Two-Phase Stratified Sampling Designs for Regional Sequencing

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

[1], Samuel Lunenfeld Research Institute of Mount Sinai Hospital, Toronto ON, Canada
[2]Department of Statistics, University of Toronto, Toronto ON, Canada
[3]Dalla Lana School of Public Health, University of Toronto, Toronto ON, Canada

ABSTRACT: By systematic examination of common tag single-nucleotide polymorphisms (SNPs) across the genome, the genome-wide association study (GWAS) has proven to be a successful approach to identify genetic variants that are associated with complex diseases and traits. Although the per base pair cost of sequencing has dropped dramatically with the advent of the next-generation technologies, it may still only be feasible to obtain DNA sequence data for a portion of available study subjects due to financial constraints. Two-phase sampling designs have been used frequently in large-scale surveys and epidemiological studies where certain variables are too costly to be measured on all subjects. We consider two-phase stratified sampling designs for genetic association, in which tag SNPs for candidate genes or regions are genotyped on all subjects in phase 1, and a proportion of subjects are selected into phase 2 based on genotypes at one or more tag SNPs. Deep sequencing in the region is then applied to genotype phase 2 subjects at sequence SNPs. We investigate alternative sampling designs for selection of phase 2 subjects within strata defined by tag SNP genotypes and develop methods of inference for sequence SNP variant associations using data from both phases. In comparison to methods that use data from phase 2 alone, the combined analysis improves efficiency. *Genet. Epidemiol.* 00:1–13, 2012.

Key words: fine-mapping; genetic association studies; two-phase design; optimal allocation; quantitative trait

Supporting Information is available in the online issue at wileyonlinelibrary.com.
Contract grant sponsor: Canadian Institutes of Health Research (CIHR).
*Correspondence to: Shelley B. Bull, Samuel Lunenfeld Research Institute of Mount Sinai Hospital, 60 Murray Street, Box No. 18, Toronto, ON M5T 3L9, Canada. E-mail: bull@lunenfeld.ca
Received 29 April 2011; Revised 29 July 2011; Accepted 17 January 2012
Published online in Wiley Online Library (wileyonlinelibrary.com/journal/gepi)
DOI: 10.1002/gepi.21624

INTRODUCTION

The population-based genetic association study is now a well-established approach to identify genetic variants that are detrimental or protective for human disease. The genome-wide association study (GWAS) attempts to comprehensively survey common variants in the entire human genome based on up to a million typed genetic markers in each individual in the sample, with imputation of another 3 million single-nucleotide polymorphisms (SNPs) based on a reference panel such as HapMap3, done without regard to any phenotypic information . This form of imputation has the advantage that different phenotypes can be tested for association without the need to redo the imputation [Li et al., 2009]. For many GWAS, single-marker association analysis of typed and imputed SNPs is the first step to identify promising regions, and associations are confirmed by more powerful and focused analysis based on replication and fine-mapping studies [Zheng et al., 2007].

In focused studies following up reasonable GWAS hits, investigators may choose to comprehensively sequence a whole region of interest using next-generation sequencing (NGS) technology, or selectively sequence the region using customized technology to genotype additional SNPs, for example, SNPs that are not imputable or are known from dbSNP [Liu and Leal, 2010]. Although most GWAS studies can impute over 3 million SNPs, the directly typed or imputed SNPs detected are not necessarily the functional variants [Fridley et al., 2010; Ioannidis et al., 2009]. Imputation coverage or accuracy may be low in the region of interest [e.g., Pei et al., 2010], particularly when the trait is influenced by multiple low-frequency/rare variants in the region rather than solely by common variants. Investigators may also use sequence data to test for association between the trait and a common variant or a gene-based summary score that incorporates information on multiple rare variants in a region.

Thus, one purpose of regional sequencing may be to discover novel, potentially functional variants in a particular region that has been detected in genome-wide association analysis or chosen as a candidate region. Due to financial constraints, however, investigators may be able to afford sequencing only a portion of the available subjects. When a covariate, such as a sequence SNP (seq SNP), is difficult or costly to measure, a two-phase stratified sampling design can dramatically reduce the cost of data collection [Breslow and Wellner, 2007]. At phase 1, measurements of the phenotype and an easily measured auxiliary variable, such as a GWAS tag SNP, are obtained for all available subjects. At phase 2, measurements on the expensive target covariate (i.e., the seq SNP) are made for a subsample drawn randomly, without replacement, from strata defined by the auxiliary variable. Loss of efficiency due to incomplete observation will be modest when the target covariate is highly correlated with the auxiliary variable. By sequencing an informative portion of the available phase 1 subjects at phase 2, sequence variants associated with one or more phenotypes can be detected efficiently. Thereafter, selected variants together with the promising GWAS tag SNPs can be examined jointly in additional larger studies.

We investigate the use of a two-phase sampling design to obtain sequence data for genetic association analysis of a quantitative trait. At phase 1, we assume that all available subjects are fully phenotyped and genotyped at an associated tag SNP within a promising region of the genome. Strata are then formed according to the genotypes of the tag SNP, and a fraction of subjects from each stratum is randomly selected for sequencing in the region of interest at phase 2. The total phase 2 sample size can be predetermined based on study budget, for example, 10% or 50% of the phase 1 sample, but the fraction of subjects selected in each stratum can differ across strata. We propose a method for the joint analysis of data from both phases and investigate strategy for the allocation of the phase 2 sample size to each stratum. The method is particularly useful in situations in which the tag SNP is in high linkage disequilibrium (LD) with a common seq SNP or with a rare variant score constructed by aggregation of multiple low frequency/rare variants. When imputation accuracy is high within a region identified by tag SNP association with a quantitative trait, the two-phase stratified design strategy we propose for a tag SNP can be similarly applied using imputed SNP data to define the sampling strata.

In most GWAS and fine-mapping studies, it is difficult to distinguish two SNPs that are in strong LD on statistical grounds without incorporating biological or other additional information. Depending on the minor allele frequencies and the strength of the LD correlation, the sample size required to conduct such fine-scale mapping and successfully distinguish two SNPs is typically one to four times larger than that required to detect the initial association [Udler et al., 2010]. We see the goal of the two-phase strategy as (1) to select a set of highly correlated polymorphisms for further evaluation, and/or (2) to identify other associated variants in the region that can be analyzed subsequently for functional consequences [Ioannidis et al., 2009]. The joint analysis method we propose here aims to efficiently detect potential association signals at SNPs that are not typed in phase 1, rather than to distinguish a causal SNP, genotyped by sequencing, from the tag SNP.

In the following sections, we develop an approach for joint analysis of phases 1 and 2 and compare it to methods of inference limited to sequenced SNP data available only in phase 2. For ease of exposition, we assume an additive model for the genetic association analysis, one that is used widely to capture the average change of the quantitative trait with each additional copy of the minor allele of an associated SNP, or with a unit increase in a rare variant summary score. In simulation studies, we quantify the relative design efficiencies across a range of possible sample allocations, considering both joint analysis and phase 2 only methods for analysis of a common variant or a rare variant score, and assess robustness to misspecification of the model used for analysis of the phase 2 data. We close with discussion of implications for studies involving multiple seq SNPs, multiple tag SNPs, or multiple traits.

METHODS

TWO-PHASE STRATIFIED SAMPLING

The basic idea of the two-phase design is to use auxiliary information available on all subjects to draw a subsample for additional, more expensive, measurements of a target variable. In genetic association analysis, the auxiliary information typically available consists of genotype data for a tag SNP, or an imputed SNP, within a candidate region. The SNP genotypes are available in all individuals in phase 1 of the study. The target covariate refers to a potentially functional seq SNP that is collected in the phase 2 subjects only.

Suppose we have N subjects that constitute a population sample, indexed by $i = 1, \ldots, N$. Let Y_i denote the quantitative trait for the ith subject, and denote the major and minor alleles of the seq SNP by D and d, respectively. Let P_d be the minor allele frequency (MAF) at the seq SNP in the population. If Hardy-Weinberg Equilibrium (HWE) holds, the population frequencies of genotypes DD, Dd, and dd are given by $(1 - P_d)^2$, $2(1 - P_d)P_d$, and P_d^2, respectively. A linear regression model under an additive genetic effect is given by

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \tag{1}$$

where X_i, the number of copies of allele d, would be potentially available on each subject if all subjects were completely observed. Let $\beta = (\beta_0, \beta_1)^{\mathrm{T}}$ be a vector of the regression parameters. The error term ϵ_i is commonly assumed to be normally distributed with mean 0 and variance σ^2.

For simplicity, we consider stratification that is based on a single tag or imputed SNP available in phase 1. Denote the major and minor alleles of the tag SNP by A and a, respectively. Correspondingly, let P_a be the MAF at this tag SNP in the population, and Z_i be the number of copies of allele a at the tag SNP observed on subject i. Again, if HWE holds for the tag SNP, the genotype frequencies $\mathrm{pr}(Z_i = j)$ are given by $(1 - P_a)^2$, $2(1 - P_a)P_a$, and P_a^2 for $j = 0, 1, 2$, respectively.

The phase 1 sample is divided into three strata according to the observed value of Z_i. Let N_j denote the number of subjects observed in stratum AA, Aa, and aa for $j = 0, 1, 2$, respectively. Under the assumption that the phase 1 sample represents the source population, $E(N_j/N)$ equals the corresponding population genotype frequency. Let ξ_i be a binary indicator such that $\xi_i = 1$ if subject i is sampled at phase 2 and $\xi_i = 0$ otherwise. Then $\pi_i = \mathrm{pr}(\xi_i = 1|Z_i)$ is the probability of such sampling.

Breslow and Wellner [2007] describe two probability models for the indicators ξ_i in two-phase stratified sampling. In the first one, known as Bernoulli sampling, ξ_i for each phase 1 subject is independently generated with probability $\pi_i = \pi_0(Z_i)$, where π_0 is a known sampling function. Under a missing at random mechanism, π_0 does not depend on the unobserved values of missing data from phase 1. This sampling scheme results in random phase 2 stratum-specific samples of size $n_j \leq N_j$ $(j = 0, 1, 2)$. For $i \neq i'$, let $\pi_{i,i'} = \mathrm{pr}(\xi_i = \xi_{i'} = 1)$ be the joint inclusion probability. Under Bernoulli sampling, $\pi_{i,i'} = \pi_i \pi_{i'}$. In contrast, under a second sampling model, known as finite population stratified sampling, the phase 2 sample size in each stratum is fixed. To be specific, at the second phase of sampling, n_j subjects are sampled at random without replacement from stratum j, with sampling for different strata conducted independently. This sampling method is of particularly interest to survey statisticians who aim to derive variances of estimates of population quantities, such as population means, totals, or quantiles, and leads to finite population joint inclusion probabilities then involved in the variance formula. In our development here we consider only Bernoulli sampling for phase 2, which is relatively simpler for the problem we are investigating.

For each subject selected into the phase 2 sample according to a promising tag SNP, a region containing the tag SNP is sequenced to identify additional, potentially functional seq SNPs. In the remainder, we assume that a functional SNP is indeed in the region containing the tag SNP and is sequenced for all phase 2 subjects. We quantify the association between the tag SNP and the seq SNP by the conditional probabilities $\alpha_{jk} = \text{pr}(X_i = k|Z_i = j)$, $j, k = 0, 1, 2$. Since $\alpha_{j0} = 1 - \alpha_{j1} - \alpha_{j2}$, $j = 0, 1, 2$, we let $\alpha = \{(\alpha_{j1}, \alpha_{j2}), j = 0, 1, 2\}^{\text{T}}$ be a vector of LD-related parameters that are to be estimated. The joint distribution of the two SNP genotypes given by $\text{pr}(X_i = k, Z_i = j) = \text{pr}(X_i = k|Z_i = j)\text{pr}(Z_i = j) = \alpha_{jk}\text{pr}(Z_i = j)$ with the correlation between X_i and Z_i defined as Pearson's correlation coefficient. The phase 1 data consist of (Y_i, Z_i), $i = 1, 2, \ldots, N$, and the phase 2 data consist of (Y_i, Z_i, X_i) for i included in the phase 2 sample.

ALLOCATION OF PHASE 2 SAMPLE SIZE AND NAIVE ANALYSIS

In this section, we discuss allocation of the phase 2 sample size to each of the phase 1 strata, in the context of fitting a standard additive linear regression model (A1) to the phase 2 data.

Let m_k be the number of phase 2 subjects carrying k copies of allele d at the seq SNP, $k = 0, 1, 2$, with a total phase 2 sample size of n. Let $\hat{\beta}_{1,\text{nai}}$ be the naive estimator for β_1 obtained from fitting model (A1) to the phase 2 data. Given variance $\text{var}(\hat{\beta}_{1,\text{nai}}) = \sigma^2\{\sum_{i\in s_2}(X_i - \bar{X})X_i\}^{-1}$, and $\bar{X} = \sum_{i\in s_2} X_i/n = (m_1 + 2m_2)/n$, it follows that

$$\begin{aligned}\text{var}(\hat{\beta}_{1,\text{nai}}) &= n\sigma^2\{m_1(n - m_1 - 2m_2)\\ &\quad +2m_2(2n - m_1 - 2m_2)\}^{-1}\\ &= n\sigma^2\left\{n(n - m_1) - (n - m_1 - 2m_2)^2\right\}^{-1},\end{aligned}$$

and minimum variance would be achieved when $m_1 = 0$ and $m_0 = m_2 = n/2$. In the context of genetic association studies, $\hat{\beta}_{1,\text{nai}}$ is most efficient when half of the phase 2 subjects have zero copies of d and the other half have two copies of d, provided that the true genetic model is additive. Since at phase 1 we observe information only on tag SNPs, a natural choice is to select subjects from the three strata defined by the tag SNP genotypes such that $E(m_1)$ is minimized while $E(m_0)$ and $E(m_2)$ are approximately equal. Because the tag SNP genotype is serving as a surrogate for the unobserved target SNP and discordance between them will reduce the chance of selecting informative subjects, the underlying correlation between the tag and the seq SNPs is important for the success of the phase 2 sample size allocation strategy.

Under Bernoulli sampling, all subjects within a stratum are sampled independently with the same probability. Let ρ_j $(j = 0, 1, 2)$ be the inclusion probability for the subjects in stratum j, that is, the stratum-specific sampling fraction. The individual sampling probability is therefore $\pi_i = \rho_j$ for all subjects in stratum j. The expected number of phase 2 observations is $E(n_j) = \rho_j N_j$, and the overall sampling fraction $\rho = \sum_{j=0}^{2} \rho_j N_j/N$ may be predetermined according to financial constraints. Therefore, the expected number of subjects carrying k copies of allele d in the phase 2 sample is $E(m_k \mid n_0, n_1, n_2) = \sum_{j=0}^{2} n_j\alpha_{jk}$, $k = 0, 1, 2$.

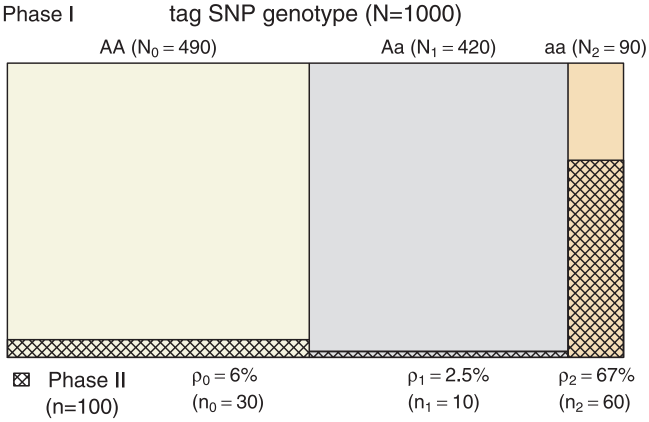

Fig. 1. An illustration of a two-phase stratified sampling ($P_a = 0.3$ under HWE, sampling fraction $\rho = 10\%$).

TABLE I. An example of realized genotype counts at the seq SNP under the sampling scheme shown in Figure 1 ($P_a = 0.3$, $P_d = 0.2$, $r = 0.75$, and sampling fraction $\rho = 10\%$. The A/a alleles are for the tag SNP and D/d are for the seq SNP)

	DD	*Dd*	*dd*	Total
AA	29	1	0	30
Aa	2	8	0	10
aa	9	23	28	60
Total	40	32	28	100

An illustration of phase 2 sampling is given in Figure 1, where the MAF at the tag SNP is 0.3, and the overall sampling fraction is 10%. Table I is an example of one realization of phase 2 sampling, where the MAF at the seq SNP is 0.2, and the correlation between the tag and seq SNPs is 0.75. The joint distribution of the SNPs can be estimated from these counts. One can see that if the two SNPs are highly correlated, the rare homozygote dd at the seq SNP will appear most frequently in the stratum with the rare homozygote aa at the tag SNP.

PHASE 2 INVERSE PROBABILITY WEIGHTED (IPW) ANALYSIS

The unweighted naive analysis just described does not take account of the sampling probabilities. In this section, we review a weighted estimation method that incorporates the stratum-specific sampling fractions. A typical method of estimation, as used by survey statisticians, is to maximize the IPW sum of log-likelihood contributions from the phase 2 observations, or equivalently, to solve an IPW version of the score equations [Manski and Lerman, 1977]. IPW is a standard approach to inference about finite population parameters (i.e., those of the entire phase 1 data), when the probability of being sampled (i.e., being included in phase 2) varies across individuals. It is usually applied to ensure that inferences are representative of the complete data, and it can limit the effects of model misspecification [Godambe and Thompson, 1986], for example, if an additive model is assumed incorrectly. We begin with the case of complete data: (Y_i, Z_i, X_i) for all $i = 1, 2, \ldots, N$, and then consider estimation with phase 2 data alone.

The likelihood contribution of subject i to estimation of (β, σ^2) is $L_i(\beta, \sigma^2) = f(Y_i \mid X_i)$, where $f(Y_i \mid X_i) = 1/\sqrt{2\pi\sigma^2}\exp[-\{Y_i - (\beta_0 + \beta_1 X_i)\}^2/(2\sigma^2)]$ is the probability density function of Y_i. If seq SNP genotypes were observed on all subjects, the likelihood of the data would be

$$L(\beta, \sigma^2) = \prod_{i=1}^{N} L_i(\beta, \sigma^2).$$

The log-likelihood contribution of subject i is $\ell_i(\beta, \sigma^2) = \log L_i(\beta, \sigma^2)$, and maximum likelihood estimation of β is equivalent to solving

$$0 = \sum_{i=1}^{N} U_i(\beta), \tag{2}$$

where $U_i(\beta) = \partial\ell_i(\beta, \sigma^2)/\partial\beta$. An estimate of the variance parameter σ^2 can be obtained from the residuals via the method of moments. In addition, estimates of the conditional probabilities in the vector α can be obtained by solving

$$0 = \sum_{i=1}^{n} Q_i(\alpha), \tag{3}$$

where $Q_i(\alpha) = ([\mathrm{I}(Z_i = j)\{\mathrm{I}(X_i = k) - \alpha_{jk}\}, k = 1, 2], j = 0, 1, 2)^{\mathrm{T}}$ and $\mathbf{I}(\cdot)$ is an indicator function.

In phase 2 data, however, Equations (2) and (3) cannot be used directly, due to missing seq SNP data for subjects with $\xi_i = 0$. Let s_2 and $\bar{s}_2$ denote the phase 2 sample and its complement, respectively. One commonly employed IPW estimation method for β, using data from phase 2 alone, applies weighted estimating equations [Skinner et al., 1989, section 3.4] which are given by

$$0 = \sum_{i \in s_2} (1/\pi_i) U_i(\beta) = \sum_{i=1}^{N} (\xi_i/\pi_i) U_i(\beta), \tag{4}$$

with $1/\pi_i$ called the sampling design weight for subject i. Under the two-phase stratified design, we have

$$E_p\left\{\sum_{i \in s_2} (1/\pi_i) U_i(\beta)\right\} = \sum_{i=1}^{N} U_i(\beta),$$

where E_p denotes expectation under the sampling scheme. Because the constructed estimating functions are unbiased for the phase 1 complete data estimating functions under expectation with respect to the sampling design, the estimator obtained by solving (4) is called a design-consistent estimator for the phase 1 sample parameter, that is, the solution to (2). Design consistency is a desirable property in randomization approaches to finite population sampling [Godambe and Thompson, 1986].

Both naive and IPW analyses of phase 2 data ignore the phenotype and tag SNP genotype data available in the phase 1 participants not included in phase 2, which leads to efficiency loss. Although the IPW estimator has attractive properties such as consistency and asymptotic normality, when the additive model (1) is correctly specified the naive estimates are not biased due to sampling, and incorporating IPW sampling weights can induce greater variation when some of the weights are large. As is evident in the simulation studies we report, the result is that the IPW estimator can be less precise than the naive estimator that ignores the sampling design.

JOINT ANALYSIS OF PHASE 1 AND PHASE 2

Although analysis of phase 2 data alone can give an estimate of β_1 at a seq SNP that is similar to the estimate that would be obtained if seq SNP data were available for all phase 1 subjects, the naive and IPW approaches are generally not powerful. In this section, we describe an alternative estimating equations method that can achieve greater power by jointly analyzing data from both phases. This approach constructs mean score functions for subjects that are not selected into phase 2. We show that the mean score function is a weighted sum of three score functions, each of which corresponds to one of the three possible seq SNP genotypes. The weight measures the likelihood of the missing seq SNP genotype given the observed trait and the tag SNP genotype.

For subject $i \in \bar{s}_2$, that is, not selected into phase 2 sample, let

$$\phi_{ik}(\alpha, \beta) = \frac{f(Y_i | X_i = k; \beta)\,\mathrm{pr}(X_i = k | Z_i; \alpha)}{\sum_{k'=0}^{2} f(Y_i | X_i = k'; \beta)\,\mathrm{pr}(X_i = k' | Z_i; \alpha)},$$

$k = 0, 1, 2$. Under the assumption that the distribution of the phenotype in the population is a mixture of three normal distributions with constant variance, $\phi_{ik}(\alpha, \beta)$ can be viewed as a weight for the possible seq SNP genotype with k copies of the minor allele. We construct estimating functions $U_i^*(\alpha, \beta) = \sum_{k=0}^{2} \phi_{ik}(\alpha, \beta) U_i(\beta; Y_i, X_i = k)$, and let $\tilde{U}_i(\alpha, \beta) = \xi_i U_i(\beta) + (1 - \xi_i) U_i^*(\alpha, \beta)$. Therefore, in parallel to Equations (3) and (4) above, the proposed estimating equations for β are given by

$$0 = \sum_{i=1}^{N} \tilde{U}_i(\alpha, \beta), \tag{5}$$

and we construct weighted estimating equations for α given by

$$0 = \sum_{i \in s_2} (1/\pi_i) Q_i(\alpha) = \sum_{i=1}^{N} (\xi_i/\pi_i) Q_i(\alpha), \tag{6}$$

with solution $\hat{\alpha}$. Under the two-phase stratified design, we have

$$E_p\left\{\sum_{i\in s_2}(1/\pi_i)Q_i(\alpha)\right\} = \sum_{i=1}^{N} Q_i(\alpha).$$

One can obtain a consistent estimator for β_1 by simultaneously solving (5) and (6) using a two-stage estimation procedure that is equivalent to an iterative Fisher scoring algorithm. Specifically, in the first step, we solve (6) for α. Let $\tilde{M}(\alpha, \beta) = \sum_{i=1}^{N} \tilde{M}_i(\alpha, \beta)$, where $\tilde{M}_i(\alpha, \beta) = \partial\tilde{U}_i(\alpha, \beta)/\partial\beta^{\mathrm{T}}$. In the second step, we replace α with $\hat{\alpha}$ and solve (5) for β via the Fisher scoring algorithm

$$\beta^{(t+1)} = \beta^{(t)} - \tilde{M}^{-1}(\hat{\alpha}, \beta^{(t)}) \sum_{i=1}^{N} \tilde{U}_i(\beta^{(t)}),$$

$t = 0, 1, \ldots,$ until convergence. Let $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)^{\mathrm{T}}$ denote the resulting limit.

Let $\theta = (\alpha^{\mathrm{T}}, \beta^{\mathrm{T}})^{\mathrm{T}}$, with estimate $\hat{\theta} = (\hat{\alpha}^{\mathrm{T}}, \hat{\beta}^{\mathrm{T}})^{\mathrm{T}}$. Large sample theory yields asymptotic properties for $\hat{\theta}$. In the Appendix, we outline the proof that $N^{1/2}(\hat{\theta} - \theta)$ is asymptotically normal with mean 0 and asymptotic covariance matrix given by $\Gamma^{-1}\Sigma(\Gamma^{-1})^{\mathrm{T}}$, where $\Psi_i(\theta) = \{(\xi_i/\pi_i)Q_i^{\mathrm{T}}(\alpha), \tilde{U}_i^{\mathrm{T}}(\alpha, \beta)\}^{\mathrm{T}}$, $i = 1, \ldots, N$, $\Gamma = E\{\Psi_i(\theta)/\partial\theta^{\mathrm{T}}\}$, and $\Sigma = E\{\Psi_i(\theta)\Psi_i^{\mathrm{T}}(\theta)\}$. As $N \to \infty$, Γ and Σ can be consistently estimated by their empirical counterparts. In the Appendix, we give the components of the approximate covariance matrix.

In concluding this section, we note the influence of sample size allocation on the efficiency of the proposed method. We wish to optimize the power to detect the association of the seq SNP with the quantitative trait by choosing a design that minimizes the variance of $\hat{\beta}_1$ obtained using both phase 1 and phase 2 data. Because the variance of $\hat{\beta}_1$ depends on the estimated LD-related conditional probabilities, α, reducing the uncertainly in α translates into reducing variability in $\hat{\beta}_1$. Therefore, allocations that improve precision of the conditional probabilities also improve precision of the genetic association estimate, although a generally valid approach is yet to be found. There are several reasons why we do not consider a design that samples only from the two tag SNP homozygote categories. First, robustness to departures from the underlying genetic model can depend on having observations from the heterozygote category. Second, the joint analysis method utilizes the correlation between the tag and the seq SNPs. If the tag SNP heterozygote stratum is not sampled at all, the phase 1 subjects in this category (usually a large proportion) will not be used in the joint analysis, which will greatly decrease the design efficiency. Third, if the MAF is low at the tag SNP but is relatively high at the seq SNP (e.g., $P_a = 0.1$, $P_d = 0.3$, positive correlation), then lack of sampling from the tag SNP heterozygote *Aa* stratum may decrease the number of phase 2 subjects carrying rare homozygote *dd* at the seq SNP, even when all subjects in the tag SNP rare homozygote *aa* stratum are sampled.

Although formulated for common seq SNPs, the proposed method can be generalized to analysis of rare variants, that is, variants with MAF < 1%. In this case, the objective of sequencing is to investigate the potential association between the phenotype and multiple rare variants within a gene or a specific genomic region. The target variable X_i in the linear model (1) now becomes the genetic score of rare variants [e.g., Morris and Zeggini, 2010]. For example, for R_{il} defined as the number of copies of minor allele at the lth rare variant, $l = 1, \ldots, K$, where K is the total number of rare variants in that region, we have $X_i = \sum_{l=1}^{K} R_{il}$. Although X_i is a count variable from 0 to $2K$, it takes far fewer values in practice due to the low MAFs of the rare variants.

TABLE II. Table of parameters for simulation design ($N = 1,000$)

Minor allele freq. (MAF)	Scenario (i): $P_a = P_d = 0.4$ Scenario (ii): $P_a = 0.3$, $P_d = 0.2$

tag SNP-seq SNP correlation	Scenario (i):
	$r = 0.95, 0.50, 0.05, -0.50$
	Scenario (ii):
	$r = 0.75, 0.25, 0.05, -0.30$

Overall sampling fraction	$\rho = 0.10, 0.25, 0.50$ ($n = 100, 250, 500$ in phase 2)
seq SNP genetic model	Additive, dominant, recessive
Genotype-specific variance	$\sigma_0^2 = \sigma_1^2 = \sigma_2^2 = 0.50$ (for all three genetic models); $\sigma_0^2 = 0.50, \sigma_1^2 = 0.75, \sigma_2^2 = 1.00$ (for additive model only)

SIMULATION STUDIES

SIMULATION DESIGN

We conducted simulation studies to investigate the relative efficiency of the proposed joint analysis method. Here, relative efficiency of the joint analysis estimator is defined as the ratio between the empirical variance of the estimator obtained when complete phase 1 data were available and the empirical variance of the joint analysis estimator. By calculating this quantity, we can investigate whether there is benefit from jointly analyzing both phases 1 and 2 data. In our simulation comparisons, we also included a naive method, which fits a standard linear model to phase 2 data ignoring the sampling design, and the IPW approach, which fits a linear model to phase 2 data weighted by the inverse of the inclusion probability.

Throughout our simulation studies, the sample size of phase 1 was fixed at $N = 1,000$. Table II summarizes the simulation design. For each combination of parameters and a specific sampling scheme, we generated 1,000 data sets. We used analysis of complete data for phase 1 as the ideal. We considered various scenarios for the genotypes of the tag SNP and the seq SNP, varying the MAFs. Two sets of MAF values were specified as (i) $P_a = P_d = 0.4$, and (ii) $P_a = 0.3$, $P_d = 0.2$. We quantified LD between the two SNPs by the correlation coefficient r, and considered a range of correlations from highly positive to moderately negative. For scenario (i) where MAF values are equal, $r = 0.95$, 0.5, 0.05, and -0.50. For scenario (ii) where MAF values differ, complete correlation between the two SNPs was not possible, and r was bounded by some value that is smaller than 1. Therefore, for scenario (ii) we set $r = 0.75$, 0.25, 0.05, and -0.30, where the highest correlation 0.75 was very close to the upper bound of possible correlation. The joint distribution of the tag SNP and the seq SNP can be inferred from their marginals and their correlation. For a given combination of MAF values and correlation, we first simulated the two haplotypes for each subject to achieve the desired frequencies of the genotypes and correlation. We then simulated a quantitative trait under an additive model given by (A1), where parameters were specified as $\beta_0 = 0.5$, $\beta_1 = 0.25$, and $\sigma^2 = 0.5$.

We considered three different values of the phase 2 sampling proportion ρ (Table II), corresponding to samples of size 100, 250, and 500. For each ρ, we investigated the influence of sample size allocation on the efficiencies of the various estimators for β_1. The sampling fractions ρ_1 and ρ_2 for the heterozygote and the rare homozygote strata were specified to reflect the extent to which the minor alleles are over sampled. The sampling proportion ρ_0 in stratum AA can be calculated from P_a, ρ_1, ρ_2, and ρ. We considered a range of phase 2 sample sizes allocated to the heterozygote stratum. For each given heterozygote count, we then varied the size allocated to the rare homozygote stratum. To avoid variance inflation in the estimated α due to sparse counts, we required a minimum of 10 for the expected number of subjects to be selected from each stratum. Therefore, there is not much freedom to allocate sample size to the two homozygote strata when most of the sample size is allocated to the heterozygote stratum. We also included two special allocations. The first one is equivalent to simple random sampling within each stratum with the same sampling fraction, and the second one allocates an equal sample size to each stratum, that is, $\rho_j N_j = \rho N/3$, $j = 0, 1, 2$.

To investigate the robustness of the methods to model misspecification, we also simulated data under various forms of departures from the additive model with constant variance specified by model (1), but then analyzed the data assuming a dosage effect and constant variance. We first considered dominant and recessive genetic effect models, both of which followed a similar form of model (A1) for the seq SNP. In the third case, we considered a true model with heteroscedastic variances across the seq SNP genotype categories. Heteroscedasticity may be encountered in situations where the phenotype is more variable in subjects with two copies of the rare allele at the seq SNP than in patients with one or zero copies. We set σ_0^2 =0.5, σ_1^2 =0.75, and σ_2^2 =1 for $X = 0$, 1, and 2, respectively. In the fourth case, we simulated genotypes at an additional seq SNP with low MAF but strong association with the phenotype, and moderate LD with the seq SNP being tested. We set the MAF at this additional SNP to be 0.01 and the regression parameter to be 0.5. This SNP was ignored in the analysis, however, yielding model misspecification. In the last case, we generated observations from an additive model in which the residuals follow a skewed distribution, a feature not uncommon in practice, but analyzed the data as if they were normally distributed. Because estimators for β_1 obtained under model misspecification are generally biased, we used mean squared error (MSE), as opposed to empirical variance, in the calculation of relative efficiency for all cases with model misspecification. That is, we calculated the ratio between two MSEs, with one from fitting model (A1) to complete phase 1 data (as if seq SNP data were available for all) and the other from applying the estimation method under the two phase design with seq SNP data only for the phase 2 sample. When the model used for analysis is misspecified, MSE is a better measure in the comparison of efficiency, since it incorporates both the bias and variability of an estimator.

In addition to studies of common seq SNPs, we conducted simulations for scenarios involving rare variants. We assumed that a rare variant score had been obtained by counting the number of rare variants across 20 loci, each of which had minor allele frequency (MAF) generated from a uniform distribution between 0.005 and 0.01, yielding a score of 0, 1, 2, or rarely 3. We then randomly selected three rare variants to be associated with the phenotype. All three causal rare variant effects are additive on the quantitative trait, with each copy of the minor allele increasing the mean trait value by 0.5, 0.75, and 1, respectively. Because most of the loci were noncausal, the overall association of the score with the trait was weak. The MAF for the tag SNP is specified as 0.3, and the correlation r with the rare variant score was around 0.5.

The primary focus of the evaluations was design efficiency based on the precision of parameter estimation. Design efficiency translates directly into power for hypothesis testing, provided the variance estimate used to construct the corresponding test statistic is accurate, and the test statistic is valid under the null hypothesis of no association. We therefore also examined the distribution of the test statistic for the proposed method under the null of no seq SNP association with the phenotype. Genotype data were simulated with the same configurations described above, and a quantitative trait was simulated with $\beta_1 = 0$. We used $Z = \hat{\beta}_1/\mathrm{se}(\hat{\beta}_1)$ as the test statistic and calculated the P-value under an asymptotic standard normal distribution. For illustration purposes we chose 5% as the threshold for type I error assessment.

RESULTS

Overview. Under HWE and correct model specification, all methods yield consistent estimators for the intercept and the additive seq SNP effect. Asymptotic normality for the estimator from the proposed method is confirmed by examination of the distribution of β_1 estimates (see Supplementary Fig. 1). We focus on the relative efficiencies of the methods under different sampling designs compared to the ideal situation where we have sequence data for all subjects in the cohort. In general, the empirical standard deviations of the naive estimate $\hat{\beta}_{1,\mathrm{nai}}$ are smaller than those of the IPW estimate. For a quantitative trait, naively fitting a linear model still leads to a consistent estimate of the additive effect even though the marginal distribution of X_i in the phase 2 sample is not the same as that in the population. When the effect size under an additive model in the phase 1 sample is of interest, the incorporation of weighting in IPW helps to guard against bias from model misspecification. The proposed joint analysis method however yields more precise estimates than the other two approaches, and illustrates improved efficiency in detecting a functional SNP within the same region as the tag SNP when we can only afford to sequence a portion of the subjects in phase 1.

Scenario (i) (MAF $P_a = P_d = 0.4$) with additive effect. Table III displays a subset of the phase 2 sample allocations we evaluated for scenario (i) with correlation $r = 0.50$ and sampling fraction $\rho = 10\%$, as well as the resulting averages of estimates and empirical standard deviations. Here, $E(n_0, n_1, n_2)$ are the expected counts in the common

TABLE III. An example of estimation efficiencies under various phase 2 sample allocations for scenario (i) with MAFs $P_a = P_d = 0.4$, correlation $r = 0.50$, effect size $\beta_1 = 0.25$, and overall sampling fraction $\rho = 10\%$ (AVE = average of the β_1 estimates over 1,000 replicates, SD = standard deviation of the β_1 estimates multiplied by 100. For the complete data, $N = 1,000$, and for the phase 2 sample, $n = 100$. Minimum and maximum standard deviations for each method are displayed in bold)

		Allocation	Complete data		Naive		IPW		Proposed	
ρ	r	$E(n_0, n_1, n_2)$	AVE	SD	AVE	SD	AVE	SD	AVE	SD
10%	0.50	(80, 10, 10)	0.250	3.27	0.248	9.62	0.246	14.19	0.247	6.15
		(45, 10, 45)			0.251	**9.43**	0.249	13.52	0.248	6.25
		(10, 10, 80)			0.254	10.01	0.252	16.02	0.251	6.86
		(60, 30, 10)			0.251	9.73	0.251	10.66	0.248	6.13
		(50, 30, 20)			0.252	9.85	0.250	10.43	0.252	**6.04**
		(35, 30, 35)			0.254	9.78	0.254	10.71	0.249	6.29
		(20, 30, 50)			0.250	10.83	0.247	12.47	0.249	6.97
		(10, 30, 60)			0.252	**12.06**	0.252	**16.41**	0.248	7.44
		(33, 34, 33)[a]			0.250	9.62	0.251	10.54	0.247	6.39
		(36, 48, 16)[b]			0.255	10.18	0.255	**10.18**	0.249	6.46
		(40, 50, 10)			0.248	10.06	0.248	10.29	0.247	6.32
		(25, 50, 25)			0.254	10.65	0.255	11.08	0.253	6.32
		(10, 50, 40)			0.252	10.54	0.254	14.03	0.252	6.80
		(20, 70, 10)			0.249	11.20	0.249	12.37	0.248	7.02
		(10, 70, 20)			0.257	11.18	0.259	14.46	0.248	**7.67**
		(10, 80, 10)			0.253	11.21	0.250	14.40	0.247	7.43

[a] Approximately equal phase 2 sample size in each stratum.
[b] Equal sampling fraction in each stratum.

homozygote, the heterozygote, and the rare homozygote strata in the phase 2 sample. We also include results for two special cases of interest: one with approximately equal phase 2 sample size in each stratum, and the other with equal sampling fraction in each stratum. The relative bias is within 2% for all methods. The SD of the proposed estimate is roughly 2–2.5 times larger than the SD for the complete data, corresponding to a relative efficiency of 40–50%.

Figure 2 shows the relative efficiencies of the three methods under various sample size allocations for scenario (i), in which the overall sampling fraction is 10% or 50% (for cases with $\rho = 25\%$ see Supporting Information). The relative efficiency depends on the strength of the correlation between the tag and the seq SNPs. When there is high LD (e.g., $r = 0.95$), the efficiency of the proposed method approaches that of fitting model (A1) to the complete data if they were available. When there is no or low LD between the two SNPs (e.g., $r = 0.05$), the phase 1 sample tag SNP does not provide much useful information for inferring the genotypes at the seq SNP for phase 1 subjects not included in phase 2. Thus, the proposed method performs no better than the naive approach. For cases where the tag and the seq SNPS are negatively correlated, the proposed method still can improve efficiency. As the overall sampling fraction ρ increases, all methods produce more precise estimates.

The phase 2 sample size allocation plays an important role in the relative efficiency of the naive method. As mentioned above, under an additive genetic association model, $\hat{\beta}_{1,\mathrm{nai}}$ achieves maximum efficiency when the seq SNP genotype is DD for half of the phase 2 sample and is dd for the other half. This strategy works well when the seq SNP and the tag SNP have similar minor allele frequencies and are highly correlated or in perfect LD. As shown in the panels with $r = 0.95$ in Figure 2, compared to other allocations, the relative efficiency for the naive method is higher when stratum Aa is sparsely sampled and the two homozygote strata AA and aa are equally heavily sampled. Compared to the other approaches, the proposed method is relatively more consistent across different allocations for cases with positive correlation as long as the sampling scheme is not extreme. For negative correlation, however, the proposed method performs better when the rare homozygote stratum is sparsely sampled. This result is as expected, since the minor allele d at the seq SNP appears less frequently with the minor allele a at the tag SNP.

Scenario (ii) (MAF $P_a = 0.3$, $P_d = 0.2$) with additive effect. Similar results are obtained for scenario (ii) with additive effect. Figure 3 shows selected results for cases with sampling fraction $\rho = 10\%$ and 50%. For cases where the correlation between the tag and the seq SNPs is positive (e.g., $r = 0.75$), both the naive method and the proposed method can achieve higher relative efficiency when the rare homozygote stratum is heavily sampled. In contrast, for cases with negative correlation (e.g., $r = -0.30$), the methods achieve higher relative efficiency when the rare homozygote stratum is sparsely sampled. Again, the proposed method performs no better than the naive method when there is weak or no correlation between the tag and the seq SNPs (e.g., $r = 0.05$) but does not do worse. See Supporting Information for more comprehensive results.

Robustness assessment. When the regression model for the genetic association of the quantitative trait with the seq SNP is misspecified, all methods produce biased estimates. Here, we focus on reporting results under scenario (i) (MAF $P_a = P_d = 0.4$) for cases where correlation between the tag SNP and the seq SNP is moderate and the

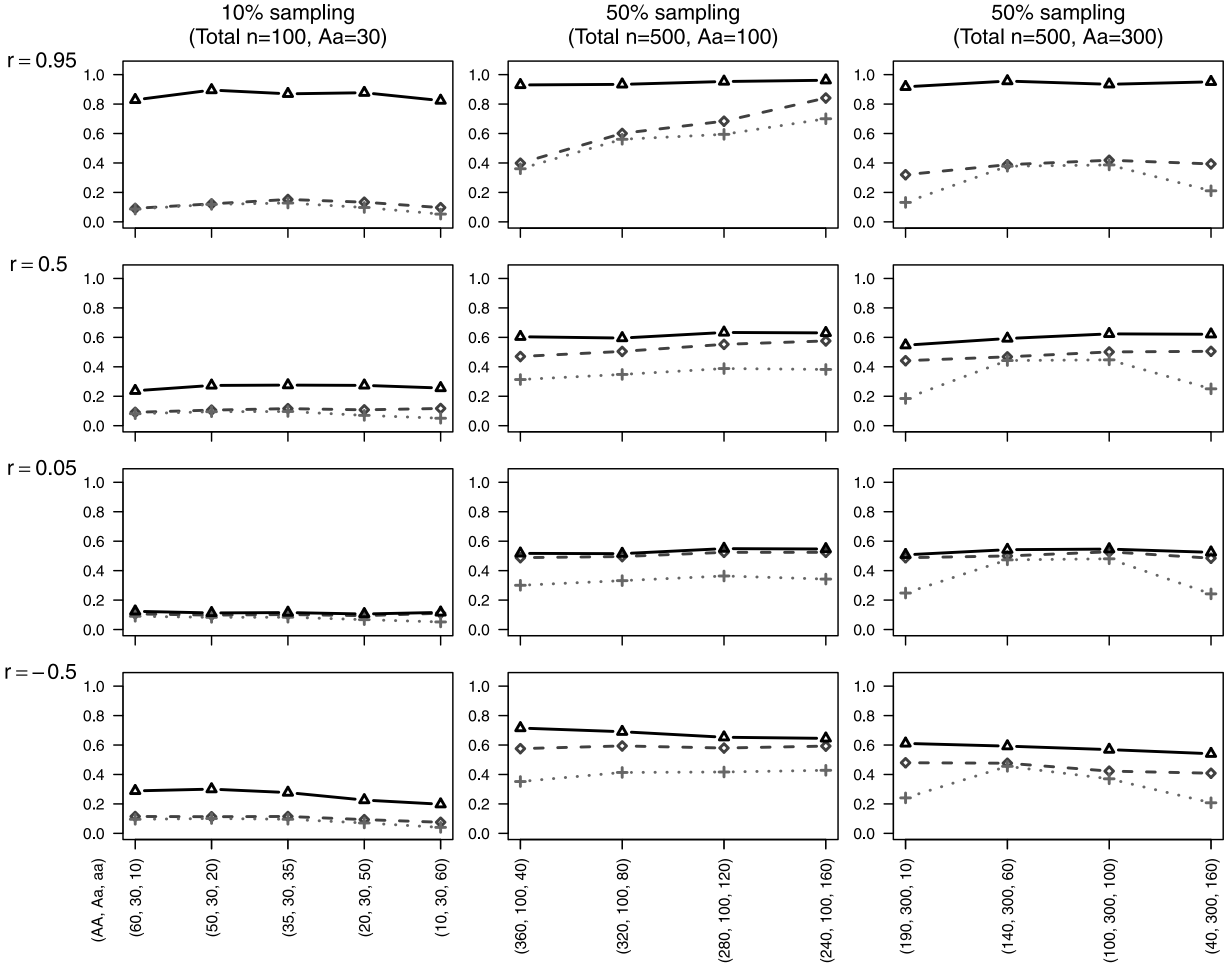

Fig. 2. Relative efficiencies of the naive (dashed line), IPW (dotted line), and proposed (solid line) methods under scenario (i) with MAF values of $P_a = P_d = 0.4$. The rows correspond to decreasing values of the tag-seq SNP correlation r. The first column corresponds to overall sampling fraction $\rho = 10\%$ with sample size = 30 allocated to stratum *Aa*. The second and third columns correspond to overall sampling fraction $\rho = 50\%$, with sample size of 100 and 300 allocated to stratum *Aa*, respectively. Within each panel, the horizontal axis indicates different tag SNP strata allocations (*AA*, *Aa*, *aa*) for fixed heterozygote (*Aa*) count. At 100% sampling, the expected counts of *AA*, *Aa*, *aa* are 360, 480, 160.

sampling fraction is $\rho = 25\%$ (Fig. 4). With the dominant model being the true model, the naive estimator $\hat{\beta}_{1,\mathrm{nai}}$ can have relative efficiency greater than 1 for some allocations. The explanation is that the complete data analysis yields biased estimates if one incorrectly fits an additive model, while the naive method can be less biased if the number with rare homozygote dd at the seq SNP in phase 2 is small. This may be achieved by oversampling from strata AA and Aa, provided that the correlation is high (see Supplementary Figs. 9– 23 for the other cases of r and ρ). For cases with negative r, the naive method performs best when stratum AA is sparsely sampled. This is due to the fact that under a dominant model, fewer cases with genotype dd in the sample leads to smaller bias in $\hat{\beta}_{1,\mathrm{nai}}$. In general, however, the proposed method performs better than the naive method for various allocations. Under a recessive model, the opposite phenomenon is observed. The relative efficiency of $\hat{\beta}_{1,\mathrm{nai}}$ increases as the expected number $E(m_1)$ of genotype DD in phase 2 sample increases. Comparison of row 1 with row 4 in Figure 4 suggests that the proposed method is reasonably robust to violation of the assumption of constant variance across genotype classes. Results for the cases with an additional causal seq SNP ignored in analysis or with a skewed phenotype distribution are very similar to the case with heteroscedasticity. Similar results are also observed for scenario (ii) (MAF $P_a = 0.3$, $P_d = 0.2$) that specifies different minor allele frequencies (see Supplementary Figs. 24–38).

Rare variant analysis. Because only a few rare variants are specified to be associated with the trait, the estimate of the overall effect size β for the aggregation score is only about 0.11. When the tag SNP and the rare variant score are moderately correlated (e.g., $r = 0.5$), the relative efficiency is consistently higher when stratum aa is oversampled than when stratum aa is undersampled. The proposed joint analysis method performed consistently better than the two methods that use phase 2 data alone (Fig. 5). Unlike common seq SNP analysis, however, undersampling of

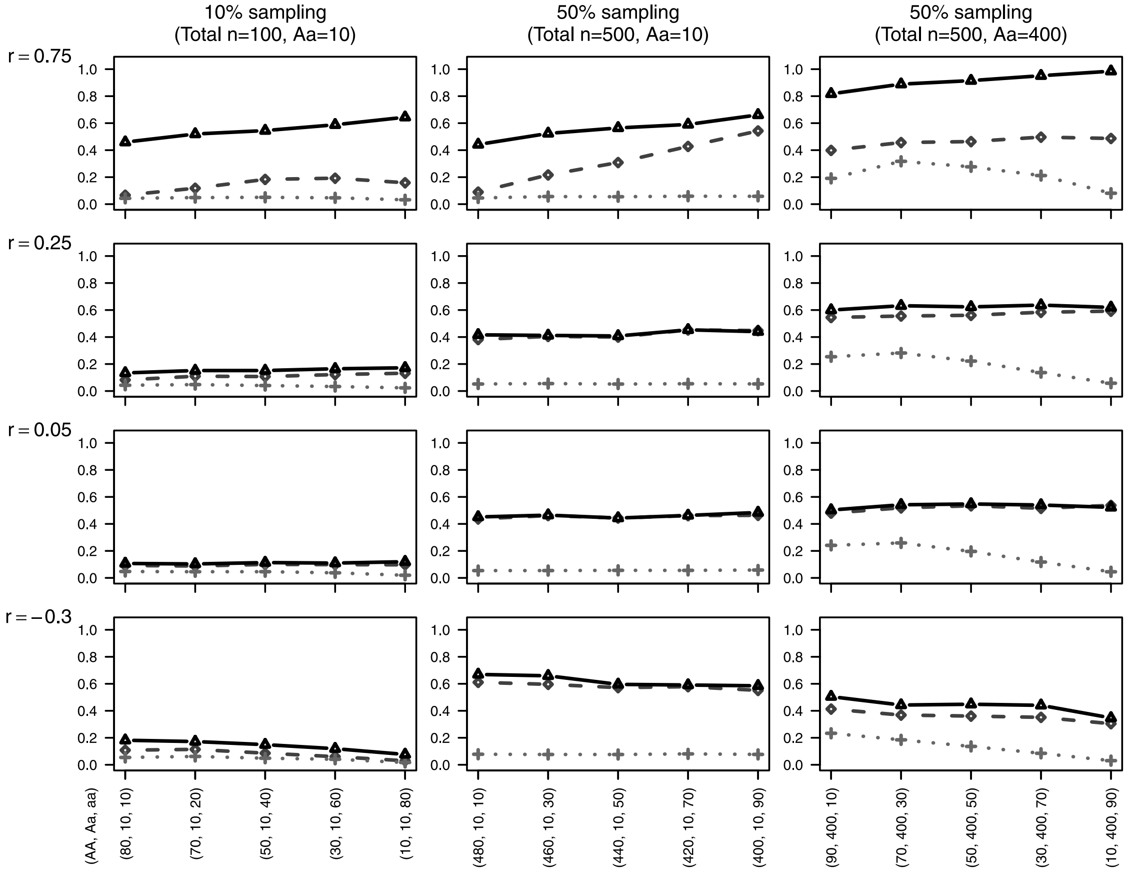

Fig. 3. Relative efficiencies of the naive (dashed line), IPW (dotted line), and proposed (solid line) methods under scenario (ii) with MAF values $P_a = 0.3$ and $P_d = 0.2$. The rows correspond to decreasing values of the tag-seq SNP correlation r; for the given MAF values, r is constrained to be less than 0.76. The first column corresponds to overall sampling fraction $\rho = 10\%$ with sample size = 10 allocated to stratum *Aa*. The second and third columns correspond to overall sampling fraction $\rho = 50\%$, with sample size of 10 and 400 allocated to stratum *Aa* , respectively. Within each panel, the horizontal axis indicates different tag SNP strata allocations (*AA*, *Aa*, *aa*) for fixed heterozygote (*Aa*) count. At 100% sampling, the expected counts of *AA*, *Aa*, *aa* are 490, 420, 90.

the tag SNP heterozygote stratum did not appear to contribute to efficiency gain, most likely due to the fact that the rare variant score distribution concentrates its mass on zero. Under the assumption that K independent rare variants are included in the rare variant score, the sum of the correlations between the tag SNP and each of the K rare variants is approximately equal to $K^{1/2}r$. For r =0.5 and K =20, the average correlation between each of the K rare variants and the tag SNP would be approximately 0.112.

Simulations under the null hypothesis of no association. When data were generated under the situation of no association between the seq SNP and the phenotype, all methods consistently estimated the genetic effect at 0 for all sampling fractions and all sampling schemes. For scenario (i) with MAF $P_a = P_d = 0.4$ and sampling fraction $\rho = 10\%$, there is a slight inflation of type 1 error rate for the proposed method under some sample size allocations. The type 1 error rate is between 5% and 10% when $r = 0.05$ (see Supplementary Fig. 39). However, type I error does not necessarily reflect accurately the statistical design efficiency. This is because the test statistic depends on the accuracy of standard error estimation (typically based on asymptotic distributions) whereas efficiency depends on the empirical precision of the estimates. This inflation may be due to the use of small phase 2 sample. For sampling fraction $\rho = 25\%$ or 50%, the type 1 error rate is close to the nominal 5% (see Supplementary Figs. 40 and 41). Similar patterns are observed for scenario (ii) where minor allele frequencies differed between the tag and seq SNPs (see Supplementary Figs. 42–44).

Summary concerning allocation. It is evident that the phase 2 sample size allocation has a major impact on methods that use only phase 2 data, which is relevant when analysis is limited to use of a linear model in phase 2 data. Although a universally optimal allocation design does not appear to exist even when the assumption of additive effect is correctly made, our simulation studies provide investigators with some guidelines for planning regional sequencing on a subset of phase 1 subjects. As illustrated by our simulations, efficiency of the naive method using only the phase 2 sample depends on the MAFs at the tag SNP and the seq SNP as well as their correlation r. When there are reasonable grounds to expect positive correlation, we recommend that the heterozygote stratum *Aa* be undersampled

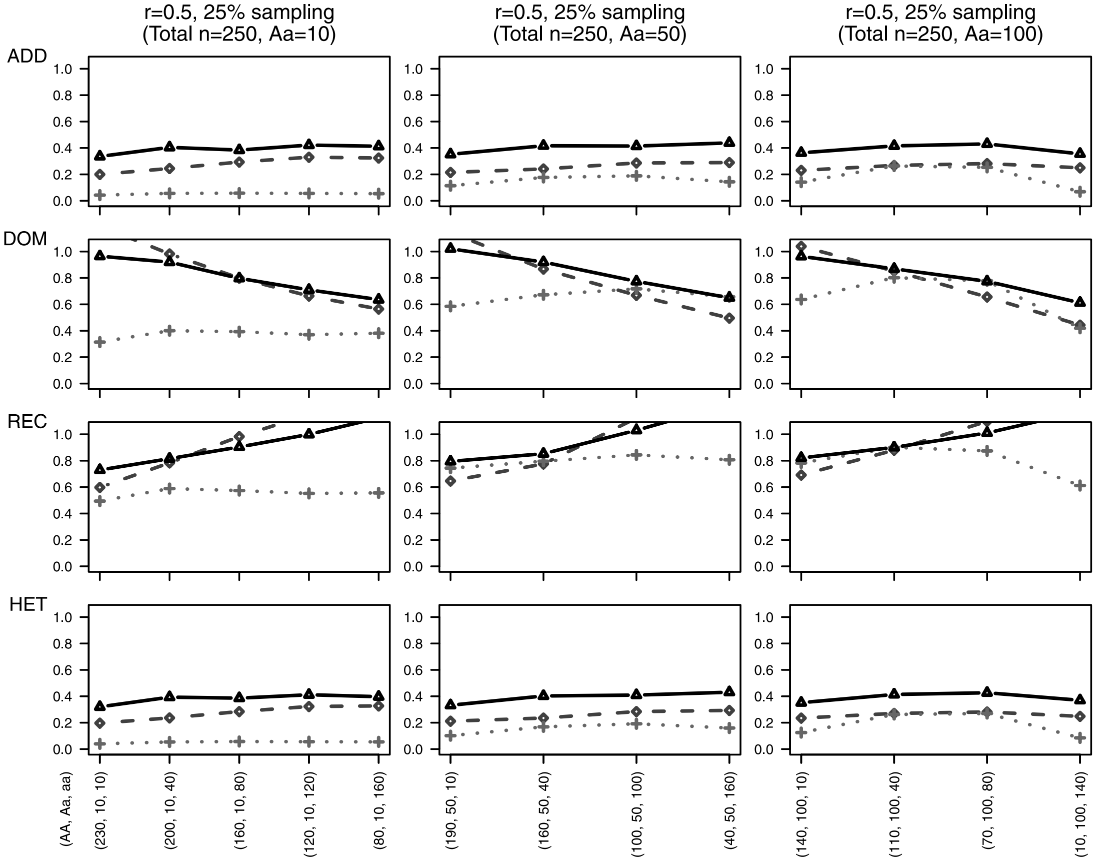

Fig. 4. Relative MSE of the naive (dashed line), IPW (dotted line), and proposed (solid line) methods under model misspecification with MAF values $P_a = P_d = 0.4$. The first row corresponds to a correctly specified additive (ADD) model. The second, third, and fourth rows correspond to cases where the true model is dominant (DOM), recessive (REC), and heteroscedastic (HET), respectively. The overall sampling fraction is $\rho = 25\%$, and the tag-seq SNP correlation is $r = 0.5$. The columns correspond to cases with sample sizes of 10, 50, and 100 allocated to stratum *Aa*, respectively. Within each panel, the horizontal axis indicates different tag SNP strata allocations (*AA*, *Aa*, *aa*) for fixed heterozygote (*Aa*) count. At 100% sampling, the expected counts of *AA*, *Aa*, *aa* are 360, 480, 160. Results for cases with an additional causal seq SNP ignored as well as for cases with skewed phenotype distribution are similar to HET, and hence are not shown here (see Supplementary Figs. 9–23).

while the two homozygote strata *AA* and *aa* be oversampled up to maximum available counts in *aa* (Figs. 2 and 3). Although similar trends of relative efficiency are observed for the proposed joint analysis of phases 1 and 2 data, it is less dependent on the sample allocation and is generally more efficient and more robust than the naive method.

DISCUSSION

The two-phase study design, widely used in epidemiological studies, focuses on collecting more detailed but expensive covariate data in a subset of the study sample based on information in auxiliary variables available in the entire sample. Although the cost of high-density genotypic data has dropped dramatically, costs of NGS are still considered high for large-scale studies that involve tens of thousands of participants. Efficient study designs are likely to remain necessary in the near future for cost-efficiency in large studies of the genetic basis of complex diseases and traits. The two-phase design is important in the sense that information from low-cost tag SNPs and imputed SNPs available in a phase 1 sample can lead to better decisions on how to select a subset of the sample to be sequenced for discovery and assessment of additional variant SNPs in phase 2. As the depth and breadth of available sequence data accumulates, for example, through initiatives such as the 1,000 Genomes Project, and lower frequency SNPs are added to GWAS SNP arrays, the number of unknown or unimputable variants within a sequenced region in a particular study may decrease. It remains to be seen, however, whether imputation accuracy will improve sufficiently for follow-up and fine-mapping studies, especially for very low frequency variants. Moreover, sequencing a portion of study participants may serve to create a very well-matched reference panel useful for imputation in the entire study [Fridley et al., 2010; Zeggini, 2011].

Two-phase stratified design and optimal sample allocation for GWAS follow-up studies have received little attention [Thomas et al., 2004, 2009]. In contrast, the well-known multistage design improves cost-efficiency in the GWAS setting by genotyping a full set of known SNP markers in a

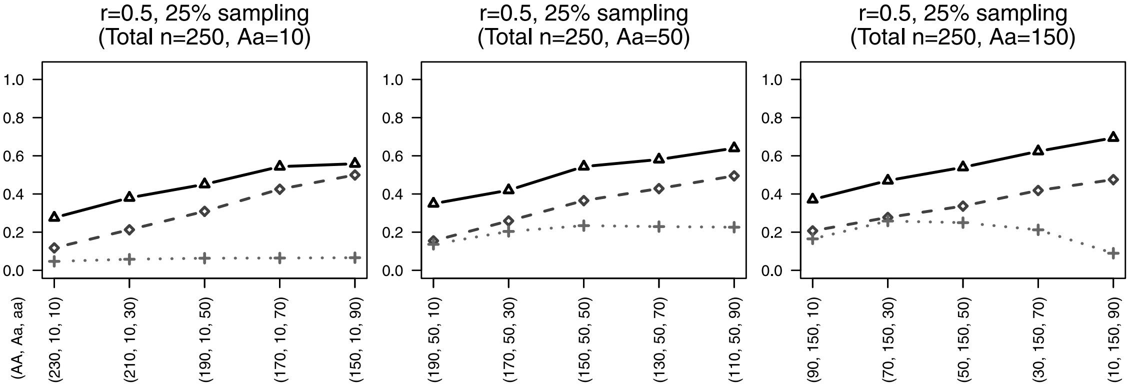

Fig. 5. Relative efficiency of the naive (dashed line), IPW (dotted line), and proposed (solid line) methods for rare variant analysis involving 20 rare variants with MAF values generated from Unif (0.005, 0.01). The overall sampling fraction is $\rho = 25\%$, and the correlation between the tag SNP and the rare variants sum score is $r = 0.5$. The panels correspond to cases with sample size of 10, 50, and 150 allocated to stratum *Aa*, respectively. Within each panel, the horizontal axis indicates different tag SNP strata allocations (*AA*, *Aa*, *aa*) for fixed heterozygote (*Aa*) count. At 100% sampling, the expected counts of *AA*, *Aa*, *aa* are 490, 420, 90.

subset of available subjects in stage 1, and then genotyping a selected subset of the SNPs in the remaining subjects in stage 2. By excluding markers that show little evidence of association in stage 1, genotyping requirements, and hence cost, can be substantially reduced while preserving much of the power of the corresponding single-stage design in which all subjects are genotyped on all markers [Skol et al., 2006]. An association detected in both stages, however, may be arising indirectly through a common variant that is in LD with a functional variant, thus motivating the need for subsequent regional fine-mapping and sequencing studies to identify additional variants.

In this report, we consider estimation of an additive genetic effect in a two-phase stratified design. As a starting point, we examine the case of one tag SNP and one seq SNP. We propose an estimating equations approach using all available data from phases 1 and 2 and study the efficiency gain compared to using only phase 2 data. We also investigate the sensitivity of estimation efficiency to allocation of the phase 2 sample size under the additive model. The main idea of the sampling design is to select fewer heterozygotes with the sequence variant while selecting more of each homozygote type. If one expects positive correlation between the tag SNP and a functional seq variant, then it is more efficient to over-sample from the tag SNP rare homozygote stratum. This strategy provides no useful information, however, if the tag SNP is not in LD with the seq SNP, which may be the case when the seq SNP is too distant from the tag SNP. Through simulation studies, we show that the correlation between a tag SNP used for stratification and the seq SNP plays an important role in estimation efficiency for the proposed joint analysis method. As the magnitude of the correlation coefficient decreases to zero, the efficiency decreases relative to complete sequencing. When the seq SNP and the tag SNP are independent, the phase 2 sample can be regarded as a random subset of the entire study sample, and stratification by the tag SNP genotype does not contribute to improved estimation of the genetic effect.

Our findings concerning sample size allocation are most directly applicable to the design of an independent replication study in which a specific region of interest has been prespecified, and typed or imputed SNP data for the region are readily available. When, however, a promising region of interest has been identified by GWAS using genome-wide significance criteria, effect estimates for tag SNPs so identified will be subject to selection bias known as the ″winner′s curse″ [Faye et al., 2011, Sun et al., 2011]. In fine-mapping conducted in the same sample, effect estimates for seq SNPs in LD with the tag SNP will also be affected indirectly by this phenomenon in a complicated manner [Faye and Bull, 2011], and similarly subject to bias. However, regardless of such complications in interpreting the results, a tag SNP in high LD with a sequenced SNP is nevertheless expected to serve well in selecting individuals enriched for informative seq SNP genotypes within the region.

For complex diseases and traits, the underlying genetic models are unknown. Thus, there is no uniformly most powerful test across all possible alternative genetic models and no single optimal phase 2 sample size allocation for all situations. We have limited consideration to a linear model for a quantitative trait with an additive effect of a genetic variant, in which the number of the copies of the minor allele is treated as dosage. Our method presumes that a potential functional variant is within the same region as the tag SNP used for stratification. In practice, with sequencing of multiple SNPs in the same region as the tag SNP, analysis of the sequence genotypes would proceed by association testing of each of the seq SNPs with the quantitative trait of interest. For any of the seq SNPs correlated with the tag SNP, a stratified sampling design will be more powerful than a simple random sample of the same size. On the other hand, for a seq SNP in the region that is uncorrelated with the tag SNP, stratified sampling will not perform worse than simple random sampling.

The development of sample allocation and joint analysis methods is based on the assumption that only a single tag SNP is available within a region. In practice, multiple SNP markers in phase 1 may be used as tag SNPs for a region or there may be tag SNPs in multiple regions of interest. When the number of genotype combinations is large, stratification using multiple tag SNPs can be problematic. One possible modification is to combine genotype categories based on the total count of minor alleles at the tag SNPs. Modeling the associations between the seq SNPs and the collapsed strata as well as designating a robust allocation, however, may not be straightforward. As a result, scope for the application of sample allocation design principles across multiple regions may be limited for purposes of cost-efficiency. For other types of analysis such as haplotype- or gene-based inference, optimal allocation depends on the specific statistical methods to be used and is very likely to employ different optimization criteria. In principle, the sample size allocation methods can be extended to situations in which haplotypes are used for stratification and/or rare variant counts are used to summarize sequence data [Price et al., 2010]. In addition, because environmental factors also play important roles in the etiology of complex traits/diseases, they can be included in the joint analysis to better explain other sources of the variation of the trait [e.g., Paterson et al., 2010].

Subject-selection strategies that depend on the quantitative trait, and associated methods of analysis, have been evaluated by a number of authors [e.g., Bacanu et al., 2011; Guey et al., 2011; Huang and Lin, 2007; Lin and Tang, 2011; Tang, 2010; Van Gestel et al., 2000; Yilmaz and Bull, 2011], and in principle could also be applied within genotype classes, as suggested by a reviewer . In cross-sectional and longitudinal study designs, however, multiple traits are often of interest, and sampling based on one trait may not improve efficiency for another. If functional variants for multiple traits are harbored in the region and are correlated with the tag SNP, then sampling based on that tag SNP can be beneficial for all traits. To the extent that the tag SNP genotype is correlated with a quantitative trait, over-sampling on the high- and low-risk genotype strata will indirectly enrich for associated extreme trait values in the selected individuals. Further work to formally evaluate additional improvements in efficiency associated with trait-dependent sampling, for example, by defining multiple strata according to genotype and phenotype, is warranted.

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. is 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. Computations were performed on the GPC supercomputer at the SciNet HPC Consortium. SciNet is funded by: the Canada Foundation for Innovation under the auspices of Compute Canada; the Government of Ontario; Ontario Research Fund—Research Excellence; and the University of Toronto. The authors thank the referees for constructive comments that significantly improved the paper.

REFERENCES

Bacanu SA, Nelson MR, Whittaker JC. 2011. Comparison of methods and sampling designs to test for association between rare variants and quantitative traits. Genet Epidemiol 35: 226–235.

Breslow NE, Wellner JA. 2007. Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. Scand J Stat 34: 86–102.

Faye, L, Bull, SB. 2011. Two-stage study designs combining GWAS tag SNPs and exome sequencing: accuracy of genetic effect estimates. BMC Proceedings 5(Suppl. 9): S64.

Faye L, Sun L, Dimitromanolakis A, Bull SB. 2011. A flexible genome-wide bootstrap method that accounts for ranking- and threshold-selection bias in GWAS interpretation and replication study design. Stat Med 30: 1898–1912.

Fridley BL, Jenkins G, Deyo-Svendsen ME, Hebbring S, Freimuth R. 2010. Utilizing genotype imputation for the augmentation of sequence data. PLoS ONE 5: e11018.

Godambe VP, Thompson ME. 1986. Parameters of superpopulation and survey population: their relationships and estimation. Int Stat Rev 54: 127–138.

Guey LT, Kravic J, Melander O, Burtt NP, Laramie JM, Lyssenko V, Jonsson A, Lindholm E, Tuomi T, Isomaa B, Nilsson P, Almgren P, Kathiresan S, Groop L, Seymour AB, Altshuler D, Voight BF. 2011. Power in the phenotypic extremes: a simulation study of power in discovery and replication of rare variants. Genet Epidemiol 35: 236–246.

Huang BE, Lin DY. 2007. Efficient association mapping of quantitative trait loci with selective genotyping. Am J Hum Genet 80: 567–572.

Ioannidis JP, Thomas G, Daly MJ. 2009. Validating, augmenting and refining genome-wide association signals. Nat Rev Genet. 210: 318–329.

Li Y, Willer C, Sanna S, Abecasis G. 2009. Genotype imputation. Annu Rev Genomics Hum Genet 10: 387–406.

Lin DY, Tang ZZ. 2011. A general framework for detecting disease associations with rare variants in sequencing studies. Am J Hum Genet 89: 354–367.

Liu DJ, Leal SM. 2010. Replication strategies for rare variant complex trait association studies via next-generation sequencing. Am J Hum Genet 87: 790–801.

Manski CF, Lerman SR. 1977. The estimation of choice probabilities from choice based samples. Econometrica 45: 1977–1988.

Morris AP, Zeggini E. 2010. An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol 34: 188–193.

Newey W, McFadden D. 1994. Large sample estimation and hypothesis testing. In: Engler R, McFadden D, eds, Handbook of Econometrics, vol. 4. Elsevier Science, Amsterdam.

Paterson AD, Waggott D, Boright AP, Hosseini SM, Shen E, Sylvestre M-P, Wong I, Bharaj B, Cleary PA, Lachin JM, Below JE, Nicolae D, Cox NJ, Canty AJ, Sun L, Bull SB, Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications Research Group. 2010. A genome-wide association study identifies a novel major locus for glycemic control in type 1 diabetes, as measured by both A1C and glucose. Diabetes 59: 539–549.

Pei YF, Zhang L, Li J, Deng HW. 2010. Analyses and comparison of imputation-based association methods. PLoS ONE 5: e10827.

Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR. 2010. Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet 86: 832–838.

Skinner, CJ, Holt, D, Smith, TMF (editors). 1989. Analysis of complex surveys. Chichester, UK: Wiley.

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

Sun L, Dimitromanolakis A, Faye L, Paterson A, Waggott D, Bull S. 2011. BR-squared: a practical solution to the winner's curse in genome-wide scans. Hum Genet 129: 545–552.

Tang Y. 2010. Equivalence of three score tests for association mapping of quantitative trait loci under selective genotyping. Genetic Epidemiol 34: 522–527.

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

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.

Udler SM, Tyrer J, Easton DF. 2010. Evaluating the power to discriminate between highly correlated SNPs in genetic association studies. Genet Epidemiol 34: 463–468.

Van Gestel S, Houwing-Duistermaat JJ, Adolfsson R, van Duijn CM, Van Broeckhoven C. 2000. Power of selective genotyping in genetic association analyses of quantitative traits. Behav Genet 30: 141–146.

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

Zeggini E. 2011. Next-generation association studies for complex traits. Nat Genet 43: 287–288.

Zheng G, Song K, Elston RC. 2007. Adaptive two-stage analysis of genetic association in case-control designs. Hum Hered 63: 175–186.

APPENDIX: CONSISTENCY AND ASYMPTOTIC DISTRIBUTION OF $\hat{\theta}$

The asymptotic behavior of $\hat{\theta}$ can be derived based on standard estimating equations theory. By Theorem 3.4 of Newey and McFadden [1994], under regularity conditions, we have that with probability approaching 1, there is a unique solution to $\sum_{i=1}^{N} \Psi_i(\theta)$, denoted by $\hat{\theta}$, that satisfies

$$\begin{aligned} 0 &= N^{-1/2} \sum_{i=1}^{N} \Psi_i(\theta) \\ &\quad + N^{-1} \sum_{i=1}^{N} \partial \Psi_i(\theta)/\partial\theta^{\mathrm{T}} N^{1/2} \left(\hat{\theta} - \theta\right) + o_p(1). \end{aligned}$$

This is equivalent to

$$\begin{aligned} N^{1/2}\left(\hat{\theta} - \theta\right) &= -\left[E\left\{\partial\Psi_i(\theta)/\partial\theta^{\mathrm{T}}\right\}\right]^{-1} \\ &\quad \times N^{-1/2} \sum_{i=1}^{N} \Psi_i(\theta) + o_p(1) \end{aligned} \tag{A1}$$

as under regularity conditions, $E\left\{\partial\Psi_i(\theta)/\partial\theta^{\mathrm{T}}\right\}$ exists and is invertible and $\mathrm{var}\{\Psi_i(\theta)\}$ is finite and positive definite. The Law of Large Numbers leads to $N^{-1}\sum_{i=1}^{N} \Psi_i(\theta) \to_p E\{\Psi_i(\theta)\} = 0$, as $N \to \infty$, and the consistency of $\hat{\theta}$ is immediate by the Slutzky theorem. By applying the Central Limit Theorem to (A1), the asymptotic distribution of $N^{1/2}(\hat{\theta} - \theta)$ can be established.

Let

$$\begin{aligned} W_i(\alpha) &= (\xi_i/\pi_i)\partial Q_i(\alpha)/\partial\alpha^{\mathrm{T}}, \\ G_i(\alpha, \beta) &= (1 - \xi_i)\partial U_i^*(\alpha, \beta)/\partial\alpha^{\mathrm{T}}. \end{aligned}$$

Note that

$$\tilde{M}_i(\alpha, \beta) = \xi_i \partial U_i(\beta)/\partial\beta^{\mathrm{T}} + (1 - \xi_i)\partial U_i^*(\alpha, \beta)/\partial\beta^{\mathrm{T}}.$$

Therefore,

$$\frac{\partial\Psi_i(\theta)}{\partial\theta^{\mathrm{T}}} = \begin{pmatrix} W_i(\alpha) & 0 \\ G_i(\alpha, \beta) & \tilde{M}_i(\alpha, \beta) \end{pmatrix}.$$

Thus,

$$\Gamma = \sum_{j=0}^{2} \mathrm{pr}(Z_i = j) \begin{pmatrix} \Gamma_{j11} & 0 \\ \Gamma_{j21} & \Gamma_{j22} \end{pmatrix},$$

where

$$\begin{aligned} \Gamma_{j11} &= E\{\partial Q_i(\alpha)/\partial\alpha^{\mathrm{T}} \mid Z_i = j\}, \\ \Gamma_{j21} &= (1 - \rho_j)E\{\partial U_i^*(\alpha, \beta)/\partial\alpha^{\mathrm{T}} \mid Z_i = j\}, \\ \Gamma_{j22} &= \rho_j E\{\partial U_i(\beta)/\partial\beta^{\mathrm{T}} \mid Z_i = j\} \\ &\quad + (1 - \rho_j)E\{\partial U_i^*(\alpha, \beta)/\partial\beta^{\mathrm{T}} \mid Z_i = j\}. \end{aligned}$$

As $N \to \infty$, $E\{\partial U_i^*(\alpha, \beta)/\partial\beta^{\mathrm{T}} \mid Z_i = j\}$, and Γ_{j11} can be consistently estimated by, respectively,

$$\begin{aligned} \hat{E}\left\{\frac{\partial U_i^*(\alpha, \beta)}{\partial\beta^{\mathrm{T}}} \mid Z_i = j\right\} &= \frac{1}{N_j - n_j} \\ &\times \sum_{i \in \{\bar{s}_2 \cup j\}} \left[\sum_{k=0}^{2} \phi_{ik}(\hat{\alpha}, \hat{\beta}) U_i(\hat{\alpha}, \hat{\beta}; Y_i, X_i = k) U_i^{\mathrm{T}}(\hat{\alpha}, \hat{\beta}; Y_i, X_i = k)\right. \\ &- \left\{\sum_{k=0}^{2} \phi_{ik}(\hat{\alpha}, \hat{\beta}) U_i(\hat{\beta}; Y_i, X_i = k)\right\} \\ &\times \left\{\sum_{k=0}^{2} \phi_{ik}(\hat{\alpha}, \hat{\beta}) U_i(\hat{\beta}; Y_i, X_i = k)\right\}^{\mathrm{T}} \\ &+ \left.\sum_{k=0}^{2} \phi_{ik}(\hat{\alpha}, \hat{\beta}) \partial U_i(\hat{\beta}; Y_i, X_i = k)/\partial\beta^{\mathrm{T}}\right], \end{aligned}$$

and

$$\begin{aligned} \hat{E}\left\{\frac{\partial U_i^*(\alpha, \beta)}{\partial\alpha^{\mathrm{T}}} \mid Z_i = j\right\} &= \frac{1}{N_j - n_j} \\ &\times \sum_{i \in \{\bar{s}_2 \cup j\}} \left[\sum_{k=0}^{2} \phi_{ik}(\hat{\alpha}, \hat{\beta}) U_i(\hat{\alpha}; Y_i, X_i = k) Q_i^{\mathrm{T}}(\hat{\alpha}; G_k, Z_i)\right. \\ &- \left\{\sum_{k=0}^{2} \phi_{ik}(\hat{\alpha}, \hat{\beta}) U_i(\hat{\alpha}, \hat{\beta}; Y_i, X_i = k)\right\} \\ &\times \left.\left\{\sum_{k=0}^{2} \phi_{ik}(\hat{\alpha}, \hat{\beta}) Q_i(\hat{\alpha}; G_k, Z_i)\right\}^{\mathrm{T}}\right]. \end{aligned}$$

Similarly, $E\left\{\partial U_i(\beta)/\partial\beta^{\mathrm{T}} \mid Z_i = j\right\}$ can be consistently estimated by

$$\hat{E}\left\{\frac{\partial U_i(\beta)}{\partial\beta^{\mathrm{T}}} \mid Z_i = j\right\} = \frac{1}{n_j} \sum_{i \in \{s_2 \cup j\}} \frac{\partial U_i(\hat{\beta})}{\partial\beta^{\mathrm{T}}}.$$

The middle term Σ in the sandwich variance matrix can be consistently estimated by its empirical counterpart

$$\hat{\Sigma} = \sum_{i=1}^{N} \begin{pmatrix} (\xi_i/\pi_i)^2 Q_i(\hat{\alpha}) Q_i^{\mathrm{T}}(\hat{\alpha}) & \xi_i/\pi_i\, Q_i(\hat{\alpha}) \tilde{U}_i^{\mathrm{T}}(\hat{\alpha}, \hat{\beta}) \\ \xi_i/\pi_i \tilde{U}_i(\hat{\alpha}, \hat{\beta}) Q_i^{\mathrm{T}}(\hat{\alpha}) & \tilde{U}_i(\hat{\alpha}, \hat{\beta}) \tilde{U}_i^{\mathrm{T}}(\hat{\alpha}, \hat{\beta}) \end{pmatrix}.$$