Am. J. Hum. Genet. 70:1480–1489, 2002

Power Calculations for Genetic Association Studies Using Estimated Probability Distributions

Nicholas J. Schork

Department of Psychiatry, University of California at San Diego, La Jolla

The determination of the power of—or of an appropriate sample size for—genetic association studies that exploit linkage disequilibrium requires many assumptions. Some of the more important assumptions include the linkage-disequilibrium strength among alleles at the observed marker-locus sites and a potential trait-influencing locus, the frequencies of the marker locus and trait-influencing alleles, and the ultimate density of the marker locus "map" (i.e., the number of bases between marker loci) necessary in order to identify, with some confidence, trait-influencing alleles. I consider an approach to assessment of the power and sample-size requirements of genetic case-control association study designs that makes use of empirically derived estimates of the distributions of important parameters often assumed to take on arbitrary values. My proposed methodology is extremely general and flexible and ultimately can provide realistic answers to questions such as "How many markers and/or how many individuals might it take to identify, with confidence, a disease gene, via linkage-disequilibrium and association methods from a candidate region or whole genome perspective?" I showcase aspects of the proposed methodology, using information abstracted from the literature.

Introduction

One crucially important issue that faces—and virtually plagues—all researchers interested in identifying or characterizing the effect that some factor (e.g., an exposure, a gene, etc.) has on some outcome (e.g., disease susceptibility, treatment response, etc.) is whether their study is designed in such a way as to maximize success while minimizing use of resources. Power and sample-size calculations are meant to address this question, but they rely on assumptions often so open to doubt that the resulting calculations are unrealistic. This issue is especially problematic for the design of genetic association studies, since they typically involve a number of parameters whose values need to be specified in advance, in order to perform the relevant calculations. These parameters include the frequency of the hypothesized trait-influencing allele, the linkage-disequilibrium strength between the trait-influencing allele and neighboring marker-locus alleles, the effect (i.e., penetrance) of the trait-influencing allele, and the frequencies of alleles at neighboring marker loci.

Although there are situations in which the values of many of these parameters are known with some confidence (e.g., when one is testing a candidate polymorphism of known frequency in a well-characterized region of the genome), they are often extremely difficult to know a priori: consider a mapping study involving many genomic regions with many marker loci and in which there is complete ignorance of the location of the trait-influencing locus relative to those marker loci—here, there is little understanding of at least the trait-influencing locus and its properties.

To avoid making unrealistic assumptions about parameters required in power calculations, one could consider the probabilities associated with the values that these parameters could take on and then, when performing the relevant calculations, consider only those parameter values that have a high probability of occurring. This intuition can be taken one step further, in that one could consider all possible values that a parameter could take on in a power or sample-size calculation and then could weight the results by the probabilities with which those values occur. This strategy is very similar to standard Bayesian statistical procedures, in which the probability distributions of unknown parameters are used to compute quantities that depend on those parameters. This is done mathematically, in Bayesian analyses, by integrating over the unknown values by using the relevant probability distributions.

The primary problem with Bayesian and related strategies is characterization or specification of the distribution of the unknown parameter values or quantities: it may be just as hard, if not more so, to specify a probability distribution for a parameter whose value is unknown a priori as it is to specify (with confidence) a

Received November 16, 2001; accepted for publication March 18, 2002; electronically published April 25, 2002.

Address for correspondence and reprints: Dr. Nicholas J. Schork, Department of Psychiatry, University of California at San Diego, 2062 Basic Science Building, 9500 Gilman Drive, La Jolla, CA 92093-0603. E-mail: nschork@ucsd.edu

0002-9297/2002/7006-0010$15.00

Table 1

Contingency-Table Format for Assessing the Relationship between a Biallelic Locus and a Disease

Allelic Status	Cases	Controls	Total Sample
+	$n_{\mathrm{d}}\,\mathrm{Pr}\,(+\,\vert\,\mathrm{d})$	$n_{\bar{\mathrm{d}}}\,\mathrm{Pr}\,(+\,\vert\,\bar{\mathrm{d}})$	n_+
−	$n_{\mathrm{d}}\,\mathrm{Pr}\,(-\,\vert\,\mathrm{d}) = n_{\mathrm{d}}[1-\mathrm{Pr}\,(+\,\vert\,\mathrm{d})]$	$n_{\bar{\mathrm{d}}}\,\mathrm{Pr}\,(-\,\vert\,\bar{\mathrm{d}}) = n_{\bar{\mathrm{d}}}[1-Pr(+\,\vert\,\bar{\mathrm{d}})]$	n_-
Total	n_{d}	$n_{\bar{\mathrm{d}}}$	N

NOTE.—For definitions, see text and Appendix.

particular value for that parameter. However, there are ways around this. If one has amassed data on a parameter (i.e., has measured its values repeatedly), then it is possible to derive an empirical estimate of the distribution of that parameter; that is, one could merely tally how often the parameter takes on certain values when it is measured and then use these counts or frequencies as estimates of the probabilities that the parameter takes on certain values.

As I describe in this article, this procedure can be pursued for assumption-laden parameter values used in power calculations for genetic association studies involving case-control samples and single-nucleotide polymorphisms (SNPs). I show how data collected on, for example, the frequencies of SNPs and linkage-disequilibrium strength, can be used to compute empirical estimates of their probability distributions, as well as how these distributions can be used to compute more-compelling and more-realistic power and sample-size requirements for genetic association studies. In this sense, it could be said that, concerning a parameter in genetic power calculations, one should pay heed to the adage "When in doubt, integrate it out." For convenience, I provide an Appendix offering a description of some of the mathematical symbols.

My results are very general, and I showcase the proposed calculations, using recently published data, on SNPs and their properties, from the population at large. In addition, the proposed strategy is not limited to genetic applications: any calculation involving an unknown parameter can benefit from consideration of that parameter's empirically derived probability distribution. Of course, the proposed strategy for genetic association studies does need some qualification, so I also will comment on a few issues and potential complications of the proposed strategy.

Material and Methods

Genetic Case-Control Study Designs

Consider a biallelic trait-influencing locus with alleles + and −. These alleles have frequencies, in the population at large, of p and $q = 1 - p$, respectively, and are assumed to be in Hardy-Weinberg equilibrium such that the frequencies of the three possible genotypes at this locus are $f_{++} = p^2$, $f_{+-} = f_{-+} = 2pq$, and $f_{--} = q^2$. The + allele increases susceptibility to a disease. Let $\mathrm{Pr}\,(+|\mathrm{d})$ be the probability that an individual carries the + allele, given that he or she has the disease; similarly, let $\mathrm{Pr}\,(-|\,\bar{\mathrm{d}})$ denote the probability that an individual carries the − allele, given that he or she does not have the disease. Assume that one has ascertained n_{d} individuals with the disease (i.e., "cases") and $n_{\bar{\mathrm{d}}} = cn_{\mathrm{d}}$ individuals without the disease (i.e., "controls"). The total sample size is thus $N = n_{\mathrm{d}} + n_{\bar{\mathrm{d}}}$. The relationship between the + and − alleles and disease status, for the cases and the controls, can be examined through the use of a simple contingency table, which, in expectation, will have the entries presented in table 1. A standard measure of the strength of the association between the trait-influencing locus alleles and disease status is the odds ratio (OR), defined as

$$\mathrm{OR} = \frac{n_{\mathrm{d}}\,\mathrm{Pr}\,(+|\mathrm{d}) \times n_{\bar{\mathrm{d}}}\,\mathrm{Pr}\,(-|\,\bar{\mathrm{d}})}{n_{\bar{\mathrm{d}}}\,\mathrm{Pr}\,(+|\,\bar{\mathrm{d}}) \times n_{\mathrm{d}}\,\mathrm{Pr}\,(-|\mathrm{d})} \ .$$

Linkage Disequilibrium (*LD*)

Assume that, concerning the sample of cases and controls, one does not have allelic or genotypic information regarding the trait-influencing locus but, rather, information regarding a biallelic marker locus near enough (i.e., linked to) the trait-influencing locus and that its alleles, denoted "M" and "m," are in LD with the trait-influencing locus alleles, + and −. In the population at large, the alleles M and m have frequencies of s and $t = 1 - s$, respectively. The alleles are also assumed to be in Hardy-Weinberg equilibrium. The strength of the LD between the + and M alleles can be expressed as the deviation that the frequency of the haplotype implicating + and M has from its expected value of ps. If this deviation is denoted by "δ," then the four possible haplotype frequencies become

$$\begin{aligned} f_{+M} &= ps + \delta \ , \\ f_{+m} &= pt - \delta \ , \\ f_{-M} &= qs - \delta \ , \\ f_{-m} &= qt + \delta \ . \end{aligned} \qquad (1)$$

It can be shown that the conditional probability, $\Pr(M|\mathrm{d})$, that an individual sampled as a case carries the M allele at the marker locus is (e.g., see Schork et al. 2000)

$$\Pr(M|\mathrm{d}) = s + \frac{\delta(\Pr(+|\mathrm{d}) - p)}{\mathrm{p}(1-\mathrm{p})} \ . \tag{2}$$

Similar equations can be derived for the probability that an individual sampled as a control will carry the M allele. These conditional probabilities can be used to investigate the strength of the association between the M and m alleles and disease status among the cases and controls, by substituting them, in table 1, for the conditional probabilities involving the + and − alleles. It should be understood that the marker alleles, when examined for association with the disease, are only "surrogate" alleles for the actual trait-influencing allele. These marker alleles will be good surrogates only to the degree that they are in strong LD with the trait-influencing alleles.

Computing Power

Schlesselman (1982) and others have considered the power of a contingency table–based association study of the type described in table 1. Define

$$\bar{p} = \frac{\Pr(M|\mathrm{d}) + c\Pr\left(M|\,\bar{\mathrm{d}}\right)}{1+c}$$

and $\bar{q} = 1 - \bar{p}$, where, again, c is the ratio of controls to cases. Further, let z_α be the quantile associated with a standard normal distribution for the assumed type I error probability, α , for the study. Define

$$z_\beta = \left\{\frac{n_\mathrm{d}\left[\Pr(M|\mathrm{d}) - \Pr\left(+|\,\bar{\mathrm{d}}\right)\right]^2}{\left[1 + \left(\frac{1}{c}\right)\right]\bar{p}\bar{q}}\right\}^{\frac{1}{2}} - z_\alpha \ .$$

Let Ω be the set of all parameter values assumed in a power calculation for a genetic association study investigating an observed marker locus and a disease outcome as discussed; that is, $\Omega = \{p,s,\mathrm{OR},\delta,\alpha\}$. The power of the proposed sample size can be then be computed as

$$g(\Omega;n_\mathrm{d},n_{\bar{\mathrm{d}}}) = \Pr(Z \leqslant z_\beta|\Omega;n_\mathrm{d},n_{\bar{\mathrm{d}}}) = \int_{-\infty}^{z_\beta} \phi(x|0,1)dx \ , \tag{4}$$

where $\phi(x\,|\,0,1)$ is the standard normal density function (i.e., with mean 0.0 and SD 1.0) evaluated at x. Thus, given assumptions about (1) a number of parameters—that is, the trait-locus allele frequencies, p and $q = 1 - p;$ (2) the effect that the + allele has on disease susceptibility as quantified by the OR; (3) the marker-locus allele frequencies, s and $t = 1 - s;$ (4) the strength of the LD between the + and M alleles, δ; and (5) the type I–error rate, α, one can compute the power of detecting the association between the M and m alleles and disease status for n_d cases and $n_{\bar{\mathrm{d}}} = cn_\mathrm{d}$ controls.

Parameter-Value Distributions

The values that certain parameters (e.g., allele frequencies, LD strength, locus effect size, etc.) take on must be specified in advance of performance of the power calculations. Assumptions about these parameter values are likely to be very arbitrary. For example, it would be difficult to know a priori the frequencies of the marker-locus alleles that are in LD with the trait-influencing alleles in, for example, a whole-genome association study, as well as the LD strength between marker and trait-influencing–locus alleles. If one knew, however, the probability distributions of these parameters, then one could consider how probable certain values that these parameters could take on might be. For example, one could assume that allele frequencies, given that they vary between 0 and 1, follow a beta distribution with a specified mean and variance. However, specifying the mean and variance for this beta distribution could be as arbitrary as specifying a particular allele-frequency value.

One could overcome this difficulty by estimating the distribution of allele frequencies and LD-strength parameters from actual data measuring these quantities. These "empirical" distributions can then be used to assess the probability that certain parameters will take on assumed values. Estimation of allele-frequency distributions is straightforward: one can sample a number of individuals from a relevant population and genotype them at a number of loci similar to (if not the same as) those loci to be used in an association study. Probabilities of certain marker-allele frequencies can then be tallied by simple counting methods. Estimation of LD-strength distributions, however, is not as straightforward, since there are a few theoretical and data-collection caveats that need to be considered, as discussed below.

Accommodating LD Strength

Consider the LD-strength parameter, δ, used in equations (1) and (2). This term can be written, in terms of haplotype and allele frequencies, as $\delta = f_{+M} - ps$. In fact, an often-used parameter for expressing LD strength is $D = \delta = f_{+M} - ps$ (Lewontin 1988; Zapata 2000) . The range of D varies according to the allele frequencies p and s. This makes modeling its distribution difficult. However, one can consider the distribution of a simple transformation of D whose range is between -1 and 1 and is virtually independent (in a theoretical sense) of allele frequencies at the loci involved (Lewontin 1988; Zapata 2000):

$$D' = \begin{cases} \dfrac{D}{\min(sq,pt)} & D > 0 \\ \dfrac{D}{\min(sp,tq)} & D < 0 \end{cases} . \qquad (5)$$

The range of D' invites the following interpretations: values of D' close to -1 suggest that the + allele is not frequently on a chromosome (or does not form a frequent haplotype) with the M allele, whereas values of D' that are closer to 1 suggest that the + allele is frequently on a chromosome with the M allele. Values of D' close to 0 suggest that the + and M alleles are not in LD and therefore that, given their frequencies, they are not associated. It is straightforward to recover D or δ from D'—that is, $\delta = D'\min(sq,pt)$ or $\delta = D'\min(sp,tq)$. Thus, models of the distribution of D' can be easily used to assess the distribution of D or δ.

Empirically Estimating the Distributions of s *and* D′

Because of the phenomenon of recombination, the strength of LD between alleles at two loci that is induced by the proximity (i.e., "linkage") of these loci on a chromosome (in contrast to LD that is induced by, say, admixture) is dictated, to a large degree, by the distance, in base pairs, between those loci. This is the fundamental phenomenon exploited in many gene-mapping studies. Therefore, in determining or estimating the distribution of LD-strength values, one should consider a specified distance between loci. To accomplish this, one could, in a relevant sample of individuals, genotype a number of loci whose interlocus distances are known, compute the LD strength, D', between alleles at those loci, and record and tally the D' values in bins reflecting interlocus distances (i.e., loci separated by 0–5,000 bases, 5,000–10,000 bases, 0–25,000 bases, etc.). A recent article by Reich et al. (2001) discusses empirical allele frequency and LD data of the type envisioned that were gathered from a sample obtained in a U.S. population. Note that this characterization or estimation of the distribution of D' is conditional on the distance between loci.

Once one has recorded a number of allele-frequency and LD values, then empirical estimation of the distributions of these quantities can proceed in a variety of ways. Consider the use of Kernel-based density-function estimators (Silverman 1986). Assume that L total allele frequencies, s, or LD values, D', have been measured. Assume that these values form a data matrix $X = [x_1, \ldots, x_L]$. These empirical values can be used to estimate the probability of any allele-frequency or LD value, denoted "x," via the function

$$\hat{f}(x) = \frac{1}{L}\sum_{l=1}^{L}\frac{1}{h}K\left(\frac{x - x_l}{h}\right) , \qquad (6)$$

where $K(x)$ is a kernel function (such as the standard normal density function) that integrates to 1 and h is a "window width" parameter that will dictate the amount of smoothing to be used in the density estimate. The value of h can be chosen via various means, such as cross-validation (Taylor 1989).

Exploiting the Distributions of s *and* D′

Once one has derived or estimated distributions for relevant parameters, these distributions can be used in the evaluation of the power of a study. First, one can assume that, for a given marker locus, the trait-influencing locus can be virtually anywhere in relation to it on a chromosome. Thus, one can assume that the distance between the marker locus and the trait-influencing locus is uniformly distributed over the chromosome. If a candidate region is studied, then one can assume that the location of the trait-influencing locus, here treated as a variable, follows a uniform distribution within the interval defined by that region. Now, let Ψ be a set of parameters whose values are not set to specific values in a power calculation but whose distributions will be exploited. For example, consider the situation in which the marker allele and LD-strength parameters will not be set in advance and have had their distributions estimated empirically. Then, following equations (4) and (5), $\Psi = \{s,D'\}$ and $\Omega = \{p,\mathrm{OR},\alpha\}$. Further, let B be the maximal distance between the marker and trait-influencing loci. The power of a sample of n_{d} cases and $n_{\bar{\mathrm{d}}}$ controls can then be computed by integrating over the possible values for the parameters in Ψ and the possible distances between the loci, by using the empirical estimates of their distributions:

$$g^*(\Omega,\Psi;n_{\mathrm{d}},n_{\bar{\mathrm{d}}}) = \int_0^1\int_0^B\int_{-1}^1 g(\Omega,\Psi = \{s = y,D' = x\};L = b,n_{\mathrm{d}},n_{\bar{\mathrm{d}}}) \times f_s(x)f_L(b)f_{D'}(y|b)dxdbdy , \qquad (7)$$

where $f_s(x)$ and $f_{D'}(y|b)$ are, respectively, the empirically derived estimate of the probability density for the marker-allele–frequency parameter s evaluated at x and the empirically derived estimate of the density for the LD-strength parameter D' evaluated at y (which, as emphasized earlier, is conditional on the distance between the loci). $f_L(b)$ is the distribution of the number of bases (or distance in some other unit) separating the marker and trait-influencing loci, assumed to be uniform, evaluated at a distance of b bases. Since one is not likely to know the position of a trait-influencing locus relative to a set marker loci, it is important to consider the possibility that the trait-influencing locus is near any of the marker loci. If one knows the positions and allele frequencies of a set of marker loci (e.g., if one is using a preestablished set of markers in a specific population), then one could simply consider the possibility that the trait-influencing locus is near any of them, by summing over the loci, accommodating the allele frequencies at each marker locus in the power calculation, and weighting the outcome by 1 over the number of marker loci (i.e., assigning uniform weight to each locus).

The integrals in equation (7) can be approximated by sums, by using discretized forms (i.e., probability mass functions) of the relevant probability densities. This may actually facilitate assumptions about the distance between the marker and trait-influencing loci. For example, one could assume that two marker loci available for study are separated by a distance of 10 kb and that the trait-influencing locus is between these loci. Then the maximal distance between the trait-influencing locus and either one of the marker loci is 5 kb. One could then assume that the trait-influencing locus is either 0–1, 1–2, 2–3, 3–4, or 4–5 kb away from a marker locus, with a probability of 1/5 for each possibility. Probability distributions for LD strength could also be computed from actual data, for loci separated at distances of 0–1, 1–2, 2–3, 3–4, and 4–5 kb. These distributions can be approximated by tallying the frequency of D' values falling into bins of, for example, $\{-1.0,-0.8\}$, $\{-0.8,-0.6\}$, $\{-0.6,-0.4\}$, $\{-0.4,-0.2\}$, $\{-0.2,0.0\}$, $\{0.0,0.2\}$, $\{0.2,0.4\}$, $\{0.4,0.6\}$, $\{0.6,0.8\}$, and $\{0.8,1.0\}$. These frequencies can then provide approximate probabilities for LD-strength values. If one further tallies allele frequencies into similar bins, of $\{0.0,0.2\}$, $\{0.2,0.4\}$, $\{0.4,0.6\}$, $\{0.6,0.8\}$, and $\{0.8,1.0\}$, then power calculations can be made by using the sums

$$g^*(\Omega,\Psi;n_{\mathrm{d}},n_{\bar{\mathrm{d}}}) = \sum_{x=1}^{5}\sum_{b=1}^{5}\sum_{y=1}^{10} g\Big(\Omega,\Psi = \{s[x],D'[y]\};L[b],n_{\mathrm{d}},n_{\bar{\mathrm{d}}}\Big) \times p(x)p(b)p(y|b)\ ,$$

where the sums are over the different bins for the allele frequency s, interlocus distance L, and LD strength D' parameters; $s[x]$, $L[b]$, and $D'[y]$ denote translation of a bin into a parameter value (e.g., when $s = 1$, the allele-frequency value is in the interval $\{0.0–0.2\}$; when $L = 5$, the interlocus-distance parameter is in the interval $\{4–5\}$, etc.); and $p(x)$ and $p(y|b)$ are the empirically derived probabilities that the allele frequency and LD strength take on the specified values (or range of values) and that $p(L)$ is a uniform discrete distribution (i.e., $p(L) = 1/5$ in this example).

Simple Extensions

The use of theoretical or empirically derived estimates of densities and probability functions as a way to consider the possible values that parameters could take on for power calculations can be extended to accommodate any (or even all) of the parameters required for such calculations. Consider that one might have intuitions about the distribution of the frequency of trait-influencing alleles or even about the effect of a trait-influencing locus. For example, Orr (1998) has recently considered the distribution of trait-influencing–allele effects from an evolutionary perspective. Alternatively, Mackay et al. (1995, 1996) have identified a number of genes and loci that influence bristle number in drosophila and whose effect sizes can be used to estimate the distribution of locus-effect sizes. Of course, just how generalizable such distributions are to cases involving human diseases is an open question.

Results

Simple Single-Locus Power Calculations

Figures 1 and 2 depict standard normal density-kernel–based estimates (i.e., see eq. [6]) of the distribution of marker-allele frequencies and LD-strength values obtained from the data discussed by Reich et al. (2001). The window-width parameter h used in this estimation procedure was chosen via cross-validation (Taylor 1989). Figure 1 makes it clear that Reich et al. (2001) studied only common SNP alleles, as they mention in their article. Figure 2 clearly shows that, for loci closely spaced (e.g., 0–5

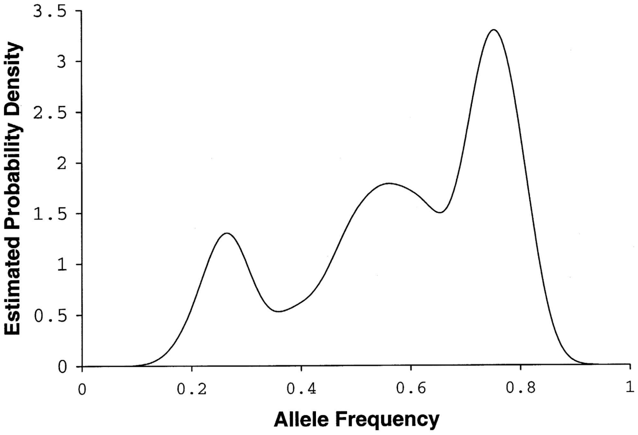

Figure 1 Estimated probability density for the frequency of SNPs, when the data reported by Reich et al. (2001) are used.

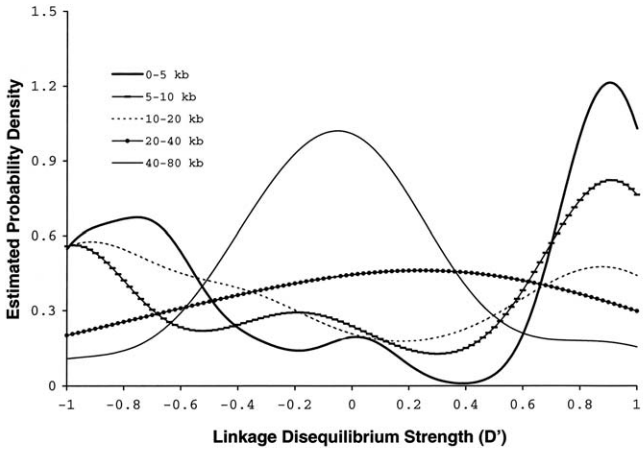

Figure 2 Estimated probability densities for LD strength between SNPs with different interlocus distances, when the data reported by Reich et al. (2001) are used. The different curves reflect different interlocus-distance bins.

kb), there is a high probability that the alleles will show near-perfect negative or positive LD. This is in stark contrast to loci separated by large distances (e.g., 40–80 kb), in which case there is a greater probability that the alleles will show weak or no LD.

For an assumed value of the locus-effect size (i.e., OR), the frequencies of the trait-influencing alleles (i.e., p and q), and the number of bases separating the marker and trait-influencing alleles, I evaluated equation (7), using the estimated distributions of marker-allele frequencies and LD values offered in figures 1 and 2, for various numbers of cases and controls assumed to be equal (i.e., $c = 1$). Figures 1 and 2 reflect the number of chromosomes to be studied (i.e., twice the number of individuals). I assumed that the trait-influencing allele, +, had a frequency of 0.25 and penetrances of $p(++) = 0.5$, $p(+-) = 0.25$, and $p(--) = 0.0$. This resulted in an OR for the disease, given that an individual carries the + allele, of 6.82. The results are described in figure 3. I offer this graph as an example of relevant calculations. A program available from the author can be used to compute power for different situations. It is clear from figure 3 that a marker closer to a trait-influencing locus will result in greater power, as expected.

Genomewide Association Studies

Now consider the use of equation (7) to evaluate the power of genomewide association studies that make use of different map densities (i.e., uniform intermarker distances, in bases). Again, make use of the data reported by Reich et al. (2001) and of the estimated distributions for marker-allele frequencies and LD values given in figures 1 and 2. One important issue in the consideration of genomewide studies of this type is the assumed type I–error rate: since the marker loci are likely to produce correlated association-test statistics, given that they might have alleles in LD, the test statistics that they produce cannot be considered independent. However, to be conservative, assume independence and invoke a simple Bonferroni correction strategy for multiple comparisons. Thus, for a map density assuming both a marker every 10 kb and an autosomal genome length of 3 billion bases, a total of 300,000 association tests would be performed. A nominal genomewide type I–error rate of 0.05 thus would require individual association-test statistic P values $<0.05/300,000$, or 0.000000167, in order to be considered statistically significant. Table 2 offers type I–error rates for different map intermarker locus distances. It should be understood that the maximal distance between the trait-influencing locus and a flanking marker locus would be half the distance between the marker loci.

I considered two scenarios for a hypothetical trait-influencing locus. The first scenario assumed that the trait-influencing allele, +, had a frequency of 0.25 and penetrances of $p(++) = 0.5$, $p(+-) = 0.25$, and $p(--) = 0.0$, for an OR of 6.82, as in the single-locus calculations discussed above. The second scenario assumed that the trait-influencing allele, +, had a frequency of 0.25 and penetrances of $p(++) = 0.20$, $p(+-) = 0.15$, and $p(--) = 0.10$, for an OR of 1.53. Figures 4 and 5 offer power curves for these two different scenarios. It is clear that either (1) a very large sample size will be necessary, for studies involving a sparse map or (2) one must use a

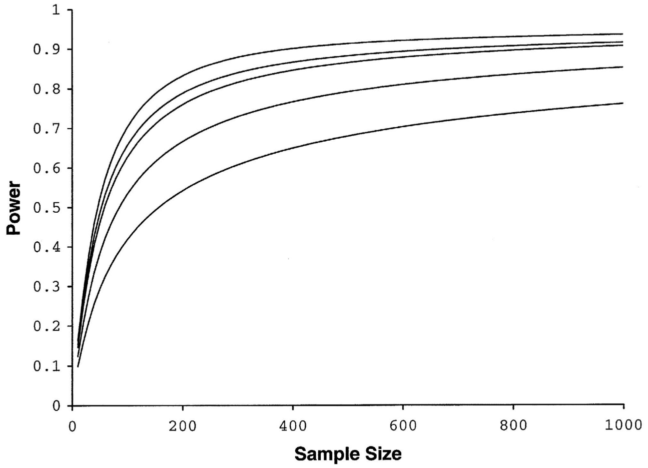

Figure 3 Power curves for association studies involving a single marker locus. A type I error of 0.05 was assumed. "Sample Size" indicates the number of chromosomes needed. The curves, from top to bottom, assume that the distance between the marker locus and the trait-influencing locus is 5, 10, 20, 40, and 80 kb, respectively. The trait-influencing allele, +, was assumed to have a frequency of 0.25 and penetrances of $p(++) = 0.5$, $p(+-) = 0.25$, and $p(--) = 0.0$.

Table 2

Type I–Error Rates for Genomewide Association Studies Involving Different Map Interlocus Distances, When a 3 Billion–bp Genome Is Assumed

Interval[a]	No. of Markers	*P*	Critical Value
10	300,000	.000000167	5.103
20	150,000	.000000333	4.971
40	75,000	.000000667	4.835
80	37,500	.000001333	4.695
160	18,750	.000002667	4.551

NOTE.—Calculations are based on the use of a simple Bonferroni correction with nominal type I–error rate of 0.05.

[a] Interlocus distance (in kb).

reasonably dense map, if a moderate sample size is used. This is especially the case for the detection of a locus with weak to moderate effect (compare figs. 4 and 5). I consider some caveats of these calculations in the "Discussion" section, below.

Discussion

The consideration and calculation of sample-size requirements and/or power calculations for genetic studies plays an extremely important role in putting such studies into perspective. It can often be the case that studies are pursued for which both the statistical-analysis methods and the assumed sample size are woefully underpowered. In determining requisite sample sizes and power, however, a researcher must make a number of assumptions about factors and parameters, such as allele frequencies, LD strength, marker density, etc., as the present article has tried to make clear. Lack of insight into the appropriate values that these parameters should take on undoubtedly contributes to the lack of success and irreconcilability of many studies. I have adopted an approach to assessing the power of a proposed genetic case-control association study that builds off the adage that is well known among Bayesians: "when in doubt, integrate it out" (where "it" is a parameter that must take on assumed or unknown values).

In the estimation of the distributions of unknown parameters from actual data, however, there are some things that need to be considered. I have listed and mentioned many of these below and encourage more inquiry into their impact and consequences.

Nature of the Data Used to Assess Parameter Distributions

I have used the data generated by Reich et al. (2001), which needs some qualification. Reich et al. clearly set out to study common alleles, as evidenced by the graph in figure 1. In addition, they studied LD, using prespecified (i.e., nonrandom) intermarker distances. These facts are likely to create some biases in my examples of power calculations (which really should be seen only as examples). The effect of the presence of common marker alleles in a map will likely reduce power for the detection of rare trait-influencing alleles, since matching of allele frequencies between marker and trait-influencing alleles leads to the greatest power (Schork et al. 2000). The effect that prespecified distances have on LD calculations will also result in downward biases in my power calculations, since, for example, it is highly conservative to use LD-strength values for markers separated by 80 kb as being indicative of LD-strength values for markers separated by 40–80 kb.

Population Dependence of Data

It is widely known that allele frequencies and LD strengths vary considerably from population to population (Goddard et al. 2000). Thus, the empirically derived probability distributions for certain parameters from data collected from these populations—and, hence, any power calculations based on them—may not be generalizable to other populations. One could conceivably use population-specific parameter distributions to assess the mapping power between and within those populations.

Genome Specificity of Loci Studied

Just as there is population specificity of allele-frequency and LD-strength data, there is genome specificity. Thus,

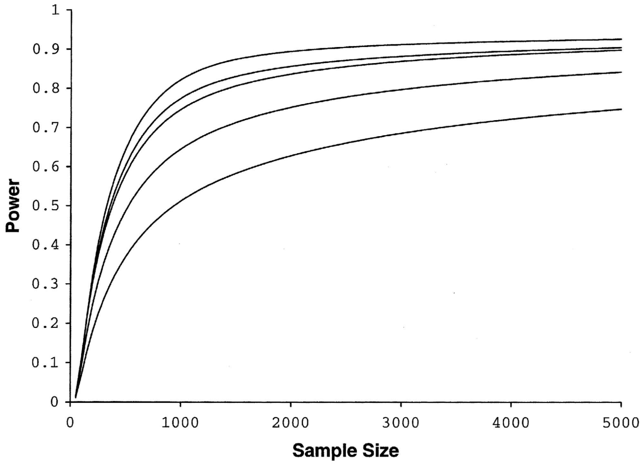

Figure 4 Power curves for a genomewide association study. Type I–error rates for the different maps are given in table 2. "Sample Size" is the number of chromosomes needed. The curves, from top to bottom, assume that interlocus map distances are 10, 20, 40, 80, and 160 kb, respectively. The trait-influencing allele, +, was assumed to have a frequency of 0.25 and penetrances of $p(++) = 0.5$, $p(+-) = 0.25$, and $p(--) = 0.0$.

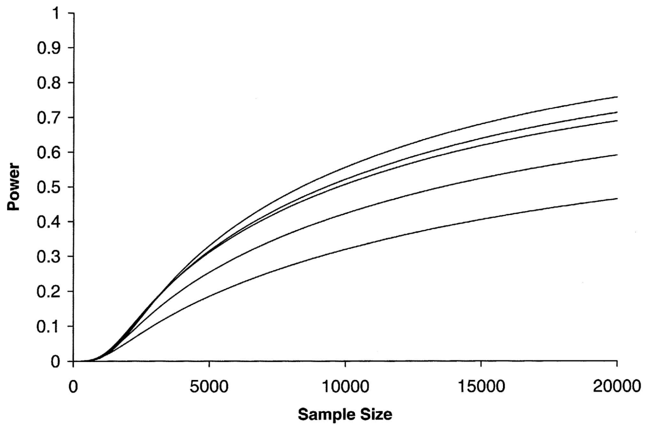

Figure 5 Power curves for a genomewide association study. Type I–error rates for the different maps are given in table 2. The curves, from top to bottom, assume that interlocus map distances are 10, 20, 40, 80, and 160 kb, respectively. The trait-influencing allele, +, was assumed to have a frequency of 0.25 and penetrances of $p(++) = 0.20$, $p(+-) = 0.15$, and $p(--) = 0.10$.

there are regions of the genome that show much greater recombination, mutation, and gene conversion than others (Abecasis et al. 2001). Thus, the loci studied to empirically assess probability distributions may not be generalizable to the genome as whole. Obviously, the more data that are used to derive empirical distributions, the better. In addition, the calculations presented in this article have assumed that the LD strength between marker alleles used to derive the LD strength distribution is the same as that between marker alleles and trait-influencing alleles. This needs empirical verification. As more associations are found, investigators will be in a position to estimate a marker/trait-influencing–locus LD-strength distribution.

Variable Intermarker Distances

The use of uniform intermarker distances is highly suspect and unrealistic. One could assess, however, the distribution of interlocus distances across the genome and incorporate this information into the power calculations.

Stratification in Case-Control Studies

I have focused on case-control association study designs. These designs are known to be negatively impacted by phenomena such as stratification (Schork et al. 2001), which I have not considered. Fortunately, there are methods for overcoming the problems created by these phenomena, especially if one has used a number of genetic markers to type the sample of individuals (see Devlin and Roeder 1999; Bacanu et al. 2000; Pritchard et al. 2000; Schork et al. 2001).

Other Designs

It would be of great value to consider how power calculations of the type proposed in this article could be used for settings involving, for example, quantitative traits, transmission/disequilibrium test (TDT) analysis settings, and family-based samples.

Multipoint and Haplotype Analysis

One of the biggest issues regarding the calculations pursued in this article is that they are based on single-locus analyses. Obviously, multipoint procedures and haplotype analyses need to be considered and undoubtedly will result in increased power to detect genetic effects (Akey et al. 2001; Fallin et al. 2001; Schork et al. 2001; author's unpublished data).

Type I–Error Rates and Corrections

My example calculations have made use of a Bonferroni correction for multiple comparisons. This correction assumes independence of the tests, which is not likely to be the case for dense-marker-map studies, since alleles at neighboring loci will be associated via LD. The problem in using an appropriate correction is not unique to the proposed methodology but, rather, plagues all studies involving multiple comparisons and correlated outcomes. More work in this area is clearly needed. One possible approach is to pursue randomization or permutation tests for evaluation of significance. Essentially, one could randomize disease status or phenotypic information (i.e., case/control status) across the subjects while preserving marker information for those subjects. Association statistics across the loci could be computed by use of the permuted data. This could be repeated a number of times, with counts of observed test statistics tallied throughout. The information that this exercise provides regarding the probability of achieving specific test-statistic values could be used to gauge the significance of the original, nonpermuted-data test statistics.

Ultimately, the proposed methodology for power calculations is very flexible and should provide more-realistic answers to questions about mapping power, especially with respect to either genomewide or large-genomic-region mapping studies. The most important caveat, however, with regard to any sample-size or power calculation is that they are only probabilistic.

Thus, there are no guarantees (*a*) that a marker, no matter how close to a trait-influencing locus, will have alleles in LD with a trait influencing allele or (*b*) that the effect size of the trait-influencing allele(s) will be large enough to be detected by use of the sample size chosen.

Acknowledgments

Aspects of this work were supported by National Institutes of Health, National Heart, Lung and Blood Institute, grants HL94-011, HL54998-01, HL69758-01, awarded to the author.

Appendix

Symbol Definitions

N	Total sample size
n_{d}	Number of diseased individuals (i.e., cases) in a sample
$n_{\bar{\mathrm{d}}}$	Number of normal individuals (i.e., controls) in the sample
c	Ratio of controls to cases in the sample
$+$	Influential genetic variant (i.e., variant that contributes to disease susceptibility)
$-$	Noninfluential genetic variant
n_{+}	Number of individuals in the sample who carry the influential allele
n_{-}	Number of individuals in the sample who carry the noninfluential allele
p	Frequency of the influential variant in the population at large
q	Frequency of the noninfluential variant in the population at large
$p_{+\mid\mathrm{d}}$	Probability that a diseased individual (i.e., case) carries the influential variant
$p_{+\mid\bar{\mathrm{d}}}$	Probability that a normal individual (i.e., control) carries the influential variant
$p_{-\mid\mathrm{d}}$	Probability that a diseased individual (i.e., case) carries the noninfluential variant
$p_{-\mid\bar{\mathrm{d}}}$	Probability that a normal individual (i.e., control) carries the noninfluential variant
$p(M\mid\mathrm{d})$	Probability that a diseased individual (i.e., case) carries the marker allele *M*
δ	LD between the marker allele *M* and the trait-influencing allele +
D'	Standardized LD
Ω	Parameters assumed in a genetic association–study power or sample-size calculation
Ψ'	Set of parameters to be "integrated out" in power calculations
B	Maximal distance between marker and trait-influencing loci

References

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

Akey J, Jin L, Xiong M (2001) Haplotypes vs single marker linkage disequilibrium tests: what do we gain? Eur J Hum Genet 9:291–300

Bacanu SK, Devlin B, Roeder K (2000) The power of genomic control. Am J Hum Genet 66:1933–1944

Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55:997–1004

Fallin D, Cohen A, Essioux L, Chumakov I, Blumenfeld M, Cohen D, Schork NJ (2001) Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease. Genome Res 11:143–151

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

Lewontin RC (1988) On measures of gametic disequilibrium. Genetics 120:849–852

Mackay TF (1995) The genetic basis of quantitative variation: numbers of sensory bristles of Drosophila melanogaster as a model system. Trends Genet 11:464–470

——— (1996) The nature of quantitative genetic variation revisited: lessons from Drosophila bristles. Bioessays 18: 113–121

Orr HA (1998) The population genetics of adaptation: the distribution of factors fixed during adaptive evolution. Evolution 52:935–949

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. Am J Hum Genet 67:170–181

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES (2001) Linkage disequilibrium in the human genome. Nature 411:199–204

Schlesselman JJ (1982) Case-control studies. Oxford University Press, New York

Schork NJ, Fallin D, Thiel B, Xu X, Broeckel U, Jacob H, Cohen D (2001) The future of genetic case/control studies. In: Rao DC, Province MA (eds) Advances in genetics. Academic Press, San Diego, pp 191–212

Schork NJ, Nath SK, Fallin, D, Chakravarti A (2000) Linkage disequilibrium analysis of biallelic DNA markers, human quantitative trait loci, and threshold-defined cases and controls. Am J Hum Genet 67:1208–1218

Silverman BW (1986) Density estimation for statistics and data analysis. Chapman & Hall, London

Taylor CC (1989) Bootstrap choice of smoothing parameters in kernel density estimation. Biometrika 76:705–712

Zapata C (2000) The *D* measure of overall gametic disequilibrium between pairs of multiallelic loci. Evolution 54: 1809–1812