J Hum Genet (2001) 46:478–482

SHORT COMMUNICATION

Jun Ohashi · Katsushi Tokunaga

The power of genome-wide association studies of complex disease genes: statistical limitations of indirect approaches using SNP markers

Reveived: March 27, 2001 / Accepted: April 21, 2001

Abstract Genome-wide association studies using a dense map of single nucleotide polymorphism (SNP) markers seem to enable us to detect a number of complex disease genes. In such indirect association studies, whether susceptibility genes can be detected is dependent not only on the degree of linkage disequilibrium between the disease variant and the SNP marker but also on the difference in their allele frequencies. These factors, as well as penetrance of the disease variant, influence the statistical power of such approaches. However, the power of indirect association studies is not well understood. We calculated the number of individuals necessary for the detection of the disease variant in both direct and indirect association studies with a case-control design. The result shows that a remarkable reduction in the statistical power of indirect studies, compared with that of direct ones, is unavoidable in the genome-wide screening of complex disease genes. If there is a large difference in allele frequency between the disease variant and the marker, the disease variant cannot be detected. Because the frequency of the disease variant is unknown, SNP markers with various allele frequencies, or a large number of SNP markers, must be used in indirect association studies. However, if the number of SNP markers is increased, the obtained *P* value may not reach the significance level due to the Bonferroni adjustment. Thus, to test a possible association between functional variants and a complex disease directly, we should identify such SNPs in as many genes as possible for use in genome-wide association studies.

Key words Case-control study · Genome-wide association studies · Sample size · SNPs · Power linkage disequilibrium · Market allele frequency

J. Ohashi (✉) · K. Tokunaga
Department of Human Genetics, Graduate School of Medicine, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
Tel. +81-3-5841-3693; Fax +81-3-5802-8619
e-mail: juno@m.u-tokyo.ac.jp

Introduction

A large number of disorders underlying a single gene mode of inheritance (MOI) have been identified by positional cloning. At present, the detection of susceptibility genes of multifactorial or complex diseases is the center of interest in human genetics. One of the significant features of a complex disease is the modest contribution of each susceptibility gene to the onset of the disease. For complex diseases, a genome-wide association study is known to be more appropriate for the detection of the susceptibility genes than a genome-wide linkage analysis (Risch and Merikangas 1996).

There are two types of whole-genome association studies: direct and indirect (Collins et al. 1997). In direct association studies, a possible association of a particular functional variant of a candidate gene with a disease is investigated directly. In indirect association studies, a particular set of genetic markers is analyzed. If a significant association between the marker and the disease is found, the disease locus is expected to be located close to the marker. Thus, the indirect approach depends on linkage disequilibrium between the disease variant and the marker. Because the degree of linkage disequilibrium decreases with increasing genetic distance, a highly dense genetic marker is necessary for indirect genome-wide association studies of complex disease genes.

Recent progress in genotyping techniques enables us to use single nucleotide polymorphisms (SNPs) as biallelic markers for the genome-wide screening of complex diseases (Collins et al. 1997). The systematic cataloging of SNPs has already been started, whereas the physical distance of separation at which two SNP markers exhibit significant linkage disequilibrium is still under consideration. Kruglyak (1999), from a theoretical study on the expected linkage disequilibrium between two SNPs, proposed the use of SNP markers separated by an average distance of 6 kb. In contrast, there are published data providing examples of linkage disequilibrium at distances of more than 50 kb (e.g., Abecasis et al. 2001). However, it has been reported that linkage disequi-

librium is undetectable even if pairs of SNPs within the same gene are examined (e.g., Goddard et al. 2000).

Although the extent of linkage disequilibrium has been studied in different populations and in different chromosomal regions, the detection of linkage disequilibrium is different from the detection of markers that are associated with a disease. Whether susceptibility genes can be detected using adjacent SNP markers showing linkage disequilibrium is not yet clear, although most researchers believe that such detection is possible. The detection of susceptibility genes depends not only on the linkage disequilibrium but also on the genotype relative risk and the difference in allele frequencies of the marker and the disease variant. Müller-Myhsok and Abel (1997) and Abel and Müller-Myhsok (1998) studied the effect of the difference in allele frequencies on the power of the transmission / disequilibrium test (TDT) by Spielman et al. (1993). We have already reported the statistical power of a case-control study (Ohashi et al. 2001). However, we have never considered linkage disequilibrium between a disease variant and a marker, or differences in their allele frequencies. In the present study, we examined these factors in order to clarify the statistical power of indirect population-based association studies of complex disease genes.

Models

We assume a large population conforming to the Hardy-Weinberg equilibrium. We also assume a disease locus with alleles A and a, and a closely linked marker locus with alleles B and b. The population frequencies of A, a, B, and b are p, $1 - p$, m, and $1 - m$, respectively. Allele A is the disease susceptibility allele, and penetrances for genotypes AA, Aa, and aa are denoted by f_{AA}, f_{Aa}, and f_{aa}, respectively ($f_{AA} \geqq f_{Aa} \geqq f_{aa}$). When a multiplicative MOI is assumed, f_{AA} and f_{Aa} are given by $\gamma^2 f_{aa}$ and γf_{aa}, respectively. The coefficient of linkage disequilibrium, δ, is defined as freq(AB) $- pm$. Throughout this study, we assume that the marker allele B is in positive linkage disequilibrium with the susceptibility allele A (i.e., $\delta > 0$). The maximum value of δ, δ_{max}, is given by min (p, m) $- pm$, where min (p, m) represents the lowest of two frequencies p and m. It should be noted here that there are only two haplotypes, AB and ab, in the population when $p = m$ and $\delta/\delta_{max} = 1$ (i.e., the marker allele is equivalent to the disease allele).

Test statistics and required number of samples

In population-based association studies, unrelated individuals are sampled as cases and controls. The aim of such studies is to detect the allele that is more frequently observed in the cases than in the controls. No population stratification is assumed here. The conditional probabilities of AA, Aa, and aa genotypes, given that the individual is affected (*case*), are given by $P(\mathrm{AA}|case) = p^2 f_{AA}/e$, $P(\mathrm{Aa}|case) = 2p(1-p)f_{Aa}/e$, and $P(\mathrm{aa}|case) = (1-p)^2 f_{aa}/e$, respectively, where $e = p^2 f_{AA} + 2p(1-p)f_{Aa} + (1-p)^2 f_{aa}$ is the disease prevalence in the population. Similarly, the conditional probabilities of each genotype, given that the individual is not affected (*control*), are given by $P(\mathrm{AA}|control) = p^2(1 - f_{AA})/(1 - e)$, $P(\mathrm{Aa}|control) = 2p(1-p)(1 - f_{Aa})/(1 - e)$, and $P(\mathrm{aa}|control) = (1-p)^2(1 - f_{aa})/(1 - e)$. By Baye's theorem, the probability of an affected individual being of the BB genotype, $P(\mathrm{BB}|case)$, is given as:

$$\begin{aligned} P(\mathrm{BB}|case) = \frac{1}{e}\{ & f_{AA}(pm+\delta)^2 \\ & + 2f_{Aa}(pm+\delta)[(1-p)m-\delta] \\ & + f_{aa}[(1-p)m-\delta]^2\} \end{aligned} \quad (1)$$

Also, $P(\mathrm{Bb}|case)$ is given as:

$$\begin{aligned} P(\mathrm{Bb}|case) = \frac{1}{e}\langle & 2f_{AA}(pm+\delta)[p(1-m)-\delta] \\ & + 2f_{Aa}\{(pm+\delta)[(1-p)(1-m)+\delta] \\ & + [p(1-m)-\delta][(1-p)m-\delta]\} \\ & + 2f_{aa}[(1-p)m-\delta][(1-p)(1-m)+\delta]\rangle \end{aligned} \quad (2)$$

$P(\mathrm{BB}|control)$ and $P(\mathrm{Bb}|control)$ are represented by:

$$\begin{aligned} P(\mathrm{BB}|control) = \frac{1}{1-e}\{ & (1-f_{AA})(pm+\delta)^2 \\ & + 2(1-f_{Aa})(pm+\delta)[(1-p)m-\delta] \\ & + (1-f_{aa})[(1-p)m-\delta]^2\} \end{aligned} \quad (3)$$

and

$$\begin{aligned} & P(\mathrm{Bb}|control) \\ & = \frac{1}{1-e}\{2(1-f_{AA})(pm+\delta)[p(1-m)-\delta] \\ & \quad + 2(1-f_{Aa})\{(pm+\delta)[(1-p)(1-m)+\delta] \\ & \quad + [p(1-m)-\delta][(1-p)m-\delta]\} \\ & \quad + 2(1-f_{aa})[(1-p)m-\delta][(1-p)(1-m)+\delta]\} \end{aligned} \quad (4)$$

respectively.

In order to test the null hypothesis of no linkage disequilibrium between the marker locus and the disease locus, the number of copies of B per individual in the cases is compared with that in the controls. We consider samples of n_1 individuals for cases and n_2 for controls. For the ith individual in each group, let $X_1(i)$ and $X_2(i)$ denote the number of copies of B, where subscripts 1 and 2 indicate the case and control, respectively, i.e., the possible values of $X_1(i)$ and $X_2(i)$ are 0, 1, and 2. Then, the mean number of copies of B per individual among the case samples, $\overline{X}_1$, and the mean number of copies of B per individual among the control samples, $\overline{X}_2$, are given as $\overline{X}_1 = \sum_{i=1}^{n_1} X_1(i) \Big/ n_1$ and $\overline{X}_2 = \sum_{i=1}^{n_2} X_2(i) \Big/ n_2$, respectively. On the other hand, the sample variances are given as $s_1^2 = \sum_{i=1}^{n_1} \left(X_1(i) - \overline{X}_1\right)^2 \Big/ (n_1 - 1)$ and

$s_2^2 = \sum_{i=1}^{n_2} (X_2(i) - \overline{X}_2)^2 \Big/ (n_2 - 1)$. If the sample size is large, $X_1(i)$ can be regarded as a random sample from the probability distribution with mean μ_1 and variance σ_1^2, and $X_2(i)$ can be a random sample from the probability distribution with mean μ_2 and variance σ_2^2. Thus, we can test the null hypothesis of no linkage disequilibrium between the alleles B and A, using the test statistic:

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

(one-sided test). Here, the difference $\overline{X}_1 - \overline{X}_2$ is assumed to be normally distributed. If n_1 and n_2 are large, sample variances, s_1^2 and s_2^2 can be regarded as the population variances, σ_1^2 and σ_2^2, respectively. Under the null hypothesis, $\mu_{1(n)} = \mu_{2(n)} = 2m$ and $\sigma_{1(n)^2} = \sigma_{2(n)^2} = 2m(1 - m)$. Under the alternative hypothesis, using equations (1) to (4), $\mu_{1(a)}$ and $\mu_{2(a)}$ are given as $\mu_{1(a)} = 2P(\mathrm{BB}|case) + P(\mathrm{Bb}|case)$ and $\mu_{2(a)} = 2P(\mathrm{BB}|control) + P(\mathrm{Bb}|control)$, respectively. Also, $\sigma_{1(a)^2}$ and $\sigma_{2(a)^2}$ are given as $\sigma_{1(a)^2} = 4P(\mathrm{BB}|case) + P(\mathrm{Bb}|case) - \mu_{1(a)^2}$ and $\sigma_{2(a)^2} = 4P(\mathrm{BB}|control) + P(\mathrm{Bb}|control) - \mu_{2(a)^2}$, respectively. Thus, to obtain a power of $1 - \beta$ for a significance level of α, the following formula should be satisfied:

$$Z_\alpha = \frac{\mu_{1(a)} - \mu_{2(a)} + \sqrt{\frac{\sigma_{1(a)^2}}{n_1} + \frac{\sigma_{2(a)^2}}{n_2}}\, Z_{1-\beta}}{\sqrt{\frac{\sigma_{1(n)^2}}{n_1} + \frac{\sigma_{2(n)^2}}{n_2}}}$$

If $n_1 = n$ and $n_2 = kn$, (i.e., the ratio of n_2 to n_1 is k), then the required sample size, n, to achieve $1 - \beta$ power for a significance level of α is given by:

$$n \geq \left[\frac{\sqrt{\sigma_{1(n)^2} + \frac{\sigma_{2(n)^2}}{k}}\, Z_\alpha - \sqrt{\sigma_{1(a)^2} + \frac{\sigma_{2(a)^2}}{k}}\, Z_{1-\beta}}{\mu_{1(a)} - \mu_{2(a)}} \right]^2$$

Results

The numbers of individuals required to obtain 80% power in the genome-wide association studies are shown in Table 1. We assume that five SNPs as biallelic markers are used to analyze one of 40,000 genes (i.e., 200,000 SNPs). In order to reduce the type I error probability for 400,000 independent association tests (400,000 one-sided tests), the significance level of α is set at 1.25×10^{-7} (corresponding to $Z_\alpha = 5.16$), and the power is determined to be 0.8 (corresponding to $Z_{1-\beta} = -0.84$). In this study, we focused on a multiplicative MOI, although all calculations presented here can be applied to a general MOI, modifying the value of penetrance. As shown in Table 1, the same numbers of individuals are sampled as cases and as controls in the case-control study. For the TDT, we assume that each family consists of an affected child and two parents. The TDT requires a larger number of affected individuals than the case-control design to obtain the same power for any parameter set, as pointed out by Morton and Collins (1998). As γ or f_{aa} increases, the difference in the power of the test between the case-control design and the TDT increases, suggesting that the case-control study is suitable for common diseases with a high population prevalence, such as non-insulin-dependent diabetes and hypertension.

Table 1 also shows that the extent of δ is not a negligible factor in genome-wide screening. The numbers for $\delta/\delta_{max} = 1$ indicate the required sample sizes in direct approaches, and the numbers for $\delta/\delta_{max} = 0.75$ and for $\delta/\delta_{max} = 0.5$

Table 1. Number of individuals necessary to obtain 80% power in genome-wide association studies

			Case-control[a]		TDT[b]
γ[c]	p and m[c]	δ/δ_{max}[c]	$f_{aa} = 0.1$[c]	$f_{aa} = 0.01$[c]	
4	0.01	1	—	972	3,114
		0.75	—	1,676	4,726
		0.5	—	3,648	8,807
			(—)	(1.1%)	
	0.1	1	—	158	426
		0.75	—	276	670
		0.5	—	612	1,310
			(—)	(1.7%)	
	0.5	1	—	172	292
		0.75	—	308	526
		0.5	—	700	1,192
			(—)	(6.3%)	
2	0.01	1	6,312	7,712	16,520
		0.75	11,084	13,526	26,975
		0.5	24,616	30,002	55,304
			(10.2%)	(1.0%)	
	0.1	1	776	986	1,972
		0.75	1,368	1,736	3,278
		0.5	3,052	3,866	6,856
			(12.1%)	(1.2%)	
	0.5	1	386	616	964
		0.75	688	1,096	1,720
		0.5	1,554	2,472	3,880
			(22.5%)	(2.3%)	
$\sqrt{2}$	0.01	1	35,390	43,006	77,062
		0.75	62,556	75,972	131,245
		0.5	139,924	169,826	282,352
			(10.1%)	(1.0%)	
	0.1	1	4,092	5,050	8,785
		0.75	7,246	8,934	15,087
		0.5	16,230	20,002	32,748
			(10.8%)	(1.1%)	
	0.5	1	1,782	2,372	3,661
		0.75	3,170	4,218	6,515
		0.5	7,136	9,496	14,668
			(14.6%)	(1.5%)	

A multiplicative mode of inheritance (MOI) (i.e., $f_{AA} = \gamma^2 f_{aa}$; $f_{Aa} = \gamma f_{aa}$) is assumed

[a] The sample size of cases is identical to that of controls (i.e., $k = 1$), and the required number of samples for $f_{aa} < 0.01$ is approximately equal to that for $f_{aa} = 0.01$ (data not shown). The disease prevalences are indicated in parentheses. For $\gamma = 4$ and $f_{aa} = 0.1$, the sample size is not given, since $f_{AA} > 1$

[b] The number of individuals for the transmission/disequilibrium test (TDT) (Müller-Myhsok and Abel 1997; Abel and Müller-Myhsok 1998) is obtained by multiplying by 3 the required number of families with affected singletons and two parents

[c] See text for explanation of these terms

represent the sizes in indirect approaches. For multiplicative MOI, the required sample size for δ/δ_{max} of 0.5 is about four times as large as that for δ/δ_{max} of 1. This tendency is not changed when a dominant, recessive, or additive MOI is assumed (data not shown).

The influence of the extent of linkage disequilibrium and the marker allele frequency on the number of individuals required in genome-wide association studies with a case-control design is shown in Fig. 1. Figure 1a represents the sample size necessary for a direct approach. When 1000 DNA samples are analyzed, we can identify the true disease variants with $\gamma > 2$ with a power of 0.8. If $m = p$ and $\delta/\delta_{max} = 0.5$ (Fig. 1b), marker alleles linked to disease variants with $\gamma > 3$ are likely to be detected with 1000 samples in an indirect approach. The influence of the difference in allele frequencies between the marker and the variant on the number of individuals required is shown in Fig. 1c, where the frequency of the marker allele B is fixed at 0.25, and $\delta/\delta_{max} = 1$. This result clearly shows the difficulty in detecting significant difference in the marker allele frequencies between cases and controls, unless the frequency of the disease variant A is close to 0.25. Moreover, when $m = 0.25$ and $\delta/\delta_{max} = 0.5$ (Fig. 1d), the ranges of p and γ in which associations can be found when 1000 samples are used are extremely restricted. From a comparison between Fig. 1a and Figs. 1b–d, we can see that indirect genome-wide association studies require a large number of individuals for the detection of disease variants with a modest effect.

In indirect genome-wide association studies, marker alleles with a large frequency will be used to identify disease variants, because we have no prior knowledge of the frequencies of the disease variants. However, this strategy has its disadvantages. For example, in the case of $p = 0.01$, $\gamma = 6$, $m = 0.25$, and $\delta/\delta_{max} = 1$, more than 10,000 individuals must be screened to detect a significant association with a power of 0.8. Furthermore, when $p = 0.01$, $\gamma = 6$, $m = 0.25$, and $\delta/\delta_{max} = 0.5$, more than 40,000 individuals are required. Such requirements are hard to satisfy. We should note that rare disease variants may be missed by SNP markers whose minor allele frequency is large.

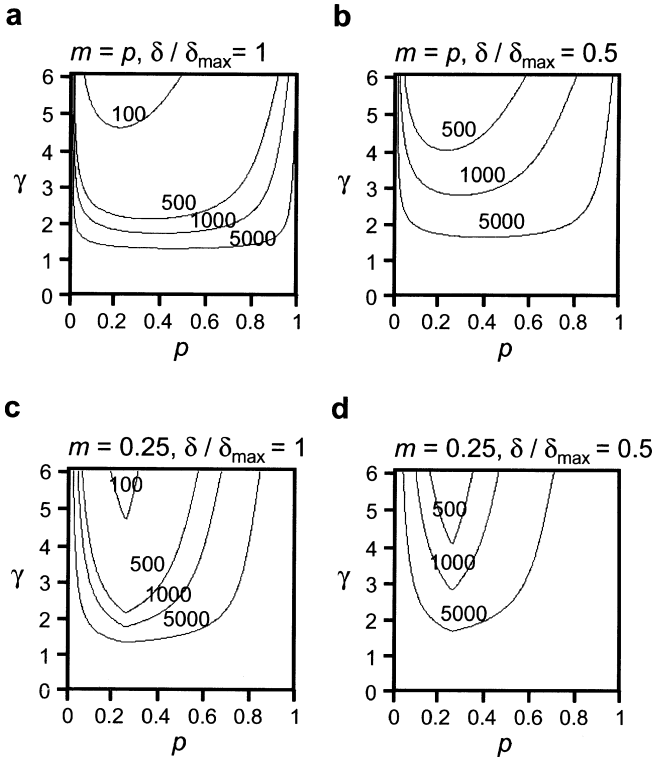

Fig. 1a–d. Number of individuals needed to detect association for a type I error of 1.25×10^{-7} and a power of 0.8. It is assumed that $f_{aa} = 0.01$, and it is assumed that there is a multiplicative mode of inheritance (MOI). The sample size, $n_1 + n_2$ ($n_1 = n_2$), is given as a function of p and γ. The curves are drawn for $n_1 + n_2 = 100$ ($n_1 = n_2 = 50$), $n_1 + n_2 = 500$ ($n_1 = n_2 = 250$), $n_1 + n_2 = 1000$ ($n_1 = n_2 = 500$), and $n_1 + n_2 = 5000$ ($n_1 = n_2 = 2500$). When the association cannot be detected with a power of 0.8 within the ranges of the parameters, the curves are not represented. **a** $m = p$ and $\delta/\delta_{max} = 1$ (the optimal case); **b** $m = p$ and $\delta/\delta_{max} = 0.5$; **c** $m = 0.25$ and $\delta/\delta_{max} = 1$; **d** $m = 0.25$ and $\delta/\delta_{max} = 0.5$. See text for explanation of all terms

Discussion

Our results lead us to conclude that there is a notable statistical limitation of indirect genome-wide association studies using SNP markers, compared with direct studies, even when ideal case-control samples are analyzed. In general, most disease variants are in positive linkage disequilibrium with major alleles at the nearby markers. When there is a large difference in the allele frequencies of the disease variant and the SNP marker, it is difficult to detect the association with a power of 0.8, using this marker (Fig. 2). This is the most serious problem in indirect genome-wide association studies. It is obvious that the optimum case for the detection ($m = p$, and $\delta/\delta_{max} = 1$) is very rare, even if an SNP marker is situated within a susceptibility gene. As the number of SNP markers with various allele frequencies is increased, the possibility that one of the markers is equivalent ($m = p$, and $\delta/\delta_{max} = 1$) or almost equivalent to the true

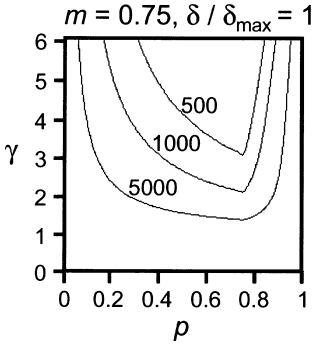

Fig. 2. Number of individuals needed to detect association for a type I error of 1.25×10^{-7} and a power of 0.8. It is assumed that $f_{aa} = 0.01$, $m = 0.75$, and $\delta/\delta_{max} = 1$; a multiplicative MOI is also assumed. The sample size, $n_1 + n_2$ ($n_1 = n_2$), is given as a function of p and γ. The curves are drawn for $n_1 + n_2 = 500$ ($n_1 = n_2 = 250$), $n_1 + n_2 = 1000$ ($n_1 = n_2 = 500$), and $n_1 + n_2 = 5000$ ($n_1 = n_2 = 2500$). Because the association cannot be detected with a power of 0.8 within the ranges of the parameters, the curve for $n_1 + n_2 = 100$ ($n_1 = n_2 = 50$) is not represented. See text for explanation of all terms

disease variant increases. However, the obtained *P* value may not reach significance level due to the Bonferroni adjustment. It is, therefore, uncertain whether indirect association studies using a very dense map of SNPs can detect the SNP marker located close to the true disease variant. Collins et al. (1999) stated that the number of SNP markers needed for genome scan could be reduced to 30,000 (1 SNP per 100 kb) or less, based on the theoretical calculation of the extent of the linkage disequilibrium in the human genome. However, as mentioned above, whether susceptibility genes can be detected is largely dependent on the difference in allele frequencies between the disease variant and the SNP marker. If we consider the difference in allele frequency, it seems that 30,000 SNP markers would be far from satisfactory in genome-wide association studies of complex disease genes.

Although most complex disease variants would be SNPs that alter amino acids in coding regions, or SNPs in regulatory regions that are involved in controlling gene expression levels, small numbers of such functional variants are currently available as candidates for association studies. Figure 1a shows that direct genome-wide association studies with reasonable sample sizes enable us to identify complex disease variants with $\gamma > 2$, when the number of candidate variants, including the true disease variants in the human genome, is less than 200,000. Thus, we should identify functional variants in as many genes as possible for use in genome-wide association studies.

Acknowledgments This study was supported by a Grant-in-Aid for Scientific Research from the Ministry of Education, Science, Sports, and Culture of Japan.

References

Abecasis GR, Noguchi E, Heinzmann A, Traherne JA, Bhattachryya S, Leaves NI, Anderson GG, Zhang Y, Lench NJ, Carey A, Cardon LR, Moffatt MF, Cookson WOC (2001) Extent and distribution of linkage disequilibrium in three genomic regions. Am J Hum Genet 68:191–197

Abel L, Müller-Myhsok B (1998) Maximum-likelihood expression of the transmission/disequilibrium test and power considerations. Am J Hum Genet 63:664–667

Collins A, Lonjou C, Morton NE (1999) Genetic epidemiology of single-nucleotide polymorphisms. Proc Natl Acad Sci USA 96:15173–15177

Collins FS, Guyer MS, Chakravarti A (1997) Variations on a theme: cataloging human DNA sequence variation. Science 278:1580–1581

Goddard KAB, Hopkins PJ, Hall JM, Witte JS (2000) Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. Am J Hum Genet 66:216–234

Kruglyak L (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet 22:139–144

Morton NE, Collins A (1998) Tests and estimates of allelic association in complex inheritance. Proc Natl Acad Sci USA 95:11389–11393

Müller-Myhsok B, Abel L (1997) Genetic analysis of complex diseases. Science 275:1328–1329

Ohashi J, Yamamoto S, Tsuchiya N, Hatta Y, Komata T, Matsushita M, Tokunaga K (2001) Comparison of statistical power between 2 × 2 allele frequency and allele positivity tables in case-control studies of complex disease genes. Ann Hum Genet 65:197–206

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517

Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506–516