Improved Minimum Cost and Maximum Power Two Stage Genome-Wide Association Study Designs

Stephen A. Stanhope[1]*, **Andrew D. Skol**[2]

[1] Department of Human Genetics, The University of Chicago, Chicago, Illinois, United States of America, [2] Program in Genetic Medicine, The University of Chicago, Chicago, Illinois, United States of America

Abstract

In a two stage genome-wide association study (2S-GWAS), a sample of cases and controls is allocated into two groups, and genetic markers are analyzed sequentially with respect to these groups. For such studies, experimental design considerations have primarily focused on minimizing study cost as a function of the allocation of cases and controls to stages, subject to a constraint on the power to detect an associated marker. However, most treatments of this problem implicitly restrict the set of feasible designs to only those that allocate the same proportions of cases and controls to each stage. In this paper, we demonstrate that removing this restriction can improve the cost advantages demonstrated by previous 2S-GWAS designs by up to 40%. Additionally, we consider designs that maximize study power with respect to a cost constraint, and show that recalculated power maximizing designs can recover a substantial amount of the planned study power that might otherwise be lost if study funding is reduced. We provide open source software for calculating cost minimizing or power maximizing 2S-GWAS designs.

Citation: Stanhope SA, Skol AD (2012) Improved Minimum Cost and Maximum Power Two Stage Genome-Wide Association Study Designs. PLoS ONE 7(9): e42367. doi:10.1371/journal.pone.0042367

Editor: Yun Li, University of North Carolina, United States of America

Received March 18, 2012; **Accepted** July 6, 2012; **Published** September 6, 2012

Funding: The authors have no support or funding to report.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: sstanhop@bsd.uchicago.edu

Introduction

Genome-wide association studies (GWAS) have become ubiquitous in complex disease genetics. While the tools to conduct these studies have improved substantially, the cost of conducting them remains expensive. This is despite the plummeting cost per genotype, and is a result of the increasing number of markers being interrogated with each successive generation of genotyping chip. Identifying efficient study designs thus remains important.

One popular efficient design for GWAS is the two stage GWAS (2S-GWAS), which has been used for investigations of a wide range of diseases, such as type 2 diabetes [1], schizophrenia [2], lupus erythematosus [3], psoriasis [4], and breast cancer [5]. In the 2S-GWAS, a full sample of cases and controls is divided between a first stage that employs a whole genome genotyping platform and tests all available markers for association with the disease, and a second stage that uses a more expensive custom genotyping platform to follow up those markers exhibiting sufficiently strong association with the disease in stage 1. The evidence of association from both stages is then considered jointly to reach a final determination of association between marker and disease. The 2S-GWAS was shown to be more efficient than one stage analyses in which all samples are genotyped on the whole genome platform by Satagopan and Elston [6] and Thomas et al [7]. Since these early investigations, continued attention has been paid to the theoretical properties of two stage designs [8,9], and efforts have been made to explicitly tie these theoretical properties to the problem of computing experimental designs [10–15]. Recent summaries of methodological and practical issues pertaining to 2S-GWAS are provided by Thomas et al [16] and Van Steen [17].

In this paper, we are concerned with computing experimental designs for 2S-GWAS. Our work is based upon that of Skol et al [10,12], who 1) demonstrated that joint analyses that combine information on case/control allele frequency differences across stages are substantially more powerful than those based on replication, although slightly less powerful than more expensive one stage analyses; and 2) developed a software package (CaTS) to compute minimum cost designs for 2S-GWAS, subject to both an explicitly stated constraint on the minimum level of acceptable study power, and an implicitly stated equality constraint on the proportion of cases and controls allocated to the stages.

We extend this work in two ways. First, we improve the cost efficiency of the 2S-GWAS by defining a procedure that allows different proportions of cases and controls to be assigned to stages, and developing software to compute minimum cost, power constrained designs for the unrestricted 2S-GWAS. We demonstrate that the unrestricted 2S-GWAS can be substantially more cost effective than designs that restrict case and control allocation proportions to be equal. In the studies we present here, which use relatively modest differences of case and control sample sizes, we achieve up to a 40% relative cost advantage as compared to the 2S-GWAS designs computed by CaTS and 80% compared to one stage designs.

Second, and based upon our success in improving the cost effectiveness of 2S-GWAS designs relative to a power constraint, we consider 2S-GWAS designs that maximize power with respect to a cost constraint. Calculating such designs may be useful for maximizing the utility of studies that are cost constrained rather than designed to meet a given level of power, or are subject to reductions in funding relative to that required to achieve a given level of power in a cost minimizing 2S-GWAS. We examine the latter case, and demonstrate that substantial amounts of power can be retained by recalculating power maximizing experimental designs subject to cost constraints, even for large reductions in planned cost.

To support our results and assist applied researchers, our 2S-GWAS design software is included with this paper (Code S1) and available at http://www.bioinformatics.org/stanhope/2SGWASdesign/.

Methods

Defining a two stage GWAS with different allocations of cases and controls to stage 1

Let the total number of cases and controls be N_1 and N_0 with $R_{cc}=N_0/N_1$ defined as the ratio of controls to cases, let M be the total number of biallelic markers to be genotyped in stage 1, and let c_1 and c_2 be stage 1 and stage 2 per genotype costs. We define the risk allele frequency at a hypothetical disease marker in cases and controls as p_1 and p_0 respectively, and we assume Hardy-Weinberg equilibrium within the population. Let π_1 and π_0 be the respective proportions of cases and controls allocated to stage 1, and let π_M be the expected proportion of markers selected for follow-up in stage 2 if no markers are associated with disease. (Note that π_M is not selected to control the type I error rate, but to reduce cost by ensuring that uninteresting markers are not genotyped in stage 2.) We suppose that the risk allele frequencies of the cases and controls assigned to stages 1 and 2 ($\{p_{1,1},p_{0,1}\}$ and $\{p_{1,2},p_{0,2}\}$ where the second term in the subscript corresponds to stage) are equal to p_1 and p_0 respectively. That is, there is no population heterogeneity.

Stage 1 of the 2S-GWAS proceeds by comparing allele frequencies at each marker, using the allocated cases and controls. For each marker showing significant differences in allele frequencies between cases and controls in stage 1 (where stage 1 significance is determined by π_M), a stage 2 test of allele frequencies is calculated using the remaining cases and controls. The stage 1 and 2 test statistics are then combined according to their Fisher informations, and a joint statistic is used to evaluate the total evidence of association with the disease. For clarity, in Fig. 1 we provide a flowchart of the steps in this 2S-GWAS. The following section provides technical details. Some of the presented results have already been established (e.g. Theorem 1). However, and for clarity, we choose to err on the side of completeness.

The stage 1 test statistic and its asymptotic behavior. In stage 1, differences between the estimated case and control allele frequencies $\hat{p}_{1,1}$ and $\hat{p}_{0,1}$ are evaluated using the statistic:

$$z_1=\frac{\hat{p}_{1,1}-\hat{p}_{0,1}}{\sqrt{\frac{\hat{p}_{1,1}(1-\hat{p}_{1,1})}{2N_1\pi_1}+\frac{\hat{p}_{0,1}(1-\hat{p}_{0,1})}{2N_0\pi_0}}} \tag{1}$$

Formally, we wish to evaluate $H_0: p_1=p_0$ vs. $H_0: p_1\neq p_0$ by comparing stage 1 case and control allele frequencies, and we do so by using the asymptotic distribution of Z_1.

Theorem 1: $Z_1 \xrightarrow{L} N(\mu_1,\sigma_1^2)$ where

$$\mu_1=\frac{p_1-p_0}{\sqrt{\frac{p_1(1-p_1)}{\pi_1 N_1}+\frac{p_0(1-p_0)}{\pi_0 N_0}}},$$

$$\sigma_1^2=\frac{p_1(1-p_1)}{2\pi_1 N_1}\left(\sqrt{\frac{1}{d}}-(p_1-p_0)\frac{(1-2p_1)}{4\pi_1 N_1 d^{3/2}}\right)^2+ \frac{p_0(1-p_0)}{2\pi_0 N_0}\left(-\sqrt{\frac{1}{d}}-(p_1-p_0)\frac{(1-2p_0)}{4\pi_0 N_0 d^{3/2}}\right)^2$$

and $d=p_1(1-p_1)(2\pi_1 N_1)^{-1}+p_0(1-p_0)(2\pi_0 N_0)^{-1}$.

(Proof provided in Appendix S1.)

Stage 1 critical value. Under H_0, $Z_1 \xrightarrow{L} N(0,1)$ follows from Theorem 1. The critical value for the stage 1 test is therefore determined by π_M, and is defined as:

$$v_1=\Phi^{-1}(1-\pi_M/2). \tag{2}$$

Note that under the null the test is expected to pass $M\pi_M$ markers from stage 1 to stage 2.

Stage 1 power. Under the alternative, the power of the stage 1 test is:

$$P_1(\pi)=\Phi\left(\frac{-v_1-\mu_1}{\sigma_1}\right)+1-\Phi\left(\frac{v_1-\mu_1}{\sigma_1}\right) \tag{3}$$

where μ_1 and σ_1 are as in Theorem 1, and we have stated P_1 as a function of $\pi=\{\pi_0,\pi_1,\pi_M\}$.

The stage 2 test statistic and its asymptotic behavior. Stage 2 analysis proceeds for markers rejecting H_0 in stage 1 by estimating case and control allele frequencies based on stage 2 genotypes, $\hat{p}_{1,2}$ and $\hat{p}_{0,2}$, and calculating the statistic:

$$z_2=\frac{\hat{p}_{1,2}-\hat{p}_{0,2}}{\sqrt{\frac{\hat{p}_{1,2}(1-\hat{p}_{1,2})}{2N_1(1-\pi_1)}+\frac{\hat{p}_{0,2}(1-\hat{p}_{0,2})}{2N_0(1-\pi_0)}}}. \tag{4}$$

The asymptotic distribution of Z_2 under either the null or the alternative is analogous to that of Z_1.

Constructing the joint test statistic. After calculating z_2, the markers under consideration are reevaluated using a joint analysis of stage 1 and stage 2 allele frequencies based on a null model Fisher information-averaged test statistic. Letting w_1 and w_2 be the weights given to z_1 and z_2, we compute the joint analysis test statistic as:

$$z=w_1 z_1+w_2 z_2, \tag{5}$$

where $w_1^2+w_2^2=1$, and w_1 is defined:

$$w_1^2=\left(\frac{(\pi_1 N_1)^{-1}+(\pi_0 N_0)^{-1}}{((1-\pi_1)N_1)^{-1}+((1-\pi_0)N_0)^{-1}}+1\right)^{-1}$$

(see Appendix S1 for details).

Stage 2 joint test critical value calculation. Let v_1 and v be critical values for the stage 1 and joint tests respectively. To be significantly associated with disease a marker must be rejected at both the first and second stages. Let $R_1=\{|Z_1|>v_1\}$ and $R=\{|Z|>v\}=\{|w_1 Z_1+w_2 Z_2|>v\}$ be stage 1 and 2 test rejection indicators; $R_1\bigcap R$ is the indicator that the marker is genome-wide significant. The probability of this event under the null can be evaluated by conditioning:

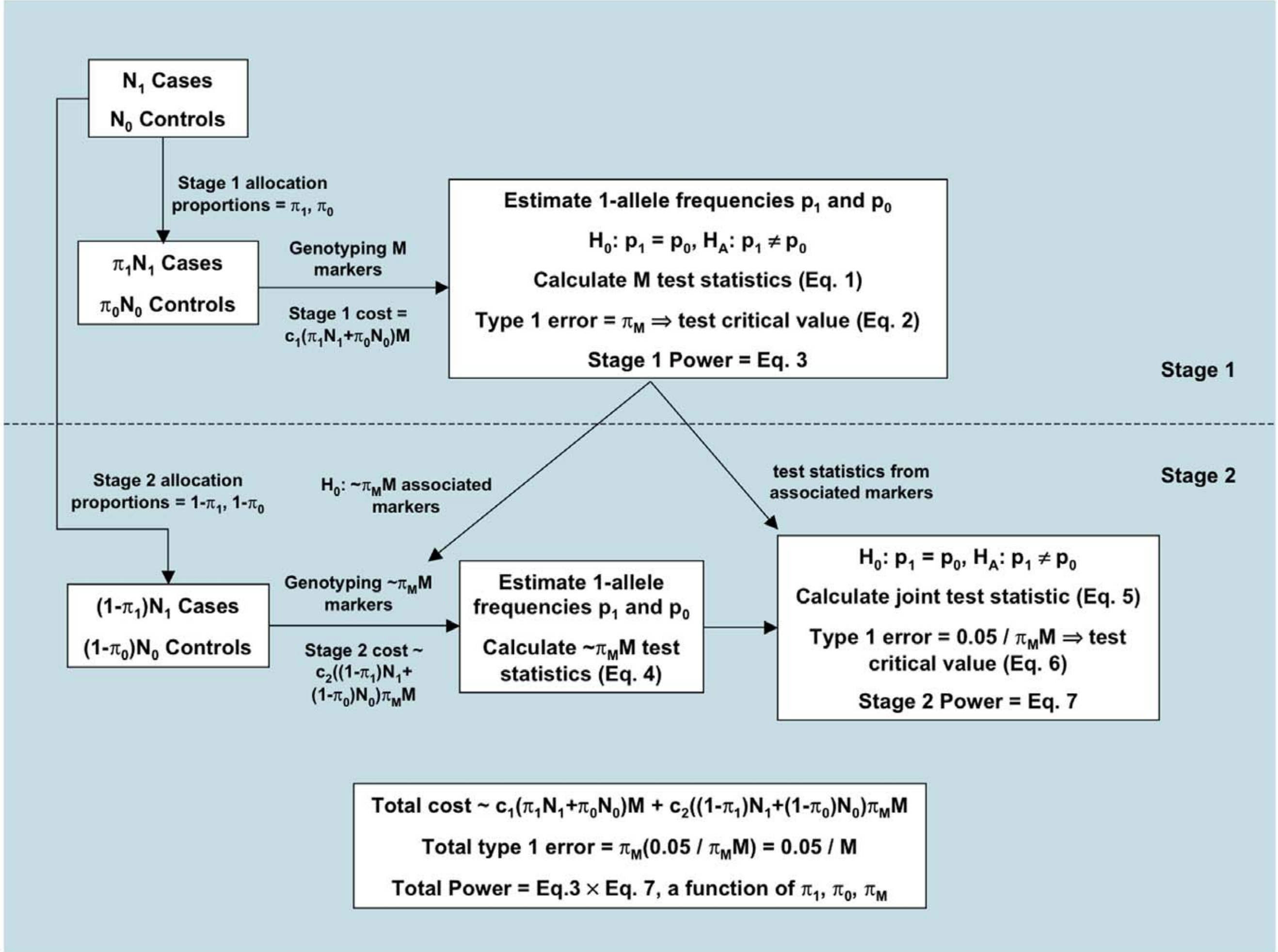
N_1 Cases
N_0 Controls
Stage 1 allocation proportions = π_1, π_0
$\pi_1 N_1$ Cases
$\pi_0 N_0$ Controls
Genotyping M markers
Stage 1 cost = $c_1(\pi_1 N_1+\pi_0 N_0)M$
Estimate 1-allele frequencies p_1 and p_0
H_0: $p_1 = p_0$, H_A: $p_1 \neq p_0$
Calculate M test statistics (Eq. 1)
Type 1 error = $\pi_M \Rightarrow$ test critical value (Eq. 2)
Stage 1 Power = Eq. 3
Stage 1
Stage 2
Stage 2 allocation proportions = 1-π_1, 1-π_0
H_0: ~π_MM associated markers
test statistics from associated markers
$(1-\pi_1)N_1$ Cases
$(1-\pi_0)N_0$ Controls
Genotyping ~π_MM markers
Stage 2 cost ~ $c_2((1-\pi_1)N_1+(1-\pi_0)N_0)\pi_M M$
Estimate 1-allele frequencies p_1 and p_0
Calculate ~π_MM test statistics (Eq. 4)
H_0: $p_1 = p_0$, H_A: $p_1 \neq p_0$
Calculate joint test statistic (Eq. 5)
Type 1 error = 0.05 / π_MM $\Rightarrow$ test critical value (Eq. 6)
Stage 2 Power = Eq. 7
Total cost ~ $c_1(\pi_1 N_1+\pi_0 N_0)M + c_2((1-\pi_1)N_1+(1-\pi_0)N_0)\pi_M M$
Total type 1 error = π_M(0.05 / π_MM) = 0.05 / M
Total Power = Eq.3 × Eq. 7, a function of π_1, π_0, π_M

Figure 1. Two stage GWAS flowchart. This flowchart describes the steps of our two stage GWAS. We begin by splitting the complete data into two groups, to be used sequentially in stages 1 and 2. In the first stage, we evaluate associations of all markers with the disease. In the second stage, we genotype only those markers shown to be associated in stage 1. We compute stage 2-specific test statistics for these markers, and then construct joint test statistics based on those from both stages. The joint test statistics are used to make final assessments of disease association.
doi:10.1371/journal.pone.0042367.g001

$$Pr^0(R_1 \cap R) = Pr^0(R \mid R_1)Pr^0(R_1),$$

where Pr^0 is the probability under the null model. To achieve a marker-wise type I error equal to α^*, $Pr^0(R \mid R_1)Pr^0(R_1) = \alpha^*$ is to be maintained. $Pr^0(R_1) = \pi_M$ by construction, therefore v must be such that $Pr^0(R \mid R_1) = \alpha^* \pi_M^{-1}$.

For example, if $\alpha^* = 0.05/M$ (e.g. a type I error equal to that of a Bonferroni-controlled 5% test), then v is to be set such that:

$$Pr^0(|Z| > v \| Z_1| > v_1) = \frac{0.05}{M\pi_M}. \tag{6}$$

We compute $Pr^0(|Z| > v \| Z_1| > v_1)$ by integrating over the conditional distribution of Z_1 and decomposing Z into its stage-specific portions, and numerically solve for v (see Appendix S1 for details).

Stage 2 power. At the susceptibility marker, Z_1 and Z_2 are $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ distributed (where μ_1 and σ_1^2 are defined in Theorem 1, and analogously for μ_2 and σ_2^2). Let

$$P_2(\pi) \tag{7}$$

be the stage 2 power stated in terms of π. Because of its complexity, we omit providing an explicit form for $P_2(\pi)$ here, but do number the equation to correspond to its reference in Fig. 1. Obtaining the power of the joint test conditional on $|Z_1| > v_1$ is done with a computation analogous to that used to compute type I error (see Appendix S1 for details and the explicit statement of the equation).

Defining constrained minimum cost and maximum power two stage designs

We define an optimal two stage design as that which achieves a specified power at the least cost or, alternatively, that which maximizes power for a given experimental cost. The genotyping cost incurred when performing a 2S-GWAS is

$$C(\pi) = c_1(\pi_1 N_1 + \pi_0 N_0)M + c_2((1-\pi_1)N_1 + (1-\pi_0)N_0)M\pi_M,$$

where c_1 and c_2 are the per marker genotyping costs for stages 1 and 2, and the total power of the 2S-GWAS is $P(\pi)=P_1(\pi)P_2(\pi)$. The optimal cost minimized design is that having power of at least P^* with the lowest cost $C(\pi)$. That is, the following constrained optimization problem is to be solved:

$$min \quad C(\pi)$$

$$s.t. \quad P(\pi) \geq P^*$$

The optimization problem determining a power maximizing, cost constrained 2S-GWAS design is defined analogously:

$$max \quad P(\pi)$$

$$s.t. \quad C(\pi) \leq C^*$$

Implementation

We developed algorithms in C to identify optimal 2S-GWAS designs as a function of π. In our methods, the integration used to calculate the joint statistic's critical value and its power are computed using rectangular cubature; stage 2 critical values are obtained using the bisection method; and the three parameter constrained cost minimization and power maximization problems are solved using grid search. These algorithms are provided in Code S1.

Evaluating the cost advantages of two stage designs allowing unequal proportions of cases and controls allocated to stage 1

We examined the cost benefits and optimal design parameters for 2S-GWAS when not restricting case-control proportions in stage 1 to be equal (i.e. $\pi_0=\pi_1$) under an array of experimental conditions. Each condition was characterized by several factors that influence the optimal design and its cost: the ratio of controls to cases (R_{cc}); stage 2 per marker genotyping cost (c_2, where stage 1 per marker genotyping cost is taken to be 1); disease prevalence (κ); and the population frequency of the risk allele (p). We considered studies with $N_1=500$ cases and $M=100000$ markers, and determined case and control disease allele frequencies such that a one stage GWAS would have 80% power under a multiplicative model of genetic effects with experiment-wise type I error rate of 5% and controlling for multiple testing with a Bonferroni correction using the CaTS software ([10], [12]). The full set of experimental conditions is outlined in Table S1.

For each experimental condition, we identified cost minimizing 2S-GWAS designs that would maintain 78% power both with and without the $\pi_0=\pi_1$ restriction, using CaTS and the unrestricted methods described in this paper respectively. (As suggested by [10] and [12], 2S-GWAS designs typically target slightly lower power levels than one stage analyses, and so we reduced our target power by 2% from the one stage baseline.) Cost minimizing designs found using the unrestricted method were determined to the nearest 1% allocation of cases and controls to stage 1 (π_1,π_0) and 0.1% proportion of markers to be passed from stage 1 to stage 2 (π_M). We compared the costs of the one stage and optimal 2S-GWAS designs and verified the powers of the 2S-GWAS designs and critical values computed by the unrestricted methods using 100000 sets of sampled risk allele data. Finally, we evaluated the power sensitivities of the cost minimizing 2S-GWAS designs determined by both CaTS and our unrestricted method to batch effects or genetic heterogeneity between stages by assuming the second stage case and control disease allele frequency to be 90% of that used to calculate design parameters. Using the new second stage disease allele frequency, we then re-computed the powers of the original 2S-GWAS designs and critical values using 100000 sets of sampled risk alleles. We repeated this process assuming the second stage disease allele frequency was 110% of that specified.

Evaluating how much power is recovered by recomputing experimental designs after reductions in study funding

To examine how much power could be recovered by recomputing experimental designs after a hypothetical reduction in the funding available to a previously planned study, we focused on the set of experimental conditions defined in Table S1 with disease prevalence (κ) of 10% and disease allele frequency (p) of 10%. For each experimental condition, we determined the minimum cost 78% power unrestricted 2S-GWAS design, and then calculated maximum power unrestricted 2S-GWAS designs that were constrained to cost 50%, 75% and 90% of that. We compared the maximum obtainable study power after cost constraint to the original 78% target, verified the power of the computed designs and critical values using 100000 sets of sampled risk allele data, and determined the performance sensitivity of the power maximizing designs to batch effects as we did in our analogous study in cost minimizing designs.

Results

Two stage designs with unequal proportions of cases and controls allocated to stage 1 are optimal when controls outnumber cases

Minimum cost experimental design parameters and performance characteristics for the full set of experimental conditions described in Table S1 are provided in Tables S2 and S3. Here, we provide plots of our results for three sets of conditions: $\{\kappa=1\%,p=10\%,c_2=100\}$, letting $R_{cc}=1$, 2, 4, and 8; $\{\kappa=1\%,R_{cc}=8,c_2=100\}$, letting $p=10$, 25 and 50%; and $\{\kappa=1\%,p=10\%,R_{cc}=8\}$ letting $c_2=1$, 10 and 100 (where κ is disease prevalence, p the population disease allele frequency, c_2 the stage 2 genotyping cost and R_{cc} the ratio of controls to cases). Figures 2 and 3 describe the cost advantages of our unrestricted methods and characteristics of its computed design parameters respectively.

As in previous work [6,7], two stage designs computed by both CaTS and our unrestricted algorithm have substantial cost advantages relative to the one-stage design (Fig. 2). More important from the perspective of this paper are the cost advantages of unrestricted 2S-GWAS designs relative to those computed by CaTS. This advantage increases as the ratio of controls to cases or the cost of stage 2 genotyping increases, and as population disease allele frequency decreases. For the experimental conditions plotted in Fig. 2, the unrestricted algorithm shows up to a 40% cost advantage in comparison to CaTS. These results are consistent with those provided in Tables S2 and S3, which show that although there is little difference in cost performance in 2S-GWAS designs when the number of cases equal that of controls ($R_{cc}=1$), when $R_{cc}=8$ there is a 10–40% cost advantage gained by using unrestricted designs (taken across all other experimental conditions). For more modest differences between the number of cases and controls ($R_{cc}=2,4$), gains in cost efficiency can be

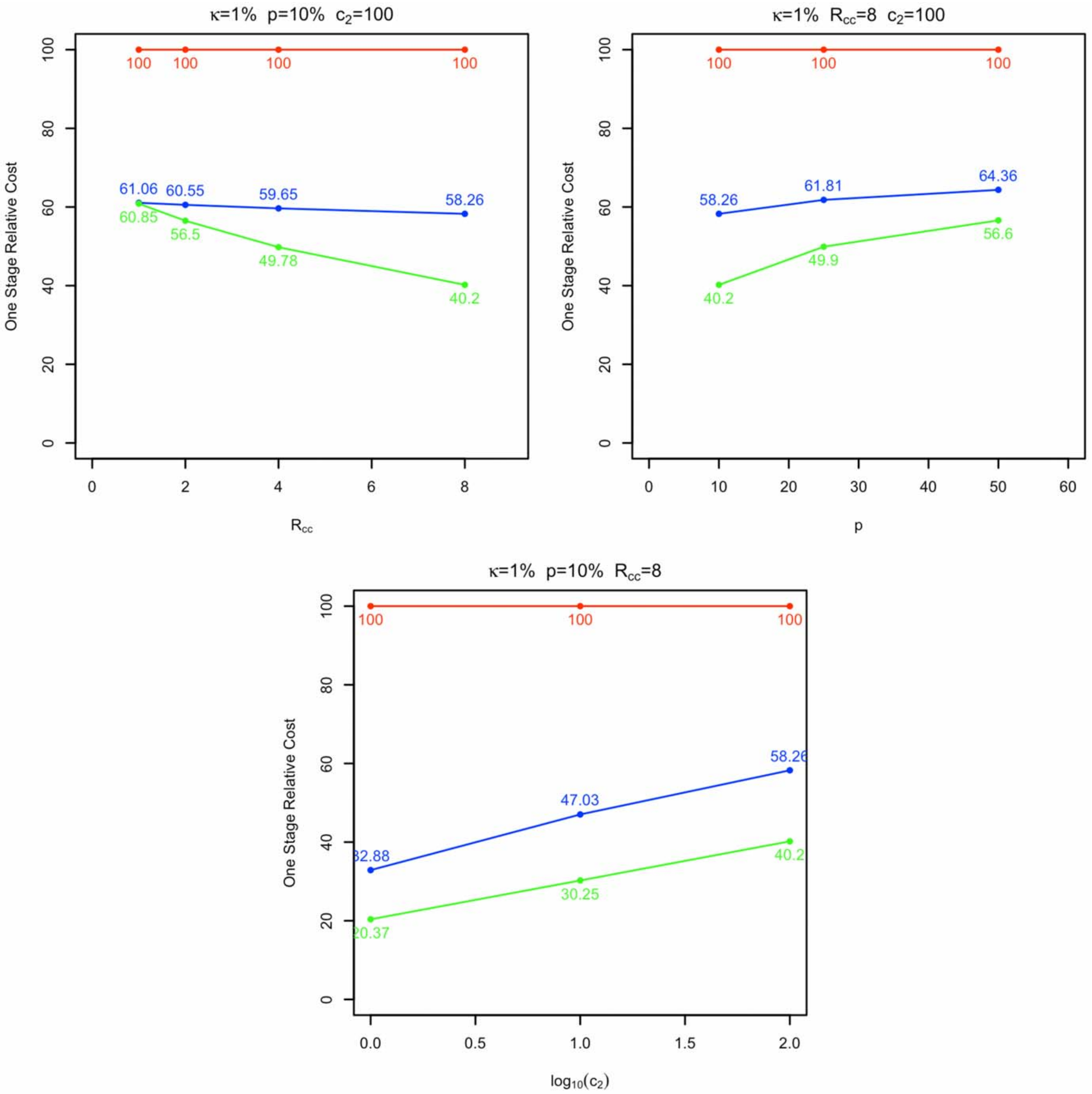

Figure 2. Relative cost curves for minimum cost designs. Cost curves for minimum cost, 78% power two stage GWAS designs calculated by CaTS (blue) and the unrestricted methods described here (green) relative to those of one stage GWAS designs (red) are provided for three experimental conditions: $\{\kappa=1\%,p=10\%,c_2=100\}$ as a function of R_{cc}; $\{\kappa=1\%,R_{cc}=8,c_2=100\}$ as a function of p; and $\{\kappa=1\%,p=10\%,R_{cc}=8\}$ as a function of c_2 (where κ is disease prevalence, p the population disease allele frequency, c_2 the stage 2 genotyping cost and R_{cc} the ratio of controls to cases). Unrestricted methods show significant cost advantages in comparison to those computed by CaTS. Cost advantages increase as R_{cc} increases, p decreases, and c_2 increases.
doi:10.1371/journal.pone.0042367.g002

observed as disease allele frequency decreases. For example, when $p=10\%$ we see a cost advantage to unrestricted designs of 10–15% relative to those of CaTS.

Given the cost advantages obtained by removing the equality constraint on the proportion of controls and cases (π_0 and π_1 respectively) allocated to stage 1, it would be expected that optimal design parameters computed using the unrestricted procedure and CaTS will differ. For the experimental conditions plotted in Fig. 3, it can be observed that designs computed with the unrestricted algorithm assign lower and higher proportions of controls and cases (respectively) to stage 1 than those computed by CaTS, and often reduce the proportion of markers passed to stage 2 (π_M). Additionally, it is clear that the degree of difference in design specification between the two methods can be influenced by each of the ratio of controls to cases, population disease allele frequency and stage 2 genotyping cost. The differences in design parameters

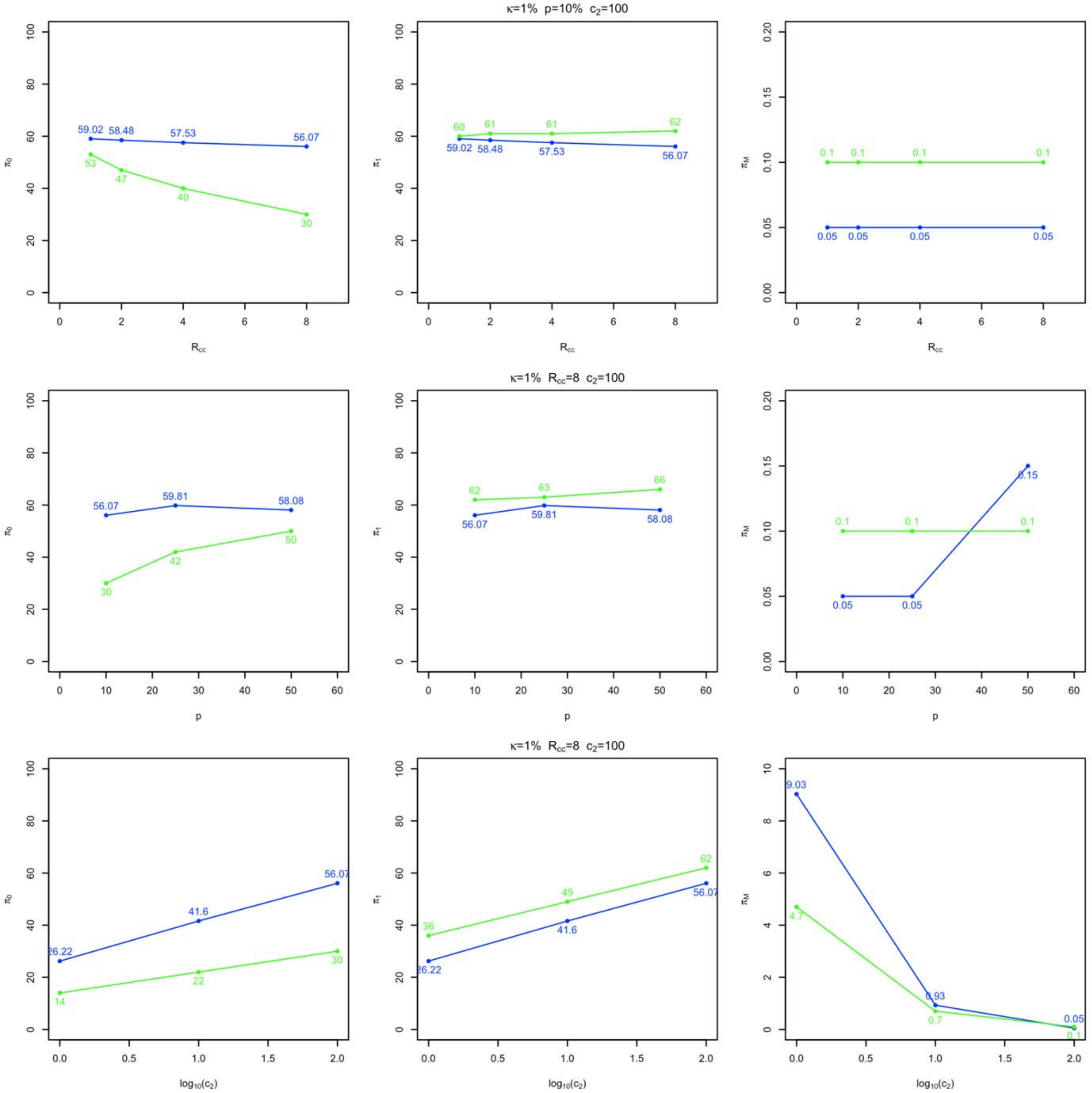

Figure 3. Design parameter curves for minimum cost designs. Two stage minimum cost, 78% power GWAS design parameters calculated by CaTS (blue) and the unrestricted methods described here (green) are provided for three experimental conditions: $\{\kappa=1\%, p=10\%, c_2=100\}$ as a function of R_{cc}; $\{\kappa=1\%, R_{cc}=8, c_2=100\}$ as a function of p; and $\{\kappa=1\%, p=10\%, R_{cc}=8\}$ as a function of c_2 (where κ is disease prevalence, p the population disease allele frequency, c_2 the stage 2 genotyping cost and R_{cc} the ratio of controls to cases). Each case is assigned a row, and design parameter plots for π_0, π_1 and π_M (the proportion of controls and cases assigned to stage 1, and the proportion of markers expected to be passed to stage 2) are displayed from left to right. Compared to designs computed by CaTS, designs computed without a $\pi_0=\pi_1$ constraint typically assign lower and higher proportions of controls and cases (respectively) to stage 1, and pass a greater proportion of markers to stage 2. The degree of difference in design specification between the two methods can be substantially influenced by each of R_{cc}, p and c_2.

doi:10.1371/journal.pone.0042367.g003

shown in Fig. 3 are again consistent with the results presented in Tables S2 and S3.

We note that the powers of the unrestricted 2S-GWAS designs and critical values are verified in our simulation studies, with no systematic deviation from the 78% target level (Tables S2 and S3). In terms of differences between stage specific powers of the design, both unrestricted 2S-GWAS designs and those proposed by CaTS generally have higher stage 1 specific power than stage 2 power. The influence of batch effects on power were similar for experiments designed with and without the sample allocation constraint; in both types of design, scaling stage 2 case and control disease allele frequencies to 90 or 110% of those in stage 1 reduces or increases the power of a proposed design by 5–10% respectively. As the cost of stage 2 genotyping increases (holding

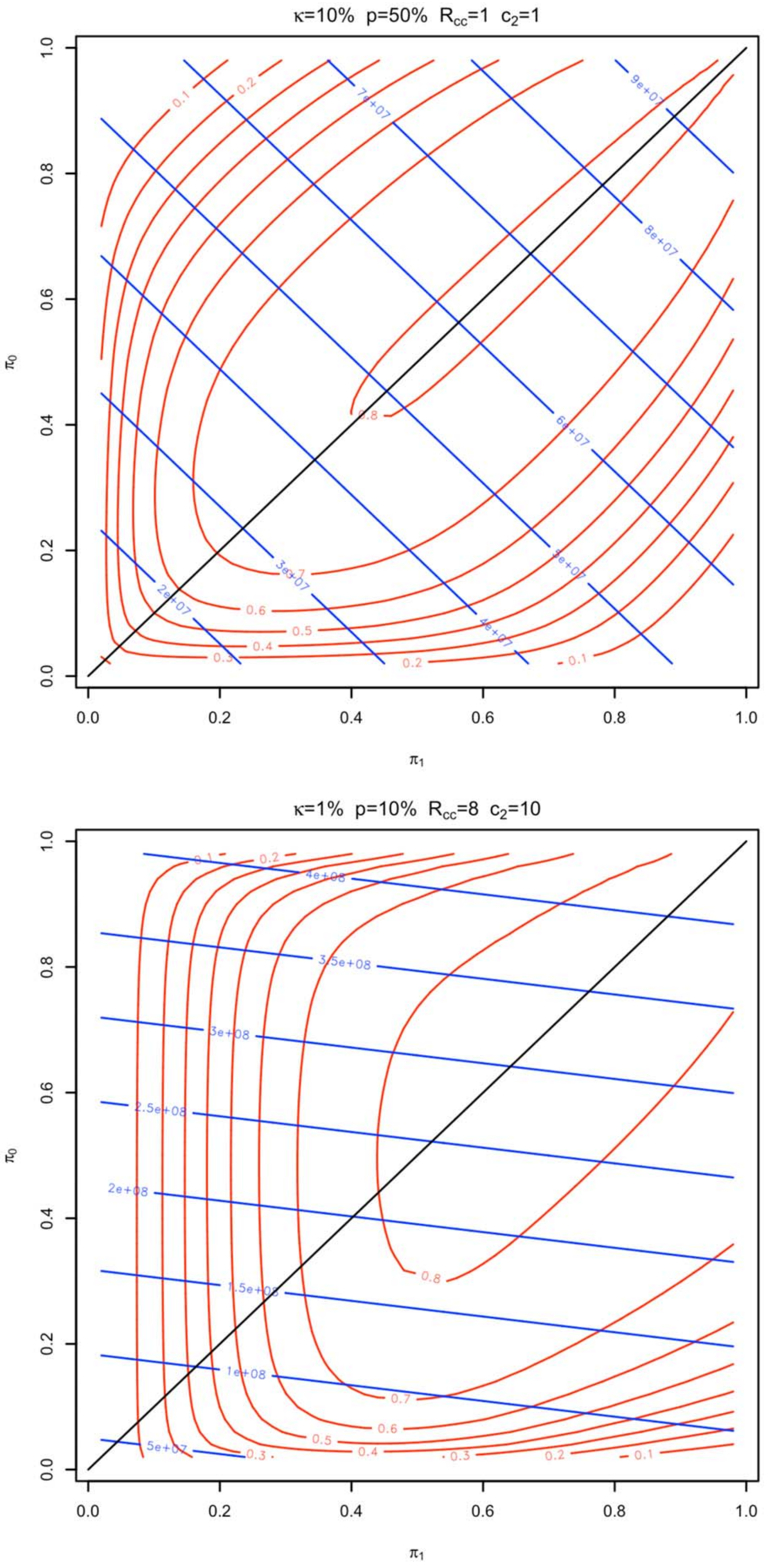
κ=10% p=50% R_{cc}=1 c_2=1
π_0
π_1
0.0
0.2
0.4
0.6
0.8
1.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
2e+07
3e+07
4e+07
5e+07
6e+07
7e+07
8e+07
9e+07
κ=1% p=10% R_{cc}=8 c_2=10
π_0
π_1
0.0
0.2
0.4
0.6
0.8
1.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
5e+07
1e+08
1.5e+08
2e+08
2.5e+08
3e+08
3.5e+08
4e+08

Figure 4. Power and cost surfaces. Power curves (red) and cost lines (blue) for the experimental conditions $\{\kappa=10\%,p=50\%,R_{cc}=1,c_2=1\}$ and $\{\kappa=1\%,p=10\%,R_{cc}=8,c_2=10\}$ (where κ is the disease prevalence, p the population disease allele frequency, c_2 the stage 2 genotyping cost, and R_{cc} the ratio of controls to cases) are plotted as a function of the proportion of controls (π_0) and cases (π_1) genotyped in stage 1, holding the proportion of markers followed up in stage 2 (π_M) at their cost minimizing values of 8.5% and 0.7% respectively. In the first case (top), power curves and cost lines are symmetric about the identify line, implying that the cost minimizing design will use equal case and control allocations. In the second (bottom) they are asymmetric, implying that the cost minimizing design will have unequal case and control allocations.
doi:10.1371/journal.pone.0042367.g004

the ratio of controls to cases constant) the sensitivity of a proposed design to batch effects is reduced.

Although it is intuitive that removing an equality constraint for the proportion of cases and controls allocated to stage 1 should improve the performance of an optimal design, it is useful to illustrate why this occurs. In Fig. 4, we plot power (red) and cost (blue) curves as functions of the proportions of controls and cases allocated to stage 1 for the experimental conditions $\{\kappa=10\%,p=50\%,R_{cc}=1,c_2=1\}$, and $\{\kappa=1\%,p=10\%,R_{cc}=8,c_2=10\}$, holding the expected proportion of markers to be passed to stage 2 at the cost minimizing values of 8.5% and 0.7% respectively (Tables S2 and S3). In these plots, the green identity line represents designs where the $\pi_0=\pi_1$ constraint holds. For the conditions with equal number of controls and cases ($R_{cc}=1$) the power surface is symmetric about and has cost constrained maxima on the identity line. That is, in this case, the optimal design should have equal proportions of cases and controls allocated to stage 1. When the number of controls increases ($R_{cc}=8$), both the power surface and cost curves become asymmetric with the optimal design having $\pi_1>\pi_0$.

Power maximizing designs with unequal proportions of cases and controls allocated to stage 1 can compensate for cost reductions

Power maximizing experimental designs and their performance characteristics are provided in Table S4. In Fig. 5, we describe the results calculated for a disease with 10% prevalence (κ), population disease allele frequency (p) of 10%, and stage 2 genotyping cost (c_2) of 10, with control/case ratios (R_{cc}) of 1 and 8 (blue and green lines respectively), as a function of the degree of relative cost restriction (50–100% of that of the cost minimizing 78% power design). As the experimental cost constraint is decreased from 100% of the minimum cost 78% power experimental design to

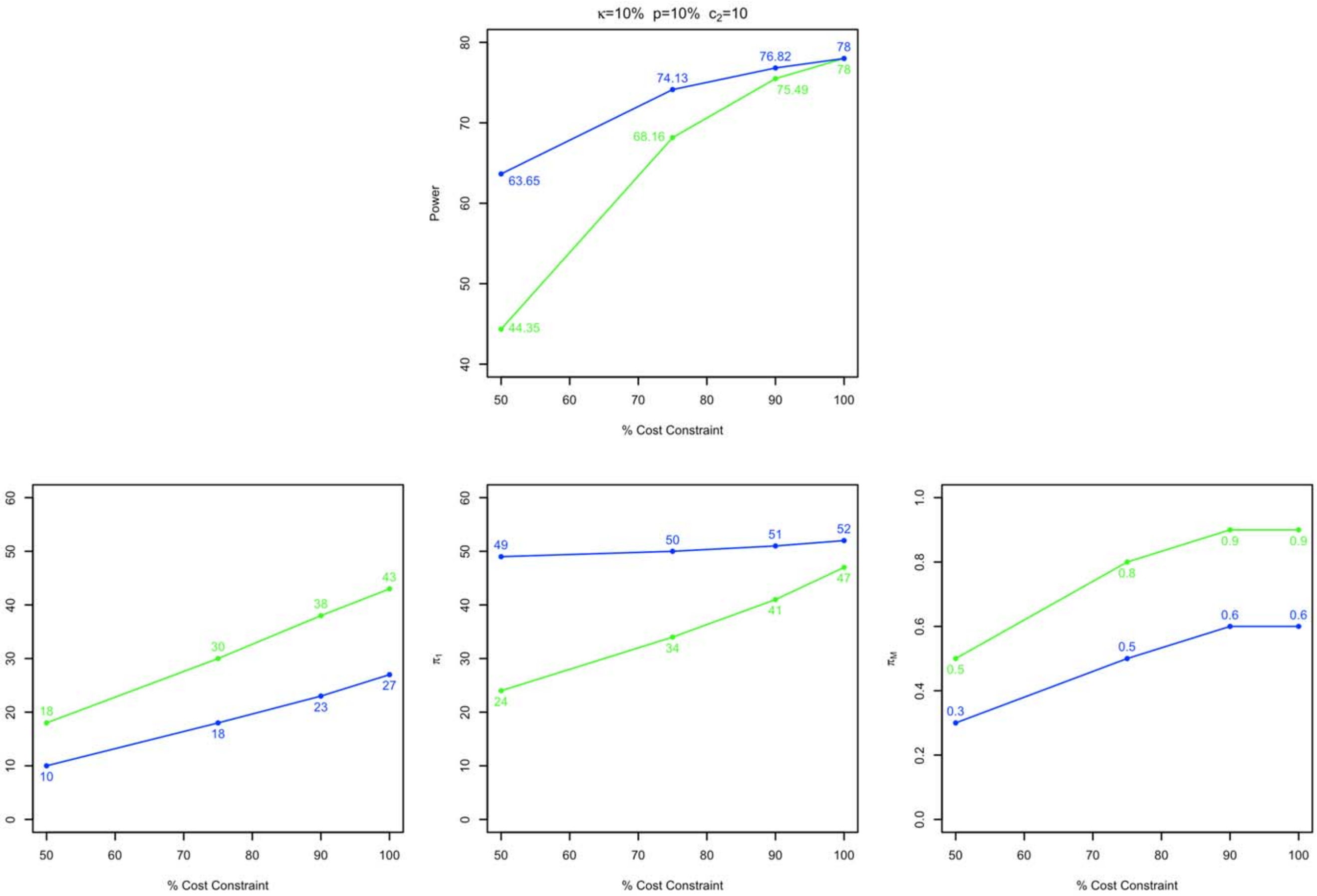

Figure 5. Performance and design characteristics of maximum power GWAS designs. The maximum achievable power and related experimental design parameters for the experimental condition $\{\kappa=10\%,p=10\%,c_2=10\}$ are plotted for control/case ratios $R_{cc}=1,8$ in blue and green respectively, as a function of percentage cost of a minimum cost, 78% power design. Although increasing the degree of cost constraint does have a negative effect on the achievable power of a 2S-GWAS, recomputing a power maximizing design can help to mitigate this. In comparison to the original cost minimized design, power maximizing cost limited designs typically pass a lower proportion of markers to stage 2, and then increase the relative power of the stage 2 test by allocating a greater proportion of both cases and controls to it.
doi:10.1371/journal.pone.0042367.g005

50% of that design, the maximum power obtainable by a cost constrained experimental design decreases from 78% to 44% and 68% for $R_{cc}=1,8$ respectively. In comparison to the baseline minimum cost 78% power design, power maximizing cost limited designs typically pass a lower proportion of markers to stage 2, and then increase the relative power of the stage 2 test by allocating a greater proportion of both cases and controls to it.

The effects of diminishing experimental cost on study design and power for the experimental conditions plotted in Fig. 5 are consistent with those for the other conditions reported in Table S4. Additionally, Table S4 reports the results from validation studies of experimental power, and analyses of the sensitivity of the performance of power maximizing designs to batch effects. The levels of power calculated for the power maximizing two stage GWAS designs were verified by simulation. We observed a reduced sensitivity of the performance of power maximizing designs to batch effects as c_2 increases (holding R_{cc} constant). This is evidenced by tightening ranges of power estimates over 90 and 110% scaling of second stage disease allele frequencies. Increases in R_{cc} (holding c_2 constant) did not have any such effect.

Discussion

In many analyses of two stage genome-wide association studies (2S-GWAS) experimental design, the proportions of cases and controls allocated to stages are implicitly constrained to be the same. In this paper we have expanded the framework for 2S-GWAS originally proposed by Skol et al [10,12] to remove this restriction. Using the expanded framework, we then demonstrated that in fact, cost-minimizing designs computed with respect to a desired level of statistical power often allocate different proportions of cases and controls to each stage. Relative to those computed under an equality constraint for proportions of cases and controls allocated to stage 1, unrestricted designs typically allocate fewer controls and more cases to stage 1, often pass fewer markers from stage 1 to stage 2, and have higher stage 1 and lower stage 2 power than those under the equality constraint. As would be expected, such designs offer substantial cost advantages relative to a 2S-GWAS that imposes equal allocation proportions. Such performance improvements become larger as the ratio of controls to cases increase, as the stage 2 per genotype cost increases, and as the population disease allele frequency decreases.

Based on this result, we extended our analysis to the problem of computing maximum power 2S-GWAS designs, subject to a cost constraint. We demonstrated that when a study budget is reduced below that of a minimum cost 2S-GWAS design meeting a targeted level of power, recomputing a maximum power 2S-GWAS design subject to the new cost constraint can retain much of the desired power. Relative to the original minimum cost design, power retention is achieved by allocating fewer cases and controls to stage 1, and passing fewer markers from stage 1 to stage 2. That is, the reduction in cost is compensated for by collecting less information in stage 1, focusing on fewer markers in stage 2, and then using a more powerful stage 2-specific test.

We note that the results achieved here are in some respects obvious - removal of a constraint from an optimization problem always weakly results in improvements in performance. However, the extent to which this is true for 2S-GWAS has not been made explicit in previous studies. Additionally, many questions pertaining to such improvements, such as why the optimal designs changed as experimental parameters changed, could only be understood by investigating the problem geometry. Related to such investigations, our studies of the sensitivity of 2S-GWAS designs to batch effects or genetic heterogeneity between stages demonstrated that our unconstrained 2S-GWAS designs are not substantially different (in that respect) from those that constrain case and control sample proportions to be equal. The gains in efficiency related to removing the sample proportion constraint do not come at the cost of higher sensitivity to batch effects.

Because most 2S-GWAS designs constrain sample allocation proportions to be the same across cases and controls, we suggest the results presented here may have implications beyond our particular study. For example, in [14], it was demonstrated that for experiments using the same numbers of cases and controls, it is in principle possible to obtain greater cost efficiencies by using three or four stages rather than two. However, the use of differential case/control allocation proportions for problems in which the numbers of cases and controls differ was not considered. Likewise, in [11], two stage GWAS designs using false discovery rate criteria were considered, again while imposing that equal numbers of cases and controls assigned to each stage. It is possible that using techniques analogous to those described here could yield greater levels of cost efficiency and power performance in such designs.

We recommend that when designing a 2S-GWAS, investigators think carefully about the relative number of controls to cases, genotyping costs, and disease allele frequencies. To assist in doing do, our programs for identifying optimal two-stage GWAS designs are provided in Code S1 or alternatively at http://www.bioinformatics.org/stanhope/2SGWASdesign/.

Supporting Information

Appendix S1 Supporting derivations. This appendix provides mathematical details omitted in the main text, including the proof of Theorem 1; the Fisher information stage weighting calculation; and the stage 2 critical value and power calculations.
(PDF)

Code S1 Supporting software. This file contains all codes necessary to calculate cost minimizing and power maximizing two-stage GWAS designs with unequal proportions of cases and controls allocated to stages. Instructions are provided for their compilation and use, as well as example calculations.
(GZ)

Table S1 Experimental conditions for 2S-GWAS design calculations. Experiments are described in terms of disease prevalences (κ); disease allele frequencies (p); ratio of controls to cases (R_{cc}); stage 2 genotyping costs c_2 ($c_1=1$ is held constant); and case/control allele frequencies (p_1,p_0). Numbers of cases and markers are constant at $N_1=500$ and $M=100000$.
(PDF)

Table S2 Cost minimizing 2S-GWAS designs and their performance characteristics, $\kappa=10\%$. For experimental conditions with $\kappa=10\%$ in Table S1, Table S2 reports two-stage 78% power designs computed from both CaTS and the unrestricted method. The costs of two-stage designs are compared to those of 80% power one-stage designs, and the costs of the unrestricted two-stage designs are compared to those of CaTS. Verification of the power levels of unrestricted 2S-GWAS designs is performed by Monte Carlo.
(PDF)

Table S3 Cost minimizing 2S-GWAS designs and their performance characteristics, $\kappa=1\%$. For experimental conditions with $\kappa=1\%$ in Table S1, Table S3 reports two-stage 78% power designs computed from both CaTS and the unrestricted method. The costs of two-stage designs are compared to those of 80% power one-stage designs, and the costs of unrestricted two-stage designs are compared to those of CaTS.

Verification of the power levels of unrestricted 2S-GWAS designs is performed by Monte Carlo.
(PDF)

Table S4 Power maximizing two stage GWAS designs and their performance characteristics, $\kappa = 10\%, p = 10\%$. For all experimental conditions with $\kappa = 10\%, p = 10\%$ in Table S1, Table S4 reports two-stage maximum power designs and the powers they attain, with respect to a cost constraint expressed as a percentage of the cost of the minimum cost designs reported in Table S2.
(PDF)

Acknowledgments

We would like to thank Mark Abney for his support.

Author Contributions

Conceived and designed the experiments: SAS ADS. Performed the experiments: SAS. Analyzed the data: SAS ADS. Contributed reagents/materials/analysis tools: SAS ADS. Wrote the paper: SAS ADS. Development of statistical software: SAS.

References

1. Scott L, Mohlke K, Bonnycastle L, Willer C, Li Y, et al. (2007) A genome-wide association study of type 2 diabetes in finns detects multiple susceptibility variants. Science 316: 1341–1345.
2. O'Donovan M, Craddock N, Norton N, Williams H, Peirce T, et al. (2008) Identification of loci associated with schizophrenia by genome-wide association and follow-up. Nature Genetics 40: 1053–1055.
3. Graham R, Cotsapas C, Davies L, Hackett R, Lessard C, et al. (2008) Genetic variants near tnfaip3 on 6q23 are associated with systemic lupus erythematosus. Nature Genetics 40: 1059–1061.
4. Nair R, Duffin K, Helms C, Ding J, Stuart P, et al. (2009) Genome-wide scan reveals association of psoriasis with il-23 and nf-κb pathways. Nature Genetics 41: 199–204.
5. Thomas G, Jacobs K, Kraft P, Yeager M, Wacholder S, et al. (2009) A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (rad51l1). Nature Genetics 41: 579–584.
6. Satagopan J, Elston R (2003) Optimal two-stage genotyping in population-based association studies. Genetic Epidemiology 25: 149–157.
7. Thomas D, Xie R, Gebregziabher M (2004) Two-stage sampling designs for gene association studies. Genetic Epidemiology 27: 401–414.
8. Kitamura N, Akazawa K, Toyabe S, Miyashita A, Kuwano R, et al. (2008) Sample-size properties of a case-control association analysis of multistage snp studies for identifying disease susceptibility genes. Journal of Human Genetics 53: 390–400.
9. Gail M, Pfeiffer R, Wheeler W, Pee D (2008) Probability that a two-stage genome-wide association study will detect a disease-associated snp and implications for multistage designs. Annals of Human Genetics 72: 812–820.
10. Skol A, Scott L, Abecasis G, Boehnke M (2006) Joint analysis is more efficient than replicationbased analysis for two-stage genome-wide association studies. Nature Genetics 38: 209–213.
11. Wang H, Stram D (2006) Optimal two-stage genome-wide association designs based on false discovery rate. Computational Statistics & Data Analysis 51: 457–465.
12. Skol A, Scott L, Abecasis G, Boehnke M (2007) Optimal designs for two-stage genome-wide association studies. Genetic Epidemiology 31: 776–788.
13. Nguyen T, Pahl R, Schäfer H (2009) Optimal robust two-stage designs for genome-wide association studies. Annals of Human Genetics 73: 638–651.
14. Pahl R, Schäfer H, Müller HH (2009) Optimal multistage designs - a general framework for genomewide association studies. Biostatistics 10: 297–309.
15. Scherag A, Hebebrand J, Schäfer H, Müller H (2009) Flexible designs for genomewide association studies. Biometrics 65: 815–821.
16. Thomas D, Casey G, Conti D, Haile R, Lewinger J, et al. (2009) Methodological issues in multistage genome-wide association studies. Statistical Science 24: 414–429.
17. Van Steen K (2010) Perspectives on genome-wide multi-stage family-based association studies. Statistics in Medicine 30: 2201–2221.