Am. J. Hum. Genet. 64:629–640, 1999

Comparison of the Power and Accuracy of Biallelic and Microsatellite Markers in Population-Based Gene-Mapping Methods

Momiao Xiong and Li Jin

Human Genetics Center, University of Texas—Houston Health Science Center, Houston

Summary

Because of their great abundance and amenability to fully automated genotyping, single-nucleotide polymorphisms (SNPs) and simple insertion/deletion are emerging as a new generation of markers for positional cloning. Although the efficiency and cost associated with the markers are important in the mapping of human disease genes, the power to detect the linkage between the marker and the disease locus, as well as the accuracy of the estimation of the map location of the disease gene, dictate the selection of the markers. Both the power and the accuracy depend not only on the type of the markers but also on other factors, such as the age of the disease mutation, the magnitude of the genetic effect, the marker-allele distribution in the population, mutation rates of marker loci, the frequency of the disease allele, the recombination fraction, and the methods for mapping the human disease genes. In this article, we develop a mathematical framework and the analytical formulas for calculation of the power and the accuracy and investigate the impact that the aforementioned factors have on the power and the accuracy, by using two population-based gene-mapping methods—likelihood-based linkage-disequilibrium mapping and the transmission/disequilibrium test, for both biallelic SNPs and microsatellites. These studies provide not only guidance in selection of the markers and in the design of the sample scheme for positional cloning but also insight into the biological bases of the mapping of human disease genes.

Introduction

Positional cloning has emerged as one of the major tools for identification of genes involved in various human diseases for which the biochemical nature is unknown (Collins 1992, 1995). It pinpoints disease genes in the human genome by testing the linkage between the markers and the tentative disease locus.

Genetic markers play an important role in the localization of human disease loci in positional cloning. Microsatellites, referred to as the "second-generation markers," have been the markers of choice since 1989, and several thousands such polymorphic markers have been developed (Litt and Luty 1989; Weber and May 1989; Edwards et al. 1992; Weissenbach et al. 1992; Gyapay et al. 1994; Dib et al. 1996). Characterized by high levels of heterozygosity and by a large number of alleles, such markers provide ideal tools for pedigree-based linkage analysis. Their applications have led to the identification of the genes involved in many monogenic and a few polygenic diseases (Collins 1995). However, this recent advent of mutation detection and highly efficient genotyping technologies (Oefner and Underhill 1995; Chee et al. 1996) has prompted the emergence of a new generation of markers mostly based on single-nucleotide polymorphisms (SNP) and simple insertion/deletion (Jin et al. 1995; Kwok et al. 1996; Wang et al. 1998). Although SNP markers, usually biallelic, are relatively less polymorphic than microsatellites, their great abundance and accessibility to high-throughput low-cost automated genotyping technologies may eventually lead to the replacement of microsatellites in positional cloning (Jin et al. 1995).

The essence of the mapping of human disease genes is to identify genomic regions that cosegregate with disease traits either in pedigrees or in populations, while excluding the rest of the genome, on the basis of the presence of meiotic recombinations between markers and diseases loci. On the basis of the source of information, statistical approaches to gene-mapping methods can be classified into two categories: pedigree-based methods and population-based methods. The former group of approaches include those using recombination information derived at the pedigree level, such as classical linkage analysis (Morton 1955; Ott 1991) and sib-pair analysis (Risch 1990*a*, 1999*b*, 1990*c*). The latter group of approaches, which includes the various types of linkage-disequilibrium method (LDM [Bodmer

Received December 15, 1997; accepted for publication December 14, 1998; electronically published January 29, 1999.

Address for correspondence and reprints: Dr. Momiao Xiong, Human Genetics Center, University of Texas—Houston, Houston, TX 77225. E-mail: mxiong@utsph.sph.uth.tmc.edu

0002-9297/99/6402-0035$02.00

1986]) and the transmission/disequilibrium test (TDT [Spielman et al. 1993]), derive recombination information at the population level.

Although the efficiency and cost of positional cloning associated with certain types of markers are important, the statistical power to detect the linkage between the markers and the diseases locus and/or the accuracy in the estimation of the map location of the disease loci dictates the selection of the markers. The former is essential in a genomewide scan, whereas the latter depicts the usefulness of markers in fine-scale mapping, and both of them depend on the statistical methods employed for gene mapping. Therefore, the comparison of biallelic SNP markers and microsatellites is only meaningful in the context of each particular method. The purposes of this report are (1) to compare the statistical power of biallelic markers in the detection of the presence of disease loci, versus that of microsatellites in two population-based gene-mapping methods, the LDM and the TDT; and (2) to compare the accuracy of biallelic markers in the estimation of the map location of disease loci, versus that of microsatellites in the population-based fine-scale mapping method, the LDM.

The high levels of variation associated with microsatellites are introduced by their high mutation rates (Edward et al. 1992; Weber and Wong 1993). This feature of microsatellite markers makes them ideal markers for pedigree-based linkage analyses, because of the abundance of heterozygotes in the population. Concerns have been raised about the application of such markers in population-based linkage analyses, including those used in whole-genome scan and in fine-scale mapping in which the mutation rate of a marker becomes comparable with the recombination rate between the marker and the disease locus (Jin et al. 1995). In contrast, biallelic SNP markers have very low mutation rates (Li 1997), and, in fact, the probability of recurrent and forward-backward mutation can generally be ignored. This group of markers, intuitively, make ideal markers for population-based mapping approaches.

Both the power and the accuracy of each population-based method depend on the age of the disease mutation, the magnitude of genetic effect of the disease allele, the type of markers, the allele-frequency distribution at the marker locus and at the disease locus in the population, and the recombination fraction between these two loci. These factors play an important role in determining the power as well as the cost of mapping projects. Therefore, in this report, studies are also conducted (1) to evaluate the impact of these factors on both the power and the accuracy of each method using certain types of markers (either biallelic SNP markers or microsatellite markers) and (2) to provide guidance to the design of disease gene–mapping projects. For convenience of presentation, throughout the paper we make the following assumptions: (1) the population is homogeneous; (2) mating is random in the population, and there is a constant population size during evolution; (3) generations are nonoverlapping; (4) all alleles at the disease locus are selectively neutral; (5) there are no phenocopies; and (6) only a single-gene disease model is considered.

Comparison of Power for Biallelic SNPs and Microsatellite Markers, in the LDM

The LDM is emerging as one of the major fine-scale tools in the mapping of genes involved in genetic diseases (Hästbacka et al. 1992, 1994; Jorde 1995; Kaplan et al. 1995; Xiong and Guo 1997). The LDM localizes disease loci when allele-frequency distributions of nearby markers differ between patients and controls. When a disease mutation is first introduced into a population, it creates a complete disequilibrium between the disease locus and its nearby marker locus. In subsequent generations, the disequilibrium reduces, because of recombinations between the markers and the disease locus. The degree of the linkage disequilibrium between a marker and the disease locus reflects the distance between the marker and disease locus and therefore can be used to map the location of the disease locus.

We consider a disease locus with two alleles, a disease-predisposing allele, $D_1 = D$, and an alternative allele, $D_2 = n$, with allele frequencies p_D and p_n, respectively, and a marker locus M with alleles M_i, $i = 1, \ldots, m$, having allele frequencies p_i ($\Sigma_{i=1}^{m} p_i = 1$). Let p_{i_d} and p_{i_n} be the frequency of the marker allele M_i on the haplotypes with allele D and on the haplotypes with allele n, respectively. Then, $p(M_iD) = p_{i_d}p_D$, $p(M_in) = p_{i_n}p_n$, and $p_i = p_{i_d}p_D + p_{i_n}p_n$,where $p(M_iD)$ and $p(M_in)$ are the frequencies of the haplotypes M_iD and M_i, respectively. If alleles M_i and D occur independently of each other, then haplotype M_iD will occur at frequency p_ip_D, and the alleles are said to be in linkage equilibrium.

Assume that the respective penetrance of the genotypes DD, Dn, and nn at the disease locus are f_{11}, f_{12}, and f_{22}, respectively, with $f_{11} \geqslant f_{12} \geqslant f_{22} \geqslant 0$ for recessive ($f_{11} = x$ and $f_{12} = f_{22} = 0$), additive ($f_{11} = x$, $f_{12} = \frac{x}{2}$, and $f_{22} = 0$), and dominant ($f_{11} = f_{12} = x$ and $f_{22} = 0$) cases ($0 < x \leqslant 1$), respectively. Let θ be the recombination fraction between the marker locus and the disease locus. Furthermore, let TM denote an allele transmitted from a parent to an affected child (A). Then (Sham and Curtis 1995),

$$p(TM = M_i|A) = p_i[1 + B(1 - \theta)(\delta_{1i} - 1)] \ , \qquad (1)$$

where

$$B = \frac{[(f_{11} - f_{22})p_D^2 + (f_{12} - f_{22})p_Dp_n]}{(f_{11}p_D^2 + 2f_{12}p_Dp_n + f_{22}p_n^2)} \ , \qquad (2)$$

and $\delta_{1i} = p(M_iD)/p_ip_D$, a measure of linkage disequilibrium between the marker locus and the disease locus. Because of evolutionary forces such as random drift and recombination, the frequency of the haplotype M_iD is a random variable. Thus we have (Xiong and Guo 1997)

$$E[\delta_{1i}] = \frac{p_{i_d}(0)}{p_i}e^{-\theta t} + 1 - e^{-\theta t} , \qquad (3)$$

where $p_{i_d}(0)$ is the frequency of the marker allele M_i in the disease population at the moment of the occurrence of the most recent disease mutation and is determined by disease-marker association patterns. For a single disease mutation, if we assume that M_1 is an associated allele with the disease mutation, we have $p_{1_d}(0) = 1$ and $p_{i_d}(0) = 0$, $i = 2, \ldots, m$. In general, however, for multiple disease mutations, the number of alleles associated with the disease mutations may be >1, and $p_{i_d}(0)$, $i = 2, \ldots, m$, may not be 0. For simplicity of presentation, the calculations throughout this report are made on the assumption that there is a single disease mutation, although the theory can be extended to multiple disease mutations. Let k_{i_d} be the observed number of the allele M_i transmitted from a heterozygous parent to the affected children and let $p(t) = [p_{1_d}, \ldots, p_{m_d}]^T$. Let $\mu_i(\theta) = E[p(TM = M_i|A)]$. Then, from equations (1), (2), and (3) it follows that $\mu_i(\theta) = p_i + B(1 - \theta)[p_{i_d}(0) - p_i]e^{-\theta t}$.

The likelihood function for an LDM using a nuclear family with two parents and an affected child is given by

$$l(\theta) = \prod_{i=1}^{m} \mu_i(\theta)^{k_{i_d}} . \qquad (4)$$

It can be seen from the this formula that the likelihood function is related to the frequencies of the marker alleles in the populations, the penetrance and the frequency of the disease allele, the recombination fraction between the marker and the disease loci, and the age of the disease mutation. This likelihood function allows us to study the power of the LDM, analytically, as will be demonstrated later in this article. The traditional likelihood function for the LDM proposed by Kaplan et al. (1995) and Xiong and Guo (1997) uses a sample of unrelated individuals. It can be shown that this likelihood function is equivalent to that proposed by Kaplan et al. (1995), when the disease mutation is recessive (for proof, see Appendix A).

The power to detect the disease gene, defined as the probability that the disease-susceptibility loci will be detected, is an important index for evaluation of the performance of mapping methods for any given type of markers. Our methods for calculating the power are based on the asymptotic distribution of the test statistic. The LDM uses the likelihood ratio as a test statistic, which is defined as

$$G(\theta) = 2\log\frac{\max_{\theta} l(\theta)}{l(\frac{1}{2})} ,$$

where all logarithms are base e unless otherwise indicated. Since only values of $\theta \leqslant \frac{1}{2}$ are admissible, under the null hypothesis $H_0{:}\theta = \frac{1}{2}$, asymptotically, $G(\theta) \sim \frac{1}{2}\chi^2_{(1)}$. Note that the only situation in which the LDM has 1 df is when all parameters (such as the age of the disease mutation and the initial distribution of the marker allele in the disease population) are known; if the parameters in the model are unknown, then the df of the aforementioned likelihood ratio–based LDM test statistic will be the number of parameters to be estimated.

Since $2G(\theta)$ is distributed as a $\chi^2_{(1)}$, the expected noncentrality parameter $\lambda_m(\theta)$, where the subscript m denotes the number of marker alleles, is given by $2G(\theta)$. To obtain the explicit formula for $\lambda_m(\theta)$, we assume that there are no mutations at the marker locus. This assumption will be released in the examples presented later.

Since $\theta = \frac{1}{2}$, even for a young disease-causing mutation—for example, $t = 20$ generations—$a = e^{-\theta t} = e^{-10} = .000045$. Thus, $\mu_i(\frac{1}{2}) \approx p_i$, $i = 1, \ldots, m$. Note that $E[k_{i_d}] = N\mu_i(\theta)$, where N is the number of parents. It follows from equation (4) that

$$E\left[\log l(\theta) - \log l\left(\frac{1}{2}\right)\right] = E\left\{\sum_{i=1}^{m} k_{i_d}[\log \mu_i(\theta) - \log \mu_i(\frac{1}{2})]\right\}$$

$$= \sum_{i=1}^{m} N\mu_i(\theta)\log\left[\frac{\mu_i(\theta)}{\mu_i(\frac{1}{2})}\right] .$$

If we assume that there is no mutation at the marker locus, the expected value of the noncentrality parameter $\lambda_m(\theta)$ can be approximated by

$$\begin{aligned}
\lambda_m(\theta) &\approx 2E[G(\theta)] \\
&= 4E\left[\log l(\theta) - \log l\left(\frac{1}{2}\right)\right] \\
&= 4N\sum_{i=1}^{m} \mu_i(\theta)\log\frac{\mu_i(\theta)}{\mu_i(\frac{1}{2})} \\
&\approx 4N\sum_{i=1}^{m} p_i\left\{1 + B(1-\theta)\frac{[p_{i_d}(0) - 1]}{p_i}a\right\} \\
&\quad \times \log\left\{1 + B(1-\theta)\frac{[p_{i_d}(0) - 1]}{p_i}a\right\} \\
&= 4NB^2(1-\theta)^2a^2\sum_{i=1}^{m}\frac{[p_{i_d}(0) - p_i]^2}{p_i} . \qquad (5)
\end{aligned}$$

If we assume that there is a single disease mutation, then $\lambda_m(\theta) \approx 4NB^2(1-\theta)^2a^2\frac{1-p_1}{p_1}$.

To evaluate the impact that the type of markers has on the power to detect a disease gene, we fix the sample size, the age of the disease mutation, and the recombination fraction between the marker locus and the disease locus. Under this condition, $\lambda_m(\theta)$ depends only on the initial values of the frequencies of the marker alleles in the disease population and in the unaffected population. For a single disease mutation, $\lambda_m(\theta)$ depends only on distribution of the associated allele in the unaffected population and is not related to the type of markers used, if we assume that there are no mutations at the marker locus.

Clearly, $\lambda_m(\theta)$ increases as p_1 decreases. Since the number of marker alleles is related, to some extent, to the marker-allele distribution in the unaffected population, the number of marker alleles will have impact on both $\lambda_m(\theta)$ and, in turn, the power to detect the disease locus. Let $f(p_1,\ldots,p_m)$ be the density of the distribution of the marker allele $M_1,\ldots,M_m$ in the population. We denote the average of $\lambda_m(\theta)$ over the distribution of m marker alleles in the population by $\Lambda_m(\theta)$.

$$\Lambda_m(\theta) = 4NB^2(1-\theta)^2a^2 \int_0^1 \int_0^{1-x_1} \cdots \int_0^{1-x_1-\cdots-x_{m-1}} \times \sum_1^m \frac{(p_{i_d}(0)-x_i)^2}{x_i} f(x_1,\cdots,x_m)dx_1\cdots dx_m \ .$$

In practice, it would be difficult to specify the density function $f(x_1,\ldots,x_m)$. However, if we assume that there is both a single disease mutation and uniform marker-allele distribution in the unaffected population—that is, $\frac{1}{m}$ for each allele—then, with the uniform distribution of $f(p_1,\ldots,p_m)$, the multiple integration in the formula above can be avoided. Thus, for a biallelic marker, $\lambda_2(\theta) \approx 4NB^2(1-\theta)^2a^2$, and, for a microsatellite marker with m alleles and with mutation being ignored, $\lambda_m(\theta) \approx (m-1)4NB^2(1-\theta)^2a^2$. Clearly, the expected noncentrality parameter increases with the number of alleles. Therefore, for the uniform marker-allele distribution, a microsatellite marker has, in general, more power than a biallelic marker, in the LDM. However, when n biallelic markers cluster together, when the recombination among markers is negligible and we assume that frequencies for all haplotype are identical, $\lambda_m(\theta) \approx 4nNB^2(1-\theta)^2$. This can provide more power than is provided by a microsatellite marker.

Figure 1 shows the expected noncentrality parameter, $\lambda_m(\theta)$ of $\chi^2_{(1)}$, as a function of the population frequency of the associated allele, under the assumption that the frequency of the disease allele is $p_D = .1$. Equation (5) is obtained by assuming no mutation at the microsatellite marker locus. This assumption is released in the computing power of microsatellites that is shown in figure 1, where $u = v = .0001$ and $t = 20$. As expected, $\lambda_m(\theta)$ decreases with the frequency of the allele associated with the disease mutation, p_1. In figure 1, $\lambda_m(\theta)$ for a biallelic marker with $p_2 = 1 - p_1$ is represented by the unbroken line, whereas that for a microsatellite marker with $m = 4$ or $m = 10$ ($p_i = \frac{1-p_1}{m-1}$, $i = 2,\ldots,m$) is represented by the dashed line and the dot-dashed line, respectively. Since the three curves are virtually indistinguishable, it is quite clear that $\lambda(\theta)$ depends on neither the number of alleles nor the frequencies of alleles not associated with the disease mutation. In other words, the power to detect a disease locus in the LDM is dictated by the frequency of the allele associated with the disease mutation, given that the mutation rate of the microsatellite is small compared with the recombination fraction between the marker and disease loci.

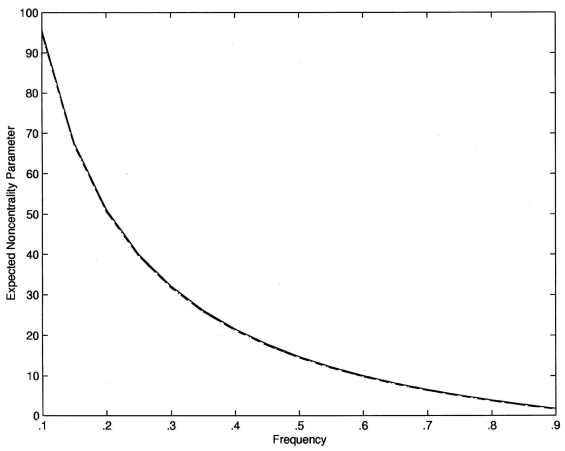

Figure 1 Expected noncentrality parameter, as a function of the population frequency of the associated allele in the LDM. The parameters are set as follows: dominant disease, $t = 20$ generations, $N = 500$, $u = v = .0001$, and $\theta = 5$ cM. Denote the population frequency of the associated allele by f. The population frequencies of other, non-associated alleles are specified as $\frac{1-f}{m-1}$, where m is the number of alleles. The unbroken, dashed, and the dot-dashed lines represent a biallelic marker, a microsatellite marker with 4 marker alleles, and a microsatellite marker with 10 marker alleles, respectively.

Figure 2 shows the power as a function of the recombination fraction between the marker and the disease locus, for biallelic and microsatellite markers in the LDM, given $t = 20$ and $p_D = .1$. In figure 2, the powers of a biallelic marker for $p_1 = .9$, $.5$, and $.1$ are represented by the dotted, dashed, and dot-dashed lines, respectively, whereas those of a microsatellite marker are represented by the left, middle, and right unbroken lines, each corresponding to three different allele frequencies. Note that p_3 is defined as frequency of the associated allele for a microsatellite marker. When the frequency of the associated allele is .5, the power of a microsatellite

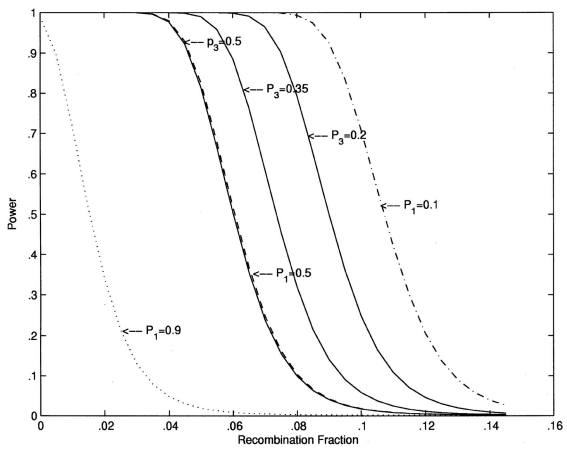

Figure 2 Power as a function of the recombination fraction in the LDM. The parameters are set as follows: recessive disease, $t = 20$, $N = 200$, and $u = v = .0001$. The dashed, dotted, and dot-dashed lines represent the power curves for the biallelic markers when the population frequency of the associated allele is $p_1 = .5$, $p_1 = .9$, and $p_1 = .1$, respectively. The left, middle, and right unbroken lines represent the power curves for the microsatellite when the population frequencies of the allele are .1, .15, .5, .15, and .1; .2, .2, .2, .2, .2, and .2 and .1, .1, .35, .35, and .1, respectively.

marker is slightly lower than that of a biallelic marker, because of the presence of new mutations at the marker locus.

In table 1, we consider both a recessive- and a dominant- disease locus, with a disease-allele frequency of $p_D = .1$ and $\theta = 5$ cM. The sample size is calculated for .8 power with significance level $\alpha = .0001$. Throughout the paper, B_i denotes the biallelic marker with population frequency $p_1 = i \times .1$ of the associated allele, and M_i denotes the microsatellite marker with population frequency $p_3 = i \times .1$ of the associated allele and with population frequency $p_i = \frac{1-p_3}{4}$ of the other, nonassociated alleles. We demonstrate that for, both biallelic and microsatellite markers, the sample size is sensitive to both the age of the disease mutation and the frequency of the associated allele in the population. For a dominant disease, the sample size dramatically increases for both biallelic and microsatellite markers. The age of the disease mutation has a large impact on the sample size. From table 1 we can see that, in all cases, for $t \geq 100$, the sample sizes necessary to achieve .8 power at $\theta = 5$ cM are impractical, for both biallelic and microsatellite markers.

Next, we examine the sample-size requirement when 1,500 biallelic markers are available ($\theta = 2$ cM). The sample sizes required to reach .8 power (significance level $\alpha = .0001$), for various scenarios, are listed at the bottom of table 1, for the three-part assumption of initial complete linkage disequilibrium between the marker and disease loci, no mutations at the marker locus, and $p_D = .01$ as the frequency of the disease allele. Interestingly, the sample size (4,900) is manageable even for the worst case in table 1: $p_1 = .5$, dominant-disease mutation, and $t = 100$.

The number of markers required for a genomewide scan can be studied by computation of the largest recombination fraction allowed between the adjacent markers, to achieve .8 power for a given sample size and marker-allele distribution. The results for the LDM are listed in table 2.

Comparison of Power for Biallelic and Microsatellite Markers, in the TDT

The TDT detects the association between the markers and the disease locus (Spielman et al. 1993) by comparing allele frequencies in patients with those in controls provided by parents. Association between the markers and disease loci may arise from either linkage (with linkage disequilibrium) or, in the absence of linkage, population stratification. The TDT exploits the internal controls provided by parents, to avoid association that is due to artifacts—for example, population stratification—and hence detects linkage between the markers and the disease locus.

For the convenience of discussion of the calculation of power, we assume that the population studied is homogeneous. Let n_{ij} be the number of parents who transmitted allele i to the affected child but did not transmit allele j. The TDT for the multiple allele is defined as

$$Z^2 = \sum_{i<j} \frac{(n_{ij} - n_{ji})^2}{\mathrm{var}(n_{ij} - n_{ji})} \approx \sum_{i<j} \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}} .$$

Let p_{ij} be the probability that alleles M_i are transmitted, conditional on parental genotype M_iM_j. Since n_{ij}, $i,j = 1, \ldots, m$, follows a multinomial distribution, we have $E[n_{ij}] = 2Np_{ij}$ and $\sigma_{ij}^2 = \mathrm{var}(n_{ij} - n_{ji}) = 2N[p_{ij} + p_{ji} - (p_{ij} - p_{ji})^2]$, where N is the number of families with parents and an affected child, including families with homozygous parents. From the formulas given by Sham and Curtis (1995), the probabilities p_{ij} and p_{ji} are calculated as $p_{ij} = p_ip_jd_{ij}$ and $p_{ji} = p_ip_jd_{ji}$, where $d_{ij} = 1 + B[(e_{1i} - 1) + \theta(e_{1j} - e_{1i})]$,

$$B = \frac{p_D[p_D(f_{11} - f_{12}) + p_d(f_{12} - f_{22})]}{p_D^2 f_{11} + 2p_D p_d f_{12} + p_d^2 f_{22}} ,$$

and $e_{1i} = \frac{p_{iD}}{p_D p_i}$ are the linkage-disequilibrium parameters between the disease allele and the marker allele i. Therefore, under the null hypothesis H_0:$\theta = \frac{1}{2}(\delta = 0)$, Z^2 is an asymptotically central χ^2 distribution with $\frac{m(m-1)}{2}$ df. Under the alternative hypothesis of the existence of linkage and linkage disequilibrium, Z^2 is an asymptotically non-

Table 1

Number of Parents Required to Achieve .8 Power with Significance Level $\alpha = .0001$, for Biallelic and Microsatellite Markers in the LDM ($u = v = 0$)

Parameters and Marker	No. of Parents Required, for					
	Recessive Disease			Dominant Disease		
	$t = 20$	$t = 50$	$t = 100$	$t = 20$	$t = 50$	$t = 100$
$\theta = 5$ cM, $p_d = .1$:						
Biallelic:						
B1	36	523	66,7000	110	1,800	238,800
B3	94	1,800	253,600	327	6,300	914,300
B5	194	4,000	590,000	711	14,000	2,000,000
B7	425	9,100	1,400,000	1,600	33,000	4,960,000
B9	1,600	35, 000	5,300,000	6,100	127,000	19,000,000
Microsatellite:						
M1	36	529	60,000	111	1,780	215,000
M5	198	4,300	1,400,000	724	15,700	5,200,000
M9	1,800	68,000	6,000,000	6,900	248,000	7,200,000
$\theta = 2$ cM, $p_d = .01$; biallelic:						
B2	43	98	307	110	220	590
B3	44	120	440	130	280	930
B5	50	170	1350	200	660	4,900

central χ^2 distribution with the following noncentrality parameter:

$$\lambda = 4N^2 \sum_{i<j} \frac{(p_{ij} - p_{ji})^2}{\sigma_{ij}^2} \quad .$$

In Appendix B, we show that, when it is assumed that there are no mutations at the marker locus, $\lambda_m(\theta)$ is given by

$$\lambda_m(\theta) \approx 4NB^2(1 - 2\theta)^2 e^{-2\theta t} \sum_{i=1}^{m} p_i \left[\frac{p_{i_d}(0)}{p_i} - 1\right]^2 \quad . \quad (6)$$

For the single disease mutation, if it is assumed that the marker allele M_1 is an associated allele, then equation (6) reduces$\lambda_m(\theta) \approx 4NB^2(1 - 2\theta)^2 e^{-2\theta t} \frac{1-p_1}{p_1}$.

Since the likelihood ratio–based LDM is a parametric method whereas the TDT is a nonparametric one, the LDM has, in general, a higher power than the TDT. This is reflected by a subtle difference between the noncentrality parameters of the two methods: $(1 - \theta)^2$ for the LDM versus $(1 - 2\theta)^2$ for the TDT. When the marker is located very close to the disease locus, these two terms are close to each other. In this case, the LDM and the TDT will have a similar noncentrality parameter. For a biallelic marker, both the LDM and the TDT have the same df. For a microsatellite marker, although the previously discussed TDT test has $\frac{m(m-1)}{2}$ df, as shown by Sham and Curtis (1995), the TDT test, based on logistic regression, has $m - 1$ df. The df for the LDM depends on the number of parameters to be estimated. If the number of unknown parameters increases, then the df for the LDM increases. In addition, if $\theta = \frac{1}{2}$, then the noncentrality parameter for the TDT is 0, but the noncentrality parameter for the LDM is not. This may imply that the TDT (but not the LDM) is still a valid test even in the presence of a "spurious" linkage disequilibrium that is created not by the linkage but by such artifacts as the substructure of population.

To evaluate the performance of biallelic and microsatellite markers in the TDT, we assume the mutation rates of $u = v = .0001$ at the microsatellite markers. Table 3 shows that the pattern of the sample sizes in the TDT is similar to that in the LDM. The younger the disease mutation, the smaller the sample sizes. For the single disease mutation, the sample sizes depend on the population frequency of the associated allele, regardless of the type of the markers. In table 3, both biallelic and microsatellite markers with the smallest population frequency (.1) of the associated allele require the smallest sample size.

Next we examine the performance of the biallelic and microsatellite markers in the TDT, for the disease models considered by Risch and Merikangas (1996)—that is,

Table 2

Distance between Adjacent Biallelic Markers, to Achieve .8 Power in the LDM ($u = v = 0$ and $p_d = .1$)

Sample Size	Distance (No. of Markers), for		
	$t = 20$	$t = 50$	$t = 100$
50	3.0 cM (1,000)	1.6 cM (1,875)	.8 cM (3,750)
100	7.0 cM (429)	3.0 cM (1,000)	1.4 cM (2,143)
500	15.0 cM (200)	6.0 cM (500)	3.0 cM (1,000)
1,000	18.0 cM (167)	8.0 cM (375)	3.6 cM (833)

Table 3

Number of Nuclear Families Required to Achieve .8 Power with Significance Level .0001, for Biallelic and Microsatellite Markers in the TDT (θ = 5 cM, $u = v$ = .0001, and p_d = .1)

	No. of Nuclear Families Required, for			
	Recessive Disease		Dominant Disease	
Marker	t = 20	t = 100	t = 20	t = 100
Biallelic:				
B1	28	75,000	119	334,000
B3	48	140,000	212	623,000
B5	94	283,000	422	1,260,000
B7	250	761,000	1,127	3,393,000
B9	1,992	6,000,000	8,981	27,000,000
Microsatellite:				
M1	35	59,000	103	193,000
M5	254	1,700,000	791	5,600,000
M9	9,300	2,000,000	28,000	7,300,000

under the assumption that the genotypic relative risk for individuals of genotype Dn is γ times greater than that for individuals with genotype nn and that the risk for individuals with genotype DD is γ^2. The sample sizes required under such disease models are summarized in table 4. The sample size increases as the complexity of the disease, the age of the disease mutation, and the population frequency of the associated allele increase.

The age of the disease mutation and the magnitude of gene effects are primary determinant of our ability to map human disease genes. On the basis of the data in table 4, we can see that, with $\gamma = 4$, the sample sizes for the TDT are still practically workable when the disease mutation is assumed to have occurred 400 years ago.

Comparison of the Accuracy of the Estimation of the Disease-Gene Location by Biallelic and Microsatellite Markers, in Fine-Scale Mapping

The purpose of fine-scale mapping is to localize the disease gene as accurately as possible. Therefore, to evaluate the performance of biallelic and microsatellite markers in fine-scale mapping, we compare the accuracy of the estimation of the disease-gene location by biallelic markers and that by multiallelic markers. Gene-mapping accuracy in the following discussion is defined as the width of the confidence interval (in centimorgans) at a certain confidence level.

The LDM can be used for fine-scale mapping. Let $\mu_i(\theta)$ be the conditional probability that allele M_i is transmitted from the parents to the child, as defined above (see Comparison of Power for Biallelic SNPs and Microsatellite Markers, in the LDM), given that the child is affected. Let $X_j = [x_{1j}, \ldots, x_{mj}]^T$, $j = 1, \ldots, N$, and

$$f(X_j,\theta) = \prod_{i=1}^{m} \mu_i^{x_{ij}}(\theta) \ ,$$

where

$$x_{ij} = \begin{cases} 1 & \text{if } j\text{th child has transmitted allele } i \\ 0 & \text{otherwise} \end{cases} .$$

The likelihood function $l(\theta)$ is defined as

$$l(\theta) = \prod_{j=1}^{N} f(X_j,\theta) \ .$$

Let $\hat{\theta}_n$ be the maximum-likelihood estimate of θ—that is,

$$l(\hat{\theta}_n) = \max_{\theta} l(\theta) \ .$$

On the basis of standard statistical theory (Serfling 1980), the level $1 - \alpha$ confidence interval of θ is given by $\hat{\theta}_n - z_{1-\frac{\alpha}{2}} n^{-\frac{1}{2}} \sigma \leqslant \theta \leqslant \hat{\theta}_n + z_{1-\frac{\alpha}{2}} n^{-\frac{1}{2}} \sigma$, where $z(p)$ denotes

Table 4

Sample Size Required to Achieve .8 Power with Significance Level .0001, for Biallelic and Microsatellite Markers in the TDT (θ = 5 cM and $u = v$ = .0001)

	Sample Size Required, for			
	r = 4		r = 2	
Marker and p_D	t = 20	t = 100	t = 20	t = 100
Biallelic:				
B1:				
.1	488	1,400,000	3,000	9,000,000
.5	75	208,000	237	675,000
B3:				
.1	888	2,600,000	5,700	17,000,000
.5	133	389,000	427	1,300,000
B5:				
.1	1,780	5,309,000	11,480	34,309,000
.5	263	785,000	853	2,544,000
B7:				
.1	4,775	14,296,000	30,850	92,123,000
.5	701	2,114,000	2,284	6,852,000
B9:				
.1	38,029	113,900,000	245,700	733,800,000
.5	5,582	16,840,000	18,190	54,580,000
Microsatellite:				
M1:				
.1	319	676,500	1,624	3,853,000
.5	72	130,740	176	351,860
M5:				
.1	1,398	4,938,000	7,825	28,194,000
.5	279	950,000	734	2,564,000
M9:				
.1	29,868	6,270,000	192,000	35,767,000
.5	4,384	1,207,000	14,280	3,257,000

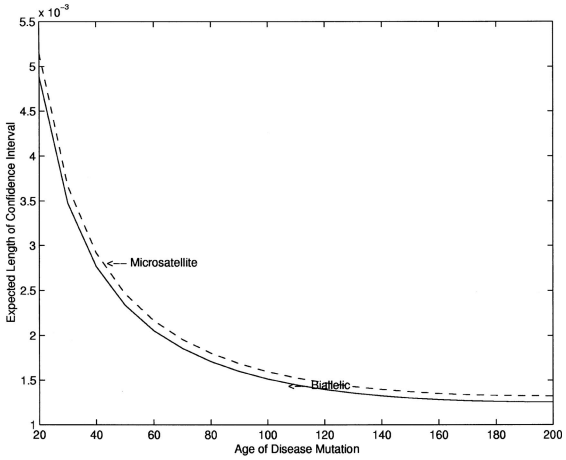

Figure 3 Expected length of the confidence interval, as a function of the age of the disease mutation, for biallelic and microsatellite markers with $\theta = 5$ cM, for a recessive disease and a given sample size: $N = 200$. The frequency of the disease allele is assumed to be $p_d = .1$. The unbroken and dashed lines represent the confidence-interval curves for the biallelic markers when the population frequency of the associated allele is $p_1 = .1$ and those of the microsatellite are .225, .225, .1, .225, and .225.

the pth quantile of the standard normal distribution and σ is calculated as in Appendix C.

For simplicity in this discussion, we assume that there is no mutation for microsatellite markers. In Appendix C, we show that

$$\frac{1}{\sigma^2} \approx B^2[1 + t(1 - \theta)]^2 e^{-2\theta t} \sum_{i=1}^{m} \frac{[p_{i_d}(0) - p_i]^2}{p_i} \ .$$

If we assume that there are a single disease mutation and an associated allele M_1, then $\frac{1}{\sigma^2}$ reduces

$$\frac{1}{\sigma^2} \approx B^2[1 + t(1 - \theta)]^2 e^{-2\theta t} \frac{1 - p_1}{p_1} \ .$$

Clearly, σ^2 depends on the age of the disease mutation, the recombination fraction between the marker and the disease loci, the initial values of the marker-allele frequencies in the disease population, the disease model, and the population's marker-allele distribution. Again, like the power in the detection of linkage, the accuracy of the LDM depends on the frequency of the associated allele, p_1, in the population but not on the type of the markers used, when other factors are identical.

Figure 3 shows that the expected length of the confidence interval decreases with the age of the disease mutation, for both biallelic and microsatellite markers, when $\theta = .005$ and $N = 200$, for a recessive disease. Figure 3 shows that the curve for biallelic markers and that for microsatellite markers are indistinguishable, which implies that the confidence interval depends on the population frequency of the associated allele, regardless of the type of markers. It is interesting to note that a younger disease mutation is associated with a higher power of being detected but with a lower accuracy in the estimation of the gene's map location. This complicates our task. Genomewide scanning and fine-scale mapping of a disease gene are two different phases of a study. A real study is less likely to collect data from a young population for a genomewide scan but data from an old population for fine-scale mapping. It seems reasonable to collect data from a young population—say, an isolated population that was founded by a small number of individuals not long ago—to start a genomewide scan. After a gene has been mapped to a specific region(s) in the genome, more markers in those candidate regions could be employed, for fine-scale mapping, to increase the accuracy.

Figure 4 illustrates the expected length of the confidence interval, as an increasing function of the recombination fraction. When markers are distant from the disease locus, the linkage disequilibrium between the marker and the disease locus decreases. The information about the location of the disease gene contained in the marker will be less, and, hence, the confidence interval will increase as the recombination fraction increases. From the data in figure 4 we can also see that the confidence interval depends on the mode of inheritance (as table 5 also illustrates). The confidence interval for a dominant disease is larger than that for a recessive dis-

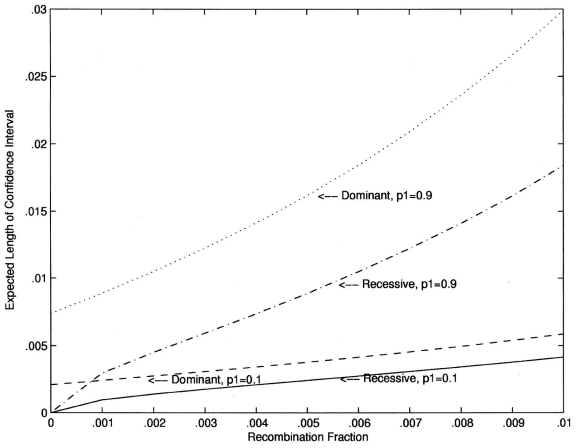

Figure 4 Expected length of the confidence interval, as a function of the recombination fraction. We assume that $t = 100$, $p_d = .001$, and $N = 200$. The unbroken and dashed lines represent the confidence-interval curves for the biallelic markers when the population frequency of the associated allele is $p_1 = .1$, for recessive- and dominant-disease models, respectively; the dot-dashed and dotted lines represent the confidence-interval curves for the biallelic markers when the population frequency of the associated allele is $p_1 = .9$ and $p_1 = .9$, respectively.

ease. The frequency of the associated allele is also a factor that influences the confidence interval. The confidence interval increases as the frequency of the allele that is associated with the disease increases.

Table 5 shows the sample sizes required to achieve the confidence interval of 200 kb with a 95% confidence level, for $t = 200$ generations and $\theta = .005$. Table 5 illustrates that the sample size increases as the population frequency of the associated allele increases, for both biallelic and microsatellite markers. Also from the data shown in table 5, it can be noted that, with the same population frequency of the associated allele, microsatellite markers require a sample size slightly larger than that required by biallelic markers.

Discussion

Although the efficiency and cost of gene mapping associated with certain types of markers are important, the power to detect the linkage between the marker and the disease loci and/or the accuracy in estimation of the map location of the disease loci dictates the selection of the markers in the mapping of the genes involved in various human diseases. In this report, we have investigated how, in the study of an isolated homogeneous population, various factors—such as the type of the markers, the population frequencies of the marker alleles, the recombination fractions between the marker and disease loci, the mode of inheritance, and the age of the disease mutations—affect (1) the power to detect the disease gene, in genomewide scanning, and (2) the accuracy of the estimation of the gene's map location, in the TDT and in the LDM. This study provides much-needed guidance for selection—both of the genetic markers and of the study design—in the mapping of complex diseases, during the process of positional cloning.

It should be noted that such a study was carried out under the simplest and probably the most favorable assumptions for the mapping of human disease genes: (1) the population considered is homogeneous; (2) mating within the population is random, and the population size remains constant during evolution; (3) generations are nonoverlapping; (4) all alleles at the disease locus are selectively neutral; (5) there are no phenocopies; and (6) only a single-gene disease model, in which a disease is caused by mutations that occur within a single gene, is considered. These assumptions may be easily violated in real studies. However, the purpose of this report is to develop some simple analytic formulas for the analysis of gene-mapping methods in the simple cases, which will provide insights into the biological bases of gene-mapping methods and will facilitate investigation of the impact that both the type of the markers and other factors have on the power and resolution of the mapping of human disease genes.

To study the power to detect linkage, we first developed simple analytical formulas for calculation of both the expected noncentrality parameter in χ^2 and the Fisher information index of the estimator of the gene's map location, on the basis of population-genetic theory assuming no mutation at the marker loci. As expected, these formulas clearly describe how the aforementioned factors affect the power and the accuracy of the TDT and the LDM. Then, the power to detect linkage was calculated, with allowance for mutations of microsatellites. These results further demonstrated that the results for the markers with no mutations also hold for the markers with mutations.

The TDT and the LDM show very similar patterns of the dependence of the power on the aforementioned factors. First, the likelihood ratio–based LDM is a parametric method, and the TDT is a nonparametric method based on linkage disequilibrium. When they use the same family structure collection of the data, as in this study, the noncentrality parameter for the LDM and that for the TDT differ only by one term; when the markers are close to the disease locus, the difference between the two methods virtually disappears. For biallelic markers, both the LDM and the TDT will have the same df. In this case, the LDM and the TDT will have similar power for the markers close to the disease locus. Furthermore, the formula of the noncentrality parameter for the TDT demonstrates that the TDT is still a valid test in the presence of the admixture and stratification of populations but that the LDM is not. This implies that the TDT is more robust than the LDM. Unlike the LDM, the TDT also does not need to specify the model. Therefore, the TDT really combines the virtue of simplicity, robustness, and elegance (Curtis 1997).

Second, the power and the accuracy in the TDT and the LDM do not depend on the type of the markers but, rather, on the population frequency of the associated

Table 5

Sample Size for Biallelic and Microsatellite Markers in Fine-Scale Mapping by the LDM ($\theta = .5$ cM and $t = 200$), for a Confidence Interval of 200 kb

	Sample Size Required, for			
Marker	Recessive Disease	Dominant Disease	$r = 4$	$r = 2$
Biallelic:				
B1	215	405	784	4,101
B3	357	760	1,677	11,654
B5	615	1,397	3,284	25,250
B7	1,214	2,884	7,035	56,973
B9	4,214	10,320	25,786	215,590
Microsatellite:				
M1	236	434	824	4,221
M5	728	1,625	3,771	28,531
M9	5,475	13,349	33,244	276,780

allele at the marker locus. A marker with a lower population frequency of associated alleles provides higher power and accuracy.

Third, the power and the accuracy in the LDM and the TDT are reduced by the presence of mutations at the marker loci. In the stepwise-mutation model describing the mutational process at microsatellite loci, the power and the accuracy of the microsatellite markers with mutation are generally lower than those of biallelic SNP markers. However, the difference between those two types of markers, in terms of their power and accuracy, are minor, given that the genetic distance between the marker and disease loci is quite large. This difference may become significant when the distance between the marker and disease loci becomes small—say, <0.1 cM, or 100 kb—when the mutation rate of the microsatellites is not negligible compared with the recombination rate (data not shown).

It is not straightforward to compare the power and the accuracy, in the LDM and the TDT, between biallelic SNP markers and microsatellite markers. For a given population frequency of the associated allele, the power of an SNP marker is slightly higher than that of a microsatellite locus, especially when the distance between the marker and disease loci is small. However, a microsatellite tends to have alleles with lower frequencies, as a consequence of having a larger number of alleles. Therefore, the probability that a microsatellite locus will have an associated allele with lower population frequency is higher than that of a biallelic SNP marker. When the population frequency of the associated allele is unknown, and when the distance between the marker and disease locus is identical and reasonably large, a microsatellite locus is probably a better choice for the population-based approaches to the mapping of human disease genes. However, when a much larger number of biallelic SNP markers can be typed efficiently, especially when several SNP markers can be used to generate haplotype information at a small genomic region of interest, SNP markers outperform microsatellite markers.

Fourth, the power and the accuracy in the LDM and the TDT depend on the age of the most recent disease mutation. Younger populations provide more power but less accuracy. In a real study, we suggest that data be collected from young populations, to start a genomewide scan. After genes have been mapped to specific regions of the chromosomes, more markers should be typed in the promising regions, for fine-scale mapping, to overcome the problem of low resolution caused by a young population.

Fifth, the power and the accuracy in the LDM and the TDT depend on the mode of inheritance. The power and the accuracy will decrease when the complexity of the disease model of inheritance increases. Therefore, the mapping of complex-trait loci is a difficult task and requires a large sample size and a dense genetic map.

Studies of the genetics of complex diseases are currently underway. Many of the alleles associated with complex diseases are likely to be very common but to have low penetrance. The large-scale discovery and scoring of SNPs represents major efforts to facilitate population-based methods for genetic studies of complex diseases (Collins et al. 1997). In this report, to simplify our analysis, we have assumed that the population is homogeneous population, and most calculations were performed for a single disease mutation. However, a population may have substructures. Both population substructure and gene flow among the subpopulations should be considered in our model. Furthermore, complex diseases with multiple disease loci should also be considered.

Acknowledgments

The authors thank Drs. Eric Boerwinkle, Chris Amos, and Julia Krushkal and two anonymous reviewers for their helpful comments on this paper, which helped to improve its presentation.

Appendix A

Clearly, for a recessive disease, from equation (1) we derive

$$\mu_i(\theta) = \theta p_i + (1 - \theta)E[p_{i_d}(t - 1)] \ , \qquad \text{(A1)}$$

where t is the age of the disease mutation in the affected child. It can be shown that

$$E[p_{i_d}(t - 1)] = (1 - \theta)^{t-1} p_{i_d}(0) + [1 - (1 - \theta)^{t-1}]p_i \ . \qquad \text{(A2)}$$

Substitution of p_{i_d} in equation (A2) into equation (A1) yields $\mu_i(\theta) = (1 - \theta)^t p_{i_d}(0) + [1 - (1 - \theta)^t]p_i$, which is the frequency of the marker allele M_i in the current disease population—that is, the probability of transmission of the marker allele M_i to the affected child is equal to the frequency of transmission of the marker allele M_i in the randomly sampled disease population. Therefore, the traditional likelihood function for a linkage-disequilibrium map (Kaplan et al. 1995; Xiong and Guo 1997) is the special case of the formulation above.

Appendix B

Note that, in the absence of mutation at the marker locus, $\frac{p_{i_d}(t)}{p_i}$ is expressible as

$$\frac{p_{i_d}(t)}{p_i} = 1 + \left[\frac{p_{i_d}(0)}{p_i} - 1\right]e^{-\theta t} .$$

Thus, after some calculations, we obtain

$$E[e_{1i} - e_{1j}] = \left[\frac{p_{i_d}(0)}{p_i} - \frac{p_{j_d}(0)}{p_j}\right]e^{-\theta t}$$

and

$$E[e_{1i} + e_{1j} - 2] = \left[\frac{p_{i_d}(0)}{p_i} - 1 + \frac{p_{j_d}(0)}{p_j} - 1\right]e^{-\theta t} ,$$

which implies that

$$E[p_{ij} - p_{ji}] = p_i p_j (1 - 2\theta) B \left[\frac{p_{i_d}(0)}{p_i} - \frac{p_{j_d}(0)}{p_j}\right]e^{-\theta t}$$

and

$$E[\sigma_{ij}^2] = 2Np_ip_j\left\{2 + B\left[\frac{p_{i_d}(0)}{p_i} - 1 + \frac{p_{j_d}(0)}{p_j} - 1\right]\right.$$
$$\left. - p_ip_jB^2(1 - 2\theta)^2\left[\frac{p_{i_d}(0)}{p_i} - \frac{p_{j_d}(0)}{p_j}\right]e^{-2\theta t}\right\} .$$

Therefore,

$$\lambda_m(\theta) \approx 4HN^2 \sum_{i<j} \frac{[E(p_{ij}) - E(p_{ji})]^2}{E(\sigma_{ij}^2)}$$
$$\approx 4HN\left[2B^2(1 - 2\theta)^2 e^{-2\theta t}\right.$$
$$\left. \times \sum_i p_i b_i^2 - B^2 e^{-3\theta t}(1 - 2\theta)^2 \sum_i p_i b_i^3\right]$$
$$\approx 4HNB^2(1 - 2\theta)^2 e^{-2\theta t} \sum_i p_i b_i^2 ,$$

where $b_i = \frac{p_{i_d}(0)}{p_i} - 1$. Similarly, we have $H_i \approx 1 - \Sigma_i p_i^2 - B\Sigma_i p_i^2 b_i e^{-\theta t}$.

Appendix C

Note that, in the absence of mutation at the marker locus, we have $\mu_i(\theta) = p_i + B(1 - \theta)[p_{i_d}(0) - p_i]e^{-\theta t}$ and $\mu_i'(\theta) = -B[1 + t(1 - \theta)][p_{i_d}(0) - p_i]e^{-\theta t}$. Thus,

$$I(\theta) = \sum_{i=1}^{m} \frac{[\mu_i'(\theta)]^2}{\mu_i}$$
$$= B^2[1 + t(1 - \theta)]^2 e^{-2\theta t} \sum_{i=1}^{m} \frac{\frac{1}{p_i}[p_{i_d}(0) - p_i]^2}{1 + B(1 - \theta)e^{-\theta t}\frac{p_{i_d}(0) - p_i}{p_i}}$$
$$\approx B^2[1 + t(1 - \theta)]^2 e^{-2\theta t} \sum_{i=1}^{m} p_i b_i^2 [1 - B(1 - \theta)e^{-\theta t} b_i]$$
$$\approx B^2[1 + t(1 - \theta)]^2 e^{-2\theta t} \sum_{i=1}^{m} p_i b_i^2 .$$

References

Bodmer WF (1986) Human genetics: the molecular challenge. Cold Spring Harbor Symp Quant Biol 51:1–13

Chee M, Yang R, Hubbell E, Berno A, Huang XC, Setern D, Winkler J, et al (1996) Accessing genetic information with high-density DNA arrays. Science 274:610–614

Collins FS (1992) Positional cloning: let's not call it reverse anymore. Nat Genet 1:3–6

——— (1995) Positional cloning moves from perditional to traditional. Nat Genet 9:347–350

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

Curtis D (1997) Use of siblings as controls in case-control associatino studies. Ann Hum Genet 61:319–333

Dib C, Faure S, Fizames C, Samson D, Drouot N, Vignal A, Millasseau P, et al (1996) A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature 380:152-l54

Edwards A, Hammond HA, Jin L, Caskey CT, Chakraborty R (1992) Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. Genomics 12:241–253

Gyapay G, Morissette J, Vignal A, Dib C, Fizames C, Millasseau P, Marc S, et al (1994) The 1993–1994 Généthon human genetic linkage map. Nat Genet 7:246–339

Hästbacka J, de la Chapelle A, Kaitila I, Sistonen P, Weaver A, Lander E (1992) Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. Nat Genet 2:204–211

Hästbacka J, de la Chapelle A, Mahtani MM, Clines G, ReevDaly MP, Daly M, Hamilton BA, et al (1994) The diastrophic dysplasia gene encodes a novel sulfate transporter: positional cloning by fine-structure linkage disequilibrium mapping. Cell 78:1073–1087

Jin L, Underhill PA, Oefner PJ, Cavalli-Sforza LL (1995) Systematic search for polymorphisms in the human genome using denaturing high-performance liquid chromatography (DHPLC). Am J Hum Genet Suppl 57:A26

Jorde LB (1995) Linkage disequilibrium as gene-mapping tool. Am J Hum Genet 56:11–14

Kaplan NL, Hill WG, Weir BS (1995) Likelihood methods for

locating disease genes in nonequilibrium populations. Am J Hum Genet 56:18–32

Kwok P, Deng Q, Zakeri H, Taylor SL, Nickerson DA (1996) Increasing the information content of STS-based genome maps: identifying polymorphisms in mapped STSs. Genomics 31:123–126

Li W (1997) Molecular evolution. Sinauer, Sunderland, MA

Litt M, Luty JA (1989) A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am J Hum Genet 44:397–401

Morton NE (1955) Sequential tests for the detection of linkage. Am J Hum Genet 7:277–318

Oefner PJ, Underhill PA (1995) Comparative DNA sequencing by denaturing high-performance liquid chromatography (DHPLC). Am J Hum Genet Suppl 57:A266

Ott J (1991) Analysis of human genetic linkage. Johns Hopkins University Press, Baltimore and London

Risch N (1990*a*) Linkage strategies for genetically complex traits. I. Multilocus models. Am J Hum Genet 46:222–228

——— (1990*b*) Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am J Hum Genet 46:229–241

——— (1990*c*) Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. Am J Hum Genet 46:242–253

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

Serfling R (1980) Approximation theorems of mathematical statistics. John Wiley & Sons, New York

Sham PC, Curtis D (1995) An extended transmission/disequilibrium test (TDT) for multiallele marker loci. Ann Hum Genet 59:323–336

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

Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, et al (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 280:1077–1082

Weber JL, May PE (1989) Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am J Hum Genet 44:388–396

Weber JL, Wong C (1993) Mutation of human short tandem repeats. Hum Mol Genet 2:1123–1128

Weissenbach J, Gyapay G, Dib C, Vignal A, Morissette J, Millasseau P, Vaysseix G, et al (1992) A second-generation linkage map of the human genome. Nature 359:794–801

Xiong M, Guo S-W (1997) Fine-scale genetic mapping based on linkage disequilibrium: theory and applications. Am J Hum Genet 60:1513–1531