent $W_{i1}/(2N)$ and $W_{i2}/(2N)$, respectively. Under the null hypothesis, X^2 is asymptotically distributed as a χ^2 distribution with *m*-1 degrees of freedom. Under the alternative hypothesis of LD or different allele frequencies between cases and controls, X^2 is asymptotically distributed as a χ^2 distribution with *m*-1 degrees of freedom and with a noncentrality parameter of $\gamma = 2NG^2 = 2N\sum_{i=1}^{m}\frac{(x_i-y_i)^2}{x_i+y_i}$. After *t* generations, a noncentrality parameter is represented by $\gamma(t) = 2NG^2(t) = 2N\sum_{i=1}^{m}\frac{(x_i(t)-y_i(t))^2}{x_i(t)+y_i(t)}$, where x_i(t) and y_i(t) are $\left(P(M_iM_i|case,\ t) + \sum_{j=1}^{m}P\left(M_iM_j|case,\ t\right)\right)/2$ and $\left(P(M_iM_i|control,\ t) + \sum_{j=1}^{m}P\left(M_iM_j|control,\ t\right)\right)/2$, respectively.

We assume that marker allele frequencies are stable at any generation and that only one of the marker alleles is in complete LD with the *D* allele at the LD-generating event or at $t=0$, as in previous studies (Chapman and Wijsman 1998; Xiong and Jin 1999). That is, $H_{Dk}(0) = p$, $H_{Di}(0) = 0$ ($i \neq k$), $H_{dk}(0) = q_k - p$, and $H_{di}(0) = q_i$ ($i \neq k$), where q_k is larger than *p*. When the frequency of *D* is small, this assumption seems to be valid. If the probability of M_k being in complete LD with *D* is proportional to q_k, the expected value of $G^2(t)$ for microsatellite markers is given by a weighted average of the conditional expected values: $G^2(t) = \sum_{k=1}^{m} q_kG_k^2(t)$

Table 1 *m*×2 contingency table (W_{ij} represents the number of observed M_i alleles in a group of *j*th column)

Allele	Group	
	Case	Control
M_1	W_{11}	W_{12}
M_i	W_{i1}	W_{i2}
M_m	W_{m1}	W_{m2}

(Chapman and Wijsman 1998), where $G^2{}_k(t)$ represents $G^2(t)$ under the condition that M_k allele is in complete LD with D allele at $t=0$.

In genome-wide LD testing, a large number of markers are examined, which causes an inflation of the type I error. To avoid this inflation, the significance level should be adjusted by the number of markers to be tested. When l markers are tested, viz., l independent association tests are performed, the significance level of α should be set to $0.05/l$. We consider the case that the LD testing is performed against a region spanning L cM. The disease locus is assumed to be exactly located at the middle of two adjacent markers, and the statistical power is calculated only for the closest marker to the disease locus. Thus, the genetic distance between the disease locus and the most adjacent marker locus is given as $L/(2l)$. By using Haldane's map function, the recombination rate between the nearest marker and the disease locus is given by $\theta = [1 - \exp(-L/l)]/2$. The total length of the human genome, L, is assumed to be 30 M. To avoid an inflation of the type I error rate attributable to multiple testing, the significance level of α is set to $0.05/l$. The asymptotic power, $1-\beta$, for the significance level of α is given as $\mathrm{Prob}[\chi^2{}_{m-1}(\gamma) \geqq \chi^2{}_{m-1,\ 1-\alpha}(\gamma)]$. The calculations were performed by using SAS software (SAS Institute, Cary, N.C.).

Results and discussion

The number of alleles differs among microsatellite markers. Because microsatellite markers with six alleles are the most frequently observed (Farrall and Weeks 1998), we examined mainly the case of $m=6$. Following the data of Farrall and Weeks (1998), the equilibrium distribution of allele frequencies for $m=6$ was set as: $q_1=0.160$, $q_2=0.199$, $q_3=0.214$, $q_4=0.205$, $q_5=0.158$, and $q_6=0.064$. The population frequency of a disease variant, p, was assumed to be 0.05 throughout this study, and the penetrance for the normal genotype f_0 was set to be 0.01 (a power for f_0 of less than 0.01 is as same as that for an f_0 of 0.01). Table 2 shows the expected power of LD testing under the condition that the marker is in complete LD with a low-frequency disease variant. In other words, this is the case of $t=0$ and the most ideal LD situation. The power was found to depend largely on sample size, genotype relative risk, and significance level. The results suggest that, in a case-control study with 100 cases and 100 controls, it would be difficult to detect a low-frequency disease variant with a genotype relative risk of 2, even at a significance level of 0.05, and that the power for the variant does not reach 0.8 even when 1000 cases and 1000 controls are analyzed. Thus, we may say that LD testing with microsatellite markers can detect a disease variant only with a large genotype relative risk when the allele frequency of the disease variant is low.

Although a high statistical power is revealed for a disease variant with a genotype relative risk of 4 in Table 2, the power is reduced by recombination between marker and disease loci and mutation at the marker. To examine the effects of recombination and mutation on the power, we calculated the expected power of LD testing at 50 and 500 generations after the LD-generating event for a study involving the analysis of 1000 cases and 1000 controls (Fig. 1). Figure 1 shows that recombination and mutation markedly influence the power of genome-wide LD testing. In the case of $m=6$ and $t=50$ (Fig. 1a), the examination of 2,000–3,000 markers provides a power of more than 0.8, whereas approximately 30,000 markers are required to gain the same power for $m=6$ and $t=500$, even when $u=10^{-5}$ (Fig. 1b). Above a certain number of markers, the addition of more markers decreases the power because of the correction of significance level based on the number of tests (Fig. 1a). However, a large number of markers needs to be analyzed to attain a high statistical power. Because the reduction in power attributable to

Table 2 Expected power of LD testing by using microsatellite markers being in complete LD with a disease variant (at $t=0$). Power was calculated for microsatellite markers with six alleles ($q_1=0.160$, $q_2=0.199$, $q_3=0.214$, $q_4=0.205$, $q_5=0.158$, and $q_6=0.064$). In the calculation, p and f_0 were set to 0.05 and 0.01, respectively

N	r	Significance level ($0.05/l$)					
		0.05	0.005	0.0005	0.00005	0.000005	0.0000005
100	$\sqrt{2}$	0.06	<0.01	<0.01	<0.01	<0.01	<0.01
	2	0.10	0.02	<0.01	<0.01	<0.01	<0.01
	4	0.49	0.19	0.06	0.02	<0.01	<0.01
500	$\sqrt{2}$	0.10	0.01	<0.01	<0.01	<0.01	<0.01
	2	0.38	0.13	0.03	<0.01	<0.01	<0.01
	4	>0.99	0.98	0.94	0.83	0.69	0.52
1000	$\sqrt{2}$	0.16	0.03	<0.01	<0.01	<0.01	<0.01
	2	0.70	0.38	0.17	0.06	0.02	<0.01
	4	>0.99	>0.99	>0.99	>0.99	>0.99	>0.99

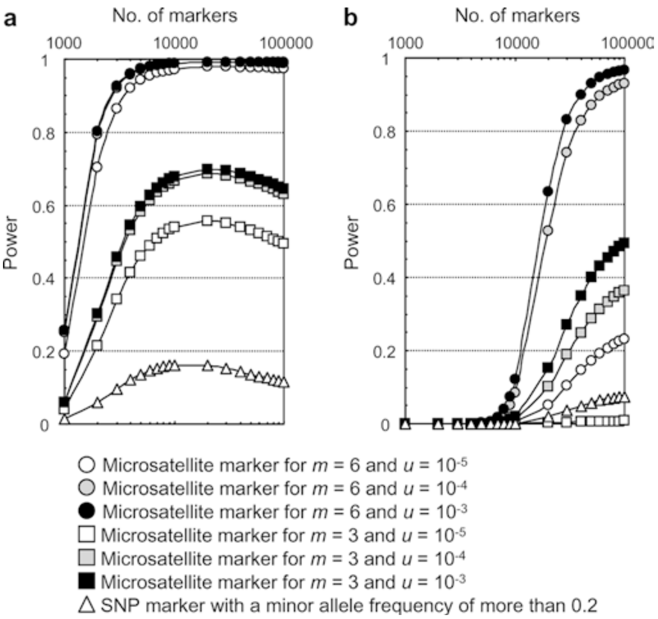

Fig. 1a, b Expected power of genome-wide LD testing for the detection of a low-frequency disease variant. **a** $f_2=0.16$, $f_1=0.04$, $f_0=0.01$, and $t=50$. **b** $f_2=0.16$, $f_1=0.04$, $f_0=0.01$, and $t=500$. The mutation rate and the number of alleles at the microsatellite marker are indicated by the symbols as explained *bottom*. The power for an SNP marker with minor allele frequency of more than 0.2 is indicated by *open triangles* (see Ohashi and Tokunaga 2002 for details). It is assumed that $p=0.05$ and $N=1000$. The significance level of α for LD testing with l markers is given by $0.05/l$

the correction of significance level is small, we recommend the analysis of as many microsatellite markers as possible in LD testing.

In Fig. 1, microsatellite markers with three alleles are also considered; the equilibrium distribution of allele frequencies for $m = 3$ was set as: $q_1 = 0.357$, $q_2 = 0.412$, and $q_3 = 0.232$. The power for $m = 3$ is clearly lower than that for $m = 6$. Thus, we should use microsatellite markers with more alleles, even though this increases the degrees of freedom in the statistical test. Note, however, that this is true only when all alleles at a marker have similar allele frequencies. If a marker locus has several alleles with very low allele frequencies, the power is markedly lower than that for markers whose allele frequencies are equal, assuming the same number of alleles. For example, when we assumed that allele frequencies at microsatellite markers with $m = 6$ were $q_1 = q_2 = q_3 = q_4 = q_5 = 0.05$, and $q_6 = 0.75$, the power was as same as that for SNPs in the same condition as in Fig. 1 (data not shown). Thus, it is necessary to use microsatellite markers with non-skewed frequent alleles to obtain a high power. To avoid a reduction in power for microsatellite markers with many low-frequency alleles, all the very low-frequency alleles (e.g., $q < 0.01$) should be regarded as one allele, and the degrees of freedom in the test should be reduced.

Although we do not consider the difference in the number of alleles among microsatellite markers in the present study, the following method allows us to deal with this problem. The proportion of microsatellite marker with m alleles in markers to be used for LD testing is denoted by s_m, and the minimum and maximum numbers of allele are denoted by a and b, i.e., $\sum_{m=a}^{b} s_m = 1$. In this case, the non-centrality parameter can be given by $\gamma * (t) = 2N \sum_{m=a}^{b} s_m H_m^2(t)$, where $H^2{}_m(t)$ represents $G^2(t)$ for microsatellite marker with m alleles. When the proportion s_m is known, this method would estimate a more reliable power.

Microsatellite markers seem to be more effective than SNPs if the genetic distance between the disease and marker loci is the same. However, the human genome contains more SNP markers than microsatellite markers. In addition, the cost of SNP typing is much lower than that of microsatellite typing, implying that more SNP markers can be tested with the same cost and resources. The use of more markers reduces the expected genetic distance between marker and disease loci. Thus, it is not immediately apparent which marker should be used in genome-wide LD testing. Figure 1 compares the expected power of LD testing by using SNP markers with that of microsatellite markers. Here, the SNPs are assumed to have a minor allele frequency of more than 0.2 (see Ohashi and Tokunaga 2002, for details). No mutation is considered for SNP markers. Microsatellite markers generally reveal a higher power than SNPs even if large numbers of SNPs are analyzed. In particular, we note that LD testing with SNP markers cannot attain a high power under the conditions of Fig. 1. Although our perspective is true only for a low-frequency disease variant and microsatellite markers with non-skewed frequent alleles, we conclude that microsatellite markers are preferable to SNPs for initial genome-wide screening, and SNPs should be used for fine-scale mapping after the screen.

However, if the analysis uses SNPs only in intragenic regions, especially those in coding and regulatory regions, a high power is expected to be attained, even in LD testing with SNPs. A disease variant may be an SNP allele. It should be noted that if a disease variant is included in the SNPs to be tested, the use of SNPs shows a higher power than that of microsatellite markers. LD defined by SNPs is likely to be structured into discrete blocks of tens to hundreds of kilobases in the human genome (Daly et al. 2001). If the pattern of LD in the human genome is clarified by being based on SNPs, the number of SNPs to be analyzed can be reduced, because only a few SNPs within the same block of LD are regarded as representative. This would increase the power of the study, because fewer SNPs can cover the entire genome. Furthermore, LD testing by using SNP haplotypes may reveal a higher power than a single-marker test. When a microsatellite marker is in an LD block, the significant association of each allele at the marker with one of two alleles at the SNP site in the same LD block is expected to be found (Omi et al. 2003), whereas it is unclear whether a microsatellite marker outside of the LD block would still show LD with SNPs or a disease variant inside of the LD block. If there is no LD between microsatellite markers outside the LD block and SNPs inside the LD block, at least one microsatellite marker is required to be in each LD block. However, it is unlikely that useful microsatellite markers are always in the LD block. Thus, the question of whether microsatellite markers are more suitable for genome-wide LD testing than SNPs remains open.

Acknowledgements I greatly appreciate two anonymous reviewers concerning their suggestions for improvements to this paper. This study was supported by a Grant-in-Aid for Scientific Research on Priority Areas (C) "Medical Science" from the Ministry of Education, Culture, Sports, Science and Technology, Japan (J.O.), the Genetic Diversity Project supported by the New Energy and Industrial Technology Development Organization (J.O.), and Health Sciences Research Grants for Research on Human Genome from Ministry of Health Labour and Welfare of Japan (J.O.).

References

Chapman NH, Wijsman EM (1998) Genome screens using linkage disequilibrium tests: optimal marker characteristics and feasibility. Am J Hum Genet 63:1872–1885

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. Nat Genet 29: 229–232

Farrall M, Weeks DE (1998) Mutational mechanisms for generating microsatellite allele-frequency distributions: an analysis of 4,558 markers. Am J Hum Genet 62:1260–1262

Ohashi J, Tokunaga K (2001) Power of genome wide association studies of complex disease genes: statistical limitation of indirect approaches using SNP markers. J Hum Genet 46:478–482

Ohashi J, Tokunaga K (2002) The expected power of genome-wide linkage disequilibrium testing using single nucleotide polymorphism markers for detecting a low-frequency disease variant. Ann Hum Genet 66:297–306

Omi K, Ohashi J, Patarapotikul J, Hananantachai H, Naka I, Looareesuwan S, Tokunaga K (2003) CD36 polymorphism is associated with protection from cerebral malaria. Am J Hum Genet 72:364–374

Xiong M, Jin L (1999) Comparison of the power and accuracy of biallelic and microsatellite markers in population-based gene-mapping methods. Am J Hum Genet 64:629–640