Genetic Epidemiology 31: 776–788 (2007)

Optimal Designs for Two-Stage Genome-Wide Association Studies

Andrew D. Skol,[1,2*] Laura J. Scott,[1] Gonçalo R. Abecasis,[1] and Michael Boehnke[1]

[1]*Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan*
[2]*Department of Medicine, Section of Genetic Medicine, University of Chicago, Chicago, Illinois*

Genome-wide association (GWA) studies require genotyping hundreds of thousands of markers on thousands of subjects, and are expensive at current genotyping costs. To conserve resources, many GWA studies are adopting a staged design in which a proportion of the available samples are genotyped on all markers in stage 1, and a proportion of these markers are genotyped on the remaining samples in stage 2. We describe a strategy for designing cost-effective two-stage GWA studies. Our strategy preserves much of the power of the corresponding one-stage design and minimizes the genotyping cost of the study while allowing for differences in per genotyping cost between stages 1 and 2. We show that the ratio of stage 2 to stage 1 per genotype cost can strongly influence both the optimal design and the genotyping cost of the study. Increasing the stage 2 per genotype cost shifts more of the genotyping and study cost to stage 1, and increases the cost of the study. This higher cost can be partially mitigated by adopting a design with reduced power while preserving the false positive rate or by increasing the false positive rate while preserving power. For example, reducing the power preserved in the two-stage design from 99 to 95% that of the one-stage design decreases the two-stage study cost by ~15%. Alternatively, the same cost savings can be had by relaxing the false positive rate by 2.5-fold, for example from 1/300,000 to 2.5/300,000, while retaining the same power. *Genet. Epidemiol.* 31:776–788, 2007.

Key words: genome-wide association; two-stage design; association; optimal design

The supplementary materials described in this article can be found at http://www.interscience.wiley.com/jpages/0741-0395/suppmat.
*Correspondence to: Andrew D. Skol, Department of Medicine, Section of Genetic Medicine, University of Chicago, 5841 South Maryland Avenue, W611A – MC6091, Chicago, Illinois 60637. E-mail: askol@uchicago.edu
Received 13 December 2006; Accepted 17 April 2007
Published online 4 June 2007 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20240

INTRODUCTION

Genome-wide association (GWA) scans represent an exciting opportunity to identify common genetic variants that predispose to human disease [Kruglyak, 1999; Risch and Merikangas, 1996; Cardon and Bell, 2001; Hirschhorn and Daly, 2005]. GWA studies require efficient study designs to examine hundreds of thousand of markers on the hundreds or thousands of samples required to detect disease predisposing variants of modest effect. Even though chip-based genotyping products have decreased genotyping costs dramatically, GWA studies are still very expensive.

In this paper, we demonstrate how to construct powerful and cost-effective GWA studies using two-stage designs. Two-stage designs gain their efficiency by excluding markers for further testing that show little evidence of association in the first stage. This can substantially reduce the genotyping requirements, and therefore cost, while preserving much of the power of the corresponding one-stage design in which all samples are genotyped on all markers [Satagopan and Elston, 2003; Satagopan et al., 2004; Thomas et al., 2004]. We consider two-stage designs that genotype all markers on a proportion of the sample in stage 1, and follow-up a small proportion of the most significantly associated markers on the remaining samples in stage 2. At the end of stage 2, the overall evidence for association is evaluated by combining the statistics from the two stages [Skol et al., 2006].

Here, we describe how to design optimal two-stage GWA studies given a fixed set of available samples. We explore how the ratio of stage 2 to stage 1 per genotype cost affects the optimal two-stage design and its cost, and how the target power and the false positive rate of the two-stage design can be used to manage study cost.

We find that given large ratios of stage 2 to stage 1 per genotype cost optimal designs genotype a larger proportion of samples in stage 1, follow up a smaller proportion of markers in stage 2, and have greater overall cost than when cost ratios of per genotype cost are lower. We also demonstrate that the genotyping cost of optimal two-stage designs can be further reduced by modestly decreasing study power while maintaining the false positive rate, or by modestly increasing the false positive rate while

maintaining power. We believe our framework for designing two-stage GWA studies will help investigators construct designs that balance the goals of high power, low false positive rate, and reasonable cost.

METHODS

IDENTIFYING AN OPTIMAL TWO-STAGE DESIGN

Four parameters define our two-stage GWA design: M, the number of markers tested for association; $2N$, the number of samples genotyped; π_{samples}, the proportion of samples genotyped in stage 1; and π_{markers}, the proportion of markers selected for follow-up in stage 2. We treat M and N as fixed, and π_{markers} and π_{samples} as values to be selected by the investigator. We define the optimal two-stage design as that with values (π_{markers}, π_{samples}) which attains the desired power at the minimum total genotyping cost. We specify the desired power as a proportion π_{power} of the one-stage design power. Typically π_{power} is not much less than 1. The optimal two-stage design is influenced by π_{power}, by the stage 2 to stage 1 per genotype cost ratio R, and by the marker-wise false positive rate α_{marker}. Our approach to designing optimal two-stage designs differs from that of Wang et al. [2005], whose goal is to attain a fixed power and allow the sample size to vary.

For simplicity, in what follows, we assume a case-control study of disease association. Generalization to family-based designs and quantitative traits is straightforward. We simplify the power calculation by assuming that all markers are in linkage equilibrium, which allows the marker-wise false positive rate to be written as $\alpha_{\text{marker}} \approx W/M$, where W is the acceptable number of false positives in the GWA study. For example, for a genome-wide scan with 300,000 markers in which one false positive is considered tolerable, we set $\alpha_{\text{marker}} = 1/300{,}000 \approx 3.3 \times 10^{-6}$.

We identify the optimal two-stage design which preserves $\pi_{\text{power}} \times 100\%$ of the corresponding one-stage study's power using Brent's algorithm [1973] to select the values of π_{samples} and π_{markers} that minimize cost when power is held constant. A unique optimal design exists for each value of π_{power}. A more detailed description of the optimization algorithm may be found in the online supplementary material (http://interscience.wiley.com/jpages/0741-0395/suppmat).

COST OF THE TWO-STAGE DESIGN

The expected genotyping cost for a two-stage study is $C_{\text{two-stage}} = M\ 2\ N\ \pi_{\text{samples}}\ c_1 + M\ \pi_{\text{markers}}\ 2\ N\ (1-\pi_{\text{samples}})\ c_2$, where c_i is the per genotype cost for stage i. The genotyping cost of the corresponding one-stage study, in which all samples are genotyped on all markers in a single stage, is $C_{\text{one-stage}} = M\ 2\ N\ c_1$. Note that, for simplicity, we assume the same per genotype cost c_1 for both the one-stage design and stage 1 of the two-stage design. In the Discussion section, we consider the impact of genotyping costs that vary according to the number of samples genotyped. Defining $R = c_2/c_1$ as the ratio of stage 2 to stage 1 per genotype cost, we can express the cost of the two-stage design as a proportion of the one-stage design cost as $\pi_{\text{cost}} = C_{\text{two-stage}}/C_{\text{one-stage}} = \pi_{\text{samples}} + \pi_{\text{markers}}\ (1-\pi_{\text{samples}})\ R$. In Appendix A, we describe how to accommodate a per genotype cost structure based on custom genotyping arrays that restrict the possible number of markers that can be genotyped in stage 2.

CALCULATING POWER FOR TWO-STAGE DESIGNS

Assume N cases and N controls are available for genotyping and that a proportion π_{samples} of these samples are genotyped in stage 1. Evidence for association at stage 1 is evaluated for each of the M markers and used to select $\pi_{\text{markers}}\ M$ markers for follow-up genotyping in the remaining $(1-\pi_{\text{samples}})$ N cases and $(1-\pi_{\text{samples}})$ N controls in stage 2.

To evaluate evidence for association at stage 1, let $\hat{p}_1'$ and $\hat{p}_1$ be the estimated risk allele frequencies in cases and controls, respectively, and define the test statistic

$$z_1 = \frac{\hat{p}_1' - \hat{p}_1}{\sqrt{[\hat{p}_1'(1-\hat{p}_1') + \hat{p}_1(1-\hat{p}_1)]/(2N\pi_{\text{samples}})}}.$$

Under the null hypothesis of no association and assuming $N\ \pi_{\text{samples}}$ is sufficiently large, z_1 follows an approximate Normal distribution with mean 0 and variance 1. The threshold T_1 for selecting markers for follow-up is determined using the quantiles of the standard Normal distribution by finding T_1 such that $P(|z_1| > T_1) = \pi_{\text{markers}}$. π_{markers} thus has two interpretations: the false positive rate for stage 1, and the expected proportion of markers followed up in stage 2, provided the number of disease related variants is small compared with the number of markers M.

In stage 2, we calculate z_2, a statistic analogous to z_1 constructed with stage 2 data only. We then compare the statistic

$$z_{\text{joint}} = \sqrt{\pi_{\text{samples}}}\, z_1 + \sqrt{1-\pi_{\text{samples}}}\, z_2 \qquad (1)$$

to a significance threshold T_{joint} chosen to control the false positive rate at the desired level, α_{marker}. Rather than combining the raw genotype data, the statistic

z_{joint} combines evidence for association summarized in the stage 1 and stage 2 statistics. This construction eliminates the risk of false positives generated by heterogeneity between the stage 1 and stage 2 sample populations. The false positive rate corresponding to thresholds T_1 and T_{joint} is $\alpha_{\text{marker}} = P(|z_1| > T_1 \text{ and } |z_{\text{joint}}| > T_{\text{joint}}) = P(|z_1| > T_1) P(|z_{\text{joint}}| > T_{\text{joint}} | |z_1| > T_1)$, and can be calculated numerically by evaluating a simple integral (see equation (2) below).

Stage 1. Power for stage 1 is the probability that a disease predisposing variant is selected for follow-up in stage 2. The statistic z_1 in large samples follows an approximate Normal distribution with mean

$$\mu_1 = \frac{p' - p}{\sqrt{[p'(1-p') + p(1-p)]/(2N\pi_{\text{samples}})}}$$

and variance

$$F(p,p') = \frac{(p' + 3p - 2p^2 - 2p'p)^2(p'(1-p')) + (p + 3p' - 2p'^2 - 2p'p)^2(p(1-p))}{4(p'(1-p') + p(1-p))^3},$$

where p' and p are the frequencies of the disease predisposing allele in the case and control populations, respectively. This formulation of $F(p', p)$ was suggested by Bukszár and van den Oord and is justified in Appendix B [Bukszar and van den Oord, 2006]. $F(p', p)$ equals one under the null hypothesis and is very close to one for most other genetic models. The probability that a marker is selected for stage 2 genotyping is

$$P_1 = 1 - \Phi\left[\frac{T_1 - \mu_1}{\sqrt{F(p',p)}}\right] + \Phi\left[\frac{-T_1 - \mu_1}{\sqrt{F(p',p)}}\right],$$

where $\Phi[x]$ is the cumulative distribution function for the standard Normal distribution evaluated at x.

Stage 2. Conditional on the observed value a for the stage 1 statistic z_1, the statistic for joint analysis z_{joint} follows an approximate Normal distribution with mean

$$\mu_{\text{joint}|z_1=a} = \sqrt{\pi_{\text{samples}}}a + \sqrt{(1-\pi_{\text{samples}})} \times \frac{p' - p}{\sqrt{[p'(1-p') + p(1-p)]/[2N(1-\pi_{\text{samples}})]}}$$

and variance $(1-\pi_{\text{samples}})\, F(p', p)$. The probability of detecting an association in stage 2 given the marker was selected for follow-up after stage 1 is

$$P_{\text{joint}} = P(|z_{\text{joint}}| > T_{\text{joint}}|I) = \int_{-\infty}^{-T_1} \left[P(z_{\text{joint}} > T_{\text{joint}}|z_1 = x) + P(z_{\text{joint}} < -T_{\text{joint}}|z_1 = x)\right] f(x|I)dx + \int_{T_1}^{\infty} \left[P(z_{\text{joint}} > T_{\text{joint}}|z_1 = x) + P(z_{\text{joint}} < -T_{\text{joint}}|z_1 = x)\right] f(x|I)dx, \quad (2)$$

where $f(x \mid I)$ is the probability density function of z_1 given event I defined as $|z_1| > T_1$. The power of the joint analysis is $P_1 P_{\text{joint}}$.

We have developed an interactive power calculator CaTS (www.sph.umich.edu/csg/abecasis/CaTS/) for two-stage GWA studies which implements these calculations and allows investigators to calculate power, thresholds T_1 and T_{joint}, and the genotyping cost for any user specified design. Our power calculator can aid investigators in designing optimal two-stage studies and also provides power calculations for one-stage designs for comparison purposes.

RESULTS

Unless otherwise noted, our examples assume a GWA study with 1,000 cases and 1,000 controls genotyped on $M = 300{,}000$ independent markers, a marker-wise false positive rate $\alpha_{\text{marker}} = 1/300{,}000 \approx 3.3 \times 10^{-6}$ (corresponding to one expected false positive per genome), a disease with 10% prevalence, and a susceptibility variant of modest effect (multiplicative genotype relative risk GRR = 1.375 and risk allele frequency in the controls $p = .35$). The one-stage design's power for this setting is 80%. Although the optimal two-stage design depends on the false positive rate, the trends we report generally vary only slightly for other false positive rates and genetic effect sizes; we note when this is not the case. Also, while we discuss results for a single variant, we would have the same power to detect any variant with the given effect size.

INFLUENCE OF STAGE 2 TO STAGE 1 PER GENOTYPE COST RATIO *R* ON OPTIMAL TWO-STAGE DESIGNS

The stage 1 per genotype cost has fallen dramatically due to competitive pricing and the economies of scale inherent in chip-based genotyping technologies which genotype standard sets of hundreds of thousands of markers. The stage 2 per genotype cost is greater since many fewer markers are genotyped and the markers selected vary between studies and so require study-specific genotyping arrays. Hence, $R > 1$.

Larger values of R result in optimal designs which shift the genotyping burden from stage 2 to stage 1 (Fig. 1). As R increases from 1 to 40, the proportion of samples genotyped in stage 1 π_{samples} increases from .37 to .63, the proportion of markers followed up in stage 2 π_{markers} decreases from .124 to .004, and total genotyping cost increases from 45 to 69% of the one-stage design cost. Interestingly, all of the increase in genotyping cost is due to increased stage 1 cost (black bars), while stage 2 cost remains nearly unchanged or decreases (gray bars). In

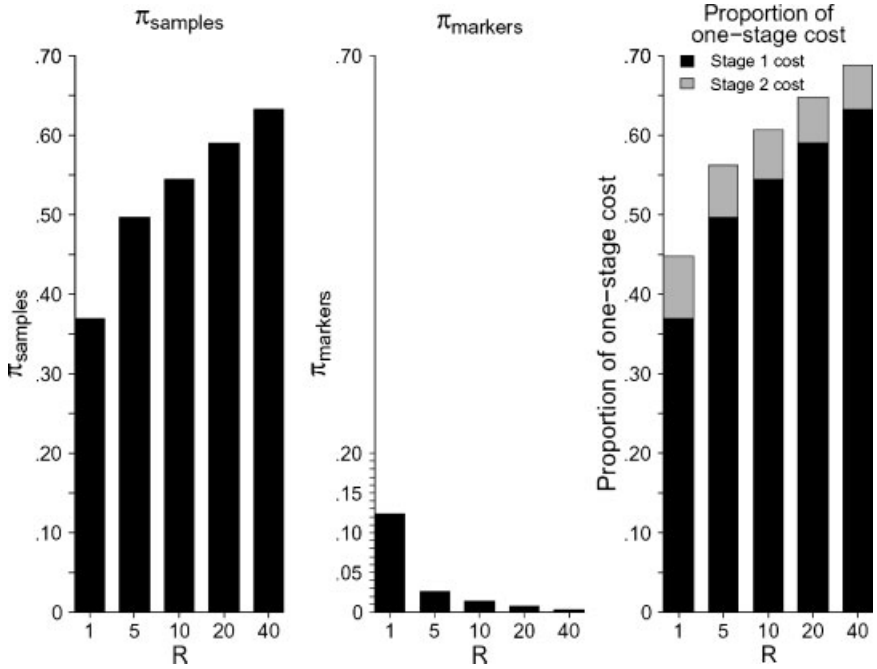

Fig. 1. The designs above employ 1,000 cases and 1,000 controls and have 79% power ($\pi_{power} = .99$) to detect a variant with GRR = 1.375 and $p = .35$ for a disease with prevalence = .10 when $\alpha_{marker} = 1/300{,}000$. The black bars in the left and center panels denote the values of $\pi_{samples}$ and $\pi_{markers}$ of the optimal two-stage design. The right panel shows the cost of the optimal two-stage design and how it is divided between stage 1 (black bar) and stage 2 (gray bar).

our example, the expense of stage 2 genotyping decreases from 7.8 to 5.5% of the one-stage study cost.

MISSPECIFICATION OF THE PER GENOTYPE COST RATIO *R*

An important consideration when designing a two-stage GWA study is the impact that misspecifying the per genotype cost ratio R has on study cost, since R may change between the time of study planning and execution. The inefficiency resulting from designing a study with R misspecified is illustrated in Figure 2. It shows the cost of two-stage designs which preserve $\pi_{power} = 99\%$ of the one-stage design power for a range of $\pi_{samples}$ values when R is 1, 5, 10, 20, and 40. The minimum of each cost curve occurs at the optimal $\pi_{samples}$ given R. If we define R_E as the cost ratio estimated at design time and R_A as the actual cost ratio when the study is carried out, then inefficiency due to misspecifying R ($R_E \neq R_A$) can be quantified as the difference between the estimated optimal design cost using R_E and the cost of that design carried out using R_A. In Figure 2, this is the length of the vertical line drawn from the lowest point on the R_E cost curve to the cost curve of R_A. An alternative measure of

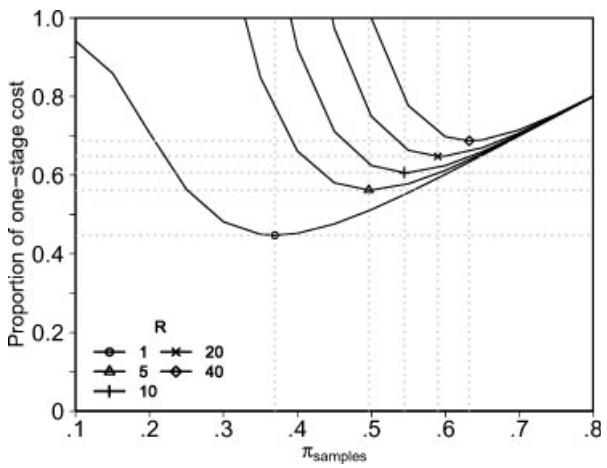

Fig. 2. The designs above employ 1,000 cases and 1,000 controls and have 79% power ($\pi_{power} = .99$) to detect a variant with GRR = 1.375 and $p = .35$ for a disease with prevalence = .10 when $\alpha_{marker} = 1/300{,}000$. The curves show how cost changes as a function of $\pi_{samples}$ for five per genotype cost ratios (1, 5, 10, 20, and 40). The value of $\pi_{samples}$ of the optimal two-stage design occurs at the minimum of each curve.

inefficiency is the difference in cost between the study above designed using R_E but carried out using R_A and the cost of a study designed and carried out

using the actual R, R_A. This difference is the vertical distance between lowest point on the R_A cost curve and the point on the R_A curve above the lowest point on the R_E curve. When R is underestimated, the study will be more costly than estimated, and will always be more expensive than if R had been correctly specified at design time. In contrast, when R is overestimated, the study will be less costly than estimated, but still will be more expensive than if R had been correctly specified at design time.

Figure 3 illustrates both of the inefficiency measures which result from misspecifying R when the two-stage study is designed expecting $R = 10$. For example, the GWA study that preserves 99% of the one-stage design power when $R = 10$ costs 61% that of one-stage design (dashed horizontal line). If upon completing stage 1 genotyping, we discover that the stage 2 per genotype cost is less than initially estimated such that the actual $R = 5$ not 10, then completing the study as designed will cost less than expected (58% of the one-stage study's cost [black bar]), but slightly more than if we initially had know that $R = 5$ (56% of the one-stage design cost [gray bar]). In contrast, if the stage 2 per genotype cost is greater than initially estimated, such that R is actually 20 not 10, then completing the study as designed will cost more than expected (67% of one-stage design cost), but only slightly more than if it had been designed knowing $R = 20$ (65% of the one-stage design cost). When R has been underestimated, increasing the proportion of samples genotyped in stage 1 is an alternative that could be considered.

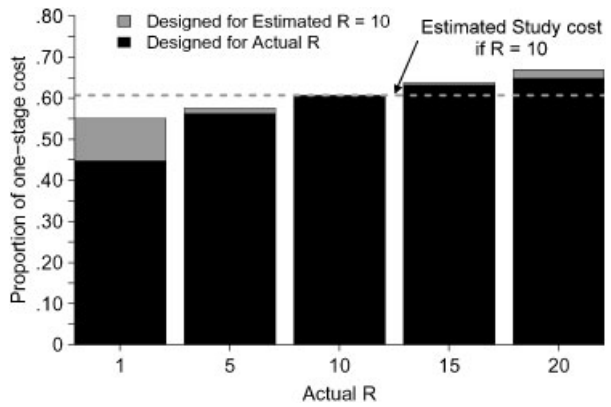

Fig. 3. The designs above employ 1,000 cases and 1,000 controls and have 79% power ($\pi_{power} = .99$) to detect a variant with GRR = 1.375 and $p = .35$ for a disease with prevalence = .10 when $\alpha_{marker} = 1/300{,}000$. The dashed horizontal gray line denotes the cost of the optimal two-stage study when $R = 10$. The gray bars show the cost of the study designed assuming $R = 10$ but carried out when R is actually the value on the x-axis. The black bars show the cost of the optimal two-stage study designed using the value of R on the x-axis.

When R is underestimated, the proposed budget may be insufficient to complete the study's genotyping and hence result in reduced power. However, unless R is grossly underestimated, power loss is minimal (data not shown).

FURTHER DECREASING THE COST OF TWO-STAGE DESIGNS

While two-stage designs can reduce the cost of GWA studies substantially, cost may still be unacceptably high. Two strategies which further reduce genotyping costs are (a) decreasing power while maintaining the false positive rate or (b) increasing the false positive rate while maintaining power.

Decreasing power reduces the genotyping cost of the optimal two-stage design (Fig. 4, Table I). Even a small decrease in π_{power} can have substantial cost benefits. These cost savings come mainly from decreasing the proportion of the sample genotyped in stage 1. Although the proportion of markers to follow-up in stage 2 also decreases as π_{power} is lowered, the number of samples genotyped in stage 2 increases, so that the total cost of stage 2 genotyping may increase (Table I). For example, when $R = 10$, decreasing π_{power} of our GWA study from .99 to .95 reduces the cost of the optimal two-stage design from 60.7 to 51.0% of the one-stage design cost (Fig. 4, Table I). This ~16% decrease in two-stage study genotyping cost results entirely from reduction in stage 1 cost (stage 1 now costs 44.7% compared with 54.5% of one-stage design cost). Stage 2 cost actually increases slightly (from 6.2 to 6.3%).

The genotyping costs saved by reducing π_{power} are nearly independent of R when cost is measured as a proportion of one-stage study cost (Table I). When cost is measured relative to the two-stage design with $\pi_{power} = 99\%$, genotyping costs saved by reducing π_{power} are greater for smaller R. For example, the proportions of the one-stage design cost saved by reducing π_{power} from .99 to .95 range from 9.4 to 9.7% for values of R between 10 and 40, and represent savings between 13.7 ($R = 40$) and 16.0% ($R = 10$) relative to the two-stage design preserving π_{power}=.99. These results also appear nearly independent of the power of the one-stage design (data not shown). Savings are slightly smaller when using a stricter false positive rate ($\alpha_{marker} = .05/300{,}000$) (see online Supplementary Tables I and II).

An alternative approach for further decreasing genotyping costs is to relax the false positive rate while keeping power constant (Fig. 5, Table II). For example, consider the optimal two-stage design for our study using $\pi_{power} = .99$ (power = 79.2%) when $R = 10$ which costs 61.0% of the one-stage design. Suppose this design is too expensive, and that sacrificing additional power is unacceptable. Relaxing

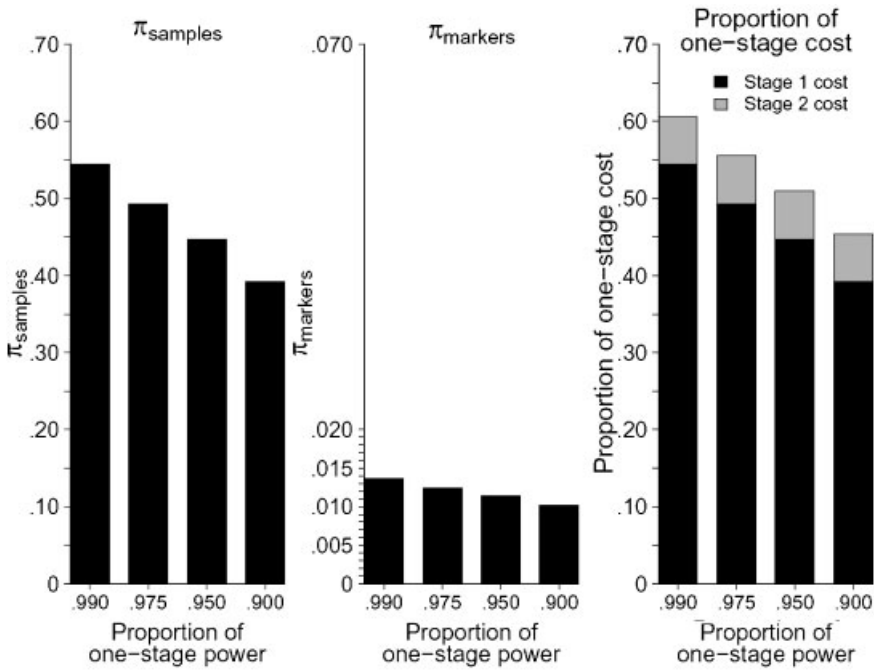

Fig. 4. The designs above employ 1,000 cases and 1,000 controls and have power of 79, 78, 75 or 72% (π_{power} = .99, .975, .95, or .90) to detect a variant with GRR = 1.375 and p = .35 for a disease with prevalence = .10 when α_{marker} = 1/300,000. The black bars in the left and center panels denote the values of $\pi_{samples}$ and $\pi_{markers}$ of the optimal two-stage design. The right panel shows the cost of the optimal two-stage design and how it is divided between stage 1 (black bar) and stage 2 (gray bar).

TABLE I. Cost saving achieved and change in optimal design when decreasing π_{power}

				Cost as % of one-stage design cost						Cost as % of reference design cost	
		Optimal design		Stage 1		Stage 2		Total		Total	
R	π_{power} (%)	$\pi_{samples}$ (%)	$\pi_{markers}$ (%)	Cost	Absolute savings*	Cost	Absolute savings*	Cost	Absolute savings*	Cost	Relative savings*
10	99	54.5	1.36	54.5	0	6.19	0	60.7	0	100	0
	97.5	49.3	1.24	49.3	5.2	6.29	−0.10	55.6	5.1	91.6	8.4
	95	44.7	1.14	44.7	9.8	6.29	−0.10	51.0	9.7	84.0	16.0
	90	39.2	1.02	39.2	15.2	6.17	0.02	45.4	15.3	74.9	25.1
20	99	59.0	0.71	59.0	0	5.85	0	64.8	0	100	0
	97.5	53.8	0.65	53.8	5.2	6.02	−0.17	59.9	5.0	92.3	7.7
	95	49.2	0.60	49.2	9.8	6.07	−0.23	55.2	9.6	85.2	14.8
	90	43.6	0.53	43.6	15.4	6.03	−0.18	49.6	15.2	76.6	23.4
40	99	63.3	0.38	63.3	0	5.53	0	68.8	0	100	0
	97.5	58.2	0.34	58.2	5.1	5.76	−0.23	63.9	4.9	92.9	7.1
	95	53.5	0.32	53.5	9.7	5.87	−0.34	59.4	9.4	86.3	13.7
	90	47.9	0.28	47.9	15.4	5.89	−0.36	53.8	15.0	78.2	21.8

Note: The designs above employ 1,000 cases and 1,000 controls and have power 80% × π_{power} to detect a variant with GRR = 1.375 and p = .35 for a disease with prevalence = .10 when α_{marker} = 1/300,000. Cost is given as a percentage either the corresponding one-stage design cost or the two-stage design cost that retains 99% of the one-stage design power.

*Absolute savings is the difference in cost when using alternative values of π_{power} instead of π_{power} = .99 (measured as a proportion of the one-stage study cost). Relative savings measures the same difference, but expressed relative to the cost of the reference design with π_{power} = .99.

the false positive rate to allow $W = 5$ (instead of 1) false positives ($\alpha_{marker} = 5/300{,}000$) while maintaining 79.2% power decreases the total study cost to 47.6% of the one-stage design (a 22% reduction in cost relative to using $\alpha_{marker} = 1/300{,}000$). We achieve this additional savings because relaxing the false positive rate (from 1/300,000 to 5/300,000) increases the power of the one-stage design (from 80 to 88%), requiring us to preserve only $\pi_{power} = 79\%/88\% = .90$ of the one-stage design power instead of .99. Further relaxing the false positive rate reduces study cost only modestly; relaxing the false positive rate from 5/300,000 to 10/300,000 decreases study cost only another 2%. This cost saving strategy is less effective when a smaller proportion of the one-stage power is preserved (initial $\pi_{power} = .95$ versus .99). In addition, the cost saving is nearly invariant to R and initial α_{marker} when cost is expressed as a percentage of the one-stage study cost. Cost savings increase when the one-stage design power is smaller and decrease when one-stage design power is larger (see online Supplementary Tables III and IV).

INCREASING POWER BY INCREASING FALSE POSITIVE RATE WHILE HOLDING COST CONSTANT

As a corollary to the previous approach, increasing the false positive rate can be used to improve power when study cost is held fixed. Such an approach may be logical if power to detect a disease predisposing variant of a certain effect size is low. However, there is no guarantee that reasonable power will be reached before the false positive rate becomes unacceptably large.

In contrast to previous strategies, the optimal design that maximizes power at a fixed cost changes little across a range of false positive rates (see online Supplementary Fig. 1) or one-stage design powers (results not shown). For example, at a set cost that is 60% of the one-stage study, a two-stage design having power of 69% when using $\alpha_{marker} = 1/300{,}000$, can be improved to 85% power by using $\alpha_{marker} = 20/300{,}000$. The optimal designs for these two studies are remarkably similar: $\pi_{samples} = .54$, $\pi_{markers} = .014$ when $\alpha_{marker} = 1/300{,}000$ and $\pi_{samples} = .53$, $\pi_{markers} = .015$ when $\alpha_{marker} = 20/300{,}000$. This suggests that if an investigator decides to use a false positive rate different from that used to design the study, little power will be lost. Indeed, the power of the two-stage study designed using $\alpha_{marker} = 1/300{,}000$ but analyzed using $\alpha_{markers} = 20/300{,}000$ would improve by less than 0.03% if designed using the larger α_{marker}. Smaller budgets limit the effectiveness of this strategy (see online Supplementary Fig. 1).

DISCUSSION

The design of two-stage studies that use a fixed number of samples is influenced primarily by three factors: the proportion of the one-stage power retained, the acceptable number of false positives, and the stage 2 to stage 1 per genotype cost ratio. Although we have limited control over the per genotype cost ratio, by making small compromises in the false positive rate or proportion of the one-stage power retained we can control study cost.

The per genotype cost ratio influences not only the cost of the optimal two-stage design but also its configuration. As the cost ratio increases, more samples are genotyped in stage 1 and fewer markers are followed up in stage 2. When the per genotype cost in stage 2 is ≤ 10 times that in stage 1, an optimal design typically genotypes as little as 30% of the sample in stage 1 and follows up $>10\%$ of the markers in stage 2. When the ratio is ≥ 20, 50–70% of the sample typically is genotyped in stage 1, and $<1\%$ of the markers are selected for stage 2 genotyping.

Although two-stage designs provide substantial cost savings over one-stage designs, at current per genotype costs GWA studies with large sample sizes that preserve almost all of the one-stage design power can still be too costly. We described two strategies that further decrease the study cost: (1) preserve a somewhat smaller proportion of the one-stage design power while maintaining the false positive rate, or (2) relax the false positive rate while maintaining power. Decreasing the percentage of the one-stage design power preserved from 99 to 97.5% (95%) reduces the cost of the study by about 5% (9%) of the one-stage design cost, and approximately 8% (16%) of the cost of the two-stage design that preserves 99% of the power. Similar cost savings can be achieved (17% of the two-stage design cost) by increasing the false positive rate 2.5-fold, for example, from 1/300,000 to 2.5/300,000, while maintaining the same power. The savings achieved by preserving less power is nearly unaffected by the stage 2 to stage 1 per genotype cost ratio, the false positive rate α_{marker}, or the one-stage power. The savings achieved by relaxing the false positive rate is also nearly unaffected by the per genotype cost ratio and false positive rate, but do depend on the power of the corresponding one-stage design.

Alternatively, relaxing the false positive rate may be used to improve power while holding study cost constant. The power that can be gained by relaxing the false positive rate is greatest when the genotyping budget is not too small ($>50\%$ of the one-stage design cost) and the per genotype cost ratio is not too great ($R \leq 10$). The optimal two-stage design that maximizes power given a fixed

TABLE II. Cost savings achieved and change in optimal design when increasing the false positive rate from $\alpha_{marker} = 1/300{,}000$ to $\alpha_{markers} = W/300{,}000$ while maintaining power

		$\pi_{power} = .99$						$\pi_{power} = .95$					
		Optimal design		Cost as % of one-stage design		Cost as % of reference design		Optimal design		Cost as % of one-stage design		Cost as % of reference design	
R	*W*	$\pi_{samples}$ (%)	$\pi_{markers}$ (%)	Cost	Absolute savings*	Cost	Relative savings*	$\pi_{samples}$ (%)	$\pi_{markers}$ (%)	Cost	Absolute savings*	Cost	Relative savings*
10	1.0	54.5	1.36	60.7	0	100	0	44.7	1.14	51.0	0	100	0
	2.5	43.6	1.15	50.1	10.6	82.6	17.4	39.9	1.06	46.3	4.7	85.2	14.8
	5.0	41.0	1.11	47.6	13.1	78.4	21.6	38.1	1.04	44.6	6.4	80.1	19.9
	10.0	39.5	1.10	46.1	14.5	76.0	24.0	37.0	1.03	43.5	7.4	76.7	23.3
20	1.0	59.0	0.71	64.8	0	100	0	49.2	0.60	55.2	0	100	0
	2.5	48.2	0.61	54.5	10.3	84.1	15.9	44.4	0.56	50.6	4.6	86.4	13.6
	5.0	45.6	0.59	52.0	12.8	80.3	19.7	42.7	0.55	49.0	6.3	81.6	18.4
	10.0	44.2	0.58	50.7	14.2	78.1	21.9	41.6	0.55	48.0	7.2	78.6	21.4
40	1.0	63.3	0.38	68.8	0	100	0	53.5	0.32	59.4	0	100	0
	2.5	52.7	0.32	58.8	10.0	85.5	14.5	48.8	0.30	54.9	4.5	87.4	12.6
	5.0	50.2	0.31	56.4	12.3	82.0	18.0	47.1	0.29	53.3	6.0	83.0	17.0
	10.0	48.8	0.31	55.1	13.6	80.2	19.8	46.1	0.29	52.4	7.0	80.3	19.7

Note: The designs above employ 1,000 cases and 1,000 controls and have 79% power ($\pi_{power} = .99$) or 76% power ($\pi_{power} = .95$) to detect a variant with GRR = 1.375 and $p = .35$ for a disease with prevalence = .10 when $\alpha_{marker} = 1/300{,}000$. Cost is given as a percentage of either the corresponding one-stage design cost or the two-stage design cost that retains 99% of the one-stage design power.

*Absolute savings is the difference in cost between the optimal design using $\alpha_{marker} = 1/300{,}000$ and the optimal design that achieves the same power of 79% (or 76%) using $\alpha_{marker} = W/300{,}000$ (expressed as a proportion of the one-stage design cost). Relative savings measures the same difference, but with cost expressed relative to the reference two-stage design using $\alpha_{markers} = 1/300{,}000$.

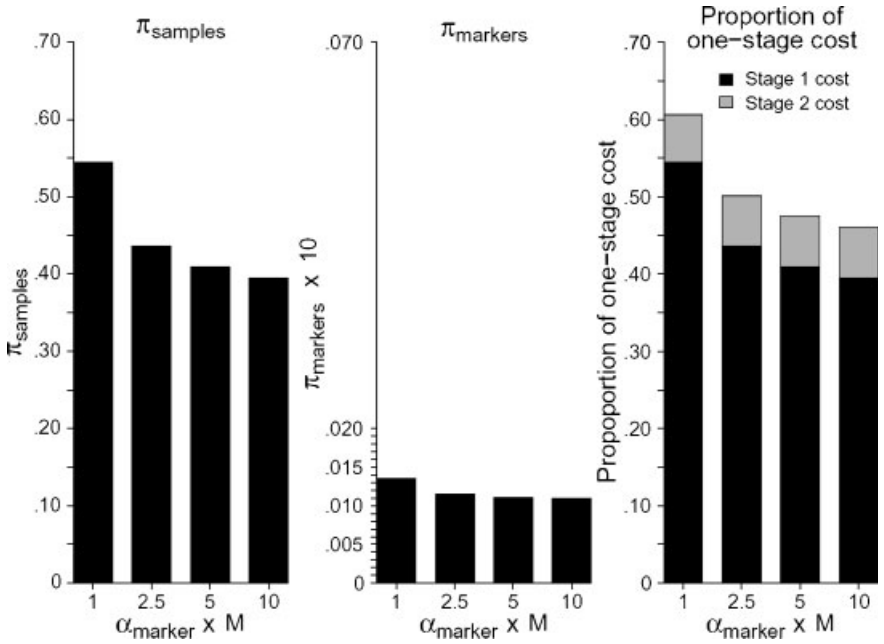

Fig. 5. The designs above employ 1,000 cases and 1,000 controls and have 79% power ($\pi_{power} = .99$) to detect a variant with GRR = 1.375 and $p = .35$ for a disease with prevalence = .10 when $\alpha_{marker} = 1/300{,}000$. The black bars in the left and center panels denote the values of $\pi_{samples}$ and $\pi_{markers}$ of the least costly designs that achieve 79% power using $\alpha_{marker} = W/300{,}000$ for $W = 1$, 2.5, 5, and 10 (x-axis) when $R = 10$. The right panel shows the cost of these designs and how it is divided between stage 1 (black bar) and stage 2 (gray bar).

budget changes little as α_{marker} increases, suggesting that relaxing the false positive rate after the study has been completed will result in very little loss of efficiency.

We have made several observations that are complementary to those of Wang et al. [2006]. Wang et al. demonstrated, as we do, that when the stage 2 to stage 1 per genotype cost ratio increases, the genotyping shifts from stage 2 to stage 1, and that the cost of the optimal design increases. Our perspectives differ, however, because of how we define the optimal two-stage design. Wang et al. sought the least expensive two-stage design with a fixed power to detect a variant of a given effect size with no restriction on sample size. We fix sample size and seek the least expensive two-stage design that preserves a given proportion of the one-stage design power. Interestingly, our observation that genotyping costs can be substantially lowered by decreasing the proportion of power preserved is consistent with Wang et al.'s tables showing that genotyping costs can be lowered by increasing the sample size while decreasing the proportion of one-stage power retained. In both instances, fewer samples are genotyped in stage 1 and fewer markers are followed up in stage 2.

Many investigators will be interested in testing marker associations with one or more disease-related traits in the same sample. If T independent traits are tested instead of just one, then the number of tests performed in stage 1 is TM. If W false positives are acceptable after testing all traits on all markers, then the appropriate marker-wise false positive rate is now approximately $W/(TM)$ instead of W/M when only one trait was tested. Thus power to detect a variant influencing any one trait is reduced. When the traits are independent, and even if they are moderately correlated, the markers selected for follow-up for each trait will tend to have little overlap. Thus, if we follow-up $\pi_{markers}$ M markers for each trait, the cost of stage 2 will be approximately T times the cost of stage 2 when testing a single trait. Thus, optimizing a two-stage design that tests for association to T independent traits is nearly equivalent to increasing the stage 2 to stage 1 per genotype cost ratio to TR. When studying a large number of traits, it may be that the most economical study is the one-stage design.

In our examples, we focus on the power to detect a single variant. For most complex diseases, there will be several susceptibility variants of varying effect sizes and our formulae can be extended to accommodate this. For example, suppose we assume K independent susceptibility alleles. The probability that a genome-wide scan will fail to detect any of

these alleles is $\Pi_{i=1}^{K}[1 - P(\text{detect variant } i)]$ and the probability that the scan will detect at least one of these alleles is $1 - \Pi_{i=1}^{K} P(\text{detect variant } i)$. Power to detect any specific number of variants can be calculated approximately using the Poisson-Binomial distribution, with parameters specifying power for each individual variant.

We have not considered interactions in our current work. Interaction tests could be performed in stage 1 data on all pairs of markers or on the subset of markers with modest marginal associations [Marchini et al., 2005]. The optimal strategy for designing two-stage GWA studies when testing for interactions remains an open research problem; nevertheless our intuition is that because of the much larger number of tests performed these designs could be quite different from studies that focus on main effects. For example, we expect that these designs will likely require a larger proportion of samples genotyped in stage 1 to ensure both that the effect of true interacting variants is great enough that the markers will be selected for follow-up in stage 2 and that the number of markers selected for follow-up is small enough to prove cost-effective.

To simplify calculations and allow us to focus on the primary factors influencing the cost of two-stage designs, we made a number of assumptions. These include using a fixed sample size, allowing any number of markers to be genotyped in stage 2, linkage equilibrium between markers, and no population substructure. We chose to focus on fixed sample sizes since many investigators will already have samples in hand when designing their GWA study. Wang et al. [2006] explore designing optimal two-stage studies when sample size is not limited.

Our assumption that any number of markers can be genotyped in stage 2 is likely unrealistic, since in practice the number of markers may be restricted by the type of custom genotyping arrays employed. These arrays may, for example, genotype 96, 384, 1,536, 10,000, or 20,000 single nucleotide polymorphisms (SNPs) per array. Additional modifications to the optimization routine are necessary when such arrays are used since there is a finite number of possible π_{markers} values, the per genotype cost for each array will differ, and there will likely be discounts when genotyping a large number of samples. Appendix A describes how this more complicated cost structure can be accommodated, and gives an example that might be typical when using a high-throughput genotyping service.

Our assumption that all stage 1 markers are in linkage equilibrium is not true of current genotyping products, and certainly will not be for future ones. When linkage disequilibrium (LD) exists between markers, using a false positive rate of $\alpha_{\text{marker}} = W/M$ to control the total number of false positives to be W still holds, although the variance of the number of false positives expected increases. However, controlling the genome-wide false positive rate by using $\alpha_{\text{marker}} = \alpha_{\text{genome}}/M$ is conservative and power is underestimated. In this case, in the design phase an appropriate marker-wise type I error rate may be estimated using simulation and genotype and LD information from a related population in the HapMap project for stage 1 markers. We have also implicitly assumed that a disease predisposing variant is in perfect LD ($r^2 = 1$) with one of the stage 1 markers. Power to detect the variant when the value of $r^2 < 1$ is less. A more accurate estimate of the one-stage design power is the power to detect the same variant, but using only Nr^2 cases and Nr^2 controls [Pritchard and Przeworski, 2001].

We have also assumed that within each stage there is no population stratification between the cases and controls. If stratification is present, any of a number of methods may be employed, and the two-stage design presents no new challenges [Pritchard and Rosenberg, 1999; Devlin et al., 2001; Reich and Goldstein, 2001; Satten et al., 2001; Price et al., 2006]. Skol et al. [2006] have explored the impact of heterogeneity between the stage 1 and stage 2 cases, and although this may affect the study power, the construction of the joint statistics ensures that an excess of false positives will not be an issue.

Our two-stage design approach is broadly applicable to GWA studies of SNPs and copy number polymorphisms, and could also be applied to large scale sequencing studies. The two-stage framework can also be used to calculate the increase in power that could be gained by genotyping additional samples when the results from existing GWA studies are treated as the first stage.

In summary, we have outlined guidelines for designing efficient two-stage designs for GWA studies. There is no single optimal two-stage GWA design; each design will depend strongly on the genotyping costs and budget, and the investigator's tolerance for false positives and false negatives. However, currently available products and pricing will likely suggest optimal two-stage designs using between 50 and 60% of the sample in stage 1 and following up between 0.1 and 1.0% of the markers per trait of interest in stage 2. If designs with these parameters are too costly, modestly increasing the false positive rate (say from 1 to 2.5 false positives per genome) while maintaining power or modestly decreasing the proportion of one-stage design power retained (say from 99 to 95%) while maintaining the false positive rate can further reduce the genotyping costs by ~15%. To help investigators balance the goal of high power, low false positive rate, and manageable cost, we have developed a graphical

power calculator CaTS to help investigators rapidly evaluate and optimize two-stage GWA designs. Our tool is freely available at www.sph.umich.edu/csg/abecasis/CaTS/.

REFERENCES

Brent R. 1973. Algorithms for Minimizatin Without Derivatives. Englewood Cliffs, NJ: Prentice-Hall.

Bukszar J, van den Oord E. 2006. Optimization of two-stage genetic designs where data are combined using an accurate and efficient approximation for pearson's statistic. Biometrics 62:1132–1137.

Cardon LR, Bell JI. 2001. Association study designs for complex diseases. Nat Rev Genet 2:91–99.

Casella G, Berger R. 2002. Statistical Inference. Pacific Grove, CA: Duxbury.

Devlin B, Roeder K, Wasserman L. 2001. Genomic control, a new approach to genetic-based association studies. Theor Popul Biol 60:155–166.

Hirschhorn JN, Daly MJ. 2005. Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6: 95–108.

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

Lehmann E, Casella G. 1998. Theory of Point Estimation. New York: Springer-Verlag.

Marchini J, Donnelly P, Cardon LR. 2005. Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet 37:413–417.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909.

Pritchard JK, Przeworski M. 2001. Linkage disequilibrium in humans: models and data. Am J Hum Genet 69:1–14.

Pritchard JK, Rosenberg NA. 1999. Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 65:220–228.

Reich DE, Goldstein DB. 2001. Detecting association in a case-control study while correcting for population stratification. Genet Epidemiol 20:4–16.

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

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.

Satten GA, Flanders WD, Yang QH. 2001. Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. Am J Hum Genet 68:466–477.

Skol AD, Scott LJ, Abecasis GR, Boehnke M. 2006. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet 38:209–213.

Thomas D, Xie RR, Gebregziabher M. 2004. Two-stage sampling designs for gene association studies. Genet Epidemiol 27: 401–414.

Wang HS, Thomas DC, Pe'er I, Stram DO. 2006. Optimal two-stage genotyping designs for genome-wide association scans. Genet Epidemiol 30:356–368.

Wang WYS, Barratt BJ, Clayton DG, Todd JA. 2005. Genome-wide association studies: theoretical and practical concerns. Nat Rev Genet 6:109–118.

APPENDIX

A. OPTIMAL TWO-STAGE DESIGNS WHEN USING ARRAY-BASED PRODUCTS IN STAGE 2

Stage 2 of a GWA study will often use custom genotyping arrays which genotypes fixed numbers of markers. For example, an array may genotype 96, 384, 768, 1536, or more markers simultaneously. Stage 2 per genotype cost structure is impacted when using these products because the cost per genotype for arrays with more markers is generally smaller. In addition, discounts may be available when genotyping large number of samples. We describe how to accommodate these features when designing optimal two-stage GWA studies.

Use of arrays that genotype a fixed number of markers actually simplifies the optimal design search. We now need to consider only the small number of values that π_{marker} can take on instead of all possible values between 0 and 1. For example, the genotyping cost structure given in Appendix Table Ia results in it being less expensive to use a 384 marker array than two 96 marker arrays. Therefore, we need not consider π_{markers} values of 192/300,000 or 288/300,000 since it would never be economical to do so—these values of π_{markers} we call inadmissible. Each admissible π_{markers} value tends to have a unique R because each array or combination of arrays typically results in a different per genotype cost.

We propose the following algorithm to identify the optimal two-stage GWA design.

Assuming discounts are available when more samples are genotyped, repeat the following for each discount cost scheme. Identify the restriction on π_{samples} for the current discount level (for example, when the costs are specified for genotyping <450 samples, we restrict attention to $(2N-450)/(2N) < \pi_{\text{samples}} \leq 1$). Values of π_{markers} for which it is impossible to achieve the desired power are not considered. For all other values of π_{markers}, identify the π_{samples} which attains the desired power using a bisection algorithm and store this value and the study cost. Repeat these steps for the next level of discounting (next range of π_{samples}). The optimal two-stage design is the least costly study identified.

An Example. Consider designing a two-stage study in which stage 2 genotyping will be conducted using a genotyping service which offers the cost

structure given in Appendix Table Ia. A subset of the possible values $\pi_{markers}$ can take on, the mix of arrays used to achieve $\pi_{markers}$, and the cost of the arrays are given in Appendix Table Ib. The admissible values of $\pi_{markers}$ are shaded. Appendix Table Ic shows all admissible values of $\pi_{markers}$ between 96/300,000 and 19,968/300,000 and their per genotype cost. Although 208 values of $\pi_{markers}$ are possible over this range, only 40 are admissible. The per genotype cost for most of the values of $\pi_{markers}$ is near \$0.05/genotype. For the GWA study using 1,000 cases and 1,000 controls with power of 80% to detect a variant of modest effect (GRR = 1.375, $P = .35$, prevalence = .10) when $\alpha_{marker} = 1/300{,}000$ and the stage 1 per genotype cost is \$.003/genotype, the optimal design that maintains 99% of the one-stage study power uses $\pi_{samples} = .536$ and $\pi_{markers} = .0154$, and costs 63.3% of the one-stage study. Following up 1.54% of the markers requires the use of three 1,536 marker arrays. Stage 2 of this study costs \$0.05/genotype, giving an $R = 13.67$. If there were no restriction on $\pi_{markers}$ and $R = 13.3$,

TABLE Ia. Hypothetical cost structure of genotyping using custom arrays

Sample size	96 SNPs	384 SNPs	1536 SNPs
<450	\$45 (47¢)	\$73 (19¢)	\$166 (11¢)
450–900	\$40 (42¢)	\$50 (13¢)	\$75 (5¢)
901–1980	\$35 (36¢)	\$47 (12¢)	\$63 (4¢)
>1980	\$35 (36¢)	\$45 (12¢)	\$55 (4¢)

Values are cost per array, with cost per genotype in parentheses.

TABLE Ib. Examples of admissible $\pi_{markers}$

No. markers	96	192	288	384	480	576	672	768
$\pi_{markers}$	.0003	.0006	.0010	.0013	.0016	.0019	.0022	.0026
96	1	2	3	0	1	2	3	0
384	0	0	0	1	1	1	1	2
1536	0	0	0	0	0	0	0	0
cost (\$)	40	80	120	50	90	130	170	100
cost/geno (¢)	41.7			13.0				

No. markers	864	960	1056	1152	1248	1344	1440	1536
$\pi_{markers}$	.0029	.0032	.0035	.0038	.0042	.0045	.0048	.0051
96	1	2	3	0	1	2	3	0
384	2	2	2	3	3	3	3	0
1536	0	0	0	0	0	0	0	1
cost (\$)	140	180	220	150	190	230	270	75
cost/geno (¢)								4.9

Note: Values of $\pi_{markers}$ that are admissible are shaded. Genotyping fewer markers is more expensive.

TABLE Ic. All admissible values of $\pi_{markers}$ and their per genotype costs when between 901 and 1,980 samples are used in stage 2 and 300,000 markers are genotyped in stage 1

$\pi_{markers}$	¢/Genotype	$\pi_{markers}$	¢/Genotype	$\pi_{markers}$	¢/Genotype	$\pi_{markers}$	¢/Genotype
.0003	41.7	.0166	5.5	.0358	4.9	.0515	5.1
.0013	13.0	.0205	4.9	.0362	5.2	.0525	5.1
.0051	4.9	.0208	5.4	.0371	5.2	.0563	4.9
.0054	7.0	.0218	5.4	.0410	4.9	.0566	5.1
.0064	6.5	.0256	4.9	.0413	5.2	.0576	5.1
.0102	4.9	.0259	5.3	.0422	5.1	.0614	4.9
.0106	6.0	.0269	5.3	.0461	4.9	.0618	5.1
.0115	5.8	.0307	4.9	.0464	5.1	.0627	5.0
.0154	4.9	.0310	5.3	.0474	5.1	.0666	4.9
.0157	5.6	.0320	5.2	.0512	4.9	.0669	5.1

then the optimal two-stage design would use $\pi_{\text{samples}} = .565$, $\pi_{\text{markers}} = .010$, and cost 62.6% that of the one-stage study design.

B. VARIANCE OF Z_1 AND Z_{JOINT}

Bukszár and van den Oord show that the distribution of

$$z_1 = \sqrt{2N\pi_{\text{samples}}}\, g(\hat{p}'_1, \hat{p}_1) = \frac{\hat{p}'_1 - \hat{p}_1}{\sqrt{[\hat{p}'_1(1-\hat{p}'_1) + \hat{p}_1(1-\hat{p}_1)]/(2N\pi_{\text{samples}})}}$$

can be found by writing

$$\hat{p}_1 = \sum_{i=1}^{2N\pi_{\text{samples}}} \frac{X_i}{2N\pi_{\text{samples}}}$$

where X_i is an indicator that is equal to 1 if the allele is a risk allele and 0 otherwise [Bukszar and van den Oord, 2006]. The Central Limit Theorem states that

$$\sqrt{2N\pi_{\text{samples}}}(\hat{p}_1 - p) = \sqrt{2N\pi_{\text{samples}}} \times \sum_{i=1}^{2N\pi_{\text{samples}}} \left(\frac{X_i}{2N\pi_{\text{samples}}} - p\right) \xrightarrow{d} N(0, p(1-p)),$$

where $N(x, y)$ denotes a Normal distribution with mean x and variance y and $\xrightarrow{d}$ signifies convergence in distribution [Casella and Berger, 2002]. In addition,

$$\sqrt{2N\pi_{\text{samples}}}(\hat{p}'_1 - p') = \sqrt{2N\pi_{\text{samples}}} \times \sum_{i=1}^{2N\pi_{\text{samples}}} \left(\frac{X'_i}{2N\pi_{\text{samples}}} - p'\right) \xrightarrow{d} N(0, p'(1-p')).$$

We can use the multivariate delta method to find the distribution of z_1 [Lehmann and Casella, 1998]. The multivariate delta method stated in terms of $\hat{p}'_1$ and $\hat{p}_1$ is

$$\sqrt{2N\pi_{\text{samples}}}(g(\hat{p}'_1, \hat{p}_1) - g(p'_1, p_1)) \xrightarrow{d} N(0, \nabla^T g(p'_1, p_1)\Sigma\nabla g(p'_1, p_1)),$$

where

$$\nabla g(p'_1, p_1) = \begin{pmatrix} \frac{\partial g(p'_1, p_1)}{\partial p'_1} \\ \frac{\partial g(p'_1, p_1)}{\partial p_1} \end{pmatrix}$$

and

$$\Sigma = \begin{pmatrix} p'_1(1-p'_1) & 0 \\ 0 & p_1(1-p_1) \end{pmatrix}.$$

Given that

$$\nabla g(p'_1, p_1) = \begin{pmatrix} \frac{p'_1 + 3p_1 - 2p_1^2 - 2p'_1 p_1}{2(p'_1(1-p'_1) + p_1(1-p_1))^{\frac{3}{2}}} \\ \frac{p_1 + 3p'_1 - 2p_1'^2 - 2p'_1 p_1}{2(p'_1(1-p'_1) + p_1(1-p_1))^{\frac{3}{2}}} \end{pmatrix},$$

the variance of z_1 is

$$\text{var}(z_1) = (p'_1 + 3p_1 - 2p_1^2 - 2p'_1 p_1)^2 (p'_1(1-p'_1)) + \frac{(p_1 + 3p'_1 - 2p_1'^2 - 2p'_1 p_1)^2 (p_1(1-p_1))}{4(p'_1(1-p'_1) + p_1(1-p_1))^3}.$$

Note that $\text{var}(z_1) = 1$ under the null hypothesis $p' = p$, and tends to be quite close to 1 unless $p' - p$ is large relative to p or p'.