Received: 7 September 2017 | Revised: 23 October 2017 | Accepted: 23 October 2017

DOI: 10.1002/gepi.22099

WILEY Genetic Epidemiology

OFFICIAL JOURNAL
INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY
www.geneticepi.org

RESEARCH ARTICLE

Two-phase designs for joint quantitative-trait-dependent and genotype-dependent sampling in post-GWAS regional sequencing

Osvaldo Espin-Garcia[1,2] | Radu V. Craiu[3] | Shelley B. Bull[1,2]

[1]Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada

[2]Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, ON, Canada

[3]Department of Statistical Sciences, University of Toronto, Toronto, ON, Canada

Correspondence
Shelley B. Bull, Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, ON, Canada.
Email: bull@lunenfeld.ca

Funding information
Canadian Institutes of Health Research, Grant/Award Numbers: GET-101831, MOP-84287; NIH, Grant/Award Numbers: U01 DK085524, U01 DK085584, U01 DK085501, U01 DK085526, U01 DK085545, P01 HL045222, R01 DK047482, R01 DK053889; National Institute of General Medical Sciences, Grant/Award Number: R01 GM031575; Canadian Foundation for Innovation; Compute Canada; Government of Ontario; Ontario Research Fund – Research Excellence; University of Toronto

ABSTRACT

We evaluate two-phase designs to follow-up findings from genome-wide association study (GWAS) when the cost of regional sequencing in the entire cohort is prohibitive. We develop novel expectation-maximization-based inference under a semiparametric maximum likelihood formulation tailored for post-GWAS inference. A GWAS-SNP (where SNP is single nucleotide polymorphism) serves as a surrogate covariate in inferring association between a sequence variant and a normally distributed quantitative trait (QT). We assess test validity and quantify efficiency and power of joint QT-SNP-dependent sampling and analysis under alternative sample allocations by simulations. Joint allocation balanced on SNP genotype and extreme-QT strata yields significant power improvements compared to marginal QT- or SNP-based allocations. We illustrate the proposed method and evaluate the sensitivity of sample allocation to sampling variation using data from a sequencing study of systolic blood pressure.

KEYWORDS

fine-mapping, Genetic Analysis Workshop 19, genetic association studies, joint outcome covariate dependent sampling, outcome-/covariate-dependent sampling

1 | INTRODUCTION

Regional sequencing to follow-up findings from a genome-wide association study (GWAS), in which a large number of single nucleotide polymorphisms (SNPs) are tested one by one, can be cost-effective for fine mapping. In the “post-GWAS” era, identifying causal variants and susceptibility genes in GWAS-identified regions of association has become an important goal for researchers. Despite decreasing costs of next-generation sequencing (NGS) technologies, sequencing all subjects in large-scale studies is still prohibitive. Thus, careful planning and innovative designs become essential to optimize the available resources.

Sequence variants at high density in the fine-mapping region of interest are typically in linkage disequilibrium (LD) with previously, strongly associated, variants from GWAS. However, the GWAS SNPs may or may not have biological function themselves. Consequently, fine-mapping variants in the selected region are tested for association to identify biologically relevant loci. A thorough review and description of the strategies utilized for fine mapping can be found in Spain and Barrett (2015).

Two-phase sampling design and analysis has proven to be an efficient technique to select and analyze a cost-effective subsample, for example, individuals to be sequenced. In their original formulation, two-phase sampling designs were developed to estimate population parameters of an expensive variable, that is, a random variable that is costly to measure (in time, materials, or personnel) (Pickles, Dunn, & Vázquez-Barquero, 1995; White, 1982). At phase 1, data for an inexpensive or surrogate variable (correlated to the expensive variable) are collected for a large random sample of the population. At phase 2, the expensive variable is measured only in a subset of the phase 1 sample; the subsample is drawn based on the information provided by the sample distribution of the inexpensive variable. In this case, the expensive variable is missing by design in individuals not selected in phase 2. Various methods have been proposed to analyze data collected under a two-phase design, including estimating functions (Breslow & Wellner, 2007; Chatterjee, Chen, & Breslow, 2003; Chen, Craiu, & Bull, 2012; Scott & Wild, 2011) and (conditional or full) maximum likelihood methods (Breslow & Cain, 1988; Breslow & Holubkov, 1997; Derkach, Lawless, & Sun, 2015; Lawless, Kalbfleisch, & Wild, 1999; Lin, Zeng, & Tang, 2013; Song, Zhou, & Kosorok, 2009; Zeng & Lin, 2014; Zhao, Lawless, & McLeish, 2009).

Under a case-control design, sampling within strata defined by both disease status and covariates is more powerful than sampling from cases and controls only (Breslow & Chatterjee, 1999; Schaid, Jenkins, Ingle, & Weinshilboum, 2013). For a quantitative trait (QT), sampling strategies that select informative individuals according to extreme-trait or genotype values are known to have good properties (Chen et al., 2012; Lin et al., 2013; Thomas, Yang, & Yang, 2013), but designs based on joint sampling are not well developed. In this effort, we evaluate phase 2 sampling according to values of a SNP genotype and a QT. Statistical analysis is based on semiparametric maximum likelihood (SPML) estimation where a GWAS-identified SNP or a candidate GWAS-SNP (Z) is a surrogate covariate used to infer the association between a sequence variant (G) and a QT (Y). We estimate parameters via an EM algorithm; our approach is novel in that it is tailored for the post-GWAS scenario by incorporating GWAS SNP data available for all individuals. As a result, efficiency is improved compared to methods that use phase 2 data alone. Specifically, we assume that the linear relationship between a causal variant, G, and a QT, Y, is of the form:

$$Y = \beta_0 + \beta_1 G + \epsilon, \tag{1}$$

where β_0, β_1 are regression parameters, $\epsilon \sim N(0, \sigma^2)$; thus $Y \sim N(\mu_G, \sigma^2)$, $\mu_G = \beta_0 + \beta_1 G$. Additionally, we hypothesize that a causal sequence variant ($G = G_1$) and the GWAS-SNP (Z) tend to occur on the same haplotype with a specified LD structure. G and Z have minor allele frequencies (MAFs) of q_G and q_Z, respectively.

The remainder of the paper is structured as follows. Section 2 describes the designs for two-phase sampling in the post-GWAS setting. Section 3 explains the model formulation and details of the statistical inference via SPML. Section 4 elaborates on the alternative phase 2 sample allocations we study. In Section 5, we report simulation studies performed to evaluate estimation efficiency, association test validity, and power. Further, we quantify improvements associated with QT-SNP joint sampling compared to QT or SNP marginal samplings under alternative sample allocations. Fine-mapping analysis of systolic blood pressure is presented in Section 6 to illustrate application of the methods and compare sampling variation of alternative allocations. A discussion focusses on limitations and extensions of the proposed method.

2 | TWO-PHASE DESIGNS IN POST-GWAS REGIONAL SEQUENCING

The main objective in the post-GWAS setting is to conduct statistical inference on the association of a sequence variant, G, that is, a potentially causal variant, with a QT of interest Y. This variant is located in a genomic region of interest, narrowed down following GWAS results. This so-called fine mapping generally begins with multiple single-variant analysis across a region; although, for analysis of low-count rare variants, the application of multivariant burden and variance component tests may be necessary.

To reduce sequencing costs, however, variants in the region are ascertained in only a fraction of individuals, making G missing by design for a potentially large subset of the first sample. Consequently, two-phase studies consist of GWAS in phase 1 and fine-mapping analysis of sequence data in phase 2, the latter obtained in a subsample of individuals from the initial GWAS. In this post-GWAS scenario, the trait data (Y) and the GWAS-SNP (Z), a surrogate for the causal variant, are observed for every subject in the study.

The design objective is to select a subset of informative subjects based on available data in phase 1, namely (Z, Y). Inference on the missing-by-design sequence variants is conducted using all available data from phases 1 and 2. We define the missing indicator $r_i = \mathbb{1}\{i \in S_2\}$, $i = 1, \ldots, N$, where N is the number in individuals measured in phase 1. S_2 represents the phase 2 sample of n subjects. We let $\bar{S}_2$ denote the set of $(N - n)$ subjects that are not in the phase 2 sample but are present in the GWAS study.

Regression analysis at the GWAS phase uses the surrogate variable, Z, which does not usually have biological function, say by fitting $Y = \gamma_0 + \gamma_1 Z$. Nonetheless, GWAS analyses serve as an efficient screening strategy to identify candidate regions in the genome. On the other hand, in fine-mapping, both causal and null variants are colocated in the region, and these null variants do not correspond to the usual model-based null hypothesis, that is, $\beta_1 = 0$ in (1). Rather, a null SNP in the region of interest (denoted by G_0) has no direct association with the QT (Y), but because it may be in LD with G_1 or Z, an indirect association with Y may be detected. This makes it difficult to distinguish among the individual contributions, even in the complete data case (see Faye, Machiela, Kraft, Bull, and Sun (2013) for reranking strategies in this scenario). In addition, we note that including an identified GWAS-SNP, Z, in strong LD in the causal variant regression model (1) would introduce collinearity and decrease precision, reducing the ability to detect genetic association of a causal variant $\boldsymbol{G}$ with the QT.

2.1 | Sampling designs

Under the sampling design, $\boldsymbol{P}(r_i = 1|Z_i, Y_i, G_i) = \boldsymbol{P}(r_i = 1|Z_i, Y_i) = \pi_i$, with π_i the inclusion probability for the ith subject. Therefore, $\boldsymbol{G}_i$ is missing at random. Observe that this definition allows for inclusion probabilities of the form $\boldsymbol{P}(r_i = 1|Z_i, Y_i) = \pi(Z_i)$ or $\pi(Y_i)$, which we call a *marginal* sampling design as opposed to $\boldsymbol{P}(r_i = 1|Z_i, Y_i) = \pi(Z_i, Y_i)$, which we refer to as a *joint* sampling design.

Among marginal sampling designs, covariate-dependent sampling, that is, $\boldsymbol{P}(r_i = 1|Z_i, Y_i) = \pi(Z_i)$, occurs when covariates of interest are used as surrogates of the expensive covariate to select individuals for the phase 2 subset. Analogously, when the trait of interest is used to select individuals this is called outcome-dependent sampling, response-dependent sampling, or trait-dependent sampling (TDS), that is, $\boldsymbol{P}(r_i = 1|Z_i, Y_i) = \pi(Y_i)$. Among TDS designs, extreme-trait sampling has arguably become the most widely used marginal sampling design. In particular it has been used in rare variant analysis (Derkach et al., 2015; Li, Lewinger, Gauderman, Murcray, & Conti, 2011; Lin et al., 2013; Yilmaz & Bull, 2011), because this approach effectively enriches for rare exposure, for example, White (1982). These methods are also applicable in the context of nonrare variants, for example, Satagopan and Elston (2003), Wang, Thomas, Pe'er, and Stram (2006), and Thomas et al. (2009) propose trait-dependent phase 2 sample allocations and discuss related issues in the context of GWAS. In the following paragraphs, we describe the classes of marginal sampling designs we examine in this work.

For marginal GWAS-SNP sampling, strata are naturally defined by the number of copies of the minor allele carried by individuals in the study. The expected counts assuming Hardy-Weinberg equilibrium (HWE) for a GWAS-SNP with MAF q_Z are given by $N \times ((1-q_Z)^2, 2(1-q_Z)q_Z, q_Z^2) = (N_{0\cdot}, N_{1\cdot}, N_{2\cdot})$. The fact that Z is a discrete covariate is important in our formulation to ensure existence of the maximum likelihood estimate (MLE) (van der Vaart & Wellner, 2001). Further extensions could also include more than one SNP genotype as stratification factors, for example, Schaid et al. (2013).

TABLE 1 Joint distribution of the GWAS-SNP, Z, and the discretized version of the QT, Y_{st}

$\boldsymbol{Z \backslash Y_{st}}$	**T1**	**T2**	**T3**	$\mathbf{M}_Z$
0	N_{01}	N_{02}	N_{03}	$N_{0\cdot}$
1	N_{11}	N_{12}	N_{13}	$N_{1\cdot}$
2	N_{21}	N_{22}	N_{23}	$N_{2\cdot}$
M_Y	$N_{\cdot 1}$	$N_{\cdot 2}$	$N_{\cdot 3}$	N

To operationalize the QT sampling, we discretize the QT (Y) into a three strata variable Y_{st} with labels (T1, T2, T3). Let (C_1, C_2) be fixed cut-off values of Y that partition the QT as follows:

$$Y_{st} \in \begin{cases} \mathrm{T1} & \text{if } Y < C_1 \\ \mathrm{T2} & \text{if } C_1 \leq Y < C_2 \\ \mathrm{T3} & \text{if } Y \geq C_2. \end{cases} \tag{2}$$

In applications, (C_1, C_2) are prespecified quantities, for example, relevant in clinical practice, such as systolic and diastolic blood pressure level categorization into hypertensive, prehypertensive, and normotensive. An important consideration for the cut points (C_1, C_2) is that these values are not determined from the data at hand; otherwise dependency among observations is introduced through the inclusion probabilities.

Joint sampling designs, on the other hand, use both response and covariate simultaneously to select individuals for the phase 2 subset. Table 1 illustrates the joint distribution of the discretized QT and the GWAS-SNP along with the marginal distributions. The goal of joint allocation in this case is to select individuals for the phase 2 subset based on the nine strata determined by the QT categories and the SNP genotype. In discretizing Y, the choice of three strata is the simplest, but is straightforward to extend. In general, for marginal and joint sampling designs, strata are determined by partitioning (Z, Y) into K groups.

3 | A SEMIPARAMETRIC MAXIMUM LIKELIHOOD APPROACH

We extend the SPML formulations described in Zhao et al. (2009) and Lin et al. (2013). Let $f_\theta(y|g)$ be the functional (parametric) relationship between $\boldsymbol{G}$ and Y indexed by $\theta = (\beta_0, \beta_1, \sigma^2)^T$; in our case, $f_\theta(y|g)$ corresponds to the probability density function of a normal distribution with parameters (μ_g, σ^2), $\mu_g = \beta_0 + \beta_1 g$. On the other hand, we denote $\mathscr{G}$, $\mathscr{Z}$ as the sets of unique observed values of $\boldsymbol{G}$ (in S_2) and Z (in $S_2 \cup \bar{S}_2$). Let $p(G, Z)$ be the joint probability function of $\boldsymbol{G}$ and Z given by the discrete probabilities $p_{g,z}$, $g \in \mathscr{G}$, $z \in \mathscr{Z}$, which we estimate nonparametrically. The proposed method differs from the formulations of Zhao et al. (2009) and Lin et al. (2013) in three aspects: conditional independence assumption between Y and Z given G, which leads to exclusion of Z in the linear predictor of the parametric trait model $f_\theta(y|g)$; use of Z in nonparametric estimation of the joint distribution of G and Z, with support for (g, z) defined by the Cartesian product; and the individual weights calculation in the E-step of the EM algorithm. We detail similarities and differences among these methods in Appendix A (supplementary material).

3.1 | Likelihood formulation

Considering the above, we define the observed data likelihood following Robins, Hsieh, and Newey (1995) and Lawless et al. (1999) by:

$$L(\theta, \boldsymbol{p}) = \prod_{i=1}^{N} \left[\pi_i f_\theta(y_i|g_i) p(g_i, z_i)\right]^{r_i} \times \left[\{1 - \pi_i\} \sum_g f_\theta(y_i|g) p(g, z_i)\right]^{1-r_i}. \quad (3)$$

Because SPML estimation of θ and observed information matrix calculation does not involve π_is, we can disregard such terms, leading to:

$$\begin{aligned} L(\theta, \boldsymbol{p}) &\propto \prod_{i=1}^{N} \left[f_\theta(y_i|g_i) p(g_i, z_i)\right]^{r_i} \times \left[\sum_g f_\theta(y_i|g) p(g, z_i)\right]^{1-r_i} \\ &= \prod_{i\in S_2} f_\theta(y_i|g_i) p(g_i, z_i) \prod_{i\in \bar{S}_2} \sum_g f_\theta(y_i|g) p(g, z_i). \end{aligned} \quad (4)$$

The loglikelihood takes the form $\ell(\theta, \boldsymbol{p}) = \sum_{i\in S_2} [\log f_\theta(y_i|g_i) p_{g_i,z_i}] + \sum_{i\in \bar{S}_2} \log \left\{\sum_{g\in\mathscr{G}} f_\theta(y_i|g) p_{g,z_i}\right\}$, where we let $p_{g,z} = p(G = g, Z = z)$. Note that subjects in $\bar{S}_2$ have incomplete data (y_i, z_i), whereas phase 2 (S_2) subjects have completely observed data (y_i, z_i, g_i). Our formulation uses the auxiliary covariate, Z, only in the nonparametric part as Y and Z are assumed to be conditionally independent given G. When Z is a non-causal surrogate for $G = G_1$, its inclusion in the regression may adversely affect model performance due to collinearity between Z and G. Direct maximization of (4) is difficult so in the next section we describe the EM algorithm used to maximize the observed-data loglikelihood.

3.2 | EM algorithm

We apply the EM algorithm to estimate θ and $p_{g,z}$, $g \in \mathscr{G}$, $z \in \mathscr{Z}$. We specify initial values in a similar fashion as Lin et al. (2013), that is, $\beta_0^{(0)} = \beta_1^{(0)} = 0$, $\sigma^{2(0)} = N^{-1} \sum_{i=1}^{N} (Y_i - \overline{Y})^2$, except for $p_{g,z}^{(0)}$ which we specify by $p_{g,z}^{(0)} = \begin{cases} 1/m & \text{if } (g, z) \in \mathscr{G} \times \mathscr{Z}(S_2) \\ 0 & \text{otherwise} \end{cases}$ for $g \in \mathscr{G}$, $z \in \mathscr{Z}$; $\mathscr{G} \times \mathscr{Z}(S_2)$ denotes the set of cardinality m containing the different pairs (g, z) observed in S_2. Further, we define $\mathscr{Z}_2$ as the set of different z's in S_2. Due to some of the allocation designs, it is possible that $\mathscr{Z}_2$ will not contain all the values of Z observed in phase 1.

E-step. Let $\ell_c(\theta, \boldsymbol{p})$ be the complete data loglikelihood, then

$$\begin{aligned} Q&\left[(\theta, \boldsymbol{p})|(\theta^{(t)}, \boldsymbol{p}^{(t)})\right] \\ &= \mathrm{E}\left\{\ell_c(\theta, \boldsymbol{p})|\mathrm{Obs}; \theta^{(t)}, \boldsymbol{p}^{(t)}\right\} \\ &= \mathrm{E}\left\{\sum_{i=1}^{N} \log\left[f_\theta(Y_i|G_i) p(G_i, Z_i)\right] |\mathrm{Obs}; \theta^{(t)}, \boldsymbol{p}^{(t)}\right\} \\ &= \sum_{i=1}^{N} \sum_{g\in\mathscr{G}} \log\left[f_\theta(Y_i|G_i = g) p(G_i = g, Z_i)\right] \\ &\quad \times Pr(G_i = g|\mathrm{Obs}; \theta^{(t)}, \boldsymbol{p}^{(t)}), \end{aligned}$$

where

$$Pr(G_i = g|\mathrm{Obs}; \theta^{(t)}, \boldsymbol{p}^{(t)}) = \omega_{i,g}^{(t)}$$

$$= \begin{cases} 1 & \text{if } i \in S_2, g_i = g \\ \dfrac{f_{\theta^{(t)}}(y_i|g) p_{g,z_i}^{(t)}}{\sum_{g'\in\mathscr{G}} f_{\theta^{(t)}}(y_i|g') p_{g',z_i}^{(t)}} & \text{if } i \in \bar{S}_2, z_i \in \mathscr{Z}_2 \\ \dfrac{f_{\theta^{(t)}}(y_i|g) \sum_{z\in\mathscr{Z}} p_{g,z}^{(t)}}{\sum_{g'\in\mathscr{G}} f_{\theta^{(t)}}(y_i|g') \sum_{z\in\mathscr{Z}} p_{g',z}^{(t)}} & \text{if } i \in \bar{S}_2, z_i \notin \mathscr{Z}_2 \\ 0 & \text{otherwise.} \end{cases}$$

M-step. We update the estimates' values as follows:

$$\boldsymbol{\beta}^{(t+1)} = \left(\sum_{i=1}^{N} \sum_{g\in\mathscr{G}} \omega_{i,g}^{(t)} X_g \otimes X_g\right)^{-1} \left(\sum_{i=1}^{N} y_i \sum_{g\in\mathscr{G}} \omega_{i,g}^{(t)} X_g\right)$$

$$\sigma^{2(t+1)} = N^{-1} \sum_{i=1}^{N} \sum_{g\in\mathscr{G}} \omega_{i,g}^{(t)} \{y_i - \boldsymbol{\beta}^{(t+1)T} X_g\}^2$$

$$p_{g,z}^{(t+1)} = N^{-1} \sum_{i=1}^{N} \omega_{i,g}^{(t)} \mathbb{1}\{z_i = z\},\ g \in \mathscr{G},\ z \in \mathscr{Z},$$

where $X_g = (1, g)^T$ and $\otimes$ is the outer product. The algorithm iterates between the E and M steps until convergence is achieved, that is, $\max(|\theta^{(t+1)} - \theta^{(t)}|) < 1 \times 10^{-5}$ and

$\max(|\boldsymbol{p}^{(t+1)} - \boldsymbol{p}^{(t)}|) < 1 \times 10^{-5}$), yielding the maximum likelihood estimates, MLEs, $(\hat{\theta}, \hat{\boldsymbol{p}}) = (\hat{\boldsymbol{\beta}}^T, \hat{\sigma}^2, \hat{\boldsymbol{p}})$.

For hypothesis testing, Wald score and likelihood ratio (LR) tests can be constructed following standard procedures (Louis, 1982). We elaborate on the construction details of these tests in Appendix B (supplementary material). Furthermore, concerning the LR, Murphy and van der Vaart (2000) point out that the usual full LR statistic fails in semiparametric models. To overcome this, they argue in favor of using the profile LR in a semiparametric framework, which is defined as $\Lambda_p = \frac{L_p(\hat{\theta})}{L_p(\tilde{\theta})} = \frac{\sup_{\theta, \boldsymbol{p}_G} L(\theta, \boldsymbol{p}_G)}{\sup_{\boldsymbol{p}_G} L(\tilde{\theta}, \boldsymbol{p}_G)}$, where $\boldsymbol{p}_G = \left\{\sum_{z \in \mathscr{Z}} p_{g,z} : g \in \mathscr{G}\right\}$. Hence, Λ_p has corresponding test statistic $D_p = 2 \ln \Lambda_p$ asymptotically distributed as χ^2_1. For the profile LR statistics, we observed in our simulations that the following simplification rendered adequate results for testing $H_0 : \beta_1 = 0$, $D'_p = 2\left\{\ell(\hat{\theta}, \hat{\boldsymbol{p}}_G) - \ell(\tilde{\theta}, \hat{\boldsymbol{p}}_G)\right\}$, which is effectively substituting the nonparametric estimates of the restricted MLEs, $\tilde{\boldsymbol{p}}_G$, by the unrestricted MLEs. Justification for this substitution under SPML estimation, in the simplest case without Z in the trait model, is that estimates for $\boldsymbol{p}_G$ were the same under the genetic association alternative and null leading to acceptable profile LR tests. A precedent on related substitutions using profile likelihoods has been discussed in Murphy and van der Vaart (2000); Pace, Salvan, and Ventura (2011).

4 | PHASE 2 SAMPLE ALLOCATION

Despite the extensive literature on two-phase designs and estimation approaches, relatively less attention has been paid to allocation of the phase 2 sample across defined strata. We first review some of the approaches that have been applied to draw phase 2 data, and then specify alternative approaches that we investigate in simulations and application.

4.1 | Background

Several authors have investigated two-phase designs that analyze QTs (continuous outcomes) in general settings with covariates. Lawless et al. (1999) explore allocations with equal phase 2 sample sizes within strata in Y-dependent sampling, that is, balanced on a discretized version of Y. Chatterjee et al. (2003) present a pseudoscore estimator and study three phase 2 allocations using a joint (discretized) outcome-covariate strata definition for dichotomous Y and X with selection probabilities $\boldsymbol{\pi} = \{\pi(Y, X)\}$, namely (a) simple random sampling with $\pi(Y, X)$ constant, (b) stratified sampling with $\pi(Y, X) > 0$ for each (Y, X), or (c) restricted sampling with $\pi(1, X) = 0$ but $\pi(0, X) > 0$. Song et al. (2009) propose semiparametric efficient inference with selection of phase 2 data via outcome-dependent sampling with the following allocations: (a) a sample taken from the trait distribution tails, or (b) all subjects from the two trait distribution tails, both augmented with a simple random sample drawn prior to the extreme-trait selection. Zhao et al. (2009) consider similar outcome-dependent sampling with an allocation based on defining three strata for Y as $Y < C_1$, $C_1 \leq Y < C_2$, and $Y \geq C_2$ such that $Pr(Y < C_1) = Pr(Y \geq C_2) = 0.05$. Then, phase 2 sampling probabilities are assigned values 1, 0.056, and 1 for the corresponding three strata, that is, all subjects from the tails of the distribution plus a random sample from the middle. Zhou, Wu, Liu, and Cai (2011) study "outcome-auxiliary-dependent sampling" for the case when there is a continuous auxiliary covariate; phase 2 data are selected from a mixture of a simple random sample of size n_0 and a sample taken from the four extreme strata derived from discretizing the outcome and auxiliary covariates in tertiles.

In the genetic analysis literature, two-phase designs have specified QT and GWAS-SNP allocations in various ways. Yilmaz and Bull (2011) consider a distance-based sampling function for extreme-trait values, that is, allocating only observations in the tails of the QT distribution, and compare it to simple random sampling; Li et al. (2011) study allocations based on extreme-phenotype sampling of the tails, that is, $Y < C_2$ or $Y > C_1$ and almost-extreme sampling, which is similar, but removes the very extremes beyond C_3 and C_4, that is, $C_4 < Y < C_2$ or $C_3 > Y > C_1$; Lin et al. (2013) examine another version of extreme sampling in which they select a predefined number of the highest and lowest values of Y while taking a random sample among the remaining values; lastly, Derkach et al. (2015) select phase 2 data under Y-dependent sampling by drawing observations from $S_1 \cup S_2$, where $S_1 = \{i : Y_i < C_l\}$ and $S_2 = \{i : Y_i > C_u\}$, with C_l and C_u determined so that $Pr(Y < C_l) = Pr(Y > C_u) = 0.3$. Notably, these approaches are mostly motivated by rare variant analysis. On the other hand, Chen et al. (2012) examine marginal GWAS-SNP sampling designs and phase 2 sample allocations, whereas Chen, Craiu, and Bull (2014) focus on a fine-mapping scenario using GWAS-SNP as a stratification factor under equal strata size and rare-homozygote stratum enriched allocations. Although there is some literature on joint trait-SNP sampling designs for case-control studies (Schaid et al., 2013; Thomas et al., 2013), to our best knowledge no previous study has focused on joint QT-genotype sampling specifically for genetic fine mapping.

4.2 | Alternative allocations

In this section, we specify marginal and joint sampling designs for various phase 2 sample allocations and focus on allocations for joint QT-SNP-dependent sampling. To make comparisons between the joint and marginal sampling designs, we define the marginal allocations as sums over the joint strata margins. Let $n_{I_z, I_y} = \#\left\{i : Z_i \in I_z, Y_i \in I_y\right\}$ be the number of subjects to be allocated in each stratum for

TABLE 2 Example sample allocations for phase two sample size $n = 2,500$, $N = 5,000$, $\mathrm{r}^2 = 0.75$, $q_G = 0.2$, $q_Z = 0.3$ under joint QT-GWAS-SNP-dependent sampling for allocations (A) proportional to size, (B) extreme on Z and Y_{st}, (C) balanced on Z and Y_{st}, and (D) balanced on Z and extreme on Y_{st}

	(A) pps			(B) extreme			(C) balanced			(D) combined		
$Z\backslash Y_{st}$	T1	T2	T3	T1	T2	T3	T1	T2	T3	T1	T2	T3
0	491	236	477	625	0	625	278	277	278	416	0	416
1	384	204	463	0	0	0	278	278	278	417	0	417
2	93	42	111	625	0	625	278	277	278	417	0	417

Adding up by row or column leads to allocations under marginal QT- and GWAS-SNP- dependent sampling, respectively.

the joint sampling design. Consequently, the marginal allocations are determined by $n_{I_z} = \sum_{I_y} n_{I_z,I_y}$ and $n_{I_y} = \sum_{I_z} n_{I_z,I_y}$, where I_z, I_y are intervals defining the strata. Let ϱ be the overall sampling fraction, where we want to draw a subsample of size $n = \varrho * N$, such that $\sum_{I_z,I_y} n_{I_z,I_y} = n$. As described in Lawless et al. (1999), variable probability sampling (VPS) and basic stratified sampling (BSS) are popular choices to select subjects in the phase 2 sample. In VPS, n_{I_z,I_y} is considered a random variable, as units are selected with specified probability p_{I_z,I_y}. BSS, on the other hand, considers n_{I_z,I_y} as fixed quantities given N_{I_z,I_y}. Thus, for each defined stratum, n_{I_z,I_y} subjects are selected with equal probability. VPS often makes more sense in the context where subjects are assigned to phase 2 dynamically as a consequence of data collection over time. In this report, we use BSS because strata can be defined a priori from the GWAS. However, full likelihood methods as well as the proposed semiparametric method are capable of handling various sampling schemes (Derkach et al., 2015).

In the following, we define four possible allocations for the regional sequencing. Table 2 illustrates possible sample allocations under the joint QT-SNP-dependent sampling. In the simulation studies, we compare statistical efficiencies for each allocation, and associated hypothesis testing properties.

4.2.1 | Proportional to stratum size

This allocation aims to preserve the strata distribution structure in the complete data. In this allocation scenario, we simply draw subsamples within strata of size $n_{jl} = \varrho * N_{jl}$, $l = 0, 1, 2;\ j = 1, 2, 3$. This is the only case in which the phase 2 sample has the same distribution (over Y and Z) as the complete data, but it may be inefficient.

4.2.2 | Extreme

As hinted by its name, extreme allocation aims to distribute the sample to the extreme values of Z and/or Y_{st}. In the joint sampling case, the samples are taken from extreme valued strata in Table 1, $(Z, Y_{st}) = \{(0, \mathrm{T1}), (0, \mathrm{T3}), (2, \mathrm{T1}), (2, \mathrm{T3})\}$. Strata T1 and T3 are defined in Equation (2). The simplest case is to allocate the same sample size across the four extreme strata, thus $n_{jl} = \begin{cases} \min\{\frac{\varrho * N}{4}, N_{jl}\}, & j = 0, 2; l = 1, 3 \\ 0 & \text{otherwise.} \end{cases}$ Observe that the min function is necessary because the number of subjects N_{jl} can be smaller than $n_{jl} = \varrho * N/4$ especially when q_Z is low. This allocation oversamples those strata believed to be most informative for the QT and the extreme genotype categories for the GWAS-SNP.

4.2.3 | Balanced

Here, allocation on Z and Y_{st} specifies the same number of subjects per strata, therefore $n_{jl} = \min\left\{\frac{\varrho * N}{9}, N_{jl}\right\}$, $j = 0, 1, 2;\ l = 1, 2, 3$. Once again, the min function is necessary to avoid empty cells but this occurs less frequently than in the extreme case. This allocation may be preferred when one still wants to oversample the most informative strata without losing all information from other strata.

4.2.4 | Combined

This allocation combines balanced selection in Z and extreme selection in Y_{st}. It reflects what has been previously reported in the literature as useful strategies to select subjects marginally (Chen et al., 2012; Derkach et al., 2015; Lin et al., 2013; Schaid et al., 2013; Thomas et al., 2013). In this case

$$n_{jl} = \begin{cases} \min\{\frac{\varrho * N}{6}, N_{jl}\}, & j = 0, 1, 2; l = 1, 3 \\ 0 & \text{otherwise.} \end{cases}$$

5 | SIMULATION STUDIES

5.1 | Design

We conduct simulation studies in which the GWAS-SNP, Z, is indirectly associated with the QT of interest, Y, through LD with a causal SNP, G_1 (Fig. 1). The correlated SNPs G_1 and Z are randomly drawn from a haplotype with fixed MAFs and LD (r^2) values assuming HWE. The generating model for the QT is given by: $Y = \beta_0 + \beta_1 G_1 + \epsilon$, $\epsilon \sim N(0, \sigma^2)$. To mimic the post-GWAS scenario, we keep only those replicates that achieve suggestive genome-wide significance, that is, P-value $< 1 \times 10^{-5}$ for the parameter γ_1 in the regression $Y = \gamma_0 + \gamma_1 Z$; we proceed until R such replicates are drawn, discarding those that do not achieve significance. Table 3 displays the simulation design parameters, and corresponding values. Genetic effects under the alternative $\beta_1 > 0$ were chosen to achieve roughly 100% power in the complete data.

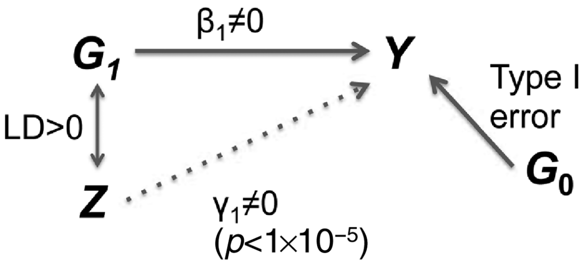

FIGURE 1 Fine-mapping scenario specified in the simulation setup

Note: A causal variant G_1 has a linear effect β_1 on the QT, Y. Z is indirectly associated with the QT; this association may be detected through a GWAS. The indirect assocation arises from the LD structure between Z and G_1. Type I error is studied through an unrelated (with any of Y, G, or Z) SNP, G_0.

TABLE 3 Simulation design parameters and values. The number of replicates yields precision of $\pm 4.27 \times 10^{-3}$ for empirical type I error (at 5%) and $\pm 2.48 \times 10^{-3}$ for empirical power (at 80%)

Design parameter	Value(s)
Replicates (R)	10,000 (studied null) and 1,000 (H_a)
Study sample size (N)	5,000
MAFs (q_G, q_Z)	0.2, 0.3; 0.2, 0.2
LD (r^2)	0.5; 0.75
β_0	2
β_1	0.25
σ^2	2.25
Phase-two sample size (n)	540; 1,000; 2,500
Phase-two sample allocation	(A) Proportional to stratum size (pps)
	(B) Extreme on both Z and Y_{st} (extreme)
	(C) Equal number in each stratum (balanced)
	(D) Balanced on Z and extreme on Y_{st} (combined)

For the case of no genetic effect, we note that trait values simulated under the null hypothesis $\beta_1 = 0$ when $G = G_1$ would lead only to false positives under GWAS screening on Z. Therefore, to evaluate type I error in the fine mapping setting, we retain the trait values generated under the alternative hypothesis, but instead generate a random null SNP, G_0 with the same MAF (q_G) as the causal G_1, but uncorrelated with Z. We then analyze the null SNP, G_0 in a similar fashion as the causal SNPs G_1.

We examine three sampling designs to select individuals into S_2, namely (1) marginal SNP-dependent sampling (M_Z), (2) marginal QT-dependent sampling (M_Y), and (3) joint QT-SNP-dependent sampling ($J_{Z,Y}$). For (2) and (3), we discretize the QT, Y, into three-strata Y_{st} (T1, T2, T3) according to fixed cut points (C_1, C_2) as the percentiles ($2/5, 3/5$) of a normal distribution with mean $\mu_Y = 2$ and variance $s_Y^2 = 2.25$, that is, under the null $\boldsymbol{P}(Y < C_1) = \boldsymbol{P}(Y > C_2) = 0.4$. These values allow for a range of sampling variability in the simulations specially for the extreme allocations. This contrasts with other extreme-trait allocations, e.g., Lin et al. (2013); Zhao et al. (2009), in which the sampling probabilities for the extreme strata are 1 (or close to 1). We explore a more extreme definition of the strata in Appendix C of the supplementary material.

We evaluate three phase 2 (S_2) sample sizes determined by overall sampling fractions of 10.8%, 20%, and 50% of the phase 1 sample. Under each design, we examine the sample allocations described in Section 4: namely (A) proportional to stratum size (pps), (B) extreme allocation on Z and Y_{st} (extreme), (C) equal numbers across Z and Y_{st} (balanced), and (D) balanced on Z and extreme on Y_{st} (combined). When sampling selection of the desired number of subjects for a particular stratum is infeasible due to an insufficient number of observations in that particular stratum, we select up to the maximum number of subjects in the small stratum, and then reallocate to the remaining strata to achieve the specified S_2 sample size.

We assess test validity and power for the Wald, score and LR statistics as described in Section 3, specifying test sizes of $\alpha = 0.05$ under the null, and $\alpha = 5 \times 10^{-8}$ under the alternative. In addition, we evaluate bias, empirical/average (model-based) standard errors and relative efficiency via root mean square error (RMSE) of $\hat{\beta}_1$. We calculate relative efficiency with the complete data RMSE as denominator.

The simulation studies are designed to address the following questions: identify allocations where $J_{Z,Y}$ shows improved power; identify design/allocation combinations with good power ($J_{Z,Y}$ vs. M_Z or M_Y); compare power of SPML versus alternative methods; and evaluate test statistics for the proposed SPML estimation (score vs. Wald vs. LR). In the next subsection, we highlight the main results on power comparisons: among allocations, between joint versus marginal designs, and between estimation methods including (1) analysis of phase 2 data only (denoted as S_2 alone) and (2) TDS described in Lin et al. (2013) (with and without Z in the formulation). We further discuss related issues including test validity under the studied S_2 sample sizes, estimation bias and variance, relative efficiency, and effect of GWAS and phase 2 sample sizes versus sampling fractions.

5.2 | Results

Under joint QT-SNP-dependent sampling, combined allocations yield higher power than alternative allocations (pps, extreme, balanced) (Table 4) while maintaining type 1 error (T1E) control (Fig. 2). We observe no differences between

TABLE 4 Empirical power (in percentages) for 1,000 replicates under the score test for a phase two sample size of $n = 2,500$, $N = 5,000$

		Comp.		(A) pps			(B) extreme			(C) balanced			(D) combined		
MAFs	LD (r^2)	Data	Method	$\mathbf{M}_Z$	$\mathbf{M}_Y$	$\mathbf{J}_{Z,Y}$	$\mathbf{M}_Z$	$\mathbf{M}_Y$	$\mathbf{J}_{Z,Y}$	$\mathbf{M}_Z$	$\mathbf{M}_Y$	$\mathbf{J}_{Z,Y}$	$\mathbf{M}_Z$	$\mathbf{M}_Y$	$\mathbf{J}_{Z,Y}$
0.2, 0.3	0.5	99.4	JTCsamp	84.4	84.8	85.7	84.6	92.9	90.3	90.3	78.1	85.1	87.8	92.0	94.3
			S_2 alone	39.7	42.8	40.9	53.8	67.3	63.2	62.1	27.3	50.1	55.6	69.7	66.1
	0.75	97.0	JTCsamp	84.2	85.6	85.7	85.0	90.7	88.1	90.2	83.0	87.7	88.7	89.1	91.8
			S_2 alone	31.8	29.3	26.6	41.4	53.7	36.8	56.5	19.8	41.4	45.4	53.9	44.4
0.2, 0.2	0.5	98.4	JTCsamp	83.4	84.5	84.9	85.8	90.1	91.8	89.6	79.3	85.3	89.3	90.7	93.8
			S_2 alone	42.8	43.3	41.0	52.3	66.0	47.3	59.7	29.8	31.6	57.8	68.1	60.9
	0.75	97.7	JTCsamp	86.7	85.0	85.1	87.0	90.2	91.5	89.3	83.4	87.8	88.8	91.7	93.0
			S_2 alone	29.9	29.8	25.9	42.3	54.6	16.5	52.0	19.9	12.1	45.1	52.5	32.4

Power is calculated at a nominal level of $\alpha = 5 \times 10^{-8}$. For sampling designs: marginal GWAS-SNP-dependent sampling ($\mathbf{M}_Z$), marginal QT-dependent sampling ($\mathbf{M}_Y$), and joint QT-GWAS-SNP-dependent sampling ($\mathbf{J}_{Z,Y}$); and allocations (A) proportional to stratum size, (B) extreme on Z and Y_{st}, (C) balanced on Z and Y_{st}, and (D) balanced on Z and extreme on Y_{st}. Column for complete data (Comp. Data) for comparison purposes.

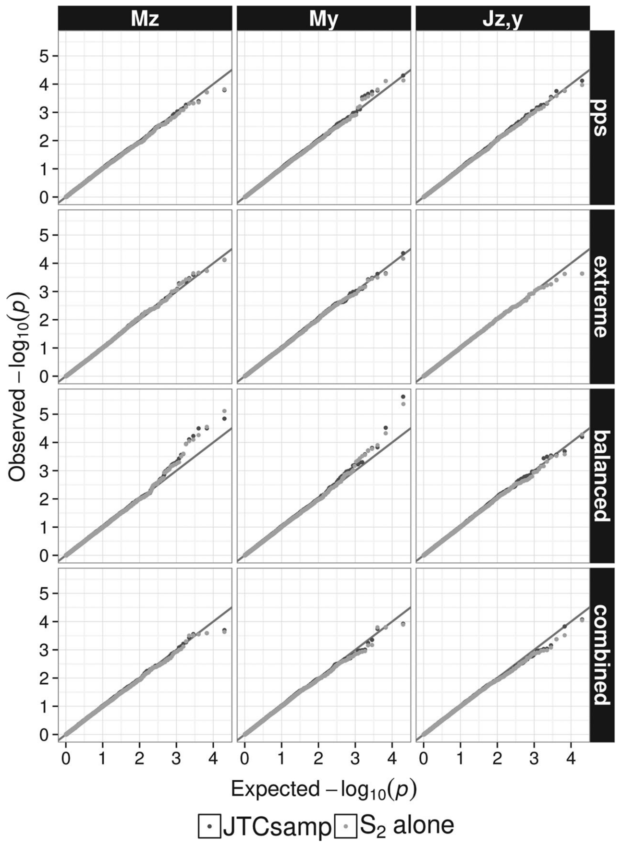

FIGURE 2 Quantile-quantile plots of $-\log_{10}$ (P values) for testing β_1 under the type 1 error scenario (G_0) across 10,000 replicates with a phase 2 sample size of 2,500; we compare analyses of JTCsamp and S_2 data alone

Notes: Each facet represents a sampling design and allocation combination. LD (r^2) is fixed at 0.75 and MAFs fixed at q_G=0.2, q_Z=0.3.

marginal and joint sampling designs within the pps allocation. Marginal dependent samplings show better power under the extreme (marginal QT) and combined (marginal SNP) allocations. Joint QT-SNP-dependent sampling consistently exhibits better power compared to marginal QT- or SNP-dependent sampling under a combined allocation (Table 4). The proposed SPML estimation under joint trait-covariate dependent sampling method (hereafter JTCsamp) compares favorably to two alternative methods: (1) S_2 alone and (2) our implementation of TDS. We observe substantial power increases for JTCsamp compared to S_2 data alone under all allocations (Table 4). TDS (with and without Z) and analysis of phase 2 data alone yield generally similar results for T1E and power with a phase 2 sample size of $n = 2,500$ (supplementary Table S1). On the other hand, as anticipated, including Z in the trait model adversely affects the power to detect association between the causal variant, $\boldsymbol{G}$, and the QT, Y (supplementary Table S1). This is not surprising mainly due to collinearity derived from the underlying LD between Z and $\boldsymbol{G}$. Overall score, Wald and LR statistics tests show similar T1E and power patterns (supplementary Figs. S1 and S2).

Regarding test validity, quantile-quantile plots of $-\log_{10}$($\boldsymbol{P}$ values) for testing $\beta_1 = 0$ in JTCsamp and S_2 alone when phase 2 sample size is $n = 2,500$ (Fig. 2) do not exhibit gross departures from the expected distribution (see also Table S2). Likewise, there are no signs of more liberal T1E rates in JTCsamp compared to the complete data case (supplementary Fig. S3). In analysis of smaller phase 2 sample sizes, we observe liberal T1E that decreases with sample size increments (see supplementary Figs. S4 and S5); we believe that small stratum-specific sample sizes are driving these results.

For pps and balanced allocations, effect estimation biases in JTCsamp and S_2 data alone are similar (supplementary Table S3), although these differences decrease with increasing LD. However, for extreme and combined allocations, these larger biases are only observed in the marginal SNP sampling, whereas marginal QT and joint QT-SNP exhibit similar biases to the complete data. This latter bias arises from GWAS screening on Z, a well-known phenomenon in genetic association studies called the “winner's curse.” In addition, in almost all cases, empirical and mean model-based standard errors for JTCsamp are smaller than those for S_2 alone (supplementary Table S3). JTCsamp has higher relative efficiency than analysis of phase 2 data alone (supplementary Table S4). Extreme and combined allocations have the largest relative efficiencies; further, under a combined allocation, joint QT-SNP-dependent sampling is more efficient than marginal SNP or QT sampling. In addition, RMSE decreases with stronger Z-G_1 LD (supplementary Table S4).

To understand to what extent the overall sampling fraction or stratum-specific sizes affect hypothesis testing, we repeated the simulations with $N = 10,000$, $n = 2,500$, and MAFs of $q_G = 0.2$ and $q_Z = 0.3$ (supplementary Table S5). These simulations show similar patterns as those in Table 4, suggesting that it is the stratum-specific sample size rather than the sampling fraction that is important in determining test validity.

6 | APPLICATION TO FINE-MAPPING OF SYSTOLIC BLOOD PRESSURE

The proposed method is intended to narrow down a candidate region by hypothesis testing of multiple variants within the region. To evaluate application under a more realistic multivariant causal model and assess sensitivity to sampling variation, we analyzed data from the Genetic Analysis Workshop 19 (GAW19). The study data included 1,943 unrelated individuals from the T2D-GENES Project 1, each with whole-exome sequencing (WES) by Illumina HiSeq2000 technology. Trait values for systolic and diastolic blood pressure (SBP and DBP) were generated from the WES data under a known polygenic model with causal variants in multiple genes including the MAP4 gene (Blangero et al., 2016) and replicated independently 200 times. Since all four generating "causal" variants are located in chromosome 3, we focus on the 38,102 chromosome 3 sequence variants. Since GWAS data per se are not available, we specify pseudo GWAS SNPs by extracting nonrare noncausal WES variants (MAF>1%) overlapping with a commercial genotype assay (Illumina Human Core Exome-12v1.0). This yields a total of 3,192 chromosome 3 SNPs (48 in the MAP4 region) in the synthetic GWAS, which serves as the phase 1 genotype data.

6.1 | Methods

For each of the 200 phenotypic replicates we perform the following.

First, we perform chromosome-wide scans on the synthetic GWAS data using complete data (N =1,943) to identify the surrogate SNPs. Linear regression of SBP is carried out for each of the genetic variants coded additively as independent covariates. The most significant SNP among the 3,192 (top SNP) is specified as the GWAS-SNP (Z) for phase 2 design and analysis (see supplementary Figs. S6A and S7). The LD structure in the MAP4 sequenced region reveals correlations among the identified GWAS-SNPs and the causal variants (supplementary Fig. S6B).

Second, to implement a two-phase design, we stratify SBP into three groups according to commonly used thresholds: normotensive (SBP<120), prehypertensive (120≤SBP<140), and hypertensive (SBP≥140). We assess three overall sampling fractions in the phase 2 data: 25%, 50%, and 75%, which roughly correspond to n =486, 971, and 1,457 respectively. The observed sample sizes can vary across SNPs in the analyzed region due to missing sequencing data. For each sampling fraction, we draw subjects for the S_2 sample using a joint QT-SNP sampling design under extreme, balanced, and combined allocations with BSS as described in Section 4 (see supplementary Fig. S8 for joint strata distributions in the complete data and the studied allocations). We do not consider pps allocation as it showed the poorest performance in our simulations.

Lastly, we conduct regional association analysis in the MAP4 gene fine-mapped through WES data. Thus, we selected sequence variants flanking 500 kb around this gene. There were 322 variants sequenced in the region (48 nonrare, that is, MAF>1%). These 48 variants constitute the analysis in the two-phase approach, which we carry out one variant at a time. Besides the three studied sampling fractions examined at four allocations, we compare SPML analysis of combined phases 1 and 2 data (JTCsamp) to analysis of phase 2 data only (S_2 alone) and to our TDS implementation (Lin et al., 2013) without Z.

6.2 | Results

Four markers, all within the MAP4 region, meet the genome-wide significant threshold, that is, P-value< 5×10^{-8} in at least one replicate (Table 5); these SNPs serve as the GWAS-SNP in their respective replicate. Region (box)plots of the 200 replicates demonstrate that a joint QT-SNP sampling design under a combined allocation achieves lower P values compared to balanced and extreme allocations for the studied sampling fractions (Fig. 3). Nevertheless, empirical power within allocations may vary across sampling designs, for example, marginal Y design has highest power under an extreme allocation, whereas marginal Z design achieves highest power under a balanced allocation. Additional results considering the best performing sampling design within the examined allocation closely agree with the results displayed in Figure 3; this suggests that among the studied phase 2 allocations, a combined allocation under a joint QT-SNP sampling design achieves lower P values compared to other strategies (supplementary Fig. S9). Summaries of the fine mapping

TABLE 5 Summaries for four SNPs identified as the top GWAS-SNP in at least one replicate

		Top	Mean (min, max) across 200 replicates			Mean (min, max) in top replicates		
GWAS-SNP	MAF	reps.[a]	Estimate	SE	$-\log_{10}(P\text{-value})$	Estimate	SE	$-\log_{10}(P\text{-value})$
chr3:47618953	0.45	136	3.45 (2.2, 4.4)	0.493 (0.47, 0.51)	11.5 (5.1, 17.0)	3.6 (2.8, 4.4)	0.492 (0.47, 0.51)	12.5 (7.5, 17.0)
chr3:47712202	0.34	18	3.06 (2.1, 4.0)	0.517 (0.50, 0.54)	8.53 (4.6, 14.0)	3.57 (3.1, 4.0)	0.515 (0.50, 0.53)	11.3 (8.5, 14.0)
chr3:48309828	0.19	45	−4.04 (−5.3, −2.9)	0.630 (0.61, 0.66)	9.84 (5.7, 16.0)	−4.45 (−5.3, −3.7)	0.630 (0.61, 0.66)	11.8 (8.1, 16.0)
chr3:48360992	0.16	1	−3.80 (−5.1, −2.7)	0.673 (0.65, 0.70)	7.83 (4.1, 13.0)	−4.12 (N/A)	0.685 (N/A)	8.68 (N/A)

[a] Denotes the number of times each SNP was detected as the most significant across 200 replicates.

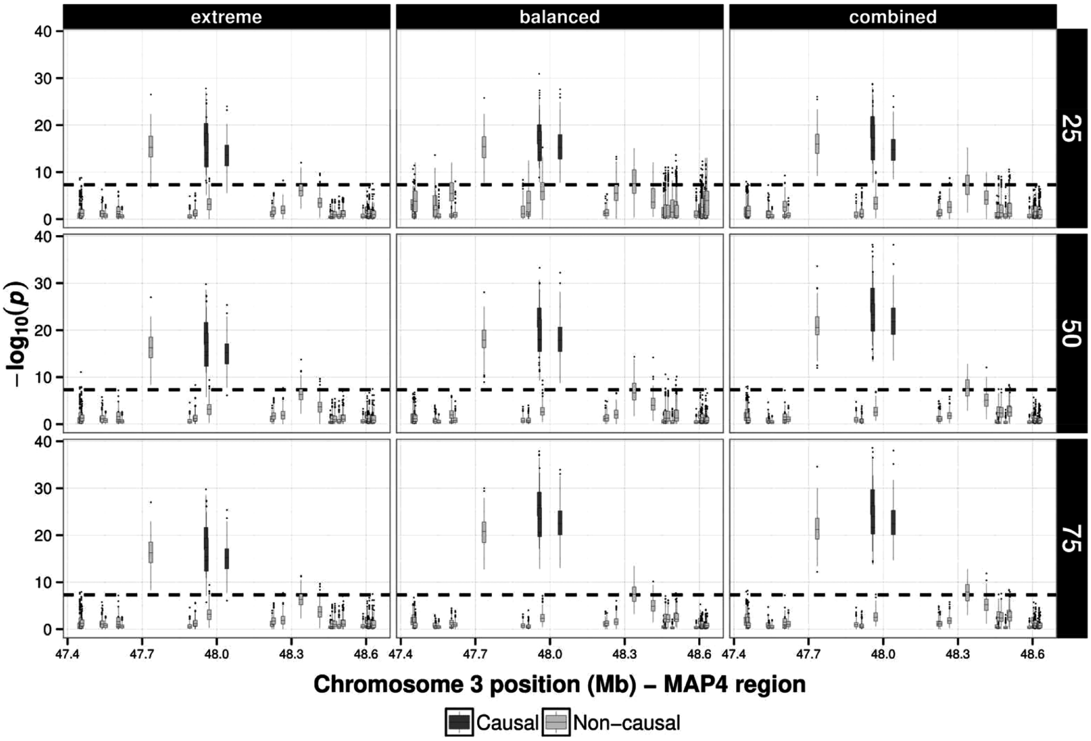

FIGURE 3 Region (box)plots of the distribution of the *P* values, in $-\log_{10}(\cdot)$ scale, across 200 replicates in the GAW19 simulated data under a joint QT-SNP design and JTCsamp analysis calculated using the likelihood ratio statistic (LRS)

Notes: Column facets denote results under extreme, balanced, and combined allocations. Row facets correspond to different overall sampling fractions (in percentages), leading to (approximate) phase 2 sample sizes of 486, 921, and 1,457, respectively. The dashed line represents the genome-wide significance threshold.

analysis of the variants that achieve genome-wide significance under the former strategy show that JTCsamp detected the four causal variants consistently across replicates (Table 6). However, an additional noncausal variant (chr3.47734700) was also detected potentially due to LD structure in the region (see supplementary Fig. S6B).

Comparisons across methods under the joint QT-SNP sampling design and combined allocation in the regional association demonstrate that JTCsamp is more powerful than S_2 and TDS and is more similar to the complete data analysis. Consequently, even at the lowest sampling fraction, JTCsamp is the only method that consistently detected the four causal variants across all replicates. Although we observe some outliers in JTCsamp for noncausal variants at the end of the region, as the sampling fraction increases, the numbers of outliers decrease considerably (supplementary Fig. S10).

Results of the WES fine-mapping application are consistent with the simulation studies: higher power for joint QT-SNP sampling and combined allocation compared to other methods (S_2 alone, TDS). Thus, the proposed method exhibits better agreement with the results of the complete data analysis. Nonetheless, further investigation is warranted in a lower powered setting. Results on sensitivity to phase 1 sampling variation suggest that the extreme and combined allocations are less sensitive to sampling variation, mainly because of the enrichment of the strata of interest that leads to sampling of most (or all) of the subjects in those categories (Appendix C supplementary material).

TABLE 6 Summaries for five sequenced variants identified using JTCsamp under a combined allocation

		Generating		Mean (min, max) across replicates			Emp. SD
Seq. variant	**MAF**	**value**	**Samp. frac**	**Estimate**	**SE**	**$-\log_{10}(P$ -value)**	**of Est.**
chr3.47734700	0.33	0.0	25	−5.57 (−7.4, −4.1)	0.656 (0.59, 0.77)	17.0 (7.6, 28.0)	0.555
			50	−5.41 (−7.1, −4.0)	0.566 (0.27, 0.61)	21.4 (12.0, 110.0)	0.473
			75	−5.41 (−7.2, −4.0)	0.561 (0.50, 0.59)	21.4 (12.0, 35.0)	0.475
chr3.47956424	0.34	−16.2	25	−6.01 (−7.7, −4.5)	0.636 (0.56, 0.73)	20.7 (12.0, 31.0)	0.592
			50	−6.00 (−7.4, −4.6)	0.553 (0.53, 0.59)	26.8 (16.0, 39.0)	0.471
			75	−6.00 (−7.3, −4.6)	0.547 (0.52, 0.57)	27.4 (17.0, 39.0)	0.475
chr3.47957996	0.02	−18.2	25	−20.2 (−29.0, −15.0)	2.28 (1.90, 2.80)	18.8 (8.6, 39.0)	2.51
			50	−18.7 (−26.0, −14.0)	2.00 (1.80, 2.20)	20.3 (12.0, 38.0)	1.91
			75	−18.7 (−26.0, −14.0)	1.97 (1.80, 2.10)	20.7 (12.0, 38.0)	1.87
chr3.47958037	0.31	-1.0×10^{-5}	25	−5.97 (−7.7, −4.3)	0.666 (0.66, 0.77)	18.7 (9.6, 30.0)	0.613
			50	−5.84 (−7.5, −4.3)	0.574 (0.47, 0.61)	23.8 (14.0, 36.0)	0.500
			75	−5.85 (−7.5, −4.3)	0.567 (0.26, 0.60)	24.9 (14.0, 160.0)	0.509
chr3.48040283	0.03	−20.7	25	−18.4 (−25.0, −13.0)	2.10 (1.70, 2.70)	18.4 (8.6, 38.0)	2.38
			50	−16.9 (−24.0, −13.0)	1.81 (1.70, 2.00)	20.1 (12.0, 37.0)	1.88
			75	−16.9 (−23.0, −13.0)	1.79 (1.70, 1.90)	20.6 (13.0, 37.0)	1.84

Generating values for the multivariate model used for SBP were provided by GAW19 data simulators (Blangero et al., 2016). Sampling fractions are in percentages. Interestingly, we identified one locus that is not in the list of "causal" variants, chr3:47734700, this can be explained by the LD structure in the MAP4 region (supplementary Fig. S6B) where we can observe that chr3:47734700 is in high LD with two of the causal variants: chr3:47956424 and chr3:47958037 (see Shin, Yi, and Bull, 2016, for details on this phenomenon).

7 | DISCUSSION

Two-phase joint QT-SNP sampling designs can be more powerful and less sensitive to sample variation than the marginal counterparts provided the proper allocation is chosen. In particular, we find that under a combined allocation, that is, extreme trait and balanced genotype, a joint QT-SNP design exhibits better power compared to marginal designs. An important advantage of the SPML estimation is that the extreme and combined sample allocations we explored can be handled under this framework unlike, for instance, an approach based on inverse probability weighted estimating functions. The EM algorithm described in Section 3 can accommodate additional nongenetic covariates in the regression model by assuming $f_\theta(y|g,w) = f_\theta(y|\mu_{g,w})$, where w is a vector of covariates with $\mu_{g,w} = \beta_0 + \beta_1 g + \gamma^T w$; and $\theta = (\beta_0, \beta_1, \gamma^T, \sigma^2)^T$. Then X_g is replaced with $X_{i,g} = (1, g, w_i^T)^T$ in the *M-step*, for $i = 1, \ldots, N$, where w_i is the ith subject's vector of covariates, these additional covariates are observed across all subjects in both phases 1 and 2 data. Note that independence between G and W is implicitly assumed. Because JTC analysis is likelihood based, credible intervals can be constructed using Bayes factors, which may be useful for comparison with other fine mapping methods (see Maller et al. (2012) for an illustration of this approach)

The simulation studies show that JTCsamp outperforms TDS in terms of power. This may be due to the fact that the latter formulation was designed for association analysis of sequencing data under marginal trait-dependent sampling focusing on set-based analysis of rare variants in hypothesis generating genome-wide analysis. The proposed method fills a gap for those situations where the missing (by design) covariate observation (G) is correlated with the surrogate (and fully observed) covariate Z. It would be of practical interest to explore the impact of low allele counts of Z in the performance of the design and the estimation. Choice of stratification criteria for the QT poses some challenges, including potential losses of information, and it is arguably a limitation of this approach despite its wide usage in two-phase designs. We examine a few reasonable joint allocations that extend existing marginal approaches that have proven useful in the literature. Further work is necessary to determine robust or optimal allocations under budget constraints.

Throughout the paper, we assume additive genetic models for the (surrogate or causal) variants and the QT as these are most often encountered. We expect that a change in the model structure, for example, dominant or recessive, may affect the performances reported here. Past studies (Chen et al., 2014) suggest, however, that the form of the genetic model has a limited impact on the efficiency of the two-phase analysis. The current specification of JTCsamp is yet to incorporate multiple G or multiple Z, which can be a desirable feature in regional sequencing; extensions in this direction are warranted.

The described method can be extended to generalized linear models (GLMs) using available estimation methods by replacing the Gaussian distribution by another member of the exponential family. To do so, the individual weights computed in the *E-step* need to reflect the chosen family distribution. Consequently, the updates for the parameters in the *M-step* cannot be obtained in closed form, requiring iteratively reweighted least squares iterations. Extensions to other location-scale or time-to-event models are warranted. In principle, these extensions can be covered by a general framework (Derkach et al., 2015). However, further considerations of sampling designs for these models will be required (see Lawless, 2016).

We emphasize that the proposed method can be used in other settings in which molecular genetic technologies are too expensive to apply to an entire large cohort. For instance, in microbiome analysis, application of initial rRNA sequencing in phase 1 could be followed in phase 2 with metagenomic whole genome shotgun sequencing, a technology that provides a deeper understanding of the functions and pathways present in the human microbiome.

ACKNOWLEDGMENTS

This research is supported by funding from the Canadian Institutes of Health Research: CIHR Operating Grant MOP-84287 (RC, SBB), CIHR Training Grant GET-101831 (OE-G). OE-G is a Fellow of CIHR STAGE (Strategic Training for Advanced Genetic Epidemiology)—CIHR Training Grant in Genetic Epidemiology and Statistical Genetics. We thank professor J. F. Lawless for valuable discussions. The Genetic Analysis Workshop (GAW) 19 sequence data were provided by the T2D-GENES Consortium, which is supported by NIH grants U01 DK085524, U01 DK085584, U01 DK085501, U01 DK085526, and U01 DK085545. Additional genetic and phenotypic data for GAW19 were provided by the San Antonio Family Heart Study and San Antonio Family Diabetes/Gallbladder study, which are supported by NIH grants P01 HL045222, R01 DK047482, and R01 DK053889. Andrew R. Wood is supported by European Research Council grant SZ-245 50371-GLUCOSEGENES-FP7-IDEAS-ERC. The Genetic Analysis Workshop is supported by NIH R01 GM031575 from the National Institute of General Medical Sciences. Computations were performed on the Lunenfeld-Tanenbaum HPC cluster, funded by the Canadian Foundation for Innovation, and on the GPC supercomputer at the SciNet HPC Consortium. SciNet is funded by: the Canadian Foundation for Innovation under the auspices of Compute Canada; the Government of Ontario; Ontario Research Fund – Research Excellence; and the University of Toronto.

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

ORCID

Osvaldo Espin-Garcia
http://orcid.org/0000-0003-2052-2626

REFERENCES

Blangero, J., Teslovich, T. M., Sim, X., Almeida, M. A., Jun, G., Dyer, T. D., … Almasy, L. (2016). Omics-squared: human genomic, transcriptomic and phenotypic data for Genetic Analysis Workshop 19. *BMC Proceedings*, *10*(7), 71–77.

Breslow, N. E., & Cain, K. C. (1988). Logistic regression for two-stage case-control data. *Biometrika*, *75*, 11–20.

Breslow, N. E., & Chatterjee, N. (1999). Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis. *Applied Statistics*, *48*(4), 457–468.

Breslow, N. E., & Holubkov, R. (1997). Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *Journal of the Royal Statistical Society, Series B*, *59*(2), 447–461.

Breslow, N. E., & Wellner, J. A. (2007). Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scandinavian Journal of Statistics*, *34*(1), 86–102.

Chatterjee, N., Chen, Y.-H., & Breslow, N. E. (2003). A pseudoscore estimator for regression problems with two-phase sampling. *Journal of the American Statistical Association*, *98*(461), 158–168.

Chen, Z., Craiu, R. V., & Bull, S. B. (2012). Two-phase stratified sampling designs for regional sequencing. *Genetic Epidemiology*, *36*(4), 320–332.

Chen, Z., Craiu, R. V., & Bull, S. B. (2014). A note on the efficiencies of sampling strategies in two-stage Bayesian regional fine mapping of a quantitative trait. *Genetic Epidemiology*, *38*(7), 599–609.

Derkach, A., Lawless, J. F., & Sun, L. (2015). Score tests for association under response-dependent sampling designs for expensive covariates. *Biometrika*, *99*(2015), 1–8.

Faye, L. L., Machiela, M. J., Kraft, P., Bull, S. B., & Sun, L. (2013). Re-ranking sequencing variants in the post-GWAS era for accurate causal variant identification. *PLoS Genetics*, *9*(8), e1003609.

Lawless, J. F. (2016). Two-phase outcome-dependent studies for failure times and testing for effects of expensive covariates. *Lifetime Data Analysis*, 1–17. https://doi.org/10.1007/s10985-016-9386-8

Lawless, J. F., Kalbfleisch, J. D., & Wild, C. J. (1999). Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society, Series B*, *61*(2), 413–438.

Li, D., Lewinger, J. P., Gauderman, W. J., Murcray, C. E., & Conti, D. (2011). Using extreme phenotype sampling to identify the rare causal variants of quantitative traits in association studies. *Genetic Epidemiology*, *35*(8), 790–799.

Lin, D.-Y., Zeng, D., & Tang, Z. Z. (2013). Quantitative trait analysis in sequencing studies under trait-dependent sampling. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(30), 12247–12252.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society*, *44*, 226–233.

Maller, J. B., McVean, G., Byrnes, J., Vukcevic, D., Palin, K., Su, Z., … Donnelly, P. (2012). Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nature Genetics*, *44*(12), 1294–1301.

Murphy, S. A., & van der Vaart, A. W. (2000). On profile likelihood. *Journal of the American Statistical Association*, *95*(450), 449–465.

Pace, L., Salvan, A., & Ventura, L. (2011). Adjustments of profile likelihood through predictive densities. *Annals of the Institute of Statistical Mathematics*, *63*(5), 923–937.

Pickles, A., Dunn, G., & Vázquez-Barquero, J. L. (1995). Screening for stratification in two-phase ("two-stage") epidemiological surveys. *Statistical Methods in Medical Research*, *4*, 73–89.

Robins, J. M., Hsieh, F., & Newey, W. (1995). Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates. *Journal of the Royal Statistical Society, Series B*, *57*(2), 409–424.

Satagopan, J. M., & Elston, R. C. (2003). Optimal two-stage genotyping in population-based association studies. *Genetic Epidemiology*, *25*(2), 149–157.

Schaid, D. J., Jenkins, G. D., Ingle, J. N., & Weinshilboum, R. M. (2013). Two-phase designs to follow-up genome-wide association signals with DNA resequencing studies. *Genetic Epidemiology*, *37*(3), 229–238.

Scott, A. J., & Wild, C. J. (2011). Fitting regression models with response-biased samples. *Canadian Journal of Statistics*, *39*(3), 519–536.

Shin, J.-H., Yi, R., & Bull, S. B. (2016). Identification of low frequency and rare variants for hypertension using sparse-data methods. *BMC Proceedings*, *10*(7), 389–395.

Song, R., Zhou, H., & Kosorok, M. R. (2009). A note on semiparametric efficient inference for two-stage outcome-dependent sampling with a continuous outcome. *Biometrika*, *96*(1), 221–228.

Spain, S. L., & Barrett, J. C. (2015). Strategies for fine-mapping complex traits. *Human Molecular Genetics*, *24*(R1), R111–R119.

Thomas, D. C., Casey, G., Conti, D. V., Haile, R. W., Lewinger, J. P., & Stram, D. O. (2009). Methodological issues in multistage genome-wide association studies. *Statistical Science*, *24*(4), 414–429.

Thomas, D. C., Yang, Z., & Yang, F. (2013). Two-phase and family-based designs for next-generation sequencing studies. *Frontiers in Genetics*, *4*(December), 1–20.

van der Vaart, A. W., & Wellner, J. A. (2001). Consistency of semiparametric maximum likelihood estimators for two-phase sampling. *Canadian Journal of Statistics/La Revue Canadienne de Statistique*, *29*(2), 269–288.

Wang, H., Thomas, D. C., Pe'er, I., & Stram, D. O. (2006). Optimal two-stage genotyping designs for genome-wide association scans. *Genetic Epidemiology*, *30*(4), 356–368.

White, J. E. (1982). A two stage design for the study of the relationship between a rare exposure and a rare disease. *American Journal of Epidemiology*, *115*(1), 119–128.

Yilmaz, Y. E., & Bull, S. B. (2011). Are quantitative trait-dependent sampling designs cost-effective for analysis of rare and common variants? *BMC Proceedings*, *5*(Suppl 9), S111.

Zeng, D., & Lin, D.-Y. (2014). Efficient estimation of semiparametric transformation models for two-phase cohort studies. *Journal of the American Statistical Association*, *109*(505), 371–383.

Zhao, Y., Lawless, J. F., & McLeish, D. L. (2009). Likelihood methods for regression models with expensive variables missing by design. *Biometrical Journal*, *51*(1), 123–136.

Zhou, H., Wu, Y., Liu, Y., & Cai, J. (2011). Semiparametric inference for a 2-stage outcome-auxiliary-dependent sampling design with continuous outcome. *Biostatistics*, *12*(3), 521–534.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Espin-Garcia O, Craiu RV, Bull SB. Two-phase designs for joint quantitative-trait-dependent and genotype-dependent sampling in post-GWAS regional sequencing. *Genet Epidemiol*. 2017; 1–13. https://doi.org/10.1002/gepi.22099