Power of Genetic Association Studies with Fixed and Random Genotype Frequencies

Julia Kozlitina[1,2], Chao Xing[1], Alexander Pertsemlidis[1],
and William R. Schucany[2]

[1]Donald W. Reynolds Cardiovascular Clinical Research Center, Eugene McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center, Dallas, TX, 75390-8591, USA. [2]Department of Statistical Science, Southern Methodist University, Dallas, TX, 75275-0332, USA

December 1, 2009

Abstract

When estimating the power of genetic association studies, the allele and genotype frequencies are often assumed to be known, and the numbers of individuals with each genotype are set equal to their expectations under Hardy-Weinberg equilibrium. In fact, both allele and genotype frequencies are unknown and random. Ambrosius et al. (2004) have demonstrated that treating these parameters as fixed can lead to inflated power expectations. To overcome the problem, they proposed averaging power estimates over the distribution of unknown parameters. We investigate their method and find that, despite theoretical appeal, it may not always improve accuracy, while significantly increasing computational time. For a given allele frequency, we show that the approach of fixing genotype counts does produce an upward bias in estimated power, but the magnitude of the bias diminishes rapidly with sample size and is completely negligible for $N > 200$. For an unknown frequency, the method of power averaging requires further assumptions about the prior distribution for the parameter, and can either overestimate or underestimate true power when the prior is misspecified. We explore the relationships between these and other assumptions of the power calculation, and propose a more economical approach to power analysis.

Key Words: Average power; sample size; allele frequency; sensitivity to model assumptions

INTRODUCTION

Estimating statistical power is a critical issue in the planning of genetic association studies. For a given study design, a power calculation generally relies on a number of assumed parameters. In particular, the allele frequencies of investigated markers are assumed to be known, and the numbers of individuals with each genotype are set equal to their expectations under Hardy-Weinberg equilibrium (HWE). In reality, the allele frequencies are rarely known prior to collecting the data, and in fact can vary considerably across different populations (e.g., The International HapMap Consortium 2005). Further, even for a fixed population allele frequency with genotypes in Hardy-Weinberg proportions, sample genotype counts are still subject to random variation. Ambrosius et al. (2004) previously showed that ignoring uncertainty in these parameters tends to produce overly optimistic power estimates, thus understating sample-size requirements for a study. To obtain more realistic estimates, they proposed averaging power over the distribution of sample genotype counts when the allele frequency is known, and placing a prior distribution on the allele frequency when the latter is unknown. The authors examined a number of examples from different study designs and found that the effect of treating unknown frequencies as fixed can range from small to substantial. Zheng et al. (2005) extended the method of power averaging to linkage studies and came to qualitatively similar conclusions.

Although Ambrosius et al. (2004) rightly point out that the uncertainty in assumed parameters can lead to incorrect power estimates, their examples cover only a limited number of parameter settings (e.g., p = 0.5 and $N = 20, 40$ in the case of association studies with quantitative traits). It is therefore not clear from their analysis whether the relationship holds mathematically and whether a similar overestimation would occur under different parameter combinations. For example, both Ambrosius et al. (2004) and Zheng et al. (2005) report much smaller errors in estimated power when the examples involve larger samples sizes ($N = 100, 500$), but they do not make an explicit connection between accuracy and sample size. Finally, the authors fail to consider the effect of other factors, such as the assumptions about the genetic model, on the quality of power estimates.

We examine the relationship between the factors affecting power more systematically, in order

to quantify the effect of ignoring variation in the unknown parameter values. Focusing on association studies with quantitative traits, we show that for a given allele frequency, the approach of fixing sample genotype counts at their expected values does lead to an upward bias in estimated power. However, the magnitude of the bias decreases dramatically with sample size and is usually negligible for $N > 200$, unless the alleles in question are extremely rare. Thus, the method of power averaging, although technically correct, may in practice achieve little in terms of accuracy while significantly increasing computational time.

When the allele frequency is unknown, Ambrosius et al. (2004) suggest using a beta prior for the unknown parameter and averaging power over the prior. To demonstrate their method, however, they choose a beta distribution that is centered at the true population frequency or estimates thereof obtained from previous studies. We note that the first approach is unachievable in practice, and the second can itself produce upward- or downward-biased results if the available estimates of the allele frequency happen by chance to be far from the population value. Zheng et al. (2005) recognize the problem and recommend performing a sensitivity analysis of averaged power with respect to the true allele frequency. We agree with Zheng et al. (2005) but observe that their procedure adds further complexity to the computation. In line with the previously stated finding, we argue that a simple power calculation over a range of allele frequencies can provide adequate information about the expected power of the study, while saving computational time and avoiding the ambiguity imposed by the subjective choice of a prior.

Lastly, we show that the assumptions about the genetic model can lead to larger differences in estimated power than varying allele or genotype frequencies. Thus, when the mode of inheritance is unknown (as is often the case), one has to examine the range of estimates obtained under different genetic models, in order to provide a realistic and accurate assessment of power for a given study design and sample size.

METHODS

In this section, we briefly review the methodology for calculating power in population-based association studies for fixed and random genotype frequencies.

Notation

Consider a diallelic marker with alleles A and B that have population frequencies p and $q = 1 - p$, respectively. Suppose that A is the minor, or less common, allele. There are three possible genotypes at this locus: BB, BA, AA (having 0, 1, and 2 copies of the minor allele, respectively), which, under the condition of HWE, have frequencies $p_0 = (1-p)^2$, $p_1 = 2p(1-p)$, and $p_2 = p^2$. Let $\mathbf{n} = (n_0, n_1, n_2) = (n_{BB}, n_{BA}, n_{AA})$ be the number of individuals with each genotype in a random sample of size N, such that $n_0 + n_1 + n_2 = N$. For a fixed N, the observed genotype counts follow a trinomial distribution, $\mathbf{n} \sim Mult(N; (1-p)^2, 2p(1-p), p^2)$. In most power and sample size calculations, one typically specifies the total sample size, N, assuming some or all components of the vector **n** are known. Specifically, when planning a genetic association study, the sample counts are often set equal to their expected values,

$$\mathrm{E}[\mathbf{n}] = (N(1-p)^2, 2Np(1-p), Np^2).$$

For a given **n**, the power of a statistical test T depends on the chosen significance level (α), the alternative hypothesis (H_A) (i.e., effect size), and the assumed genetic model (M). Adopting the notation given in Ambrosius et al. (2004), we will write this as $\pi(\mathbf{n}|\alpha, T, H_A, M)$. When the main interest is in the effect of random variation in **n** (while holding α, T, H_A, and M constant), we will denote it by $\pi(\mathbf{n})$ for short. Under the assumptions of HWE, the distribution of **n** depends on a single parameter, p. Hence power is often viewed as a function of the minor allele frequency, p, and calculated by setting $\mathbf{n} \equiv \mathrm{E}[\mathbf{n}]$, i.e.,

$$\pi(p|N) = \pi(\mathrm{E}[\mathbf{n}]). \tag{1}$$

However, this is clearly wrong because **n** is a random function of p. The correct approach would be to estimate $\mathrm{E}_{\mathbf{n}}[\pi(\mathbf{n})|p]$, since in general $\mathrm{E}_{\mathbf{n}}[\pi(\mathbf{n})|p] \neq \pi(\mathrm{E}[\mathbf{n}])$ unless the power function is

linear. In particular, for the case of genetic studies with quantitative traits it will be shown that $\pi(\mathrm{E}[\mathbf{n}]) \geq \mathrm{E}_{\mathbf{n}}[\pi(\mathbf{n})|p]$.

Further, the population allele frequency, p, itself is rarely known in advance and can be viewed as a random variable from some probability distribution. Ambrosius et al. (2004) introduce a beta prior for p and calculate $\pi(N) = \mathrm{E}_p\{\mathrm{E}_{\mathbf{n}}[\pi(\mathbf{n})|p]\}$. Their numerical examples show that $\pi(\mathrm{E}[\mathbf{n}])$ is generally greater than $\mathrm{E}_{\mathbf{n}}[\pi(\mathbf{n})|p]$, the difference ranging from small ($< 1\%$) to substantial (4%), and $\mathrm{E}_{\mathbf{n}}[\pi(\mathbf{n})|p]$ is in turn greater than $\mathrm{E}_p\{\mathrm{E}_{\mathbf{n}}[\pi(\mathbf{n})|p]\}$. They conclude, therefore, that ignoring random variation in $\mathbf{n}$ and p can lead to a considerable overestimation of power for a given study design and sample size (N). In the following sections we examine their claim and explore the dependence of $\pi(\mathbf{n})$ on the various parameters of the analysis.

Power of Association Studies with Quantitative Traits

In studies of quantitative traits, the association between genotype and phenotype is commonly tested using the general linear model. Let μ_i, $i = 0, 1, 2$, denote the mean trait value of individuals with i copies of the A allele. If the mode of inheritance for the trait is known, then the problem is one of linear regression,

$$\mu_i = \mu_0 + \beta x_i, \tag{2}$$

where $(x_0, x_1, x_2) = (0, d, 1)$, and $d = 1, 1/2, 0$ under the dominant ($\mu_1 = \mu_2$), additive ($\mu_1 = (\mu_0 + \mu_2)/2$), and recessive ($\mu_0 = \mu_1$) models, respectively. The corresponding F-ratio for testing the equality of means ($H_0 : \beta = 0$ versus $H_A : \beta \neq 0$) in this case follows a non-central F distribution with 1 and $N - 2$ degrees of freedom and a non-centrality parameter,

$$\lambda = \frac{\left(\sum_i n_i(x_i - \bar{x})(\mu_i - \mu)\right)^2}{\sum_i n_i(x_i - \bar{x})^2\sigma^2}, \tag{3}$$

where $\bar{x} = \sum_i n_i x_i \Big/ \sum_i n_i$, $\mu = \sum_i n_i \mu_i \Big/ \sum_i n_i$ and σ is the within-group standard deviation. Note that $\lambda = 0$, when all means are equal, and is strictly greater than zero, when at least one mean is different. Let $F_{\text{crit}} = F_{1-\alpha,1,N-2}$ be the $(1 - \alpha)100$ percentile of the central F distribution (for

which $\lambda = 0$) with 1 and $N - 2$ degrees of freedom. Then the power is given by

$$\pi = Pr(F_{1,N-2,\lambda} > F_{\text{crit}}), \tag{4}$$

which is a monotonically increasing function of λ. For details on testing the general linear hypothesis and the non-central F-distribution see, for example, Searle (1971), or Graybill (1976).

From the form of the non-centrality parameter it becomes clear that the power of an association test depends on the means $\{\mu_i\}$ (i.e., genetic model and effect size), the assigned scores $\{x_i\}$, and the underlying allele frequency, p, through the observed vector $\mathbf{n} = \{n_i\}$. In particular, when the scores x_i are chosen correctly (i.e., the assumed model is true), the expression (3) is maximized and reduces to

$$\lambda = \frac{\sum_i n_i(\mu_i - \mu)^2}{\sigma^2}. \tag{5}$$

If the scores are misspecified, on the other hand (i.e., the true model is different from the one assumed), the actual power may fall below the optimal value. Thus, when the mode of inheritance is unknown, the test for association is often performed using the one-way analysis of variance. The ANOVA statistic has 2 and $N - 3$ degrees of freedom and a non-centrality parameter (5). For the same value of λ, the 1-df test is more powerful than the 2-df test, but may be less powerful than the ANOVA whenever a wrong set of scores is used. For both tests, however, the power depends on the layout of the means under the true model. We illustrate the relationship between these parameters with an example (see Results).

Averaging Power

Here we briefly review the formulae for computing the average power, as described in Ambrosius et al. (2004). For a given p the expected power is the proper weighting of the estimates $\pi(\mathbf{n})$ by the probabilities of multinomial counts, $\mathbf{n}$,

$$\mathrm{E}_{\mathbf{n}}[\pi(\mathbf{n})|p] = \sum_{\sum n_i = N;\ n_i \geq 0\ \forall i} \pi(\mathbf{n}) \frac{N!}{n_0! n_1! n_2!} 2^{n_1} p^{2n_0 + n_1} (1-p)^{n_1 + 2n_2}. \tag{6}$$

When the allele frequency is unknown, one may use a beta prior to describe the uncertainty in p. For a beta prior with parameters γ and δ ($\gamma, \delta > 0$), the expected power is then given by

$$\mathrm{E}_p\{\mathrm{E}_{\mathbf{n}}[\pi(\mathbf{n})|p]\} = \frac{\Gamma(\gamma+\delta)}{\Gamma(\gamma)\Gamma(\delta)} \sum_{\sum n_i = N;\ n_i \geq 0\ \forall i} \pi(\mathbf{n}) \frac{N!}{n_0! n_1! n_2!} 2^{n_1} \times$$

$$\frac{\Gamma(2n_0 + n_1 + \gamma)\Gamma(n_1 + 2n_2 + \delta)}{\Gamma(2N + \gamma + \delta)}. \tag{7}$$

A beta prior constitutes a natural choice for modeling a frequency (i.e., a quantity varying between 0 and 1), and leads to a tractable solution for the expected power in (7). The mean of a beta(γ, δ) distribution is $\gamma/(\gamma+\delta)$, which can be chosen to take any value between 0 and 1 by varying γ and δ. For each specified mean, one can modify the variance by multiplying the two parameters by any positive constant, provided that the resulting $\gamma, \delta \geq 1$. The density gets more concentrated about the mean for larger γ and δ.

A reasonable question is which values of γ and δ to choose in order to get a good estimate of power. Ambrosius et al. (2004) discussed two methods for specifying the prior. The first is to perform a search of the literature or public databases (e.g., the HapMap), and take the allele frequency estimates reported in previous studies as the mean of the prior. The second approach is to make use of any available genotyping information obtained, for example, from a pilot study. Specifically, assuming a flat prior for p (where no information on the frequency is available before the initial data collection), the posterior distribution of p, conditional on observing x copies of the A allele in n individuals ($2n$ chromosomes), is $p|x, n \sim \mathrm{beta}(1+x, 1+2n-x)$. This updated distribution could be used as a prior for p in the power calculation. Ambrosius et al. (2004) found that the average power $\mathrm{E}_p\{\mathrm{E}_{\mathbf{n}}[\pi(\mathbf{n})|p]\}$ was generally lower than $\mathrm{E}_{\mathbf{n}}[\pi(\mathbf{n})|p]$, although somewhat close to it for more concentrated priors. Their examples, however, involved a prior distribution that was centered at the true population frequency. It is not clear, therefore, that such a method of averaging would produce a more conservative and realistic estimate of power if the mean of the prior were chosen incorrectly (that is if the power were averaged around the wrong value).

Computational Details

All computations were performed in the R statistical language and environment (see Web Resources), using built-in routines for the gamma function and the non-central F-distribution. Note that for a given N, there are $(N+1)(N+2)/2$ distinct outcomes for $\mathbf{n}$ to consider, so the computational complexity is $O(N^2)$. We calculate exact expectations for $N \leq 100$. For $N = 1000$ the expectation is approximated by a Monte Carlo method, drawing 10,000 random samples from the corresponding trinomial distribution for each p and averaging power over the observed samples. Programs in R are available from the authors upon request.

RESULTS

Power as a Function of Genetic Model and Allele Frequency

Our study examines the effects of various factors on the power function. We define the allelic effect (θ) as the standardized difference between the means (measured in standard deviations) of the two homozygous samples ($\theta = (\mu_0 - \mu_2)/\sigma$), assuming common variance (σ^2) within each genotype. For a fixed effect size, the means follow one of the three genetic models: additive ($\boldsymbol{\mu}_{ADD} = (0, \theta/2, \theta)$), dominant ($\boldsymbol{\mu}_{DOM} = (0, \theta, \theta)$), and recessive ($\boldsymbol{\mu}_{REC} = (0, 0, \theta)$). The effect size is controlled so that the power is not identically 1 over the entire allele frequency range. We shall assume for the moment that the sample sizes are fixed, i.e., $\mathbf{n} \equiv E[\mathbf{n}|p]$, and calculate power at $\alpha = 0.05$, using Equations (1) and (4), for p in the interval $(0, 0.5]$. As we see later, the estimates are essentially the same when the power is averaged over $\mathbf{n}$, as in Equation (6).

Figure 1(a) summarizes the relationship between power and minor allele frequency (p) when no a priori assumption about the genetic model is made and the 2-df test is used. The results illustrate that the power function varies substantially over the range of minor allele frequencies. In our specific example, the power estimates range from 11% up to 81% under the dominant model as the allele frequency changes. Secondly, the power varies widely among the three models. For instance, when $p = 0.25$, the power is 80.8% under the dominant, 39.1% under the additive, and 25.8% under the recessive model. Thus, if one had assumed the dominant model when estimating

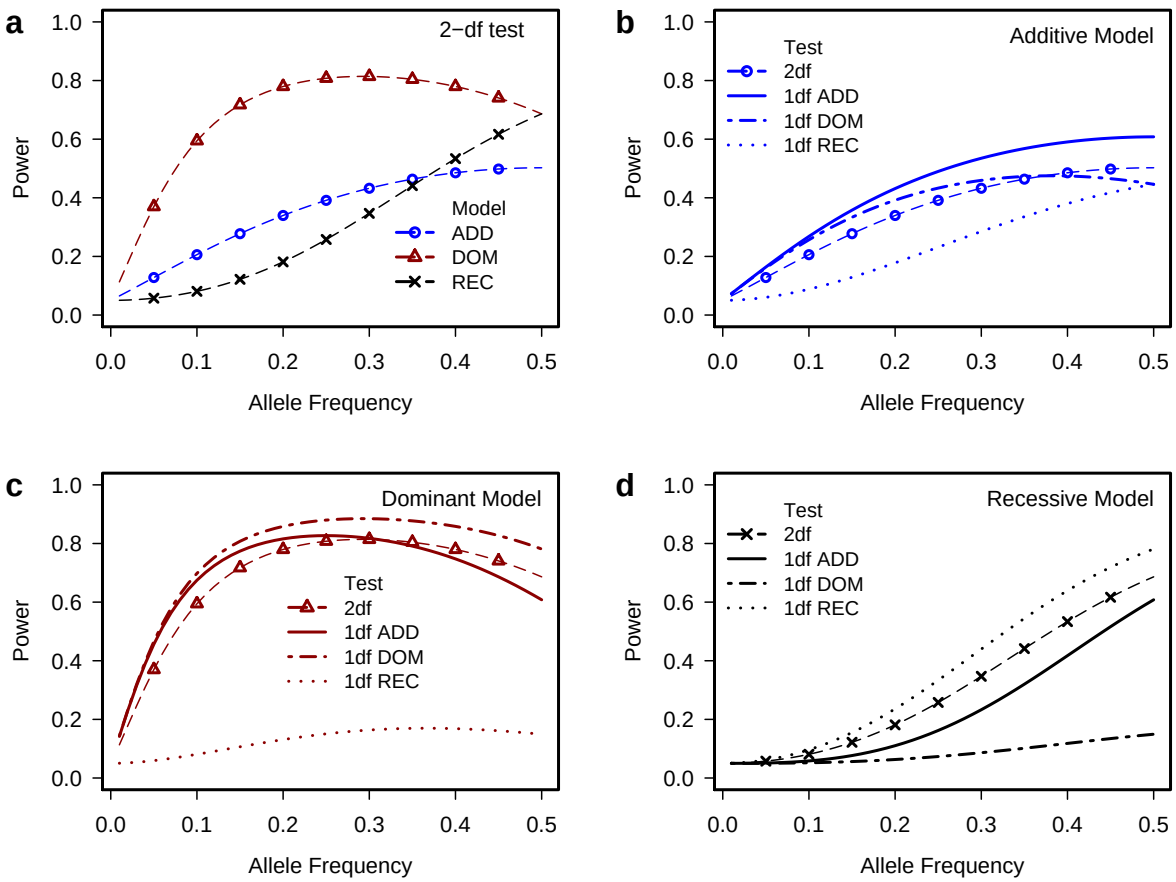

Figure 1: Power as a function of allele frequency under the three genetic models: dominant (DOM), additive (ADD), recessive (REC). $\boldsymbol{\mu}_{DOM} = (0, 0.2, 0.2)$, $\boldsymbol{\mu}_{ADD} = (0, 0.1, 0.2)$, $\boldsymbol{\mu}_{REC} = (0, 0, 0.2)$, $\sigma = 1$, $N = 1000$. True model is (a) unknown and 2-df test is used, (b) additive, (c) dominant, (d) recessive. Tests for association: - - - 2-df, model unknown; —— 1df assuming additive model; $- \cdot -$ 1-df assuming dominant model; $\cdots$ 1-df assuming recessive model.

power, while the means followed the additive pattern, the true power would be overestimated by about 41.6% (55% for the recessive model). Further, we observe that the power curves have a different shape and peak, depending on the true configuration of the means. Specifically, in the case of the additive model, the power is maximized when $p = 0.5$. For the dominant model the maximum is at $p = 0.29$, when the expected number of carriers is equal to the number of non-carriers ($n_{AA} + n_{AB} = n_{BB}$). Finally, the power is always higher under the dominant model, and is lowest under the recessive, unless the allele in question is very common ($p > 0.35$).

Figures 1(b)-(d) illustrate the result of performing the 1-df test assuming a particular model and set of scores. Whenever the scores do not agree with the true pattern of the means, there is a loss in power, which can be small or large depending on the model (true and supposed) and the allele frequency. Thus, if one had assumed a wrong model when testing the association, there would be a further difference between the estimated and true power due to misspecification of the model scores.

Example 1

To assess the effect of averaging power over random genotype frequencies, we re-examine an example given in Ambrosius et al. (2004) and extend the result to $p \in (0, 1)$. Assume the dominant mode of inheritance, in which genotypes AA and AB predispose individuals to one (e.g., disease) phenotype and genotype BB to another (e.g., normal) phenotype. Suppose that we have a total of eight subjects, $\mu_{AA} = \mu_{AB} = 1$, $\mu_{BB} = 3$, and the within-group variance (σ^2) is 1. The association is tested using the 1-df regression model, which is equivalent to comparing the combined sample of AA and AB individuals to BB individuals. For a total sample size $N = 8$, there can be anywhere from zero to eight carriers of the A allele. For each outcome, we first calculate power using the standard method (listed in Table 1 in the column headed $\pi(\mathbf{n})$). Next, rather than assuming a single allele frequency, as was done in the original example, we determine the maximum likelihood estimate of the allele frequency for each outcome, by $\hat{p} = 1 - \sqrt{n_{BB}/N}$, and evaluate the expected power, $\mathrm{E}_{\mathbf{n}}[\pi(\mathbf{n})|\hat{p}]$. The last column in the table shows the difference between the two sets of estimates. For each $\hat{p}$, the average power over the distribution of multinomial counts

Table 1: Power for a Dominant Model, $N = 8$.

$n_{AA} + n_{AB}$	n_{BB}	$\hat{p}$	$\pi(\mathbf{n})$	$\mathrm{E}_{\mathbf{n}}[\pi(\mathbf{n})\|\hat{p}]$	$\pi(\mathbf{n}) - \mathrm{E}_{\mathbf{n}}[\pi(\mathbf{n})\|\hat{p}]$
0	8	0	0.00	0.00	0.00
1	7	0.06	0.3507	0.2859	0.0648
2	6	0.13	0.5373	0.4653	0.0721
3	5	0.21	0.6295	0.5633	0.0662
4	4	0.29	0.6569	0.5945	0.0624
5	3	0.39	0.6295	0.5633	0.0662
6	2	0.50	0.5373	0.4653	0.0721
7	1	0.65	0.3507	0.2859	0.0648
8	0	1	0.00	0.00	0.00

is indeed several percentage points lower than that calculated for fixed **n**. Figure 2(a) presents this relationship graphically. We note that for $\hat{p} > 0.5$, the model should technically be termed 'recessive', however we demonstrate the relationship over the entire range of allele frequencies for completeness. It is clear from the graph that the power function, $\pi(\mathbf{n})$, is concave in the sample counts. Hence, for any p, $\pi(\mathrm{E}[\mathbf{n}]) \geq \mathrm{E}_{\mathbf{n}}[\pi(\mathbf{n})|p]$, by Jensen's inequality (see for example Casella & Berger 2001). That is, evaluating the power at the expected sample counts will systematically overestimate the expected power for each given allele frequency.

However, as we repeat the analysis for $N = 20$, 100, and 1000, we see that as N increases, the size of the bias decreases dramatically (Figure 2(b)-(d)). For each N the effect size, θ, is scaled appropriately so that the power is bound away from 100%. In the examples shown, the bias in estimated power is close to 7% for $N = 8$; it ranges between $2 - 3\%$ for $N = 20$, generally stays between $0.2 - 1\%$ when $N = 100$, and does not exceed 0.2%, except at the very boundaries, for $N = 1000$ (Supplementary Tables 1-3). The reason for larger bias at the boundaries is, perhaps, that for rare alleles the expected number of carriers remains very small, even with large sample

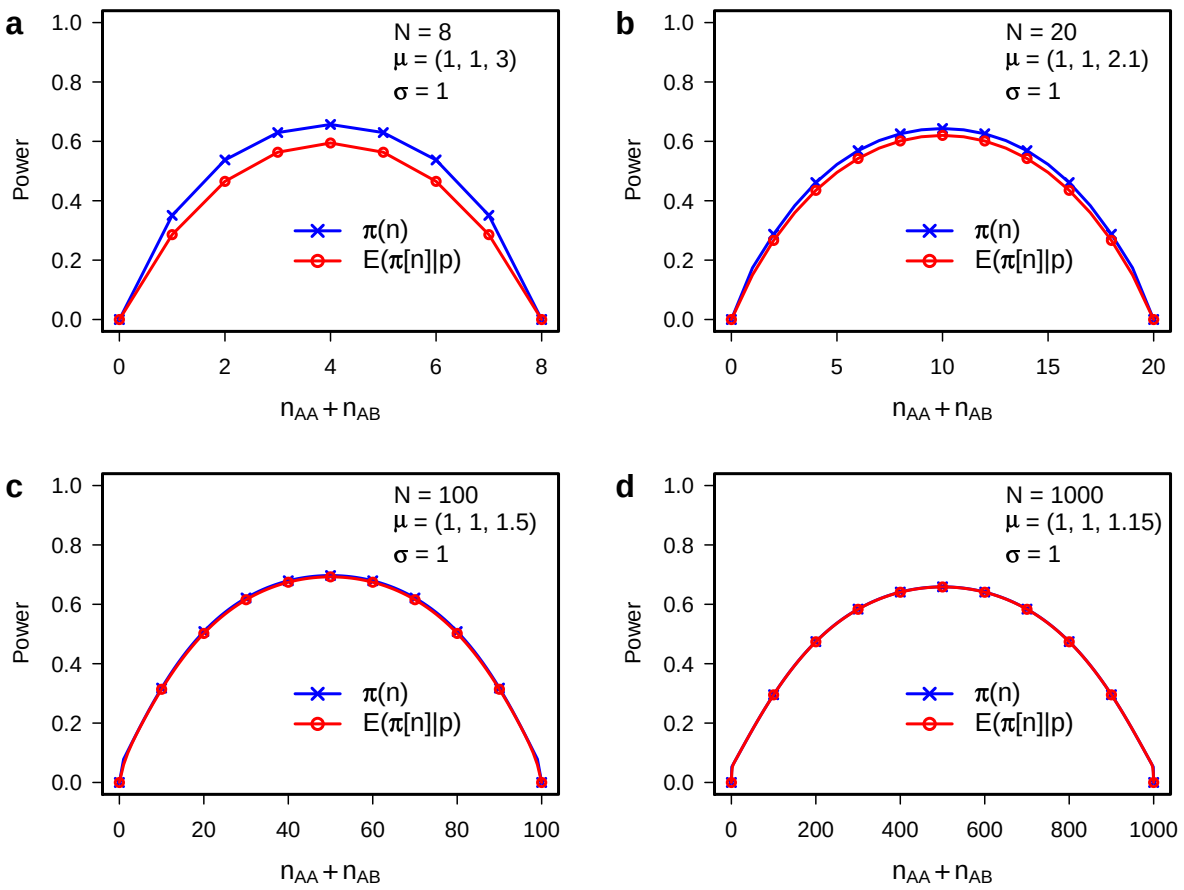

Figure 2: Estimated power under a dominant model with fixed and random genotype counts. (a) $N = 8$, (b) $N = 20$, (c) $N = 100$, (d) $N = 1000$.

sizes. As long as there are at least 3 people in the smaller group, however, the bias approaches zero for large N.

To demonstrate the effect of uncertainty in the allele frequencies, Ambrosius et al. (2004) assume a beta$(5, 5\sqrt{2}+5)$ prior for p, which implies $\mathrm{E}[p] = .29$. The expected power using this prior is $\mathrm{E}_p\{\mathrm{E}_{\mathbf{n}}[\pi(\mathbf{n})|p]\} = .5494$. This is several percentage points lower than both $\mathrm{E}_{\mathbf{n}}[\pi(\mathbf{n})|p] = .5945$ and $\pi(\mathrm{E}[\mathbf{n}]) = .6569$ for $p = .29$. A plot of the beta$(5, 5\sqrt{2}+5)$ density reveals that the central 95% probability interval for the distribution extends from about .1 to .52 (Figure 3). Given the potential range of variation in the power function, as exemplified in Figures 1 and 2, averaging power over such a wide interval will clearly result in a much lower estimate of power. If more information about the allele frequency can be assumed, and the variance of the prior reduced so that the 95% probability interval lies between .2 and .4, as with a beta$(25, 25\sqrt{2}+25)$, the expected power becomes $\mathrm{E}_p\{\mathrm{E}_{\mathbf{n}}[\pi(\mathbf{n})|p]\} = .5850$, which is only 1% less than the power calculated for a fixed parameter value of $p = .29$. On the other hand, if the true allele frequency in the new study population is $p = .18$ (which is well within the range of values allowed by the beta$(5, 5\sqrt{2}+5)$ prior), the expected power for this p is $\mathrm{E}_{\mathbf{n}}[\pi(\mathbf{n})|p] = .5345$. In this case, averaging power using the above prior would actually result in an overestimate of power. Thus, the approach of averaging power for an unknown allele frequency requires that the mean of such a prior be specified correctly in order for it to produce an accurate and conservative estimate.

Example 2

In practice, when planning a genetic study, rather than estimating power for a list of possible genotype counts (which becomes difficult when $\mathbf{n}$ is 3-dimensional, as it is under the additive model), one could construct a power curve for a full range of allele frequencies. We compute and compare such power curves for the dominant (recessive) and additive genetic models, with fixed and random sample genotype counts. Letting $p = 1\%, \ldots, 99\%$, we determine the expected genotype counts, $\mathrm{E}[\mathbf{n}]$, and evaluate $\pi(\mathrm{E}[\mathbf{n}])$ and $\mathrm{E}_{\mathbf{n}}[\pi(\mathbf{n})|p]$ for each point in the range, for $N = 20, 100$, and 1000. We note that a dominant model for the A allele when $p < 0.5$, becomes a recessive for the B allele when $p \geq 0.5$. The alternative hypotheses are chosen so that the power

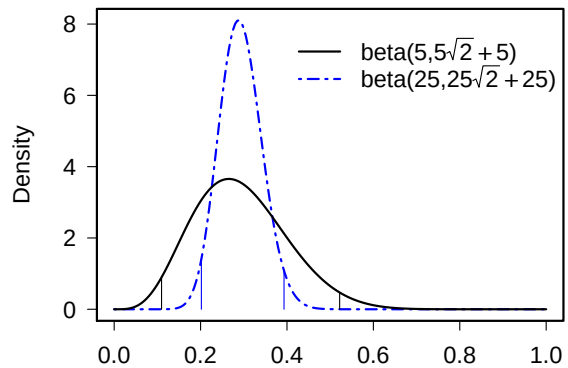

Figure 3: Beta densities used in Example 1. Vertical segments mark the limits of a central 95% probability interval.

function reaches approximately the same maximum value for each N. The results are presented in Figure 4. Note that when the sample size is small ($N = 20$), several distinct allele frequencies lead to the same expected sample counts due to rounding error, giving the graph a step-like appearance. The rounding means that for small sample sizes, the power calculated for a fixed set of genotype counts can be either lower or higher than the average power. Plotting the power surface as a function of two of the three sample counts (e.g., n_0 and n_2), however, reveals that π remains concave in the counts, so that theoretically $\pi(\mathrm{E}[\mathbf{n}]) \geq \mathrm{E}_{\mathbf{n}}[\pi(\mathbf{n})|p]$ for any p. As the sample size increases, the steplike pattern disappears and the two curves become essentially identical. As demonstrated in Figure 4(c)-(d), for N greater than 100 the difference between the two sets of estimates is generally less than 1%.

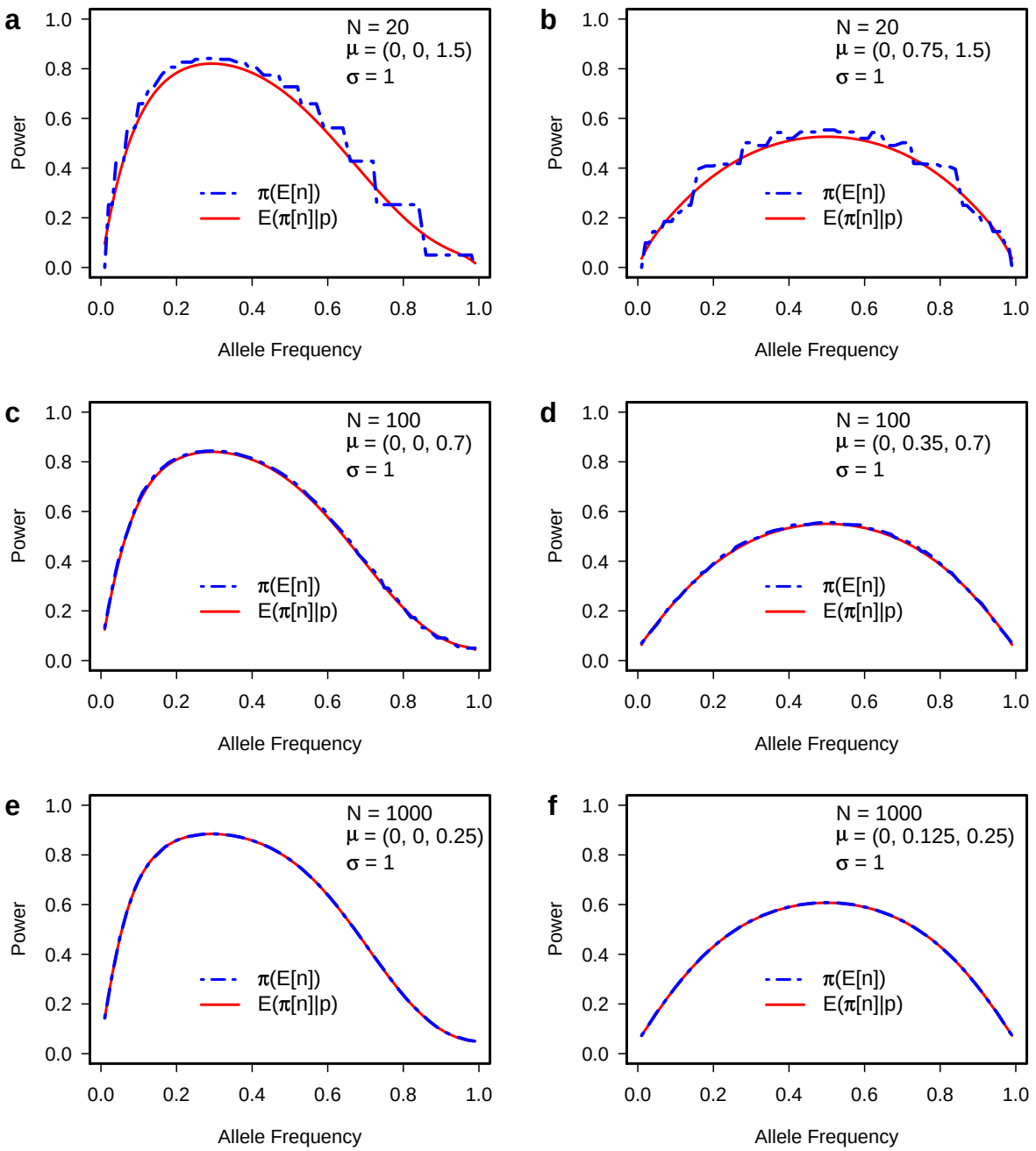

Figure 4: Power and expected power as a function of allele frequency for (a), (c), (e) - dominant/recessive model, (b), (d), (f) - additive model.

DISCUSSION

A power calculation for a genetic association study requires many assumptions, such as allele and genotype frequencies, allelic effect size, and the mode-of-inheritance model for the trait of interest. Since most of these parameters are unknown, one must properly account for the uncertainty in the assumed values in order to obtain an accurate estimate of power for a given design and sample size. In this work we have examined the effect of random variation in, and the relative importance of, the specified factors in determining power.

Our results show that for any given allele frequency, the effect of ignoring sampling variation in the genotype counts diminishes rapidly with sample size. Although the approach of fixing genotype frequencies at their expected values tends to overestimate power, the size of the estimation bias becomes negligible as the sample size grows. Therefore, in most practical situations (for N as small as 100, and most certainly for $N > 200$), using expected counts for calculating power seems to be a reasonable approach. While for very low allele frequencies ($p \leq .01$) the bias did not approach zero equally fast (and remained close to 2% in our examples), we note that for rare alleles the power to detect an association is generally low, regardless of the calculation method used, and is unlikely to be 'inflated', even when the variation in the genotype counts is ignored.

The uncertainty in the allele frequency, on the other hand, can result in considerable errors in estimated power. To overcome this problem, one possible solution is to place a prior distribution on the unknown frequency and average these power estimates over the prior. Ambrosius et al. (2004) modeled allele frequencies with a beta distribution, using the estimates obtained from previous studies to specify the parameters of the prior. However, they always compared their average power to that calculated for the mean of the prior, implicitly assuming that the mean was identical to the true population allele frequency in the new study. We note that in practice the mean is likely to differ from the true allele frequency, since the estimates obtained from previous studies are only estimates themselves of an unknown population quantity. In the extreme situation, when the sample estimates happen to be sufficiently far from the true population value (i.e., the prior is misspecified), we have shown in Example 1 that it is possible to overestimate power even after integrating over a

fairly diffuse prior, albeit to a smaller degree than by using a single frequency estimate. Thus, even though the method of averaging can mitigate the consequences of misspecifying an unknown allele frequency, it still relies on assumptions about the prior distribution and gives only an estimate of the expected power.

Other authors considered the idea of integrating over the unknown parameter values when estimating the power of association studies. Schork (2002) proposed using empirically estimated distributions of allele frequencies as a prior in the power calculation. Although his approach appears similar in principle to that discussed in Ambrosius et al. (2004), we note that the two articles are targeting quite different sources of randomness. Schork (2002) was concerned with assessing the average power for a study with multiple markers, while Ambrosius et al. (2004) focused on estimating the expected power for a particular candidate polymorphism. Clearly, very different prior distributions would be required in these two situations to obtain a reasonable estimate of power. In the first case, the expectation should be taken over all parameter values that would be observed in, say, a genome-wide association study (GWAS), so a very diffuse prior would have to be specified. In the second, the goal of averaging is to account for sampling variability in allele frequency estimates, obtained from previous studies, and a much more concentrated prior may be warranted.

Regardless of the purpose of the power calculation, we argue that it is more instructive to perform a power analysis over a range of parameter settings, rather than present a single average number (see Gordon and Finch 2005 for a similar discussion). For example, when planning a whole-genome investigation one can expect to observe a wide range of allele frequencies. Since one does not know in advance the properties of the causative allele, a more conservative approach might be to select a sample size that would ensure an adequate level of power for all tested markers, including those with some minimum allele frequency (e.g., 5%). Averages can certainly be useful, but they can also conceal the range of variation in the quantity of interest. We believe that presenting an entire power curve for a given effect size is computationally simpler and actually provides the researcher with more information about the expected power of a study.

Finally, the averaging methods discussed in this article fail to address another important source of uncertainty in estimating the power of association studies - the mode of inheritance for the

trait in question. We have shown that the power estimates vary vastly between the different genetic models, especially for moderately common alleles. In particular, by assuming a dominant model at the design stage, the power will almost always be overestimated unless the true effect is dominant. To overcome the problem, we recommend plotting power curves for each of the three genetic models and a fixed effect size, similar to the way presented in Figure 1.

While we have focused on association studies of quantitative traits, we believe that similar conclusions would apply in the case of other study designs. Both Ambrosius et al. (2004) and Zheng et al. (2005) observed the convergence of power estimates for fixed and random genotype counts whenever the sample sizes were large enough (see examples with binary traits and case-control studies in Ambrosius et al. 2004), although they did not study the relationship systematically. It would be valuable, however, to explore the dependence of power on variation in unknown parameters under different study designs in more detail.

One issue that has not been addressed in the current work is the sensitivity of power to deviations of genotype frequencies from Hardy-Weinberg proportions. While the assumption of HWE is reasonable in cohort studies, it may be seriously suspect in most case-control comparisons (see, for example, Wittke-Thompson et al. 2005). We note in conclusion that the assumption of HWE does not necessarily overestimate power. In fact, one can show that certain deviations (e.g., absence of heterozygotes under the additive model) can lead to higher power to detect an association. However, a different set of genotype frequencies would then be required for estimating the power for a given allele.

Acknowledgements

We thank Drs. Helen H. Hobbs and Jonathan Cohen for helpful discussions and general support over the course of this project.

Web Resources

http://www.R-project.org

References

Ambrosius, W.T., Lange, E.M., Langefeld, C.D. (2004) Power of genetic association studies with random allelle frequencies and genotype distributions. *Am. J. Hum. Genet.* 74, 683-693.

Casella, G. & Berger, R.L. (2001) *Statistical inference* (2nd ed.). California: Duxbury Press.

Gordon, D. & Finch, S.J. (2005) Factors affecting statistical power in the detection of genetic association. *J. Clin. Invest.* 115, 1408-1418.

Graybill, F.A. (1976) *Theory and application of the linear model.* California: Duxbury Press.

Schork, N.J. (2002) Power calculations for genetic association studies using estimated probability distributions. *Am. J. Hum. Genet.* 70, 1480-1489.

Searle, S.R. (1971) *Linear models.* New York: Jonh Wiley & Sons.

The International HapMap Consortium (2005) A haplotype map of the human genome. *Nature* 437, 1299-1320.

Wittke-Thompson, J.K., Pluzhnikov, A., Cox, N.J. (2005) Rational inferences about departures from Hardy-Weinberg equilibrium. *Am. J. Hum. Genet.* 76, 967-986.

Zheng, G., Joo, J., Ganesh, S.K., Nabel, E.G., Geller, N.L. (2005) On averaging power for genetic association and linkage studies. *Hum. Hered.* 59, 14-20.

Supplementary Materials

Supplementary Table 1. Power for a dominant model, $N = 20$. $\mu_{AA} = \mu_{AB} = 1, \mu_{BB} = 2.1, \sigma = 1$.

Supplementary Table 2. Power for a dominant model, $N = 100$. $\mu_{AA} = \mu_{AB} = 1, \mu_{BB} = 1.5, \sigma = 1$. Selected values of **n** are shown.

Supplementary Table 3. Power for a dominant model, $N = 1000$. $\mu_{AA} = \mu_{AB} = 1, \mu_{BB} = 1.15, \sigma = 1$. Selected values of **n** are presented.