RESEARCH ARTICLE

Genetic Epidemiology

OFFICIAL JOURNAL
INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY
www.geneticepi.org

A Note on the Efficiencies of Sampling Strategies in Two-Stage Bayesian Regional Fine Mapping of a Quantitative Trait

Zhijian Chen,[1] Radu V. Craiu,[2]* and Shelley B. Bull[1,3]

[1]*Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, Toronto, Ontario, Canada;* [2]*Department of Statistical Sciences, University of Toronto, Toronto, Ontario, Canada;* [3]*Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Ontario, Canada*

Received 21 December 2013; Revised 12 June 2014; accepted revised manuscript 16 June 2014.
Published online 1 August 2014 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21845

ABSTRACT: In focused studies designed to follow up associations detected in a genome-wide association study (GWAS), investigators can proceed to fine-map a genomic region by targeted sequencing or dense genotyping of all variants in the region, aiming to identify a functional sequence variant. For the analysis of a quantitative trait, we consider a Bayesian approach to fine-mapping study design that incorporates stratification according to a promising GWAS tag SNP in the same region. Improved cost-efficiency can be achieved when the fine-mapping phase incorporates a two-stage design, with identification of a smaller set of more promising variants in a subsample taken in stage 1, followed by their evaluation in an independent stage 2 subsample. To avoid the potential negative impact of genetic model misspecification on inference we incorporate genetic model selection based on posterior probabilities for each competing model. Our simulation study shows that, compared to simple random sampling that ignores genetic information from GWAS, tag-SNP-based stratified sample allocation methods reduce the number of variants continuing to stage 2 and are more likely to promote the functional sequence variant into confirmation studies.

Genet Epidemiol 38:599–609, 2014.

KEY WORDS: Bayes factor; genetic models; multistage; next-generation sequencing

Introduction

When large samples have been recruited for genome-wide association study (GWAS) but whole genome sequencing is still not a viable option for fine-mapping despite the decreasing cost of next-generation sequencing (NGS) [Hedges et al., 2011], targeted sequencing or dense genotyping of all variants in a candidate region is an attractive alternative [Almomani et al., 2011]. For example, the Wellcome Trust Case Control Consortium (WTCCC) investigated regions identified in GWASs for three diseases by dense genotyping of variants across these regions, and defined, using Bayes factors, credible sets of variants that were likely to contain the causal disease-associated variants [Wellcome Trust Case Control Consortium et al., 2012]. Additional savings can be gained when the fine-mapping phase incorporates a two-stage design, analogous to that previously developed in the GWAS setting [e.g., Skol et al., 2007; Thomas et al., 2009]. In stage 1, a subset of the original GWAS subjects is selected and densely genotyped, examining all variants in the target region using expensive regional sequencing technology. In stage 2, selected variants identified in stage 1 are typed in the remaining subjects using cost-effective genotyping technologies. Subsequently, association of these variants with the quantitative trait can be evaluated using the combined data from both stages. Figure 1 illustrates a two-stage fine-mapping design based on an existing GWAS sample.

The purpose of a fine-mapping study for a complex quantitative trait is to identify a few variants, if not a single one, that are potentially responsible for the variation in the trait, estimate genetic effect sizes, and characterize genetic association at the gene level. The information provided by GWAS tag SNPs can be useful in the selection of subjects for sequencing [Chen et al., 2012; Schaid et al., 2013]. As opposed to a simple random-sampling (SRS) procedure, a good sample-size allocation in a properly stratified sample (involving under- or oversampling of strata) may improve efficiency of effect size estimation at a functional sequence variant. One approach stratifies the GWAS sample according to the three tag SNP genotype categories: common homozygote, heterozygote, and rare homozygote. For a quantitative trait, Chen et al. [2012] found that estimation efficiency can be gained when the frequency of sampling the homozygote strata is higher than one would expect under SRS and also when the frequency of samples from the heterozygote stratum is lower than under SRS, provided that the additive genetic model is correctly specified and the tag-seq linkage disequilibrium (LD) is reasonably high, for example, above 0.80. In a case-control setting, Schaid et al. [2013] showed that stratified sampling based on both tag genotypes and case-control status is not likely to have lower power than stratified sampling

Supporting Information is available in the online issue at wileyonlinelibrary.com.

*Correspondence to: Radu V. Craiu, Department of Statistical Sciences, University of Toronto, Toronto, Ontario M5S 3G3, Canada. Email: craiu@utstat.toronto.edu

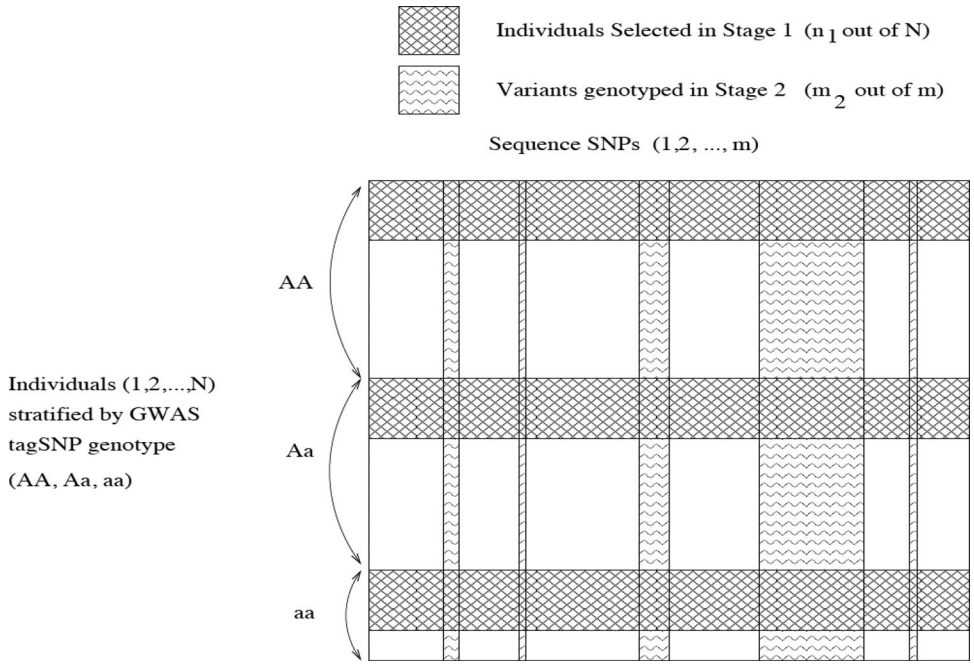

Figure 1. Illustration of a two-phase two-stage design, with the GWAS phase examining 1∼3 millions of tag SNPs in a total of N subjects and the fine-mapping phase focusing on m SNPs within a specified region identified by GWAS. The rows of the matrix correspond to N individuals stratified by the GWAS tag SNP genotype (e.g., N = 5,000, with expected strata sizes of N_{AA} = 2450, N_{Aa} = 2100, N_{aa} = 450 for a tag SNP with MAF = 0.30), and the columns correspond to m sequenced SNPs ordered by chromosome position. The fine-mapping phase consists of two-stages: in Stage 1, n_1 individuals are sampled for sequencing of all variants in a region surrounding the tag SNP (e.g., n_1 = 1,000, with random sampling of an equal number of individuals from each of the three strata: n_{AA} = 334, n_{Aa} = 333, $n_{aa,}$ = 333 corresponding roughly to sampling fractions of 1/8, 1/4, and 3/4, respectively). A subset of m_2 promising sequence SNPs is identified (e.g., m_2 equal to 30% of the m variants); the selected SNPs are not necessarily contiguous, although their distribution within the region will depend strongly on the local LD structure. In Stage 2, the m_2 variants are genotyped in the $N - n_1$ remaining subjects.

based only on case-control status, and can sometimes have substantially greater power. Both these studies considered analysis under an additive or log-additive model for a functional sequence variant, an assumption that may be violated in practice.

Genetic model specification in genetic analysis is a very long-standing problem [for discussion see Joo et al., 2010; Stephens and Balding, 2009; Strauch et al., 2003; Vukcevic et al., 2011]. In our context, model misspecification may have a negative impact on the choice of variants for stage 2. Although the additive model has been widely used in the discovery stage for GWASs of many complex traits and diseases, genetic effect size estimates at the sequence variant are biased when the underlying genetic model is nonadditive. For a nonfunctional sequence variant, the impact of model misspecification depends on LD with the functional variant. Furthermore, the correct genetic model for the sequence variant may be difficult to identify when few heterozygotes at the sequence variant are observed. Several authors have explored the nature of the relationship between a GWAS tag SNP, used to identify the region of interest, and a functional sequence variant within the region, examining the impact of the LD correlation on the association estimate, the ability to identify a genetic model, and the accuracy of localization [Faye et al., 2013; Spencer et al., 2011; Vukcevic et al., 2011]. Characterizing the genetic model, i.e., the mode of inheritance, for a putative functional variant, even approximately, is of substantial interest in this fine-mapping process, and may be informative for ongoing study design.

In this article, we consider a Bayesian approach for regional fine-mapping with selection of a credible set of variants similar to that of Wellcome Trust Case Control Consortium et al. [2012], but here we incorporate a two-stage sampling procedure in the fine-mapping phase. The Bayesian approach enables comparison among variants and the identification of a credible set that is analogous to, but more directly interpretable than, a confidence interval in a frequentist approach. By comparisons among genetic models using the stage 1 sample, as well as among variants, we aim primarily to improve knowledge about the position of the functional sequence variant and secondarily, to learn about the genetic model. Selected potentially functional variants are then evaluated in stage 2 by genotyping the remaining samples using a more cost-effective technology. We focus on features associated with the stage 1 design and analysis, and the value of sampling and genetic model information, including reduction in the size of the credible set, genetic model identification, as well as the probability of selecting a function sequence variant for stage 2.

The rest of this report is organized as follows. In the following section, we propose a Bayesian method for two-stage

stratified design and the selection of a credible set of variants for confirmation. To demonstrate how one can implement the proposed method in practice, we simulate data from 1000 Genomes Project and evaluate three sample allocation schemes under various settings, including cost-efficiency considerations. We conclude with discussion.

Methods

Two-Stage Stratified Design

Consider a fine-mapping study of a complex quantitative trait denoted by the random variable Y. Let G be the genotype of a tag SNP, with major and minor alleles A and a, that has drawn attention to a region potentially harboring a functional sequence variant. Each of the individuals in the GWAS sample will yield a pair of observed values (G_i, Y_i), $i = 1, \ldots, N$. We define three strata in the GWAS sample (Fig. 1) according to the three categories of the tag SNP genotype, i.e., common and rare homozygote and heterozygote, within which we will select individuals into a stage 1 fine-mapping sample. Let n_1 and n_2 be the number of individuals in the stage 1 and stage 2 samples, respectively, $N = n_1 + n_2$. Ideally, the stage 1 sample proportion n_1/N should be chosen to maximize the ability to correctly select the functional variant for typing in stage 2, subject to financial constraints on the genotyping costs. Practically, the choice of n_1 is dictated by more modest goals, e.g., financial constraints and/or the ability to reduce the scope of the fine-mapping in stage 2 to a feasible number of variants that includes the functional variant with high probability. For individuals selected into the stage 1 sample, we genotype all sequence variants found in the fine mapping region. Let Y_{S1} and X_{S1} denote the response and sequence genotype data in the stage 1 sample. We assume that among all sequence variants examined in the fine-mapping region there is at least one functional sequence variant.

At stage 1 sampling, a good sample allocation is expected to maximize the evidence for the functional sequence variant and minimize the number of variants selected for evaluation in subsequent stages. In some cases, the region may contain a large number of variants that are in high LD with the unknown functional sequence variant such that they account for a large portion of the credible set (we call them hitchhiker variants). We are interested in three representative sampling schemes. The first scheme uses SRS, which ignores information from the tag SNPs (or imputed SNPs) provided in the GWAS phase. The second scheme samples an equal number of individuals from each of the three strata defined by the GWAS tag genotypes (equal strata [ES]). The third uses a stratified sampling strategy in which the relative number of samples from the rare homozygote stratum is larger than would be expected under SRS (HO). Similarly, the relative number of samples from the heterozygote stratum is smaller than expected under an SRS scheme. The third sampling scheme is expected to lead to better efficiency when the genetic model for the functional variant is additive and the tag-seq LD correlation is high, (e.g., $r^2 > 0.8$). On the other hand, if the genetic model is not additive and/or the tag-seq LD is low, then increasing the relative frequency of samples from the homozygous strata does not necessarily lead to improved efficiency, because the sampling scheme becomes more like SRS.

Model Formulation for a Quantitative Trait

We assume there is a functional sequence variant with a minor allele frequency (MAF) of 1% or greater. Let X be a variable that counts the number of copies of the minor allele at this functional variant for an individual. Without loss of generality we consider a simple linear regression model, but in practice a set of relevant nongenetic covariates can be specified and included in the regression models. Although the additive model is frequently used in GWAS for discovering association at tag SNPs, the underlying genetic model for a functional variant close to a promising tag SNP may not be truly additive. Following Spencer et al. [2011], a general three-parameter model is

$$Y = \beta_0 + \beta_1 X + \gamma 1_{X=1} + \varepsilon, \quad (1)$$

which encompasses additive, dominant, and recessive models. Here β_1 measures the increase or decrease in the value of the trait with each additional copy of the minor allele, $1_{X=1}$ is an indicator function that takes value 1 for heterozygotes and 0 for the two homozygotes, the dominance parameter γ measures deviation from additivity, and ε follows a normal distribution with mean 0 and variance σ^2 independently across individuals. Each of the genetic models can be recovered by setting the dominance parameter in (1) to a specific value: $\gamma = 0$ gives the additive model, and $\gamma = \beta_1$ and $\gamma = -\beta_1$ correspond to the dominant and recessive models, respectively. Under additive, dominant, and recessive models, the conditional mean $E(Y|X)$ of the trait value is $\beta_0 + \beta_1 X$, $\beta_0 + \beta_1(X + 1_{X=1})$, and $\beta_0 + \beta_1(X - 1_{X=1})$, respectively. As these three models involve the same number of regression parameters, we focus on inference about the association parameter β_1, simplifying the model comparison procedure described in the following subsections.

Bayesian Inference

In this section, we consider Bayesian inference with a stage 1 sample of n_1 subjects. Assume that a total of L variants in the region are typed using targeted sequencing technology. We analyze the stage 1 sample phenotype and sequence genotype data (Y_{S1}, X_{S1}) to narrow down the set of sequence variants that are potentially functional, and to identify the underlying genetic model. To develop the methods, we first specify priors for genetic models and regression parameters. Then, for a functional sequence variant we derive the model-specific posterior for the regression parameters and compare genetic models using Bayes factors. We specify selection of the underlying genetic model for the sequence variant by calculating the posterior probability for each of the three genetic models. Finally, we analyze all sequence variants in the region, and by making comparisons among variants,

we compute the posterior probability of each variant being functional and select variants for the 95% credible set.

Prior Specification

Let $\theta = (\beta, \sigma^2)^T$ be a vector of the parameters in the quantitative trait model, where $\beta = (\beta_0, \beta_1)^T$ are regression coefficients. Let M_1, M_2, and M_3 denote the additive, dominant, and recessive genetic models, respectively. Let $p(M_j)$ be the prior probability for M_j and $p(\theta|M_j)$ be the prior distribution of θ under model M_j, $j = 1, 2, 3$. In the absence of a priori information on the genetic model, we assume that $p(M_j) = 1/3$ and $p(\theta|M_j) = p(\theta)$, i.e., the prior for θ is independent of the underlying genetic model. We specify a conjugate prior $p(\theta)$ given by $p(\beta, \sigma^2) = p(\beta|\sigma^2)p(\sigma^2)$, where $p(\beta|\sigma^2) = Normal(b_0, \sigma^2 B_0)$ and $p(\sigma^2) = Inv-Gamma(\nu_0/2, \sigma_0^2\nu_0/2)$, for some $\nu_0 > 2$. That is, the prior joint density is normal-inverse-gamma $NIG(b_0, B_0, \nu_0/2, \sigma_0^2\nu_0/2)$. Here σ_0^2 is a prior guess at the variance and ν_0 measures the strength of belief in that guess. The matrix B_0 is assumed diagonal. In cases of no strong a priori belief concerning the magnitude of the genetic effect, we specify the prior to be reasonably flat over the range of plausible effect values. Such vague prior distributions do not favor any particular value, letting the posterior depend largely on the data alone.

Genetic Model Selection for a Sequence Variant

In this section we temporarily suppress the index for variants and present the model selection method for the functional variant; but the same analysis is applied to each variant in the region. The posterior for θ under genetic model M_j is

$$p\left(\theta|Y_{S1}, X_{S1}, M_j\right) = p\left(Y_{S1}|X_{S1}, \theta, M_j\right) p\left(\theta|X_{S1}, M_j\right)/c_j,$$

where $c_j = p(Y_{S1}|X_{S1}, M_j) = \int p(Y_{S1}|X_{S1}, \theta, M_j)p(\theta|X_{S1}, M_j)d\theta$ is the normalizing constant of the posterior distribution. The posterior mean and variance of θ are $E(\theta|Y_{S1}, X_{S1}, M_j) = \int \theta p(\theta|Y_{S1}, X_{S1}, M_j)d\theta$ and $\text{var}(\theta|Y_{S1}, X_{S1}, M_j) = \int\{\theta - E(\theta|Y_{S1}, X_{S1}, M_j)\}\{\theta - E(\theta|Y_{S1}, X_{S1}, M_j)\}^T p(\theta|Y_{S1}, X_{S1}, M_j)d\theta$, respectively.

The posterior distribution of θ can be derived analytically when the prior distribution is specified as normal-inverse-gamma. This is desirable as it allows for fast processing of the data and, because it does not rely on Monte Carlo methods for analyzing the posterior distribution, it does not require additional computational effort. For genetic model M_j, let $X_{S1,j}$ be the corresponding design matrix. Define $B_{S1,j} = [B_0^{-1} + \{X_{S1,j}\}^T X_{S1,j}]^{-1}$, and $\beta_{S1,j} = B_{S1,j}[B_0^{-1}b_0 + \{X_{S1,j}\}^T Y_{S1}]$. Then the posterior for θ is also a normal-inverse-gamma distribution $NIG(\beta_{S1,j}, B_{S1,j}, \nu_{S1}/2, \sigma_{S1,j}^2\nu_{S1}/2)$, where $\nu_{S1} = \nu_0 + n_1$ does not depend on the genetic model, and $\sigma_{S1,j}^2 = [\sigma_0^2\nu_0 + Y_{S1}^T Y_{S1} + b_0^T B_0^{-1} b_0 - \beta_{S1,j}^T\{B_{S1,j}\}^{-1}\beta_{S1,j}]/\nu_{S1}$. The posterior marginal densities of β and σ^2 as well as the marginal likelihood are analytically tractable. Specifically, the posterior marginal for β is a multivariate t-distribution with ν_{S1} degrees of freedom. It can be shown that the posterior marginal mean of β is $\beta_{S1,j}$, which is essentially a weighted average of the prior guess b_0 and the maximum-likelihood (ML) estimate. The posterior marginal variance of β is $\nu_{S1}/[\{\nu_{S1} - 2\}\sigma_{S1,j}^2 B_{S1,j}]$. The posterior marginal for σ^2 is inverse gamma with parameters $\nu_{S1}/2$ and $\sigma_{S1,j}^2\nu_{S1}/2$.

We note that conditional on second-stage data, the Bayesian analysis is independent of the sampling weights established in stage 1. This follows because the stratifying variable from the GWAS (tag SNP), although correlated with the target sequence variant, is conditionally independent of the response given the information on the functional sequence variant. In online Supplementary Appendix A we outline a proof that the inclusion probability and the distribution of the sequence genotype do not enter into the calculation of the posterior for θ.

Typically, Bayesian selection of a genetic model for a sequence variant involves computation of the posterior weight of each model. Assuming all three sequence genotype categories are observed, we first compare the three genetic models using the Bayes factor [e.g., Stephens and Balding, 2009; Wakefield, 2009]. For instance, the Bayes factor comparing M_j to $M_{j'}$, with $M_{j'}$ being the reference genetic model, is

$$BF_{jj'} = p\left(Y_{S1}|X_{S1}, M_j\right) p\left(M_j\right) / \left\{p\left(Y_{S1}|X_{S1}, M_{j'}\right) p\left(M_{j'}\right)\right\}.$$

In the case of each model being equally likely a priori, i.e., $p(M_{j'}) = p(M_j)$, this is equivalent to computing the ratio of the normalizing constants for the posteriors obtained under M_j and $M_{j'}$. In general, $3 \le BF_{jj'} \le 10$ (or equivalently $1/10 \le BF_{j'j} \le 1/3$) suggests substantial evidence that the data are in favor of genetic model M_j over $M_{j'}$, whereas $BF_{jj'} \ge 10$ (or equivalently $BF_{j'j} \le 1/10$) suggests strong evidence in favor of M_j over $M_{j'}$. The conjugate prior specification makes the calculation of Bayes factors straightforward.

The posterior probabilities, or posterior weights, of genetic models are defined by $w_j = p(M_j|Y_{S1}, X_{S1})$, $j = 1, 2, 3$, which summarize evidence for the underlying genetic mechanism at the seq variant after incorporating the observed data and prior belief. These weights can be calculated as $w_j = 1/\{1 + \sum_{j' \ne j} BF_{j'j}\}$. The genetic model with the largest posterior weight is deemed to be the underlying model. This is equivalent to selection based on normalizing constants such that the best genetic model corresponds to the maximum normalizing constant $c_{\max} = \max\{c_j\}_{j=1}^3$. For variants where only two genotype categories are observed, we assume an additive genetic model and set $c_{\max} = c_1$. Using the Bayes factor criteria, we would conclude that there is strong evidence that the effect at the sequence variant follows genetic model M_j whenever $w_j > 83.3\,\%$, $j = 1, 2, 3$.

Selection of the 95% Credible Set

Having first determined a genetic model for each sequence variant, we can compare any two variants in the region by computing their associated Bayesian factor. But to compare all sequence variants, we use the normalizing constant corresponding to the selected genetic model for each variant

and compute the posterior probability of each variant being functional. Following the approach of Wellcome Trust Case Control Consortium et al. [2012] for case-control genetic studies, we define a credible set of variants that is 95% likely to contain the functional variant for the quantitative trait, using the posterior probabilities computed below. Our definition of the credible set is similar to Wellcome Trust Case Control Consortium et al. [2012] except that we allow for differences in genetic models among the variants.

Assuming that the variant being examined is a functional variant, we compute $p^k = P(\text{variant } k \text{ is functional} | Y_{S1}, \{X^l_{S1}\}^L_{l=1})$, the posterior weight of variant k being functional given the quantitative trait and the genotype data of the L variants in the stage 1 sample, where X^l_{S1} is the genotype data of variant l. Specifically, the posterior weight of variant k being functional is given by $p^k = c^k_{\max} / \sum_{l=1}^{L} c^l_{\max}$. We sort these posterior weights in descending order $p^{(1)} \geq p^{(2)} \geq \ldots \geq p^{(L)}$. Then, the 95% credible set is defined such that the variants in the set have posterior weights $p^{(1)}, \ldots, p^{(K)}$, where K is the minimum integer that satisfies $\sum_{k=1}^{K} p^{(k)} \geq 0.95$. With a fixed number of subjects in the stage 1 sample, a smaller value of K implies that the sampling design performs better in narrowing down the credible set that is likely to contain the functional variant. In particular, a good design amounts to (1) higher probability that the correct genetic model will be identified for the functional seq variant, (2) higher probability that the functional seq variant will be selected into the credible set, and (3) fewer seq variants assigned to the credible set (i.e., smaller set size).

Evaluation of Stratified Sampling Designs

Design of Simulation Studies

To demonstrate application of the Bayesian two-stage fine-mapping design, we conducted simulation studies using data derived from known population haplotypes. We evaluated and compared two stratified sampling designs to a nonstratified sampling strategy and examined the impact of various factors on the performance of the designs. We considered the following factors: stage 1 sample size, MAFs of the tag SNP and the functional sequence variant, tag-seq LD correlation, and the noise-to-effect ratio (σ/β_1).

Using the 1000 Genomes Project Phase 1 dataset of 381 European subjects, we simulated the genotypes of 201 common variants (MAF $\geq$ 0.05) within a 100 kb genomic region surrounding the APOE gene, i.e., 45,400–45,500 kb on chromosome 19, for a total of 5,000 subjects. The GWAS sample size was selected in such a way that when the noise-to-effect value was 4.9 and seq MAF = 0.2, the GWAS tag SNP genotype correlated with the functional sequence genotype in the region would have identified association with the quantitative trait that reached GWAS significance using a frequentist approach (see Supplementary Table S1 for required sample sizes in other settings). We selected rs5117 to be the functional variant, not typed in the GWAS phase, and rs75627662 to be the tag SNP, which was typed in GWAS and drew attention to the region. The observed MAFs were 0.20 and 0.18 for rs5117 and rs75627662, respectively. These two variants were selected to be the tag SNP and functional variant such that the tag-seq LD correlation was $r^2 = 0.80$ ($r = 0.894$), based on the suggestion of Vukcevic et al. [2011] that the value of r would be high when the tag SNP was identified in a GWAS of the same sample, and larger than if the tag SNP had been identified in a previous independent study. In addition to the tag rs75627662, another two nonfunctional common variants in the same region were found to be in high LD with the functional variant. These two "hitchhiker" variants, rs483028 and rs438811, were in perfect LD with each other and had a correlation $r^2 = 0.94$ ($r = 0.97$) with the functional variant (Fig. 2) and a correlation of $r^2 = 0.87$ ($r = 0.93$) with the tag SNP.

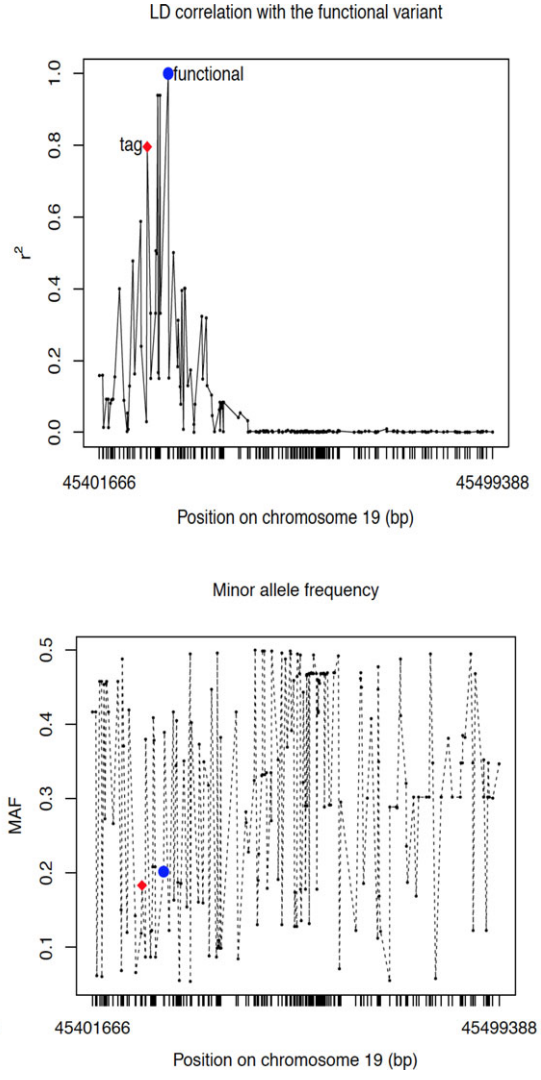

Figure 2. Common variants in the 100 kb region surrounding the APOE gene (Source: 1000 Genomes Project Phase 1). In simulation C1, rs75627662 (red diamond) and rs5117 (blue dot) were chosen to be the tag and functional variants, respectively.

Table 1. Summary of simulation study scenarios with a common variant set ($m = 201$) or with a low-frequency and common variant set ($m = 332$)

Simulation scenario	Functional variant MAF	GWAS tag SNP MAF	Tag-functional correlation (r)	Number of hitchhikers ($r > 0.8$)	*Size (% of m)*	*Size (m_2)*	*P* (select)	*P* (rank)
Common variant set ($m = 201$) with a single functional variant								
C1	0.200	0.180	0.894	2	35%	70	0.95	0.35
Low-frequency and common variant set ($m = 332$) with a single functional variant								
L1	0.022	0.142	0.372	6	65%	216	0.94	0.01
L2	0.046	0.055	0.880	9	58%	193	0.95	0.15
L3	0.087	0.099	0.936	5	58%	193	0.92	0.17
L4	0.142	0.120	0.905	3	45%	149	0.96	0.18

In scenarios C1 and L1–L4, each with a single functional variant, model parameters were specified by $\beta_0 = 5$, $\beta_1 = 0.25$, and $\sigma^2 = 0.1,\ 0.5,\ 1.5$. (See Supplementary Table S3 for the specific SNP variant "rs numbers" in the APOE gene). The last four columns compare performance of the fine-mapping method under selected values: $n_1 = 500$, $\sigma/\beta_1 = 4.9$, additive (ADD) genetic model, and equal strata sampling (ES). *Size* is the number m_2 of sequence SNPs in the 95% credible set, expressed as a percentage of m or as a count. *P* (select) is the *probability* that the functional variant is selected into the credible set. *P* (rank) is the *probability* that the functional variant is top-ranked in the credible set. (A complete set of plots across a range of values is provided as Supplementary information).

We designed additional simulations to investigate the impact of the functional variant MAF and the number of hitchhikers on the efficiency of each sampling design. Table 1 provides details on four additional simulation scenarios, L1–L4. These scenarios feature low-frequency and common variant sets ($m = 332$) and between three and nine hitchhikers. Supplementary Table S3 specifies the MAFs for the functional SNP and the tag SNP for each scenario. For instance, in simulation L1, the tag SNP is common (rs429358 with MAF 0.142) and the functional variant is low frequency (rs1081105 with MAF 0.022). A total of six hitchhikers have correlation with the functional variant greater than 0.8.

We generated quantitative trait data following (1) under additive, dominant, and recessive genetic models. For each genetic model, the regression coefficients were specified by $\beta_0 = 5$, and $\beta_1 = 0.25$, with three values of the normal residual variance $\sigma^2 = 0.1,\ 0.5,\ 1.5$, corresponding to a noise-to-effect ratio of 1.3, 2.8, and 4.9, respectively. This specification of noise-to-effect ratio values covers a range of sampling design efficiencies, and is equivalent to the specification of fixed σ^2 and varying β_1. To investigate how the stage 1 sample size influences the performance of the design, we varied the value of n_1 from 100 to 1000 with an increment of 100.

Three sample allocation schemes were considered: (1) SRS, (2) equal sample size for each tag genotype category (ES), and (3) tag homozygote increased relative frequency (HO). For the two tag-stratified sampling schemes, however, the rare homozygote tag stratum may contain fewer subjects than the allocated sample size when n_1 is large. In such cases, we first sampled all subjects in the rare homozygote stratum. Then, for the ES scheme we allocated the rest of the rare homozygote sample size equally to the common homozygote and heterozygote strata, and for the HO scheme, we allocated the rest of the rare homozygote sample size to the common homozygote only. An illustration of sample sizes allocated to each stratum under the three sampling schemes is provided by Supplementary Table S2 for a tag SNP MAF of 0.2.

Because diffuse priors are used when one has little prior knowledge of the genetic association, we specified equal genetic model prior probabilities, i.e., $p(M_j) = 1/3, j = 1, 2, 3$. For the regression parameters, we chose a relatively flat conjugate prior $NIG(b_0, B_0, \nu_0/2, \sigma_0^2\nu_0/2)$, where the hyperparameters were specified by $b_0 = (0, 0)^T$, $B_0 = diag(10^6, 10^6)$, $\sigma_0^2 = 0.5$, and $\nu_0 = 10$.

Based on 1000 simulations, we evaluated the performance of the three sample size allocation schemes according to the characteristics of the credible set obtained in stage 1. Although a complete two-stage study implementation would typically incorporate joint analysis of stage 1 and stage 2 data, we did not include stage 2 data collection and joint analysis as part of our simulation studies.

Simulation Study Results

Under the common variant scenario C1, when $\sigma/\beta_1 = 4.9$ and the stage 1 sample size was small, e.g., $n_1 = 100$, the percentage of variants selected into the credible set appeared to be similar (at around 70% of 201) across all three sampling schemes (Fig. 3A). The size of the credible set decreased as n_1 increased, particularly for ES and HO sampling schemes when the underlying genetic model was additive or recessive, with HO performing slightly better than ES. When $n_1 = 600$, for instance, the percentage size under an additive model was 18% and 24% for HO and ES, respectively, but was 55% for SRS. This suggests that informative tag-stratified sampling can dramatically reduce the number of credible set variants, which means reduced genotyping cost in stage 2. The three sampling schemes had similar performance for the dominant genetic model. As expected, the size of the credible set decreased as the noise-to-signal ratio decreased for each given stage 1 sample size and underlying genetic model (Supplementary Figs. S1.1 and S1.2). With a stage 1 sample of size $n_1 = 200$ and a σ/β_1 value of 1.3, for instance, ES and HO schemes resulted in credible sets as small as 1% of the total variants (2/201) in the region. Similar findings were obtained for the scenarios with the low-frequency variant set (< 5%) and various LD structures (Table 1), with the ES scheme nearly always performing better than SRS in reducing the size of the credible set, and often better than HO (Fig. 3B, Supplementary Figs. S2.2.1–3, S2.3.1–3, and S2.4.1–3).

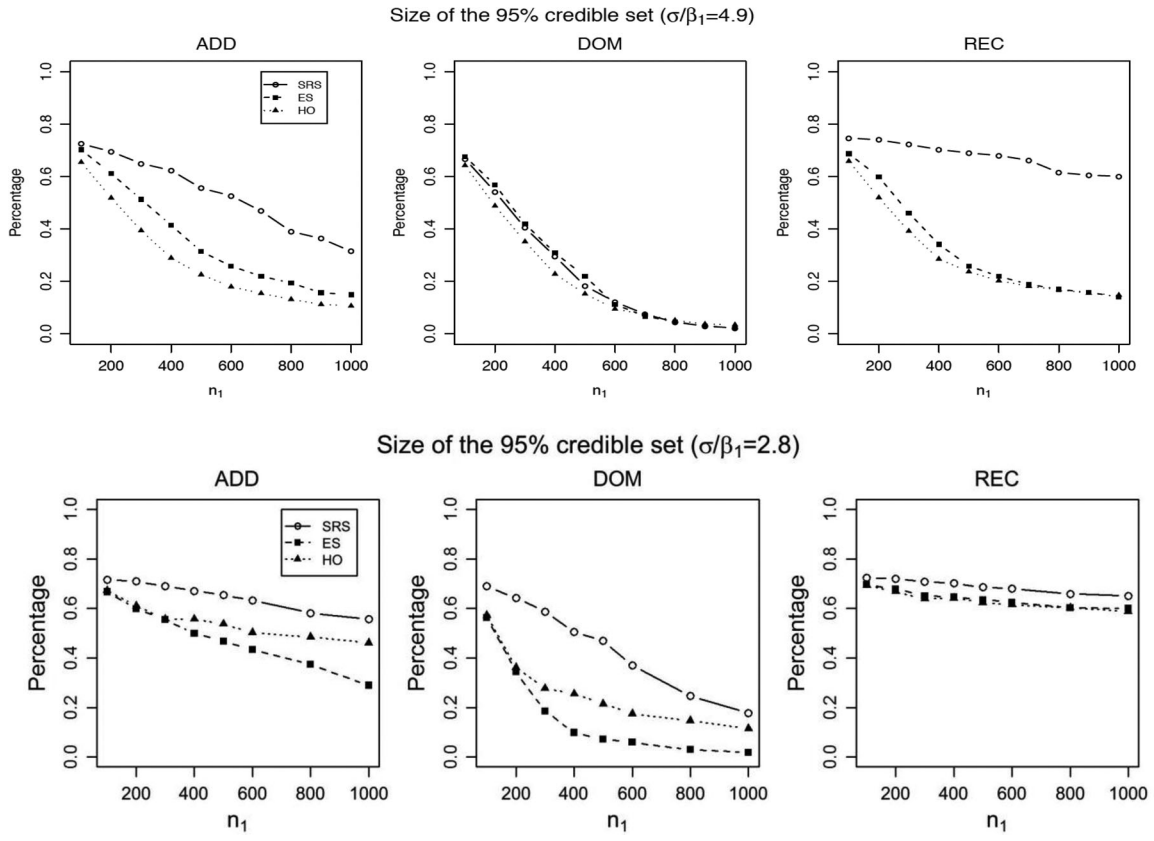

Figure 3. The size of the selected 95% credible variant set under three sampling schemes: simple random sampling (SRS), equal sample size for each tag genotype category (ES), and tag homozygote oversampling (HO). From left to right: Data were simulated under additive, dominant, and recessive genetic models, respectively. Results are based on 1000 simulations. (A) Upper panels are common variant scenario C1 (m = 201) with σ/β_1 = 4.9. (B) Lower panels are low-frequency variant scenario L1 (m = 332) with σ/β_1 = 2.8.

When the stage 1 sample size was small, e.g., n_1 = 100, the probability of selecting the functional variant into the credible set under the SRS scheme was higher than that under ES and HO (Fig. 4A, see also Supplementary Fig. S2.2.6). For modest values of n_1, SRS performed similarly to ES and HO under additive and dominant models but underperformed under a recessive model. When n_1 = 1,000, the three schemes had similar rates for successfully selecting the functional variant, all close to 95% (Fig. 4A and B), likely because the variants that had high correlation with the functional variant could be well separated from those that had low correlation. As the noise-to-signal ratio decreased, the rate was above 95% in general and close to 100% for a large stage 1 sample size (Supplementary Figs. S1.3 and S1.4).

For ranking the functional variant as the top variant, the ES and HO schemes performed better than SRS under the additive and recessive models for the common variant set (Fig. 5A). However, the probability of ranking rs5117 as the top variant was below 5% for all three schemes when n_1 = 100 and σ/β_1 = 4.9, regardless of the underlying genetic model. When n_1 increased to 1,000 (of a GWAS sample of 5,000), this probability increased to about 50% for ES and HO and 40% for SRS under the additive model, and increased to 60% for ES and HO and 20% for SRS under the recessive model. The low probability here can be explained by the existence of hitchhiker variants that compete with the functional variant [Faye et al., 2013]. The SRS and ES schemes performed slightly better than HO under the dominant model, with the probability of ranking rs5117 as the top being 75% for SRS and ES and 60% for HO, for n_1 = 100. As the noise-to-signal ratio decreased, the probability increased for all sampling schemes and all underlying genetic models (Supplementary Figs. S1.5 and S1.6). For the low-frequency variant set, the probability of the functional SNP being ranked first depends on the MAF of the functional variants, being very low for MAF = 0.022 under all three models (Fig. 5B). However, for higher MAF (scenarios L2–L4, Supplementary Figs. S2.2–2.4), these probabilities were generally moderate to high with ES sampling often performing best.

Finally, to illustrate the value of genetic model selection in the determination of the 95% credible set we conducted an additional simulation that compared analyses with and without genetic model selection. When one is not interested in selecting the genetic model (without model selection), γ is set to 0, $-\beta_1$ or $+\beta_1$ for additive, dominant, and recessive models, respectively (according to (1)). Otherwise, when the

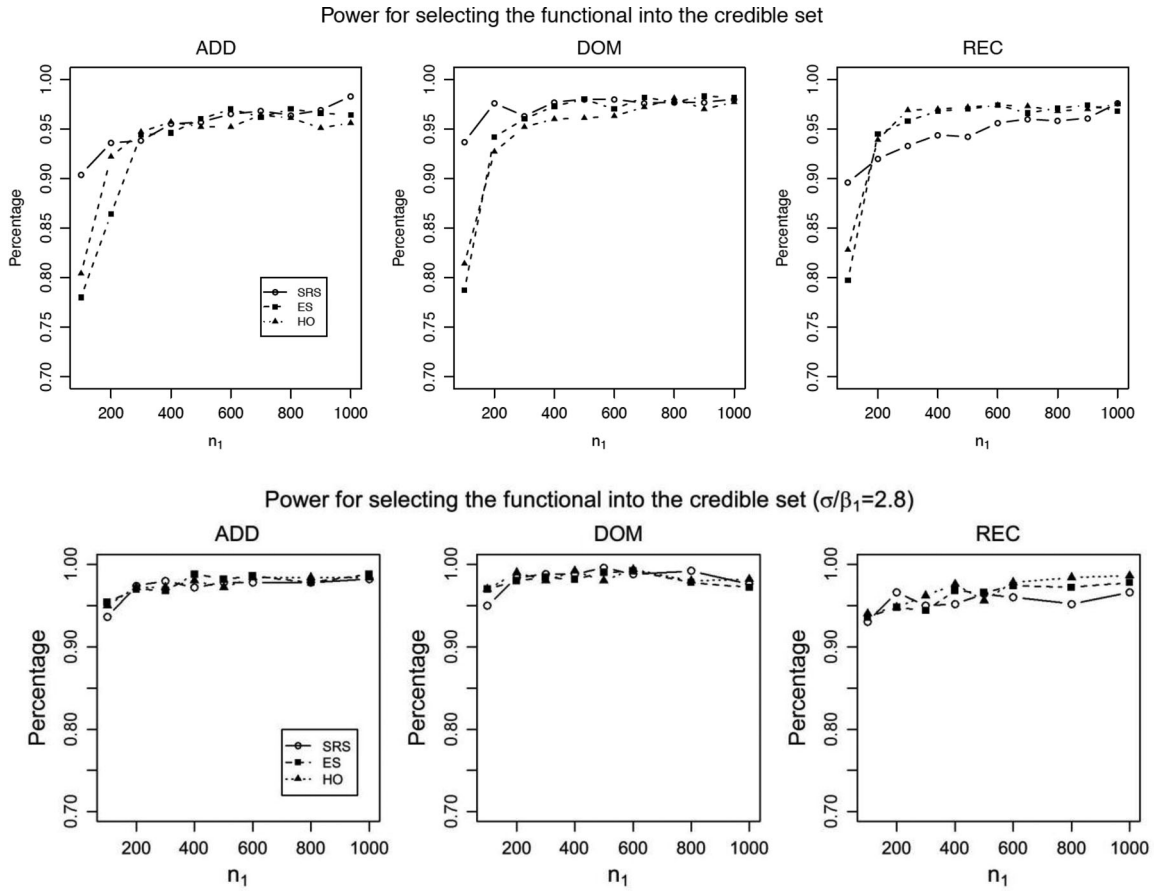

Figure 4. Empirical probability of selecting the functional variant into the 95% credible set under three sampling schemes: simple random sampling (SRS), equal sample size for each tag genotype category (ES), and tag homozygote oversampling (HO). From left to right: Data were simulated under additive, dominant, and recessive genetic models, respectively. Results are based on 1000 simulations. (A) Upper panels are common variant scenario C1 ($m = 201$) with $\sigma/\beta_1 = 4.9$. (B) Lower panels are low-frequency variant scenario L1 ($m = 332$) with $\sigma/\beta_1 = 2.8$.

genetic model is uncertain (with model selection), we compute parameter estimates under the model most favored by the data in terms of its posterior distribution. The parameters were specified by $\beta_0 = 5$ and $\sigma/\beta_1 = 4.9$, with a stage 1 sample size of $n_1 = 600$. For all three sample allocation schemes, including genetic model selection consistently reduced the size of the credible set compared to analysis without model selection, particularly for cases in which the additivity assumption was incorrect (Table 2). When the true underlying model was dominant and ES allocation was used, for example, the percentage size of the credible set was 22.1% assuming additivity but was only 11.2% with genetic model selection. The results in Table 2 and the additional results for the simulations with low-frequency variants (Supplementary Figs.) show that with model selection, ES often dominates HO in terms of the probability of including the functional variant in the credible set *and* in terms of the probability of ranking the functional variant as first. This may be explained largely by the better ability of ES to identify the underlying genetic model (Supplementary Figs. S1.7–1.9). The reduced size of the credible set means reduced cost for typing the selected variants in stage 2.

The two-stage approach proposed here reveals a subtle trade-off between the costs involved in each of the two stages of the analysis. Our simulations show that as we increase the stage 1 sample size we reduce the size of the credible set and thus reduce the costs involved in the second stage. In the next section we assess this trade-off from a cost-efficiency (CE) perspective.

CE Considerations

Assuming the sequencing cost is $\$c_1$ per individual, obtained by targeted sequencing for example, the stage 1 cost is $n_1 \times c_1$. In stage 2 of the two-stage design, custom genotyping is conducted for m_2 markers at a cost of $\$c_2$ per individual per marker, and the stage 2 cost is n_2 ($m_2 \times c_2$). The n_2 individuals are those not selected for sequencing in stage 1 ($n_2 = N - n_1$). (Although in some circumstances, it may be desirable to

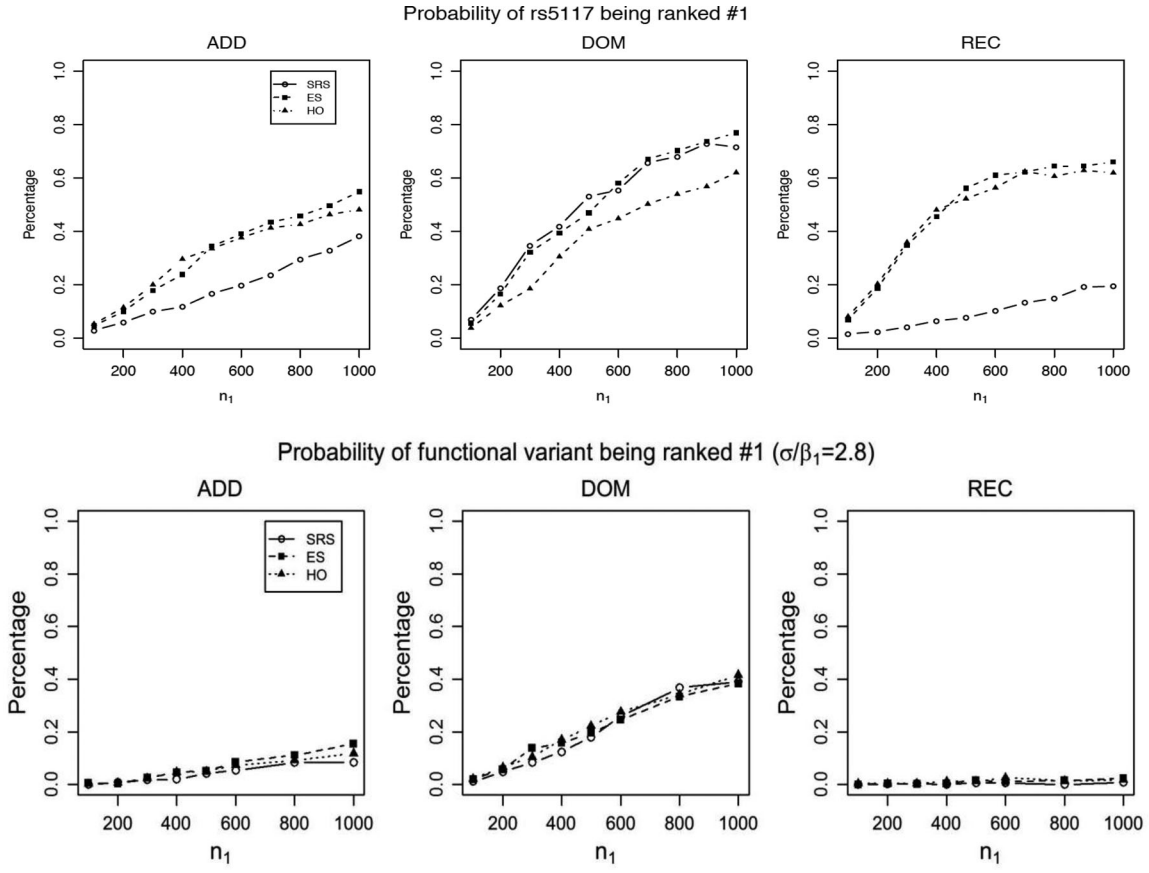

Figure 5. Empirical probability of the functional variant being ranked as the top variant in the region under three sampling schemes: simple random sampling (SRS), equal sample size for each tag genotype category (ES), and tag homozygote oversampling (HO). From left to right: Data were simulated under additive, dominant, and recessive genetic models, respectively. Results are based on 1000 simulations. (A) Upper panels are common variant scenario C1 ($m = 201$) with $\sigma/\beta_1 = 4.9$. (B) Lower panels are low-frequency variant scenario L1 ($m = 332$) with $\sigma/\beta_1 = 2.8$.

genotype everyone, $n_2 = N$, to confirm the sequencing or to create a reference panel.) For example, if $N = 5000$, $n_1 = 500$, $c_1 = \$1000$, $n_2 = 4500$, $m_2 = 100$, and $c_2 = \$0.50$, then the total two-stage cost is \$500,000 + \$225,000 = \$725,000 compared to a one-stage cost of \$5 million.

For fixed values of c_1 and c_2, the stage 1 cost increases linearly in n_1, whereas the stage 2 cost decreases monotonically in n_1, because the expected size of the credible set is nonincreasing in n_1 (as seen in Fig. 3). However, the total cost can be increasing or decreasing in n_1, depending on the values of (c_1/c_2) and the rate of decline in m_2. In some cases, the total cost can be minimized (see Supplementary Appendix B for details). As n_1 increases, the probability that a functional variant falls within the 95% credible set also increases (as seen in Fig. 4), which we loosely refer to as "power." We define CE to be Power divided by Cost, with higher values corresponding to higher power and/or lower cost. Other definitions, notably the ARCE = {(variance of lnRR)/cost} of Thomas et al. [2013] has the advantage that the numerator is not bounded above by 1.0. Nevertheless, it is desirable to require high power in stage 1, so that the functional variant is unlikely to be left out of stage 2. The trends in CE are then determined largely by trends in the total cost. In Figure 6, we calculate CE values based on empirical estimates of the proportion of variants retained in the credible set (m_2/m) and the probability that the functional variant is included therein, which we obtained from the simulations for SRS and ES sampling under the additive model (Figs. 3A and 4A). As might be expected, high cost ratios, $(c_1/c_2) > 2{,}000$, favor smaller n_1, and increasing stage 2 costs reduce CE (Fig. 6). Except for the smallest stage 1 sample size (where SRS has higher power), CE is consistently higher under the ES sample allocation, in accordance with more efficient reduction in the number of variants to be genotyped in stage 2. Under ES, CE is maximized at $n_1 =$ 500 and 600, for $c_2 = 1.0$ and 2.0, respectively, when $c_1 = 1000$ (Fig. 6A), and at $n_1 = 800$ when $c_1 = 100$. Our observation that ES (and often HO) allocations for low-frequency variants and other LD structures also reduce the proportion (m_2/m) more than SRS suggests that CE patterns will be similar in these settings.

Table 2. Comparison of sampling design performance with and without genetic MS under the common variant simulation scenario C1 with the three sample allocation schemes: SRS, equal sample size for each tag genotype category (ES), and tag homozygote oversampling (HO)

	ADD		DOM		REC	
Allocation scheme	No MS	MS	No MS	MS	No MS	MS
Size of the credible set (%)						
SRS	55.5%	52.6%	13.0%	12.1%	83.2%	68.0%
ES	29.8%	25.8%	22.1%	11.2%	38.8%	21.8%
HO	19.3%	17.9%	12.5%	9.4%	27.0%	20.2%
P (functional variant being selected into the credible set)						
SRS	0.984	0.965	0.972	0.980	0.952	0.956
ES	0.986	0.970	0.980	0.970	0.986	0.974
HO	0.964	0.952	0.958	0.963	0.964	0.974
P (functional variant being ranked #1)						
SRS	0.238	0.198	0.542	0.554	0.034	0.102
ES	0.396	0.390	0.396	0.580	0.380	0.610
HO	0.392	0.376	0.306	0.448	0.404	0.563

Data were generated under each of additive (ADD), dominant (DOM), and recessive (REC) models with $n_1 = 600$, $\sigma/\beta_1 = 4.9$ in a simulation independent of that reported in Figures 3 to 5. For cases without MS (no MS), the analysis assumed an additive genetic model.
MS, model selection; SRS, simple random sampling.

Discussion

Two-stage designs for quantitative traits or complex diseases that minimize cost and maximize power are of interest as strategies to narrow down the set of variants associated with the traits [Stanhope and Skol, 2012]. Using a Bayesian approach we considered a two-phase two-stage design for fine-mapping a region detected by a GWAS of a quantitative trait. The two-phase approach relies on high power in the phase I GWAS tag SNP to correctly identify a region for phase II fine-mapping that includes at least one functional SNP. Under the two-stage design, all variants in the region are first examined in a subset of subjects, using an expensive sequencing technology, and promising potentially functional variants are selected into a 95% credible set. In stage 2 the selected variants are then typed in the remaining subjects using a relatively cost-effective genotyping technology. In the absence of knowledge of the genetic model for a candidate functional variant, additive coding can lead to loss of efficiency when the additivity assumption is incorrect. In the stage 1 analysis, we employed genetic model selection for each candidate variant using the Bayes factor, and compared all variants within the region to select a credible set for further evaluation.

As opposed to random selection of individuals into the stage 1 sample, our simulations show that a more informative sample can be obtained by stratified sample allocation, such as sampling an equal number from each tag SNP stratum or a proportionately larger number from the homozygote tag SNP strata. A well-chosen sample size allocation scheme has the potential to improve functional variant localization and reduce the size of the 95% credible variant set for evaluation in a subsequent stage. Various factors, including stage 1 sample size and tag-seq correlation can influence the performance of a two-stage fine-mapping design. Our simulations confirm the intuition that the efficiency of the tag-stratified sampling strategy increases with tag-seq correlation. When the true genetic model is correctly identified, notably by Bayesian genetic model selection, an informative sample size allocation scheme can effectively reduce the posterior variance of the association parameter. Across the various scenarios evaluated in the simulation studies, the ES sampling strategy generally performed equivalently or better than the HO sampling strategy, particularly with respect to decreasing the size of the credible set.

The number of variants that will be included in a credible set is a random variable that depends mainly on the stage

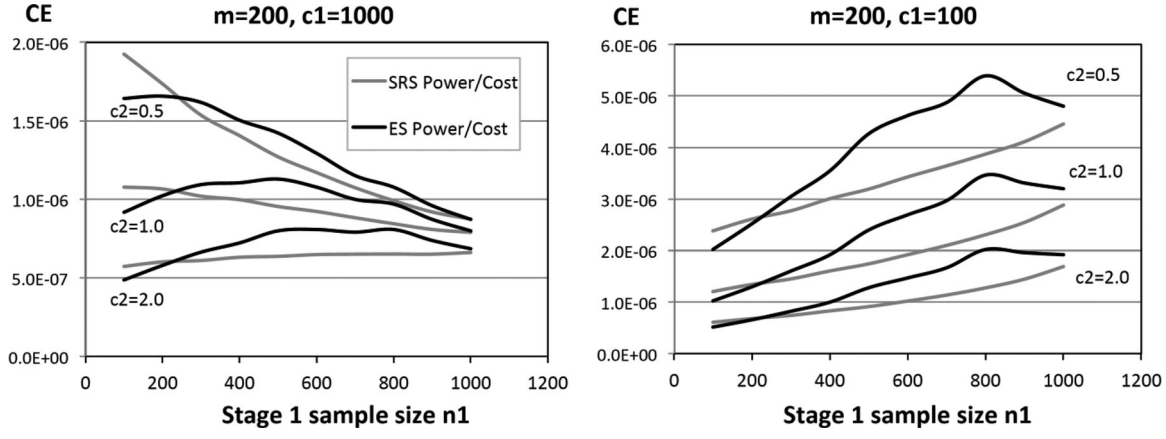

Figure 6. Trends in cost efficiency (CE) defined as Power/Cost with stage 1 sample size n_1 increasing from 100 to 1,000 for the SRS (gray lines) and ES (black lines) sample allocation schemes. The stage 2 sample size is $n_2 = N - n_1$ with $N = 5{,}000$. A total of m sequence variants are identified in stage 1, and a proportion $q = (m_2/m)$ are genotyped in stage 2. Empirical values for q and for power used to calculate CE as a function of n_1 were obtained from the common variant simulation study under an additive model. Cost depends on c_1, the stage 1 per individual sequencing cost, and on c_2, the stage 2 per individual per marker genotyping cost. For a fixed value of c_1, CE decreases as c_2 increases from 0.50 to 2.00.

1 sample size and the LD structure in the region, as well as the sampling design. This differs from the two-stage GWAS design considered by Skol et al. [2007], where the proportion of variants to be typed at stage 2 can be pre-specified. If there is more than one functional variant in the fine-mapping region, then the probability that a secondary functional variant is included in the credible set will depend on whether it is tagged by the GWAS SNP used for stratification, and on its effect size and MAF relative to the primary functional variant. We suggest that a fine-mapping region be determined such that there is a dominant tag SNP. With complete information available on the credible set of variants, one can jointly analyze stage 1 and 2 data and update the posterior probabilities of these variants being functional. Efficient stage 1 sampling and analysis for identifying a credible set of variants for stage 2, however, remain crucial in a two-stage fine-mapping design.

In our Bayesian analysis we focused on common and low-frequency variants for association with the trait. In practice, a large portion of the variants discovered in a region will be low-frequency (MAF between 0.01 and 0.05) and rare (MAF < 0.01). For sufficiently large samples, single variant analysis for credible set construction is feasible even for uncommon variants, but the incorporation of aggregated rare-variant statistics into a two-stage design is likely to be of interest, provided the pooled rare variants comprise a meaningful unit of analysis. In circumstances where these aggregation methods select a large number of rare variants into the credible set, the reduction of genotyping costs in stage 2 may be too modest compared to performing dense genotyping/sequencing, undermining the original goal of a two-stage fine-mapping design.

In addition to tag-SNP-based two-phase sampling designs, there are other approaches to design for sequencing studies. Trait-dependent sampling designs, for example, have been implemented for sequencing studies of quantitative traits, in which subjects in the two extremes of the trait distribution are selected with the hope that minor alleles at the functional variant are enriched in the selected sample [e.g., Lin et al., 2013; Yilmaz and Bull, 2011]. Sequencing data, obtained on the selected subjects, can be analyzed for association with the quantitative trait either using methods that correctly adjust for the sampling design, or simply using binary regression models by treating the selected subjects as cases and controls. Although trait-dependent sampling is more suitable for WGS studies in which multiple regions are examined, the work of Schaid et al. [2013] suggests it may be worthwhile to combine trait- and SNP-based sampling.

Acknowledgments

This research was supported by funding from the Canadian Institutes of Health Research: CIHR Operating Grant MOP-84287 (R.C., S.B.B.), CIHR Training Grant GET-101831 (Z.C.). Z.C. was a CIHR Fellow in Genetic Epidemiology and Statistical Genetics with CIHR STAGE (Strategic Training for Advanced Genetic Epidemiology) – CIHR Training Grant in Genetic Epidemiology and Statistical Genetics. We thank Laura Faye and Andrew Paterson for helpful discussions.

References

Almomani R, van der Heijden J, Ariyurek Y, Lai Y, Bakker E, van Galen M, Breuning MH, den Dunnen JT. 2011. Experiences with array-based sequence capture; toward clinical applications. *Eur J Hum Genet* 19:50–55.

Chen Z, Craiu RV, Bull SB. 2012. Two-phase stratified sampling designs for regional sequencing. *Genetic Epidemiol* 36:320–332.

Faye LL, Machiela MJ, Kraft P, Bull SB, Sun L. 2013. Re-ranking sequencing variants in the post-GWAS era for accurate causal variant identification. *PLoS Genet* 9(8):e1003609.

Hedges DJ, Guettouche T, Yang S, Bademci G, Diaz A, Andersen A, Hulme WF, Linker S, Mehta A, Edwards YJ and others. 2011. Comparison of three targeted enrichment strategies on the SOLiD sequencing platform. *PLoS One* 6:e18595.

Joo J, Kwak M, Chen Z, Zheng G. 2010. Efficiency robust statistics for genetic linkage and association studies under genetic model uncertainty. *Stat Med* 29:158–180.

Lin DY, Zeng D, Tang ZZ. 2013. Quantitative trait analysis in sequencing studies under trait-dependent sampling. *Proc Natl Acad Sci USA* 110:12247–12252.

Schaid DJ, Jenkins GD, Ingle JN, Weinshilboum RM. 2013. Two-phase designs to follow-up genome-wide association signals with DNA resequencing studies. *Genet Epidemiol* 37:229–238.

Skol AD, Scott LJ, Abecasis GR, Boehnke M. 2007. Optimal designs for two-stage genome-wide association studies. *Genet Epidemiol* 31:776–788.

Spencer C, Hechter E, Vukcevic D, Donnelly P. 2011. Quantifying the underestimation of relative risks from genome-wide association studies. *PLoS Genet* 7(3):e1001337.

Stanhope SA, Skol AD. 2012. Improved minimum cost and maximum power two stage genome-wide association study designs. *PLoS One* 7:e42367.

Stephens M, Balding D. 2009. Bayesian statistical methods for genetic association studies. *Nat Rev Genet* 10:681–690.

Strauch K, Fimmers R, Baur MP, Wienker TF. 2003. How to model a complex trait, 1. General considerations and suggestions. *Hum Hered* 55:202–210.

Thomas DC, Casey G, Conti DV, Haile RW, Lewinger JP, Stram DO. 2009. Methodological issues in multistage genome-wide association studies. *Stat Sci* 24: 414–429.

Thomas DC, Yang Z, Yang F. 2013. Two-phase and family-based designs for next-generation sequencing studies. *Front Genet* 4:276.

Vukcevic D, Hechter E, Spencer C, Donnelly P. 2011. Disease model distortion in association studies. *Genet Epidemiol* 35:278–290.

Wakefield J. 2009. Bayes factors for genome-wide association studies: comparison with P-values. *Genet Epidemiol* 33:79–86.

Wellcome Trust Case Control Consortium, Maller JB, McVean G, Byrnes J, Vukcevic D, Palin K, Su Z, Howson JM, Auton A, Myers S and others. 2012. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat Genet* 44:1294–1301.

Yilmaz YE, Bull SB. 2011. Are quantitative trait-dependent sampling designs cost-effective for analysis of rare and common variants? *BMC Proc* 5(Suppl 9):S111.