Topic Introduction

Power Calculations in Genetic Studies

David M. Evans and Shaun Purcell

The power of a statistical test is the probability that it will yield a statistically significant result given that the null hypothesis is false. In other words, it represents the chance that the study will be successful in detecting a true effect and is dependent on a number of factors, including the magnitude of the effect, the sample size and study design, and the specified false-positive rate. Power calculations are primarily performed during the planning stages of a study, most typically in determining the sample size required. Consideration of statistical power can also sometimes shed light on the results of completed studies, particularly in the interpretation of negative results. In this article, we review the fundamentals of statistical power, discuss how power is calculated (using a genetic case/control study as an example), and consider the most pertinent factors that influence power in genetic studies. Finally, we focus on power in the context of modern whole-genome association studies, in which issues of coverage, multiple testing, and staged designs are paramount.

INTRODUCTION

To make the concept of statistical power concrete, consider the following example: a genetic case/control study to test the association with a functional locus (i.e., a locus that actually does confer risk of disease). We sample repeatedly from some hypothetical population, phenotype and genotype each sample, calculate the association odds ratio, and then perform a significance test for whether the odds ratio was different from 1.0 (i.e., to reject the null hypothesis of “no association”). Purely because of the random sampling process, each replicate sample differs slightly in its estimated odds ratio. Because of sampling error, some proportion of tests would be significant, and some not; the exact ratio would depend on a number of factors, including the true effect size, so it is possible that all replicates might be significant. The proportion of replicates in which a significant result is obtained is an estimate of the power of the test, and it tells us how likely the study would be to succeed if it were repeated once more under exactly the same conditions. In practice, we do not calculate power by performing the experiment multiple times. Instead, we rely on analytic theory or use computer simulations to perform the repeated set of experiments. Power calculations can then be used to make studies more efficient, by indicating the smallest possible study that still has good power to detect all effects of a particular size.

Thinking about statistical power (the chance of detecting true effects) immediately brings two types of inferential error into focus: namely, not detecting true effects (type II error) and “detecting” nonexistent effects (type I error). A “false-positive” or type II error corresponds to failing to reject the null hypothesis (i.e., obtaining a nonsignificant test result) when the alternative hypothesis is true (i.e., the specific, non-null alternative hypothesis is in fact true). The probability of committing a type II error, denoted by the Greek letter β , is equal to one minus the statistical power: If the alternative hypothesis is true, we either detect the effect (at rate $1 - \beta$) or fail to detect it (at rate β).

Adapted from *Genetic Variation* (ed. Weiner et al.). CSHL Press, Cold Spring Harbor, NY, USA, 2007.

© 2012 Cold Spring Harbor Laboratory Press

Cite this article as *Cold Spring Harb Protoc*; 2012; doi:10.1101/pdb.top069559

FIGURE 1. Distribution of the test statistic under both null (*left*) and alternative (*right*) hypotheses. The vertical line is the critical value of the test. The black and gray areas under the alternative hypothesis represent the power of the test. The shaded gray area refers to the type I error rate of the test. The unshaded area under the alternative hypothesis is the probability of committing a type II error.

In contrast, if the null hypothesis of no effect is in fact true, then we can either correctly fail to reject it (i.e., if the test is not significant), or incorrectly reject the null (i.e., if the test is significant, at rate α). Typically, the experimenter controls α by deciding at what threshold the test will be called significant. In this way, the sensitivity and specificity of a statistical test can be described and controlled. Naturally, there is a trade-off between these two properties: Lowering the threshold increases sensitivity (i.e., increases power, reduces type II errors) but also decreases specificity (i.e., increases type I error rate). Historically, for many small-scale studies, investigators have by convention adopted values such as $\alpha = 0.05$ and $\beta = 0.20$ (80% power) as representing a realistic and adequate trade-off.

The relationship between these quantities is displayed in Figure 1 (for ease of representation we assume a one-sided test of significance). Figure 1 shows the distribution of the test statistic X under both the null hypothesis (left-hand side) and the alternative hypothesis (right-hand side). The vertical line represents the critical value of the test so that values lying to the right of this line are declared significant, whereas values to the left are not. The area to the right of the critical value under the alternative hypothesis represents the power of the test (i.e., the sum of the black and gray areas), whereas the area to the left of the critical value under the alternative hypothesis represents the probability of committing a type II error. The probability of a type I error is represented by the gray area under the null hypothesis to the right of the critical value.

Besides altering the critical value, there are two other ways to increase the power of a statistical test: increasing the size of the sample, and increasing the effect size. As described below, the expected test statistic is a function of both sample size and effect size. Sample size is typically clearly under the investigator's control. Less obviously, the effect size can also be increased; for example, by reducing the measurement error with more accurate phenotyping, or through genotyping a region of interest more densely.

If we know the distribution of the test statistic under the null and alternative hypotheses, it is possible to calculate the expected power of the statistical test analytically (see Sham et al. 2000). For example, many tests of genetic association follow a chi-squared distribution. The shape of this distribution is determined by the “noncentrality parameter” (NCP). When the null hypothesis is true, the NCP is zero and the resulting chi-squared distribution is called a “central chi-squared distribution.” Under the alternative hypothesis, the NCP will be greater than zero and the resulting distribution is called a “noncentral chi-squared” distribution. This means that the distribution of the test statistic under the alternative hypothesis changes in shape and is shifted to the right by an amount quantified by the NCP. The expected test statistic equals the NCP plus the degrees of freedom. Because the shapes of the central and noncentral chi-squared distributions are known, it is possible to deduce the areas under the curve in Figure 1 (and therefore the power) if we know the expected NCP, degrees of freedom of the test, and the critical value used.

PARAMETERIZING THE GENETIC MODEL

An investigator conducting a power analysis must decide in advance which values certain parameters will take. These include quantities that are known or under the control of the investigator (e.g., sample

size, type I error rate, disease prevalence). The investigator must also make assumptions about parameters that are unknown, including the size of the genetic effect and underlying mode of inheritance, the level of linkage disequilibrium between marker and trait loci, and the allele frequencies at these loci. For example, the relative risk parameters R_{AA} and R_{Aa} quantify the effect size of the gene of interest and represent the number of times more likely an individual with genotype AA (or an individual with Aa genotype) is to be affected with disease than an individual with a lower risk aa genotype (arbitrarily assuming that aa is the low-risk, reference genotype). Given these parameters and knowledge of the disease prevalence (K), it is possible to calculate the absolute risk that an individual with the aa genotype is affected with disease(g_{aa}):

$$g_{aa} = K / (P_A^2 R_{AA} + 2p_A p_a R_{Aa} + p_a^2)$$

It is generally accepted that common complex diseases are unlikely to be the result of loci of large effect. Rather, genotypic relative risks are more likely to be in the range of 1–1.5. In the case of quantitative phenotypes, the measure of effect size is the proportion of the total phenotypic variance that is caused by the quantitative trait locus. Typically, this value will be on the order of a few percent (0%–5%) except perhaps in situations where there is evidence for a major locus. In such a case, a larger effect size might be assumed.

The underlying mode of inheritance is specified by the relationship between the relative risk parameters R_{AA} and R_{Aa} . For example, relative risks of $R_{AA} = 4$ and $R_{Aa} = 1$ indicate a recessive mode of inheritance for the A allele, because the Aa heterozygote contributes no more risk than the recessive aa homozygote. In contrast, relative risks of $R_{AA} = 16$ and $R_{Aa} = 4$ imply a multiplicative mode of inheritance, because the risk of the AA genotype is the square of the risk associated with the heterozygote.

There is also the very real possibility that the marker typed may not be the functional variant responsible for disease, but rather a marker in linkage disequilibrium with the true functional locus. The degree of linkage disequilibrium between trait and marker loci can be quantified by Lewontin's (1964) D' coefficient. Values of $D' = 1$ indicate an absence of ancestral recombination between marker and disease loci and thus complete disequilibrium. In contrast, a $D' = 0$ indicates independence between marker and trait loci. If one is interested in typing a small genomic region or a number of candidate genes thoroughly, one could assume that at least one marker will be in high linkage disequilibrium with a putative functional variant (assuming, of course, that a common variant actually exists in the region). In contrast, this assumption is unrealistic in genome-wide-association studies, because the degree to which currently available commercial panels capture common variation in the genome is less than perfect. For example, it has recently been estimated that the Affymetrix 111K and 500K chips and the Illumina HumanHap300 panel capture around 31%, 65%, and 75%, respectively, of the common genetic variation in Western European populations through linkage disequilibrium (Barrett and Cardon 2006).

Finally, it is necessary to specify the allele frequencies at the trait and marker loci. Unless the investigator is only interested in a specific single-nucleotide polymorphism (SNP) whose frequency is known a priori, one would typically conduct power calculations across a range of possible allele frequencies. If using a commercial product that specifically tags common variants (e.g., the Illumina 300K panel of markers), one might have to assume the existence of a “common” disease allele (i.e., minor allele frequency >5%), because the tag SNPs in these panels do not capture rare variation well (Barrett and Cardon 2006). In contrast, if the focus of research is on nonsynonymous SNPs (i.e., SNPs that produce an amino-acid change in a protein), one might assume that the frequency of the disease-causing allele is less (because nonsynonymous SNPs tend to have lower minor allele frequencies).

CALCULATING POWER TO DETECT ASSOCIATION ANALYTICALLY

Given assumptions about the underlying genetic model, it is possible to calculate the expected NCP and hence the power of tests for genetic association. Closed-form expressions for the NCPs of several

tests of association have been described in the literature, including the transmission disequilibrium test (TDT; McGinnis et al. 2002), case/control tests of genetic association (Schork 2002), and the quantitative trait disequilibrium test (Sham et al. 2000). With these formulae, an investigator can quickly and easily determine power under a variety of different scenarios without having to perform time-consuming data simulations. In addition, the NCP of a statistical test is a useful quantity in itself in that it increases linearly with sample size, unlike power to detect association (Witte et al. 2000). A number of utilities are available for helping investigators calculate the expected power to detect genetic association for a variety of different study designs, including QUANTO (<http://hydra.usc.edu/GxE>), power for association with error (PAWE; <http://linkage.rockefeller.edu/pawe>), and Purcell's on-line Genetic Power Calculator (<http://pngu.mgh.harvard.edu/~purcell/gpc/>) (Gauderman 2002; Gordon et al. 2002; Purcell et al. 2003).

To illustrate how the power of a genetic test of association can be calculated analytically, we describe a procedure for calculating the NCP for the Pearson chi-square test of allelic association. The Pearson chi-square test for allelic association compares the frequency of marker alleles in cases and controls:

$$\chi^2 = \sum_{i=0,1} \sum_{j=A,U} \frac{(n_{ij} - E[n_{ij}])^2}{E[n_{ij}]}$$

where the summation is over alleles (i) and affection status (j), n_{ij} is the cell containing the observed number of alleles of type i from individuals with affection status j , and

$$E[n_{ij}] = \frac{n_{i.}n_{.j}}{n..}$$

is the expected cell count given the totals at the margins. Implicit in this allelic test is the assumption of a multiplicative mode of inheritance and that Hardy–Weinberg equilibrium holds in the general population (i.e., such that it is appropriate to treat an individual's two alleles as independent observations). That is, this test collapses across genotypes to form a 2×2 contingency table, counting alleles as opposed to individuals. If these assumptions do not hold, a genotypic test of association or the Armitage test of trend may be more appropriate (Sasieni 1997).

To illustrate the derivation of the NCP for the allelic test of association, consider a single biallelic disease-causing locus with alleles A and a in n cases and m control individuals (so that the ratio of controls to cases is r). We denote the prevalence of disease by K , the frequency of the high-risk allele as p_A , and the genotypic relative risks for the Aa and AA genotypes as R_{Aa} and R_{AA} , respectively. Given these quantities, it is simple to calculate the risk that an individual possessing the low-risk genotype aa is affected:

$$g_{aa} = \frac{K}{R_{AA}p_A^2 + 2R_{Aa}p_A(1 - p_A) + (1 - p_A)^2}$$

and, consequently, the absolute risk of disease in individuals with Aa and AA genotypes:

$$g_{Aa} = R_{Aa}g_{aa}$$

$$g_{AA} = R_{AA}g_{aa}$$

From these quantities we can calculate the expected proportions of genotypes in cases using Bayes' theorem:

$$c_{aa} = \frac{g_{aa}(1 - p_A)^2}{K}$$

$$c_{Aa} = \frac{2g_{Aa}p_A(1 - p_A)}{K}$$

$$c_{AA} = \frac{g_{AA}p_A^2}{K}$$

and similarly in controls:

$$u_{aa} = \frac{(1 - g_{aa})(1 - p_A)^2}{1 - K}$$

$$u_{Aa} = \frac{2(1 - g_{Aa})p_A(1 - p_A)}{1 - K}$$

$$u_{AA} = \frac{(1 - g_{AA})p_A^2}{1 - K}$$

It is now a simple matter to calculate the expected allelic frequencies in case (c_A) and control (u_A) samples:

$$c_A = c_{AA} + \frac{c_{Aa}}{2}$$

and

$$u_A = u_{AA} + \frac{u_{Aa}}{2}$$

These expected allelic frequencies are summarized in the contingency table (Table 1). It is now possible to use these values to calculate the NCP (denoted by the Greek letter λ) as first derived by Mitra (1958):

$$\lambda = \frac{2nr(c_A - u_A)^2(1 + r)}{(c_A + ru_A)(1 + r - c_A - ru_A)}$$

TABLE 1. Expected allele frequencies for case/control allelic test of association

Allele	Case	Control
A	$2c_A n$	$2u_A m$
a	$2(1 - c_A)n$	$2(1 - u_A)m$
	$2n$	$2m$

The power of the allelic test of association is found by integrating under the noncentral chi-square distribution:

$$1 - \beta = \int_{\chi^2_{\alpha}(v, 0)}^{\infty} d\chi'^2(v, \lambda)$$

where $\chi'^2_{\alpha}(v, 0)$ is the $100(1 - \alpha)$ percentage point of the central χ^2 distribution with v degrees of freedom and $\chi'^2(v, \lambda)$ is a noncentral χ^2 distribution with noncentrality parameter λ .

This basic procedure assumes that the investigator is lucky enough to genotype the actual trait locus. Of course, in reality this is seldom the case, except perhaps when attempting to replicate a known association. It is therefore important to extend these calculations to situations where the trait locus has not been typed, but rather is a marker in linkage disequilibrium. Consider a biallelic marker locus with alleles M and m and corresponding allele frequencies p_M and p_m . We quantify the amount of linkage disequilibrium between the trait and marker locus by Lewontin's (1964) D' coefficient:

$$D' = \frac{p_{AM} - p_A p_M}{\min(p_A p_m, p_a p_M)} = \frac{\delta}{\delta_{\max}} \text{ when } \delta > 0$$

and

$$D' = \frac{p_{AM} - p_A p_M}{\min(p_a p_M, p_a p_m)} = \frac{\delta}{\delta_{\max}} \text{ when } \delta < 0$$

where p_{AM} is the frequency of the haplotype with A and M alleles, and δ quantifies the difference between observed and expected (under independence) haplotype frequencies. For $\delta > 0$, $\delta_{\max}$ is the most positive value of δ possible given the specific allele frequencies at the marker and disease loci, whereas for $\delta < 0$, $\delta_{\max}$ is the most negative possible value of δ (Devlin and Risch 1995; McGinnis et al. 2002). The haplotype frequencies at the marker and trait loci are expressed in terms of δ in Table 2.

To incorporate linkage disequilibrium into calculation of the NCP, we simply determine the risks associated with each marker genotype. This can be obtained by multiplying the risks at the trait locus by the probability of the trait genotype given the marker genotype and then summing over all possible trait genotypes. For example, the probability that an individual is affected given they have an MM genotype at the marker locus is calculated thus:

$$\begin{aligned} P(D = \text{Affected} | G_M = MM) &= \sum_{G_T} P(D = \text{Affected} | G_T) P(G_T | MM) \\ &= \sum_{G_T} P(D = \text{Affected} | G_T) \frac{P(G_T, G_M = MM)}{P(G_M = MM)} \\ &= \frac{g_{AA}(p_A p_M + \delta)^2 + 2g_{Aa}(p_A p_M + \delta)(p_a p_M - \delta) + g_{aa}(p_a p_M - \delta)^2}{p_M^2} \end{aligned}$$

TABLE 2. Haplotype frequencies at marker and disease loci

		Trait allele	
		A	a
Marker allele	M	$p_A p_M + \delta$	$p_a p_M - \delta$
	m	$p_A p_m - \delta$	$p_a p_m + \delta$

where G_T and G_M denote the genotypes at the trait and marker loci, respectively. The risk associated with other marker genotypes is derived in a similar fashion. The procedure for calculating the NCP of the test at the marker locus proceeds exactly as before, except now the frequencies and risks at the marker locus are used in the calculation instead of at the disease locus.

CALCULATING POWER TO DETECT ASSOCIATION THROUGH SIMULATION

As shown, closed-form expressions for the NCP of a statistical test for genetic association can be used quickly and easily to calculate expected power under a range of scenarios. However, such formulae are useful only for the simplest of study designs. In reality, data sets often involve much more complicated situations. For example, a data set may consist of a number of different types of pedigree structures, or individuals may have been ascertained on the basis of an extreme score on a quantitative trait. Web-based tools do not exist for calculating power in complicated situations such as these, and it may not be a simple matter for an investigator to derive the NCP and calculate power analytically on his own. In these situations, power calculations can still be performed with the aid of data simulation.

The basic idea behind data simulation is to generate thousands of replicate data sets under the alternative hypothesis of association for a given genetic model. In other words, each replicate data set is like drawing a random sample from the same underlying population. Data generation may involve simulating a number of different processes (e.g., transmitting alleles from parents to offspring) or perhaps ascertaining individuals on the basis of some simulated trait value. Each replicate data set is then tested for genetic association, and the proportion of replicates in which the evidence of association exceeds the critical value is an estimate of the power of the test. Simulation is therefore a very flexible strategy for estimating the power to detect genetic association. The downside is that it is computationally intensive; thousands of replicate samples need to be generated for each specific set of population parameters.

FACTORS INFLUENCING THE POWER TO DETECT ASSOCIATION

A number of factors can affect the power to detect association. These include the size of a genetic effect, its mode of inheritance, the prevalence of disease, the ratio of case to control individuals, allelic frequencies, and the extent of linkage disequilibrium between marker and trait loci. For all situations discussed below, we determined the expected power to detect association using Purcell's on-line Genetic Power Calculator (<http://pngu.mgh.harvard.edu/~purcell/gpc>; Purcell et al. 2003).

Mode of Inheritance

Table 3 displays the effects that varying the mode of inheritance, genotype relative risk, and disease allele frequency have on the power to detect association. Specifically, the table displays the number of parent-child trios (case/control pairs) required to obtain 90% power to detect association at a significance threshold of either $\alpha = 0.05$ or $\alpha = 5 \times 10^{-7}$. A significance level of $\alpha = 0.05$ might be appropriate in the analysis of a single polymorphism as in a candidate gene study, whereas a significance level of $\alpha = 5 \times 10^{-7}$ would be more typical of that required in a genome-wide association study. In both cases, we assume a relatively rare disease ($K = 0.001$), complete linkage disequilibrium between trait and marker loci (i.e., $D' = 1$), and that marker and disease loci have equal allele frequencies (which is in effect saying that we are directly genotyping and testing the disease locus itself). For case/control pairs, the test of association is Pearson's chi-squared test of allelic independence (Sasieni 1997). For parent-child trios, the relevant test is Spielman's TDT (Spielman et al. 1993).

Not surprisingly, the power to detect association is greatest when the effect size is large. Both case-control and parent-offspring designs have considerable power to detect common alleles that have moderate-to-large effects (i.e., genotypic relative risks > 2). In most scenarios, only tens or hundreds of

TABLE 3. Number of (A) case individuals or (B) parent–child trios required to obtain 90% power to detect association at a significance threshold of either $\alpha = 0.05$ or $\alpha = 5 \times 10^{-7}$

A. Case Individuals		$\alpha = 0.05$				$\alpha = 5 \times 10^{-7}$			
GRR _{Aa}	p _A	multiplicative	additive	recessive	dominant	multiplicative	additive	recessive	dominant
4.0	0.01	301	312	1.195×10^6	318	1137	1179	4.525×10^6	1201
	0.10	40	53	1531	64	150	201	5796	242
	0.50	27	60	55	195	101	225	207	736
	0.80	60	153	71	2611	225	577	267	9888
2.0	0.01	1602	1602	1.065×10^7	1656	6065	1305	4.032×10^7	6270
	0.10	190	190	1.007×10^4	263	718	718	4.678×10^4	995
	0.50	92	92	260	512	348	348	983	1936
	0.80	175	175	233	6099	661	661	881	2.31×10^4
1.5	0.01	5319	5179	3.462×10^7	5465	2.014×10^4	1.961×10^4	3.462×10^7	2.069×10^4
	0.10	609	483	3.913×10^4	800	2306	1826	3.913×10^4	3032
	0.50	260	126	691	1266	983	477	3207	4795
	0.80	457	192	527	1.406×10^4	1728	725	2447	5.325×10^4
B. Parent–Child Trios		$\alpha = 0.05$				$\alpha = 5 \times 10^{-7}$			
GRR _{Aa}	p _A	multiplicative	additive	recessive	dominant	multiplicative	additive	recessive	dominant
4.0	0.01	304	315	1.197×10^6	321	1150	1192	4.534×10^6	1214
	0.10	43	56	1537	67	160	212	5819	253
	0.50	30	62	58	198	111	235	217	748
	0.80	63	156	74	2619	235	588	278	9917
2.0	0.01	1608	1608	1.067×10^7	1662	6089	6089	4.04×10^7	6294
	0.10	193	193	1.238×10^4	266	730	730	4.688×10^4	1007
	0.50	95	95	263	515	359	359	995	1950
	0.80	178	178	236	6114	672	672	893	2.315×10^4
1.5	0.01	5333	5193	3.469×10^7	5479	2.02×10^4	1.967×10^4	1.612×10^8	2.075×10^4
	0.10	613	486	4.811×10^4	805	2321	1840	1.822×10^5	3048
	0.50	263	129	852	1271	995	488	3223	4814
	0.80	460	195	651	1.409×10^4	1741	736	2462	5.337×10^4

Results are shown for multiplicative, additive, recessive, and dominant models. An equal number of case and control individuals and a disease prevalence of $K = 0.001$ are assumed.

cases/trios are required. In contrast, to detect loci of small effect (e.g., a genotypic relative risk of about 1.5, which is probably indicative of the size of effects underlying most complex traits and diseases), sample sizes need to be much larger. This is particularly true for genome-wide association where sample sizes will need to contain thousands of individuals to have appreciable power to detect alleles of small effect. In this respect, it is interesting to note that although the required level of significance has changed by five orders of magnitude, the required sample size only increases around five times (i.e., the required sample is roughly proportional to the logarithm of the alpha level [Witte et al. 2000]). In general, power is greatest to detect association when the frequency of the disease allele is intermediate, although this depends to some extent on the underlying mode of inheritance. For example, in the case of a dominant disease, many more individuals are required when the frequency of the disease allele is high. The power to detect rare variants is low in general.

Table 3 also shows that for a broad array of disease models, when the disease is rare, the number of trios required to detect association is approximately equal to the number of case individuals for a given

level of power (assuming equal numbers of case and control individuals). In other words, the total number of individuals requiring genotyping in a case/control study is around two-thirds that required for a study of trios in the case of a rare disease (Bacanu et al. 2000). This reduction is especially relevant for genome-wide association studies where the number of individuals and markers analyzed (and hence the cost of the study) is very large. Nevertheless, it must be remembered that family-based tests of association do have certain advantages over a case/control design. For example, the TDT controls for the effect of population stratification that can produce spurious associations in case/control designs and also allows for the detection of parent-of-origin effects and genotyping error.

Disease Prevalence

As the prevalence of disease increases, the power to detect association using the case/control design also increases. This is because the frequency of the high-risk allele decreases in control individuals (assuming that control individuals are selected on the basis of not having the risk allele). Hence, there is a greater divergence in the frequency of alleles between case and control groups. Thus, for common diseases, the number of individuals required is far larger for the TDT than for the case/control test of association, assuming the same underlying genetic model. However, prevalence of disease does not affect the power of the TDT, provided that both the genotypic relative risks and the mode of inheritance remain constant. This is because the TDT uses family-based controls that are not screened for the presence/absence of disease (i.e., the genetic composition of the parents is not expected to change as a function of disease prevalence).

Ratio of Case to Control Individuals

The power of the case/control test of association also changes as a function of the relative proportion of affected and unaffected individuals in the sample. In general, for a given number of individuals, the most powerful ratio is an equal number of cases and controls. However, in some situations, it may be difficult or expensive to ascertain affected individuals. In these situations, recruiting more control individuals can still increase power. McGinnis et al. (2002) showed that in the hypothetical situation of an infinite number of controls, one would only need to recruit half the number of cases to obtain the same power as a study with equal numbers of cases and controls. However, most of this potential gain in power can be achieved by recruiting three to five times the number of control individuals. In the end, the optimal ratio of cases and controls will depend on the relative cost of acquiring each (McGinnis et al. 2002).

Linkage Disequilibrium and Marker Allele Frequency

Power to detect association is greatest when linkage disequilibrium is high (i.e., there is little ancestral recombination between marker and trait loci), and when trait and marker loci have similar allele frequencies. This poses a problem for the detection of rare alleles via genetic association, as most of the SNPs found in the dbSNP (Sachidanandam et al. 2001) and HapMap databases (Altshuler et al. 2005), and consequently those available on commercial marker panels, are common variants (i.e., minor allele frequency >5%). Therefore, association studies that use markers from these sources may have little power to detect rare variants (remember that the power to detect rare disease alleles is already very low). This limitation can be addressed somewhat by forming haplotypes of adjacent SNPs that might have a closer frequency to the disease allele. Such an approach is not entirely satisfactory (Lin et al. 2004).

GENOME-WIDE ASSOCIATION

Advances in high-throughput technology and decreased genotyping costs (Matsuzaki et al. 2004), as well as the publication of the International Haplotype Map of the human genome (Altshuler et al.

2005), make it possible to search for complex disease genes by screening thousands of individuals on hundreds of thousands of polymorphisms across the genome. The first genome-wide association studies appeared in the literature with encouraging results (Ozaki et al. 2002; Cheung et al. 2005; Klein et al. 2005; Maraganore et al. 2005). Whereas these designs hold great promise for dissecting the genetic basis of complex traits and diseases, there are a number of statistical issues pertaining to this sort of study design. One issue that we examine here is that of using a two-stage approach to decrease the costs associated with genotyping.

To decrease the cost of genotyping thousands of subjects with hundreds of thousands of markers, a more cost-effective strategy might be to genotype subjects in stages. Initially, a proportion of individuals ($\pi_{\text{individuals}}$) would be genotyped on all markers (N_{markers}). Subsequently, a proportion of markers displaying the most promising results (π_{markers}) would then be typed in the remaining individuals (Sobell et al. 1993). Compared to one-stage approaches where all markers in all individuals are genotyped, two-stage designs can lead to appreciable savings in genotyping costs while still maintaining power (Satagopan et al. 2002, 2004; Satagopan and Elston 2003; Maraganore et al. 2005; Thomas et al. 2005; Skol et al. 2006).

Recently, Skol et al. (2006) examined the performance of two different approaches to analyzing the data from two-stage genome-wide association scans. In the “replication-based” strategy, association was only tested on markers and individuals genotyped in the second phase of the procedure. To maintain the genome-wide error rate below $\alpha_{\text{genome}} = 0.05$, a Bonferroni corrected significance level of $\alpha_{\text{genome}}/(N_{\text{markers}} \times \pi_{\text{markers}})$ was used. In the “joint analysis” strategy, test statistics in stages one and two were combined and compared against an approximate significance level of $\alpha_{\text{genome}}/N_{\text{markers}}$. The study found that despite the more stringent significance level, jointly analyzing the data from both stages almost always resulted in increased power to detect association as compared to a replication-based strategy. With appropriately chosen thresholds, the power provided by this method was comparable to that of the one-stage design. The power of the joint analysis decreased as the proportion of samples typed in stage one decreased, presumably because variants that predispose to disease were less likely to be selected for genotyping in stage two. Similarly, taking too few markers through to stage two also decreased the power of the joint analysis because of the decreased probability of taking a true risk variant through to the next stage. In contrast, power increased when fewer markers were selected for follow-up in the replication-based strategy, because fewer statistical tests were performed and hence there were fewer penalties because of multiple testing. In fact, the only situation in which the joint analysis was not more powerful than a replication-based strategy was when the strength of association was far greater in stage two than in stage one. It is therefore recommended that two-stage genome-wide association scans be analyzed using a joint analysis strategy, that a large proportion of samples be genotyped in stage one (i.e., $\pi_{\text{individuals}} > 30\%$), and that a relatively large proportion of markers be selected for follow-up in stage two ($\pi_{\text{markers}} > 1\%$). The Center for Statistical Genetics provides a useful web-based tool (<http://csg.sph.umich.edu>) that can be used by researchers to calculate the power to detect association in two-stage designs.

CONCLUSIONS

We conclude with some basic guidelines investigators might find useful when performing power calculations in either the design or analysis phase of genetic association studies:

1. During the design phase of a study, use a web-based utility such as Purcell’s GPC (<http://pngu.mgh.harvard.edu/~purcell/gpc/>) to calculate the expected number of individuals who will need to be genotyped given what is known about the disease (e.g., mode of inheritance, size of effect). If one is performing a candidate gene study and is willing to assume that a functional variant lies somewhere in the region being typed (e.g., a previous linkage analysis might indicate that a disease locus lies in the region), then it might be safe to assume high levels of linkage disequilibrium between marker and disease loci. In contrast, if a genome-wide association study is being planned,

then the possibility that a causal variant may not be in appreciable linkage disequilibrium with any of the markers must be considered. The ability of commercially available marker panels to capture common variation in the genome has been quantified in Barrett and Cardon (2006).

2. If using a complicated study design, it may be necessary to perform data simulations to calculate the expected power.
3. If a genome-wide association study is being considered, a two-stage design is an efficient and cost-effective method of genotyping subjects while maintaining power to detect association. Generally, a large proportion of samples should be genotyped in the first stage (>30%), a relatively large proportion of markers should be selected for follow-up (>1%), and the data should be analyzed using a joint analysis strategy. A web-based tool that can be used to calculate the power to detect association in two-stage designs is located at <http://csg.sph.umich.edu>.

REFERENCES

- Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P. 2005. A haplotype map of the human genome. *Nature* 437: 1299–1320.
- Bacanu SA, Devlin B, Roeder K. 2000. The power of genomic control. *Am J Hum Genet* 66: 1933–1944.
- Barrett JC, Cardon LR. 2006. Evaluating coverage of genome-wide association studies. *Nat Genet* 38: 659–662.
- Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT. 2005. Mapping determinants of human gene expression by regional and genome-wide association. *Nature* 437: 1365–1369.
- Devlin B, Risch N. 1995. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* 29: 311–322.
- Gauderman WJ. 2002. Sample size requirements for association studies of gene-gene interaction. *Am J Epidemiol* 155: 478–484.
- Gordon D, Finch SJ, Nothnagel M, Ott J. 2002. Power and sample size calculations for case-control genetic association tests when errors are present: Application to single nucleotide polymorphisms. *Hum Hered* 54: 22–33.
- Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, et al. 2005. Complement factor H polymorphism in age-related macular degeneration. *Science* 308: 385–389.
- Lewontin RC. 1964. The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* 49: 49–67.
- Lin S, Chakravarti A, Cutler DJ. 2004. Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. *Nat Genet* 36: 1181–1188.
- Maraganore DM, de Andrade M, Lesnick TG, Strain KJ, Farrer MJ, Rocca WA, Pant PV, Frazer KA, Cox DR, Ballinger DG. 2005. High-resolution whole-genome association study of Parkinson disease. *Am J Hum Genet* 77: 685–693.
- Matsuzaki H, Loi H, Dong S, Tsai YY, Fang J, Law J, Di X, Liu WM, Yang G, Liu G, et al. 2004. Parallel genotyping of over 10,000 SNPs using a one-primer assay on a high-density oligonucleotide array. *Genome Res* 14: 414–425.
- McGinnis R, Shifman S, Darvasi A. 2002. Power and efficiency of the TDT and case-control design for association scans. *Behav Genet* 32: 135–144.
- Mitra SK. 1958. On the limiting power function of the frequency chi-square test. *Ann Math Stat* 29: 1221–1233.
- Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, Tsunoda T, Sato H, Sato H, Hori M, Nakamura Y, Tanaka T. 2002. Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nat Genet* 32: 650–654.
- Purcell S, Cherny SS, Sham PC. 2003. Genetic Power Calculator: Design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* 19: 149–150.
- Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL. 2001. (International SNP Map Working Group). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409: 928–933.
- Sasieni PD. 1997. From genotypes to genes: Doubling the sample size. *Biometrics* 53: 1253–1261.
- Satagopan JM, Elston RC. 2003. Optimal two-stage -genotyping in population-based association studies. *Genet Epidemiol* 25: 149–157.
- Satagopan JM, Venkatraman ES, Begg CB. 2004. Two-stage designs for gene-disease association studies with sample size constraints. *Biometrics* 60: 589–597.
- Satagopan JM, Verbel DA, Venkatraman ES, Offit KE, Begg CB. 2002. Two-stage designs for gene-disease association studies. *Biometrics* 58: 163–170.
- Schork NJ. 2002. Power calculations for genetic association studies using estimated probability distributions. *Am J Hum Genet* 70: 1480–1489.
- Sham PC, Cherny SS, Purcell S, Hewitt JK. 2000. Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. *Am J Hum Genet* 66: 1616–1630.
- Skol AD, Scott LJ, Abecasis GR, Boehnke M. 2006. Corrigendum: Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet* 38: 390.
- Sobell JL, Heston LL, Sommer SS. 1993. Novel association approach for determining the genetic predisposition to schizophrenia: Case-control resource and testing of a candidate gene. *Am J Med Genet* 48: 28–35.
- 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.
- Thomas DC, Haile RW, Duggan D. 2005. Recent developments in genome-wide association scans: A workshop-summary and review. *Am J Hum Genet* 77: 337–345.
- Witte JS, Elston RC, Cardon LR. 2000. On the relative sample size required for multiple comparisons. *Stat Med* 19: 369–372.

Cold Spring Harbor Protocols

Power Calculations in Genetic Studies

David M. Evans and Shaun Purcell

Cold Spring Harb Protoc; doi: 10.1101/pdb.top069559

Email Alerting Service

Receive free email alerts when new articles cite this article - [click here](#).

Subject Categories

Browse articles on similar topics from *Cold Spring Harbor Protocols*.

[Genetic Variation](#) (86 articles)

[Genome Analysis](#) (191 articles)

[Genome Wide Association Studies \(GWAS\)](#) (11 articles)

To subscribe to *Cold Spring Harbor Protocols* go to:
<http://cshprotocols.cshlp.org/subscriptions>