Statistics
in Medicine WILEY

Two-phase sample selection strategies for design and analysis in post-genome-wide association fine-mapping studies

Osvaldo Espin-Garcia[1,2] | **Radu V. Craiu[3]** | **Shelley B. Bull[2,4]**

[1]Department of Biostatistics, Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada

[2]Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada

[3]Department of Statistical Sciences, Faculty of Arts and Sciences, University of Toronto, Toronto, Ontario, Canada

[4]Lunenfeld-Tanenbaum Research Institute, Sinai Health, Toronto, Ontario, Canada

Correspondence
Osvaldo Espin-Garcia, Department of Biostatistics, Princess Margaret Cancer Centre, University Health Network, Toronto, ON, Canada.
Email: osvaldo.espin-garcia@uhnresearch.ca

Funding information
CIHR Training Grant, Grant/Award Number: GET-101831; CIHR Operating Grant, Grant/Award Number: MOP-84287; CIHR Project Grant, Grant/Award Number: PJT-159463

Post-GWAS analysis, in many cases, focuses on fine-mapping targeted genetic regions discovered at GWAS-stage; that is, the aim is to pinpoint potential causal variants and susceptibility genes for complex traits and disease outcomes using next-generation sequencing (NGS) technologies. Large-scale GWAS cohorts are necessary to identify target regions given the typically modest genetic effect sizes. In this context, two-phase sampling design and analysis is a cost-reduction technique that utilizes data collected during phase 1 GWAS to select an informative subsample for phase 2 sequencing. The main goal is to make inference for genetic variants measured via NGS by efficiently combining data from phases 1 and 2. We propose two approaches for selecting a phase 2 design under a budget constraint. The first method identifies sampling fractions that select a phase 2 design yielding an asymptotic variance covariance matrix with certain optimal characteristics, for example, smallest trace, via Lagrange multipliers (LM). The second relies on a genetic algorithm (GA) with a defined fitness function to identify exactly a phase 2 subsample. We perform comprehensive simulation studies to evaluate the empirical properties of the proposed designs for a genetic association study of a quantitative trait. We compare our methods against two ranked designs: residual-dependent sampling and a recently identified optimal design. Our findings demonstrate that the proposed designs, GA in particular, can render competitive power in combined phase 1 and 2 analysis compared with alternative designs while preserving type 1 error control. These results are especially evident under the more practical scenario where design values need to be defined a priori and are subject to misspecification. We illustrate the proposed methods in a study of triglyceride levels in the North Finland Birth Cohort of 1966. R code to reproduce our results is available at github.com/egosv/TwoPhase_postGWAS.

KEYWORDS
genetic algorithm, post-GWAS targeted sequencing, practical two-phase studies, statistical fine-mapping, two-phase study design and analysis

1 | INTRODUCTION

Genome-wide association studies (GWASs) have become well-established untargeted approaches for identifying genetic loci that influence the etiology of complex diseases and traits. Single-nucleotide polymorphisms (SNPs) genotyped using GWAS arrays typically lack any known biological function. Consequently, in post-GWAS studies, identifying causal variants and susceptibility genes in GWAS-identified regions of association is the next important step for researchers. Identified variants and genes can become instrumental in personalized medicine from diagnosis and intervention to drug development and other forms of therapy.

Recent advances in next-generation sequencing (NGS) technologies allow investigators to sequence the entire human genome at the base-pair level, but, the costs of whole genome sequencing are relatively high in comparison to GWAS analysis. Targeted sequencing, which identifies all variants in a region with high-confidence, can be cost effective when fine mapping a genetic region identified at GWAS stage. Indeed, high-density sequence variants in the targeted region are typically in linkage disequilibrium (LD) with strongly associated SNPs from GWAS, making the latter good candidates as auxiliary covariates for subsample selection. Thus, two-phase sampling design and analysis[1,2] emerges as a suitable cost-reduction technique in the post-GWAS context. The main goal of this strategy is to make inference on incompletely observed sequencing data. At phase 1, GWAS data are collected for everyone in the study. At phase 2, sequencing data are collected only in a subsample of the phase 1 sample. The subsample is selected based on phase 1 information alone (outcome, auxiliary SNPs), making the sequence data missing-by-design in the nonselected individuals.

While the majority of the literature in two-phase sampling designs concentrates on effect estimation and hypothesis testing, relatively less attention has been paid to phase 2 sample selection. Specifically, most of the work examining optimal designs has focused on case-control studies,[3,4] in which for example, a *balanced* design (equal sample distribution across strata) has been recommended as near optimal.[5] Typically, in the design of case-control studies, optimization is performed to determine sampling fractions across predefined strata subject to a budget constraint on the phase 2 sample size.[6-8] Another approach, described in Zhao et al, seeks to optimize the sampling fraction (ρ) under simple random sampling considering asymptotic relative efficiency of the maximum likelihood estimators from the one- vs two-phase designs.[9] More recently, Tao et al derived general optimal designs of two-phase studies paying special attention to continuous, binary, and time-to-event outcomes.[10] Specifically, Tao et al demonstrate the relationship between their optimal design (hereafter referred to as TZL) and previously proposed (ranked) designs such as outcome-dependent and residual-dependent sampling (ODS and RDS, respectively).

In this report, we propose two approaches for two-phase sample selection in post-GWA fine-mapping studies. Our previously described methods,[11] when paired with the resulting sample designs, preserve type I error control and are applicable for all distributions in the exponential family. The first approach, LM, extends and adapts previous work primarily developed for case-control studies by solving a constrained optimization problem via Lagrange multipliers using numerical methods. The second approach, GA, exploits the advantages of genetic algorithms (GAs) for discrete optimization with fixed-subsets. To the best of our knowledge, this work introduces a novel usage of GAs in the context of selecting phase 2 designs.

In the next section we introduce a maximum likelihood framework for design and analysis of two-phase studies, and define the two approaches to select a phase 2 subsample. In addition, we contrast the proposed designs (LM and GA) with ranked designs (ODS, RDS, and TZL). In Section 3 we conduct simulation studies of a quantitative trait (QT) to evaluate the performance of the proposed designs against ranked designs under the ideal scenario in which all design quantities are known in advance. In Section 4, we assess a more practical scenario where the design values are misspecified using simulated data with realistic LD patterns from the 1000 Genomes Project. Our results show that the proposed designs, GA in particular, achieve competitive power against alternative designs under various scenarios. Additionally, in Section 5, we illustrate these methods in an application to the North Finland Birth Cohort of 1966. We conclude with a discussion of the advantages and challenges of the studied approaches as well as potential avenues of future research.

2 | PHASE 2 SAMPLE SELECTION UNDER MAXIMUM LIKELIHOOD

2.1 | Two-phase designs in fine-mapping studies

Let Y be the trait of interest and G be a (potentially causal) sequence variant located in a genomic region identified by GWAS test results. By design, variants in the region of interest are ascertained in only a fraction of individuals.

Consequently, two-phase studies consist of a GWAS in phase 1 from which a subsample of individuals is selected; in phase 2, fine-mapping sequence data are collected for the subsample and combined analysis is performed using information from phases 1 and 2. In this post-GWAS setting, the trait data (Y) and the GWAS-SNP (Z), are observed for every subject in the study. The two-phase design aims to select a subset of informative subjects based on available data in the GWAS, namely, (Y, Z). Of note, Z can be either an observed or imputed genotype, in the latter case the purpose might be to verify the association with sequencing data. Inference on the missing-by-design sequence variants is conducted using all available data. We define the missing indicator $R_i = \mathbb{1}\{i \in S_2\}, i = 1, \dots, N$ where N is the number of individuals in the entire phase 1 cohort and S_2 represents the set of $n = \sum_{i=1}^{N} R_i$ subjects selected into the phase 2 subsample. We let $\bar{S}_2$ denote the set of ($N - n$) subjects in the GWAS study who were unselected for phase 2. We specify the selection model for the ith subject as $\pi_i(\boldsymbol{\psi}) = \pi(Y_i, Z_i; \boldsymbol{\psi}) = \Pr(R_i = 1 | Y_i, Z_i; \boldsymbol{\psi})$, where $\boldsymbol{\psi}$ is a vector that characterizes the distribution of the inclusion probabilities. To operationalize the selection, (Y, Z) can be stratified into K disjoint groups, $\{1\}, \dots, \{K\}$, such that $\pi_i(\boldsymbol{\psi}) = \pi_k(\boldsymbol{\psi})$ for all (Y_i, Z_i) $\in \{k\}$; that is, all subjects in the kth stratum have equal selection probabilities. R_i is designed to be conditionally independent of G_i given Y_i and Z_i, that is, the phase 2 selection mechanism dictated by R_i is completely determined by Y_i and Z_i, making G_i missing at random.[12]

2.2 | Maximum likelihood formulation

Let $f_\beta(y|g, z)$ be the parametric relationship between (G, Z) and Y indexed by $\boldsymbol{\beta}$. Here, $f_\beta(y|g, z)$ corresponds to a probability function in the exponential family with $E[Y|g, z; \boldsymbol{\beta}] = \mu(g, z; \boldsymbol{\beta}) = h^{-1}(\beta_0 + \beta_1 g + \beta_z z)$, where $h(\cdot)$ denotes the link function. We denote $\mathcal{G}, \mathcal{Z}$ as the sets of uniquely observed values of G (in S_2) and Z (in $S_2 \cup \bar{S}_2$). Let $\Pr(G, Z)$ be the joint probability function of G and Z given by the discrete probabilities $p_{g,z}, g \in \mathcal{G}$ and $z \in \mathcal{Z}$, which is left unspecified and define $\boldsymbol{p} = \{p_{g,z}\}_{g \in \mathcal{G}, z \in \mathcal{Z}}$. We consider here the nonparametric estimation of the joint distribution of G and Z, with support on the Cartesian product between $\mathcal{G}$ and $\mathcal{Z}$. Considering the above, we define the observed-data likelihood following previous literature[13-15] as

$$L(\boldsymbol{\beta}, \boldsymbol{p}) = \prod_{i=1}^{N} \left[\pi_i f_\beta(y_i|g_i, z_i) \Pr(G = g_i, Z = z_i) \right]^{R_i} \left[\{1 - \pi_i\} \sum_{g \in \mathcal{G}} f_\beta(y_i|g, z_i) \Pr(G = g, Z = z_i) \right]^{1-R_i}$$

$$\propto \prod_{i=1}^{N} \left[f_\beta(y_i|g_i, z_i) p_{g_i, z_i} \right]^{R_i} \left[\sum_{g \in \mathcal{G}} f_\beta(y_i|g, z_i) p_{g, z_i} \right]^{1-R_i}. \tag{1}$$

In (1), the proportionality arises since estimation of ($\boldsymbol{\beta}, \boldsymbol{p}$) does not involve π_i's. Thus the log-likelihood is

$$\ell(\boldsymbol{\beta}, \boldsymbol{p}) \propto \sum_{i=1}^{N} \left[R_i \times \left(\log \left\{ f_\beta(y_i|g_i, z_i) \right\} + \log \left\{ p_{g_i, z_i} \right\} \right) + (1 - R_i) \times \log \left\{ \sum_{g \in \mathcal{G}} f_\beta(y_i|g, z_i) p_{g, z_i} \right\} \right].$$

For simplicity, the formulation above considers G as a single variable, however, this can be extended to a vector with the respective considerations as illustrated in Section 4 below. We also note that additional phase 1 covariates $\boldsymbol{X}$ can be introduced into the parametric model, that is, $f_\beta(y_i|g_i, z_i, \boldsymbol{x}_i)$ with corresponding $\mu(g, z, \boldsymbol{x}_i; \boldsymbol{\beta}) = h^{-1}(\beta_0 + \beta_1 g + \beta_z z + \boldsymbol{\beta}_x' \boldsymbol{x}_i)$, by assuming G and $\boldsymbol{X}$ are conditionally independent given Z. In the case of covariates such as genomic principal components to account for population stratification, conditional independence is a working assumption, which does not negatively affect the model performance as shown in Section 5.

If we denote $\theta = (\boldsymbol{\beta}', \boldsymbol{p}')'$, then, under regularity conditions, the limiting distribution of the maximum likelihood estimator ($\hat{\theta}$) follows asymptotically $\sqrt{N}(\hat{\theta} - \theta) \sim \mathcal{N}(0, \mathbb{J}(\Phi)^{-1})$,[16] where $\Phi = (\boldsymbol{\beta}', \boldsymbol{p}', \boldsymbol{\psi}')'$ and $\mathbb{J}(\Phi)$ is the expected information matrix, which is a function of the full parameter set Φ as the expectation is taken with respect to (R, Y, G, Z); note that G is observed in S_2 and missing by design in $\bar{S}_2$.[8] The derivation of the expected information matrix is shown in Appendix A.

2.3 | Post-GWAS analysis under maximum likelihood

Note that the likelihood in Equation (1) is most useful at the design stage when no phase 2 subsample has been identified nor have any data been collected. However, once these items are available, the following reexpression is typically used:

$$L(\theta) \propto \prod_{i \in S_2} f_\beta(y_i|g_i, z_i)p(g_i, z_i) \prod_{i \in \bar{S}_2} \sum_g f_\beta(y_i|g, z_i)p(g, z_i). \tag{2}$$

The formulation above has been amply studied.[11,14,15,17] Estimates can be obtained via the EM algorithm[18-20] and the corresponding asymptotic variance covariance matrix is computed via the Louis' method.[21] In fine-mapping, the aim is to identify and prioritize potential causal variants in a genetic region of interest to allow for follow-up replication and functional studies.[22] This can be achieved under the proposed maximum likelihood (ML) as follows: first, the genetic effect of each variant in the targeted region is estimated and tested individually (single-variant analysis); second, genetic effects are estimated and tested in multivariable models (conditional on strongest single-variant signals). Conditional analysis serves to identify independent signals in the region and to unmask associations that may have been missed in single-variant analysis. These steps are detailed in Section 4.

2.4 | Selecting phase 2 designs

In post-GWAS fine-mapping studies that target an identified genomic region, the costs of sequencing can make it unfeasible or inefficient to sequence all subjects available in phase 1, restricting the number of individuals in S_2 (n). Here we propose two approaches to select a phase 2 design under a budget constraint and flexible optimality criteria using Lagrange multipliers or genetic algorithms. In addition, we discuss the specification of such optimality criteria and compare the proposed methods against another class of widely used phase 2 sample selection strategies, the so-called ranked designs.

2.4.1 | Lagrange multipliers (LM)

Following previous ideas,[8,23,24] we first propose to obtain a phase 2 design for regional fine-mapping studies by minimizing the following expression

$$\Lambda\left(\mathbb{J}(\Phi)^{-1}\right) - \lambda N^{-1}\left[\sum_{k=1}^{K} \pi_k(\psi)N_k - n\right], \tag{3}$$

where $\Lambda(\cdot)$ is an optimality criterion, λ is a Lagrange multiplier accounting for the budget constraint, and N_k is the number of subjects in phase 1 belonging to the kth stratum.

Here, we formulate the approach specifically for the ML framework described in Sections 2.2 and 2.3. For our purposes, β and p are design quantities, thus, they need to be specified a priori leaving the πs to be determined from phase 1 data alone. Note that this approach aims to find selection probabilities, π_k^*, that minimize equation (3) for allocating the phase 2 sample across strata $\{k\}$, $k = 1, \ldots, K$. The vector p can be interpreted in terms of the (joint) genotyping distribution between G and Z, which can be easily specified according to well-established genetic principles, for example, Hardy-Weinberg equilibrium (HWE), or by external data such as the 1000 Genomes Project. Thus, the expected effect size of the sequence variant, β_1, becomes the primary parameter to specify.

2.4.2 | Genetic algorithms (GA)

Genetic algorithms are designed to mimic nature's evolutionary process, in which the fittest members of a population are selected to pass on their genetic information. GAs are powerful tools to optimize a fitness measure/objective function, $\Lambda(\cdot)$; overviews can be found in Holland[25] and Whitley.[26] This optimization technique is suitable for a discrete solution space and is performed through a stochastic search by building an initial population of candidate solutions that evolves generationally through pairing, mating, recombining and mutating the candidate solutions. In our case, these candidate solutions correspond to vectors of the form $R = (R_1, \ldots, R_N)$ with $\sum_{i=1}^{N} R_i = n$ and $R_i \in \{0, 1\}$ for all i, that is, vectors of indicator variables denoting whether the ith subject is selected for phase 2. The reasoning behind GA implementation in the context of phase 2 sample selection is twofold: 1) it provides a suitable framework for discrete optimization, and 2) it has proven to be an efficient strategy to find a fittest member in large search spaces (2^N possibilities in this case).[26] These appealing features of GAs come along with some challenges, namely, that there are no clear convergence criteria,

tuning parameters need to be specified, and they can be computationally expensive when the objective function is hard to calculate.

Nevertheless, the GA approach brings novelty to the field as it nullifies the uncertainty brought by the sampling variability introduced when utilizing stratum-specific selection probabilities. This is achieved by selecting a vector $\boldsymbol{R}^*$ that characterizes a unique phase 2 subsample and optimizes the fitness measure. Furthermore, GA can forgo strata definition since the search can be agnostic to specific strata configurations. In GA, the budget constraint can be introduced by a so-called cardinality constraint, which consists of selecting a subset of a required size (n).[27] This constraint guarantees that the phase 2 sample is exactly of size n as opposed to methods that depend on selection probabilities, which introduce some variation into the achieved phase 2 sample size.

To date, there are several implementations for GAs available in the R statistical language[28] namely packages *GA*, *genalg*, *kofnGA*, *mcga*, *mco*, and *NMOF*. Of these, only *kofnGA* is specifically designed for fixed-size subset selection with a flexible specification of the objective/fitness function, see Wolters for a detailed explanation of the package.[29] In any GA, it is important to consider the set of control parameters necessary to implement decision rules at each step. Specifically, *kofnGA* requires the user to specify the objective/fitness function (ie, $\Lambda(\cdot)$), the subset size (n), and the number of candidates (N) while additional control parameters related to algorithmic design have default (but adjustable) settings. These control parameters are: *population size* ($\mathcal{M}$), *number of generations* ($\mathcal{H}$), *size of selection tournament* ($\mathcal{T}$), *mutation rate* ($\mathcal{r}$), and *number of elites* ($\mathcal{E}$). A pseudo-algorithm relating standard GA terminology with the two-phase design along with a description of the steps where the control parameters are used is presented in Algorithm 1.

Algorithm 1. Pseudo-algorithm of the implemented genetic algorithm

---

procedure GENETIC ALGORITHM DESCRIPTION (under *kofnGA*)

    Generate an initial population of size $\mathcal{M}$, i.e. $\text{pop}^0 = (\boldsymbol{R}^0_{(1)}, \ldots, \boldsymbol{R}^0_{(\mathcal{M})})$, a set of vectors of size N, each representing a single design

    Compute fitness for each member of pop^0 ($\Lambda\{\mathbb{J}^{-1}(\boldsymbol{R}^0_{(m)})\}, m = 1, \ldots, \mathcal{M}$)

    for $\hbar = 1, \ldots, \mathcal{H}$ (the number of generations) **do**

      Selection (draw pairs of candidate solutions from $\text{pop}^{\hbar-1}$)

        ▷ this is achieved through a tournament strategy in which members with higher fitness have a higher chance to be chosen into one of the two sets of size $\mathcal{T}$. Pairs of candidate solutions are then taken at random from these sets

      Crossover (combine paired candidate solutions at random → $\text{pop}^{\hbar}$)

          ▷ after combining unique elements of both parents, n of them are selected at random

      Mutation (swap indices from elements in $\text{pop}^{\hbar}$ at random with probability $\mathcal{r}$)

      Fitness (compute $\Lambda\{\mathbb{J}^{-1}(\boldsymbol{R}^{\hbar}_{(m)})\}, m = 1, \ldots, \mathcal{M}$)

      Elitism (replace the $\mathcal{E}$ least fit members from $\text{pop}^{\hbar}$ with the $\mathcal{E}$ fittest members from $\text{pop}^{\hbar-1}$)

    end for

    Select the phase 2 subsample corresponding to $\min_{m \in 1, \ldots, \mathcal{M}} \Lambda\{\mathbb{J}^{-1}(\boldsymbol{R}^{\mathcal{H}}_{(m)})\}$

end procedure

The population is the pool of candidate solutions at each iteration from which the fittest members, that is, members with optimal $\Lambda(\cdot)$ values, will be ultimately selected. The number of generations denotes the number of iterations the algorithm will run for. The size of selection tournament determines the number of members of the population selected to produce the next generation. The mutation rate determines the probability at which random swaps in the indexes of the candidate solutions occur in the population. Lastly, elites are the fittest members in a given generation that get to be kept in the next generation.

The algorithm parameters can be tuned by the user in accordance with the problem at hand. For simplicity, parameters $\mathcal{H}$ and $\mathcal{E}$ can be reformatted as proportions of the population size ($\mathcal{M}$). Because *kofnGA* does not implement a stopping rule, the algorithm iterates for as many times as specified by the provided number of generations ($\mathcal{H}$). To accelerate convergence, we set $\mathcal{T}$ and $\mathcal{E}$ at high levels ($= 0.90 \times \mathcal{M}$) as suggested by Walters.[29] This approach may diminish the search improvements derived from mutations, relying more heavily on the initial population and number of generations. Therefore, instead of setting a completely random initial population (pop^0 in Algorithm 1), we initialize it with an equal number of samples with top 20 performers (based on $\Lambda(\cdot)$) out of 100 draws of each of the balanced, combined, and LM

TABLE 1 Description of the three optimality criteria evaluated. $\mathbb{J}(\Phi)^{-1}$ denotes the variance-covariance matrix.

$\Lambda(\cdot)$	Formula	Description
A-optimality	$\sum \text{diag}(\mathbb{J}(\Phi)^{-1})$	Minimizes the average variance of the parameter estimates
D-optimality	$\det(\mathbb{J}(\Phi)^{-1})$	Minimizes the product of the variances for diagonal matrices
Parameter-specific	$\mathbb{J}(\Phi)^{-1}_{[\beta_1,\beta_1]}$	Minimizes the variance of a particular entry in the VCM

sampling designs plus the RDS design. This strategy guarantees that GA has at least the same performance as the RDS design.

Additional considerations for the optimization strategies in LM and GA as well as further details on the balanced and combined designs can be found in Sections S2 and S3, respectively (Online Supplementary Material). It is also worth noting that the proposed approaches are feasible in any context where $Y|G, Z$ can be modeled within the exponential family, and both G and Z are discrete (or easy to discretize) covariates.

2.4.3 | Specifying an optimality criterion

There are several ways to define a functional $\Lambda(\cdot)$ as the optimality criterion/fitness measure, mostly grounded in experimental design.[30] In this report, we explore three criteria: A-optimality, D-optimality, and parameter-specific. Each criterion focuses on different features of the variance-covariance matrix (VCM), $\mathbb{J}(\Phi)^{-1}$ (Table 1). The parameter-specific criterion is optimal to identify designs with minimum variance when testing a single parameter. Similarly, A- and D-criteria would be optimal to identify designs with minimum average variance across all parameters and minimum volume of the confidence ellipsoid, respectively.

In the outlined post-GWAS setting, the focus lies on testing a single parameter, β_1. Thus, the parameter specific criterion is the most natural choice. However, when multiple parameters are of interest, that is, β_1 is a vector, A-optimality may be preferred if the parameters of interest are loosely correlated, whereas D-optimality may be preferred when the parameters are strongly correlated. This makes intuitive sense when there are two (or more) estimators, for example, $\hat{\beta}_{11}$ and $\hat{\beta}_{12}$ but $\hat{v} = G_1\hat{\beta}_{11} + G_2\hat{\beta}_{12}$ is of interest. Then $V(\hat{v}) = G_1^2 V(\hat{\beta}_{11}) + G_2^2 V(\hat{\beta}_{12}) + 2G_1 G_2 Cov(\hat{\beta}_{11}, \hat{\beta}_{12})$, thus, involving off-diagonal elements of the VCM. We evaluate potential differences in the choice of optimality criterion in Section 4.

2.4.4 | Ranked designs

Recently, Tao et al proposed general optimal designs for phase 2 studies.[10] In this section, we aim to describe their approach, summarize their findings, and draw comparisons with the proposed designs: LM and GA.

Despite their names, the outcome-dependent sampling (ODS) and residual-dependent sampling (RDS), as defined by Tao et al, are not sampling designs in the classical sense given the fact that their specification is independent of any sampling mechanism.[10] Indeed, this is also true for the TZL design. We refer to them as ranked designs because they are defined in terms of ordered quantities: outcome/residuals/scaled residuals for ODS, RDS, and TZL, respectively. Tao et al show that the scaling factor in TZL is given by $\text{Var}(G|Z)^{1/2}$, which is unknown at design stage and thus needs to be specified prior to phase 2. An intuition on why this scaling factor is important for the optimal design, provided by Tao et al, is that G is harder to be retrieved by Z when $\text{Var}(G|Z)$ is large and thus these subjects need to be oversampled.[10]

The ranked designs achieve a given phase 2 sample size by selecting an equal number of subjects from each of the top and bottom rankings of the outcome/residuals/scaled residuals. This particularity makes them appealing for a few reasons: 1) the ranked designs are unique, 2) the selection is intuitive and can be performed quickly, and 3) for QTs, no stratification on the outcome is required. Tao et al show that the TZL design reduces to the RDS design when G and Z are independent (and RDS reduces to ODS when Y and Z are independent); in Sections 4 and 5, we investigate the effect of misspecifying $\text{Var}(G|Z)$.

There are five main underlying differences between the proposed designs (LM and GA) and the ranked designs, particularly TZL:

1. LM depends on the stratification strategy undertaken for the outcome while none of the ranked design depends directly on outcome stratification for QTs. On the other hand, GA can, in principle, be performed without defining any stratification; however, selection of initial values may depend on values drawn from LM or other designs to accelerate convergence, which could introduce some dependency on a chosen stratification.

2. LM provides optimal sampling fractions and thus a sample must be drawn accordingly, subjecting this design to sampling variability. GA selects a unique solution, R^*, with optimal $\Lambda(\cdot)$ value, but, given the stochastic nature of the genetic algorithm, this solution is approximate and varies at each run unless a random seed is specified. In contrast, the ranked designs are not subject to sampling variation.

3. TZL and the proposed designs (under the parameter-specific criterion) seek to minimize the variance of β_1. In the case of TZL, this is achieved by maximizing the inverse of the efficiency bound for estimating β_1 with one observation in theorem 1 of Tao et al. Note that the proof of this theorem relies on the assumption that Y and G are approximately independent given Z, which is justified when the effect of G on Y is small, that is, $\beta_1 = o(1)$. LM and GA, on the other hand, do not depend on the small β_1 assumption and can be thus implemented in more general settings.

4. Related to the point above, LM and GA rely on an empirical approximation to the information matrix whereas the variance considered in TZL uses an exact expected information under the working assumption of $\beta_1 = o(1)$. This defines a trade-off between the generality of LM and GA and the increase in efficiency of TZL when the assumption is justified.

5. LM and GA can optimize general functions beyond $\mathbb{J}(\Phi)^{-1}$ through $\Lambda(\cdot)$, whereas the results in TZL are mostly concerned with $\text{Var}(\hat{\beta}_1)$ (or functions thereof).

It remains unclear what constitutes a good stratification for LM; intuitively, LM should approximate TZL as the number of strata approaches the phase 1 sample size. However, a theoretical proof is beyond the scope of this paper. To circumvent the sampling variability issue in LM, one could draw a predetermined number of subsamples and select the one with optimal $\Lambda(\cdot)$ value. Regarding whether restricted/unrestricted values of θ are preferred, Tao et al show that TZL performs well for alternatives close to the null. Although these alternatives are typical for genetic studies, a more comprehensive comparison for alternatives farther away for the null is warranted. Lastly, LM and GA can, in fact, approximate the optimization strategy in TZL by utilizing $\mathbb{V}_1^{-1}$ instead of $\mathbb{J}(\Phi)^{-1}$ in the objective function, where $\mathbb{V}_1 = \mathbb{F}_{11} - \mathbb{F}_{10}\mathbb{F}_{00}^{-1}\mathbb{F}_{01}$ and $\mathbb{F} = \mathbb{J}(\Phi)$ is partitioned with respect to β_1 as $\begin{bmatrix} \mathbb{F}_{11} & \mathbb{F}_{10} \\ \mathbb{F}_{01} & \mathbb{F}_{00} \end{bmatrix}$. It is worth noting that the dimension of $\mathbb{V}_1$ corresponds to that of the subspace determined by the null hypothesis of interest. In the simplest case of β_1 being a scalar, $\mathbb{V}_1$ is also a scalar. In LM and GA, higher dimensions can be easily accommodated by specifying a different $\Lambda(\cdot)$ on $\mathbb{V}_1$, to obtain say A- or D-optimal designs.

3 | SIMULATION STUDIES

In this section, we describe the data generation steps, analysis plan, and report the results of an initial set of simulations. The main objective is to compare the statistical power of the proposed phase 2 designs, LM and GA, in a post-GWAS fine-mapping scenario by testing for the effect of G (the missing-by-design variable), that is, $H_0 : \beta_1 = 0$. In addition, we compare LM and GA against two ranked designs: TZL and RDS. Of note, we exclude ODS from these studies given the known indirect association between Z and Y at GWAS stage. Estimates and standard errors are constructed following Equation (2) in Section 2.3. For comparability with RDS and TZL, we use $\mathbb{V}_1^{-1}$ in the parameter-specific optimality criterion for LM and GA as described in Section 2.4.4 given that this variance estimate does not depend on the assumption of $\beta_1 = o(1)$. Additional numerical studies comparing LM and GA against alternative heuristic designs utilizing $\mathbb{J}(\Phi)^{-1}$ in the optimization are found in Section S3, Online Supplementary Material.

3.1 | Data generation

We assume a data generating mechanism similar to Espin-Garcia et al.[11] Briefly, for a phase 1 sample size (N), and given values for minor allele frequencies (MAFs), q_G and q_z and the linkage disequilibrium (LD), quantified through the Pearson correlation coefficient, r, we simulate two variants on the same haplotype under Hardy-Weinberg equilibrium (HWE): G_1 and Z. Here, q_G and q_Z are the frequencies of the less common allele in the population for G_1 and Z, respectively, whereas LD is the level of correlation between them. Notably, since the actual allele frequencies cannot be

negative and the additive linkage disequilibrium coefficient D is constrained, not all combinations of r, q_G and q_Z can occur.

The trait of interest is then generated as $Y = \beta_0 + \beta_1 G_1 + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. We note that in this setting, as opposed to Section 2.2, Z is assumed to be conditionally independent of Y given G. This simulation setup aims to resemble a more realistic scenario in which the GWAS-SNP, Z, is not causal itself but rather is in linkage disequilibrium with the causal variant, G.

To imitate the GWAS setting in each dataset, we test γ_z, the genetic effect of Z, for association in the regression model $Y = \gamma_0 + \gamma_z Z$ and only keep replicates that meet a suggestive genome-wide significance criterion of $p < 1 \times 10^{-5}$ for the hypothesis $H_0 : \gamma_z = 0$. Lastly, to study type 1 error (T1E) under this data-generation mechanism, we simulate another SNP, G_0 independently from Z and G_1 with MAF q_G.

Strata for Y are defined by discretizing the trait values into a three-category variable, $Y_{st} = \{T_1, T_2, T_3\}$, according to fixed cut points $(C1, C2)$ as the percentiles $(40, 60)$ of a normal distribution with mean $\mu = 2$ and variance $\sigma^2 = 1$, so that under the null, $\Pr(Y < C1) = \Pr(Y > C2) = 0.4$. Strata for the biallelic GWAS SNP, Z, are defined by considering Z as a three-category variable corresponding to genotypes, (AA, Aa, aa) and coded by the number of copies of the minor allele (a), that is, $Z = 0, 1, 2$ (additive association).

Of note, stratification by Y_{st} and Z is only employed for optimization in LM and for visualization to compare the distribution of selected individuals under other designs.

3.2 | Assessing the phase 2 designs

The first set of evaluations consists of the following. We specify a phase 1 sample size of $N = 5000$ and a phase 2 sample size of $n = 540, 1500, 2500$, that is, 0.108, 0.25, and 0.50 of the phase 1 data, respectively. We draw 1250 replicates for each combination of simulation parameters $q_G = 0.2$, $q_Z = 0.3$, $r = 0.75$, $\beta_0 = 2$, $\sigma^2 = 1$, and $\beta_1 \in \langle 0.1 + 0.2j | j = 0, \dots, 3 \rangle$. We evaluate the performance of the proposed designs against two ranked designs, RDS and TZL, across three statistical tests (Wald, likelihood ratio [LR] and score). Since β_1 is a scalar, the comparison against ranked designs only examines a parameter-specific criterion as a consequence from considering $\mathbb{V}_1^{-1}$ in the objective function, as mentioned in Section 2.4.4.

To compare power, we assess the ratio of the empirical power of each design over that of the complete data case (relative empirical power, rEP). In addition, estimation efficiency is compared via relative asymptotic and empirical standard error (rASE and rESE, respectively) of $\hat{\beta}_1$ for each design over that of the complete data. We deem these measures to provide a better reflection of the design performance compared with studies that benchmark against simple random sampling. Note that the closer these ratios are to 1 (100%), the better the studied designs are able to recover the performance of the complete data analysis.

The specification of $\boldsymbol{\beta}_{des}$, the design regression parameters, corresponds to $(\hat{\gamma}_0, 0, \hat{\gamma}_z)'$. Here $\hat{\gamma}_0$ and $\hat{\gamma}_z$ denote the maximum likelihood estimates (MLEs) from GWAS, that is, the MLEs for $Y = \gamma_0 + \gamma_1 Z$. Similarly, we specify $\boldsymbol{p}_{des}$, the design haplotype distribution between G and Z, under HWE by estimating q_Z from the phase 1 sample and designating q_G and r to be the equal to their generating values. These design values are used to determine LM, GA, and TZL designs but not RDS, which is agnostic to these design quantities.

Although correct specification of the design quantities is hardly ever attainable, the settings above allow us to evaluate the true type 1 error/power of the studied designs. We discuss in the next section how to proceed in practice when the true design values are unavailable. Moreover, by specifying the regression parameters under the null hypothesis, that is, $\boldsymbol{\beta}_{des} = (\hat{\gamma}_0, 0, \hat{\gamma}_z)'$, the design problem greatly simplifies to only specify values for $\boldsymbol{p}_{des}$.

3.3 | Results

The ranked designs can be specified without strata definitions, however, for visualization and comparison purposes, we plot the distribution of RDS and TZL according to the predetermined strata. When comparing LM and GA against the ranked designs for the smallest and largest studied genetic effects ($\beta_1 = 0.1$ and $\beta_1 = 0.7$), we observe: (1) LM, GA and TZL vary considerably across values of β_1 and n, (2) LM displays more unstable strata distribution when compared against GA and ranked designs, especially for the smaller phase 2 sample sizes ($n = 540, 1250$), (3) GA follows closely the RDS design especially when $n = 540, 1250$, (4) LM and GA reach an approximately equal strata distribution when $n = 2500$,
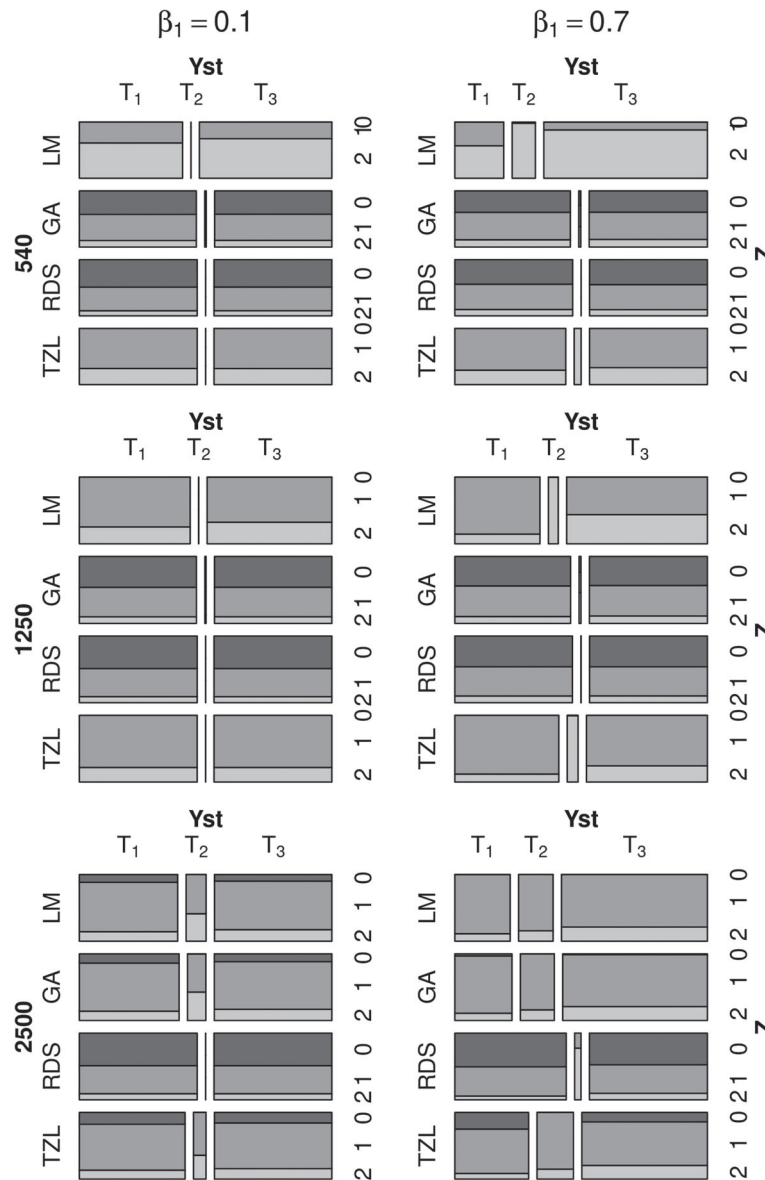
FIGURE 1 Mosaic plots with the average strata sizes across replicates for the proposed designs against ranked designs across phase 2 sample sizes, $n = 540, 1250, 2500$, under the parameter-specific criterion. Averages were taken from the resulting designs in the main simulation study for the two most extreme values of β_1 (0.1, 0.7)

and (5) as expected, the strata distribution of the RDS design remains practically unchanged between genetic effect sizes and phase 2 sample sizes (Figure 1).

At $\alpha = 1\%$, type 1 error (T1E) rates demonstrate well controlled values across the three tests in most cases. For LM, the observed T1E of the Wald test are slightly anti-conservative when $n = 540$ and stabilize around the nominal rate as n increases (Table 2). Closer inspection of the p-value distribution under LR displays no gross departure from the expected uniform distribution (Figure S1). Overall, we observed the LR statistics showed better behavior compared with score and Wald statistics even under small sample sizes specially under LM design (Table 2). Empirical bias ($\hat{\beta}_1 - \beta_1$) is well centered around zero overall and decreases as n increases for all designs when the true value for β_1 is small (< 0.3) (Figure 2). However, for larger values of β_1 (≥ 0.5), LM and TZL show biased estimates when $n = 540, 1250$ and deteriorate as β_1 increases (Figure 2). All designs show relatively close agreement between (r)ASE and (r)ESE across values of β_1 and n (Tables S1 and S2). TZL shows values of rASE and rESE closer to 1, with GA second, RDS third, and LM coming last. GA, RDS designs exhibit adequate coverage while the coverage for LM and TZL worsens as β_1 increases for $n = 540, 1250$ (Table S3).

TABLE 2 Type 1 error (T1E) rates along with their corresponding 99% Clopper-Pearson confidence intervals ($\alpha = 1\%$) across studied designs, phase 2 sample sizes ($n = 540, 1250, 2500$) and statistical tests under a parameter-specific criterion

n	Test	LM	GA	RDS	TZL
540	Wald	1.45 (1.23-1.70)	1.17 (0.97-1.40)	1.10 (0.91-1.32)	1.01 (0.83-1.22)
	LR	1.02 (0.84-1.24)	1.14 (0.94-1.36)	1.10 (0.90-1.32)	0.97 (0.79-1.17)
	Score	0.94 (0.76-1.14)	1.10 (0.90-1.32)	1.07 (0.88-1.29)	0.95 (0.77-1.15)
1250	Wald	1.06 (0.87-1.28)	1.15 (0.95-1.37)	1.12 (0.93-1.34)	1.02 (0.84-1.24)
	LR	0.94 (0.76-1.14)	1.14 (0.95-1.37)	1.12 (0.93-1.34)	0.99 (0.81-1.20)
	Score	0.90 (0.72-1.10)	1.13 (0.94-1.35)	1.10 (0.90-1.32)	0.96 (0.78-1.17)
2500	Wald	1.00 (0.82-1.21)	1.08 (0.89-1.30)	1.08 (0.89-1.30)	1.09 (0.89-1.30)
	LR	0.97 (0.79-1.18)	1.06 (0.87-1.27)	1.07 (0.88-1.29)	1.07 (0.88-1.29)
	Score	0.96 (0.78-1.17)	1.05 (0.86-1.27)	1.07 (0.88-1.29)	1.06 (0.87-1.27)

Note: Each entry represents 17 500 replicates pooled across empirical null scenarios. The rest of the simulation parameters correspond to $q_G = 0.2$, $q_Z = 0.3$, $r = 0.75$, $\beta_0 = 2$, $\sigma^2 = 1$, $N = 5000$. The complete data T1E rate is 1.16 (0.97-1.37) for Wald/LR tests and 1.14 (0.95-1.35) for the score test. To further evaluate test validity under the studied sample sizes, we plot histograms of the observed LR test p-values in Figure S1.

TABLE 3 Relative empirical power (rEP), calculated as the ratio of the empirical power of each studied design over that of the complete data, across studied designs, phase 2 sample sizes, and effect sizes under the LR test ($\alpha = 1 \times 10^{-8}$) for the ideal scenario of correctly specifying Var$(G|Z)$

n	β_1	Complete (EP)	Relative empirical power (rEP)			
			LM	GA	RDS	TZL
540	0.225	58.1	2.9	9.2	6.7	32.5
	0.250	80.6	4.5	14.8	11.7	40.4
	0.300	98.2	15.0	38.7	35.5	74.2
	0.400	100.0	69.2	94.2	91.8	99.6
	0.500	100.0	97.1	100.0	99.9	100.0
1250	0.225	58.1	36.0	44.1	45.5	82.1
	0.250	80.6	44.4	57.5	56.0	87.4
	0.300	98.2	77.0	85.7	84.9	97.0
	0.400	100.0	99.6	99.9	99.9	100.0
	0.500	100.0	100.0	100.0	100.0	100.0
2500	0.225	58.1	96.0	96.1	86.4	95.9
	0.250	80.6	96.8	97.1	90.8	98.6
	0.300	98.2	99.3	99.2	98.2	99.6
	0.400	100.0	100.0	100.0	100.0	100.0
	0.500	100.0	100.0	100.0	100.0	100.0

Note: Column "Complete" corresponds to the estimated power of the complete data (ie, the denominator of the rEP ratio). Phase 1 sample size is $N = 5000$ whereas phase 2 sample size is $n = 540, 1250, 2500$. These results exclude values of β_1 lower than 0.225 and greater than 0.5 since the power of the complete data was less than 50% for the former and had already reached 100% for latter. The rest of the simulation parameters correspond to $q_G = 0.2$, $q_Z = 0.3$, $r = 0.75$, $\beta_0 = 2$, and $\sigma^2 = 1$.
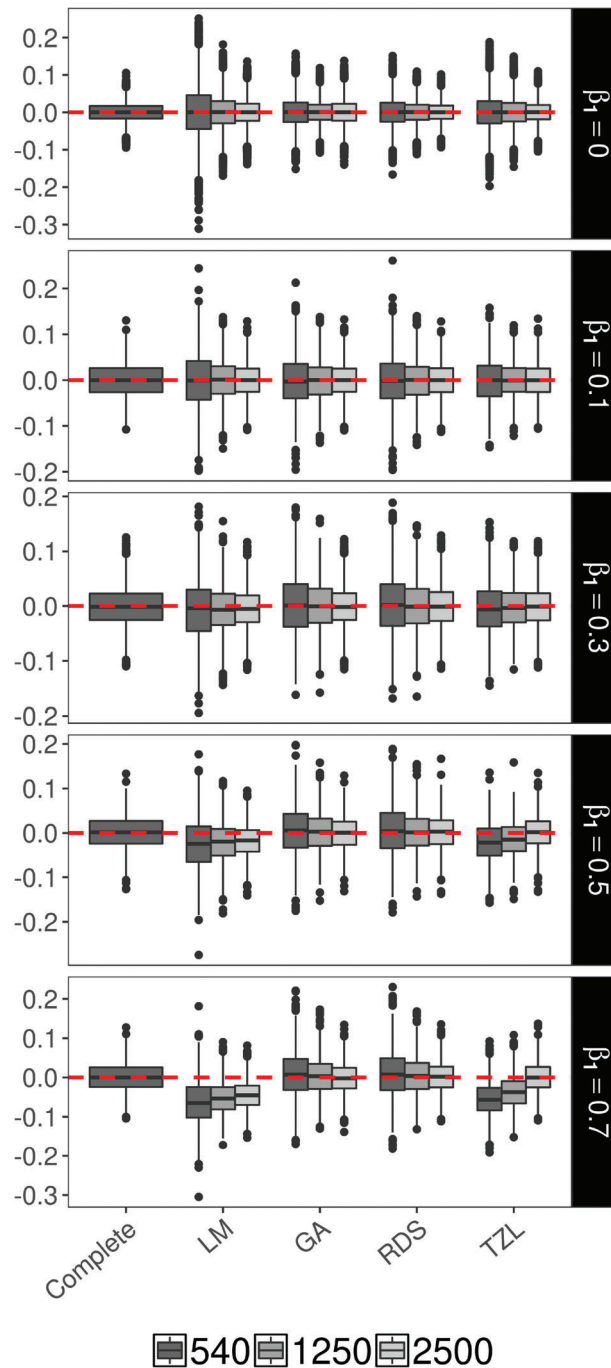
FIGURE 2 Boxplots for the distribution of the bias across genetic effect estimates ($\hat{\beta}_1 - \beta_1$) in the studied designs under a parameter-specific criterion. Row facets denote different true β_1 values (0, 0.1, 0.3, 0.5, 0.7) [Colour figure can be viewed at wileyonlinelibrary.com]

Power curves under the LR test at $\alpha = 1 \times 10^{-8}$ level, show that TZL consistently demonstrates the highest power across values of n with GA second, RDS in the third place and LM having the lowest power. Notably, all designs reach similar power when $n = 2500$ (Figure S2). Interestingly, not all methods show power increases at the same rate due to the differences in efficiency across designs. Additional simulations for larger phase 1 sample size ($N = 10\,000$) and similar selection fractions ($n/N = 0.10, 0.25, 0.50$) result in analogous type 1 error and power results among designs, suggesting that testing performance is contingent upon sampling fraction and not phase 2 sample size (Section S4.1, Online Supplementary Material). Under the LR test, the rEP is highest for the TZL across values of β_1 and n. GA shows higher power than RDS across virtually all scenarios while LM comes last when $n = 540, 1250$ but reaches similar power to GA when $n = 2500$ (Table 3).

Besides the additional simulations on different phase 1 and 2 sample sizes, we also studied the influence of different specifications of the joint distribution of G and Z. In summary, these results are analogous to what was reported above with GA showing competitive power when compared against alternative designs (Section S4.2, Online Supplementary Material).

4 | TWO-PHASE STUDY DESIGN IN PRACTICE

For LM, GA, and TZL designs, specifying different design quantities, $\theta_{des} = (\boldsymbol{\beta}'_{des}, \boldsymbol{p}'_{des})'$, will lead to different phase 2 subsamples. Little attention has been paid to the practical considerations entailed in choosing a study design. One practical strategy is to make an educated guess for the design quantities; another is to consider a range of plausible values. Though adaptive/sequential designs may be feasible in some circumstances,[31] in the post-GWAS setting processing data by batch may be operationally inefficient. In addition, although the sequential strategy will provide more precise design parameters, it will not necessarily aid in solving the withstanding issue of selecting a unique phase 2 sample given that multiple more precise estimates will be potentially identified. Therefore, we propose a strategy that relies only on phase 1 data to select a unique phase 2 sample when a range of effect sizes, allele frequencies, and LD values can be considered at design stage.

4.1 | A grid search procedure to select a unique phase 2 subsample

It is likely that there will be uncertainty about the specification of the effect size ($\boldsymbol{\beta}_{des}$) and haplotype distribution ($\boldsymbol{p}_{des}$) at design stage, so we must consider a range of probable values and define a grid of intermediate points inside this range. Let $\{\theta_h\}$, $h = 1, \ldots, H$ be the set of probable values or design quantities of interest. Each design quantity θ_h will yield an optimal phase 2 subsample, P2S$^{(h)}$, for the second stage. Thus, to select an unique design under the set $\{\theta_h\}_{h=1,\ldots,H}$ we propose the following procedure, which is motivated by robustness considerations.

1. Given an optimality criterion, $\Lambda(\cdot)$, for each h
   - obtain a phase 2 subsample, namely, P2S$^{(h)}$, via LM/GA or otherwise by optimizing $\Lambda(\theta_h)$.
   - given P2S$^{(h)}$, calculate $\Lambda_{h'}^{(h)} = \Lambda^{(h)}(\theta_{h'})$ for $h' = 1, \ldots, H$
   - compute $\kappa^{(h)} = \underset{h' \in 1, \ldots, H}{\kappa} \left\{ \Lambda_{h'}^{(h)} \right\}$, where κ is a summary function, for example, mean or median
2. select the P2S$^{(h)}$ with minimum $\kappa^{(h)}$

This procedure will identify a unique design from the ones generated using alternative specifications $\{\theta_h\}$. To better understand the proposed procedure, let us assume that we are interested in comparing two designs, PS2$^{(1)}$ and PS2$^{(2)}$ which are optimal when the design values are θ_1 or θ_2, respectively. In order to select the best design, we adopt a criterion based on robustness. In other words, we are interested in determining which one of these two designs exhibits an overall superior performance when θ differs from its generating design value. To this end, we compute the fitness function for each design and each $\theta_{h'}$ with $h' \neq h$ and select the design that achieves the best average (or median) performance. The formal description simply extends this principle to comparing H designs and selecting the most robust one. Ultimately, regional sequencing data will be collected for the subsample from the resulting design alone. Once data are collected, statistical fine-mapping can be conducted following the analytic strategy described in Section 2.3.

4.2 | Simulation under realistic LD patterns

The purpose of this simulation study is to evaluate the studied designs when the values of $\theta = (\boldsymbol{\beta}', \boldsymbol{p}')'$ are unknown and a range of values for θ_{des} is considered instead. In this study, we generate data under a scenario where multiple "causal variants" and a realistic LD structure from a targeted region were considered. Specifically, we select four loci in chromosome 16 as causal variants, $\boldsymbol{G} = (G_1, \ldots, G_4)$ with corresponding effect sizes $\boldsymbol{\beta}_1 = (\beta_{1G_1}, \ldots, \beta_{1G_4})' = (-0.200, 0.125, 0.250, -0.150)'$ in hg19 positions 56989830, 56993324, 56994990, 56995236, and designate rs247617 (pos.

56990716) as the GWAS SNP (hereafter all positions are truncated to the last five digits). We then generate a QT, Y, across 500 replicates following $Y = \beta_0 + \boldsymbol{\beta}_1' \boldsymbol{G} + \varepsilon$, where $\beta_0 = 2$, and $\varepsilon \sim \mathcal{N}(0, \sigma^2 = 1)$. Details of the data generation are provided in Section S4.5, Online Supplementary Material.

4.3 | Selecting phase 2 samples under prespecified sets of design quantities

Since multiple causal variants are assumed in this section, we ascertain the performance of the studied designs under alternative optimization criteria in addition to the parameter-specific criterion, specifically under A- and D-optimality. This allows β_1 to be treated as a vector at design stage, that is $\boldsymbol{\beta}_{des} = (\hat{\gamma}_0, \boldsymbol{\beta}_1 = \boldsymbol{0}, \hat{\gamma}_z)'$, with $\boldsymbol{0}$ being a zero vector. As before, $(\hat{\gamma}_0, \hat{\gamma}_z)'$ correspond to the MLEs of the regression model $Y = \gamma_0 + \gamma_z Z$, that is, GWAS MLEs. In the simulated data, $\bar{\hat{\gamma}}_0 = 1.73$ range (1.68-1.78), $\bar{\hat{\gamma}}_z = 0.116$ range (0.097-0.176) across the 500 replicates.

As mentioned in Section 2.4.4, it is straightforward to modify the optimality criterion for LM and GA. For TZL, no specific results were provided under alternative optimality criteria. However, Tao et al discussed that since $\text{Var}(\boldsymbol{G}|Z)$ is a matrix when $\boldsymbol{\beta}_1$ is a vector, it was sufficient to replace the scaling factor $\text{Var}(G|Z)^{1/2}$ (when β_1 is a scalar) with $\Lambda[\text{Var}(\boldsymbol{G}|Z)]$.[10]

Additionally, for LM and GA, the optimization is performed using $\mathbb{V}_1^{-1}$, as this approach showed best performance in the first simulation study when the true values are close to the null hypothesis (Section 3.2). A unique design is then selected for LM, GA, and TZL considering multiple (misspecified) values of $\boldsymbol{p}_{des}$ following Section 4.1. We also considered RDS in this simulation study, however, since RDS does not depend on θ_{des} in any way, it was not determined using the outlined procedure 4.1. For each replicate, we select a phase 2 data of size $n = 1250, 2500$. The specification of $\boldsymbol{p}_{des}$ under parameter-specific, A-, and D-optimality criteria is described below.

4.3.1 | Parameter-specific criterion

Under this criterion, β_1 is a scalar. Thus, we specify $\boldsymbol{p}_{des} = P(G, Z) = p_{gz}$ assuming each of the resulting combinations between the following:

-    $\boldsymbol{q}_G = \left\{ q_G^{Q_1 1}, q_G^{Q_2}, q_G^{Q_3} \right\}$, and $\boldsymbol{r}_{Z,G} = \{ r^{Q_1}, r^{Q_2}, r^{Q_3} \}$,

where Q_1, Q_2, Q_3 denote the first, second, and third quartiles of q_G (MAF) or $r_{Z,G}$ (LD between Z and G) across the 29 sequence variants in the region, for example, $q_G^{Q_2}$ is the median MAF value across seq-SNPs in the fine-mapped region while $r_{Z,G}^{Q_1}$ denotes the 25th percentile across correlation values between the GWAS-SNP, Z, and the seq-SNPs, G.

4.3.2 | A- and D-optimality criteria

Under these criteria, $\boldsymbol{\beta}_1$ is assumed to be a vector of size 2. Thus, we specify $\boldsymbol{p}_{des} = P(G_1, G_2, Z) = p_{g_1 g_2 z}$ assuming each of the resulting combinations between the following:

-    $\boldsymbol{q}_{G_1} = \left\{ q_G^{Q_1}, q_G^{Q_3} \right\}$, $\boldsymbol{q}_{G_2} = \left\{ q_G^{Q_1}, q_G^{Q_3} \right\}$, $\boldsymbol{r}_{Z,G_1} = \left\{ r_{Z,G}^{Q_1}, r_{Z,G}^{Q_3} \right\}$, $\boldsymbol{r}_{Z,G_2} = \left\{ r_{Z,G}^{Q_1}, r_{Z,G}^{Q_3} \right\}$, and $\boldsymbol{r}_{G_1, G_2} = \left\{ r_{G,G'}^{Q_1}, r_{G,G'}^{Q_3} \right\}$,

as before, Q_1 and Q_3 denote the first and third quartiles of q_G (MAF), $r_{Z,G}$ (LD between Z and G), or $r_{G,G'}$ (LD between G and G') across the 29 sequence variants in the region.

In the simulated data, $\boldsymbol{q}_G = \{ q_G^{Q_1}, q_G^{Q_2}, q_G^{Q_3} \} = \{ 0.167, 0.232, 0.294 \}$; $\boldsymbol{r}_{Z,G} = \{ r_{Z,G}^{Q_1}, r_{Z,G}^{Q_2}, r_{Z,G}^{Q_3} \} = \{ -0.229, -0.173, 0.942 \}$; and $\boldsymbol{r}_{G_1, G_2} = \{ r_{G,G'}^{Q_1}, r_{G,G'}^{Q_2}, r_{G,G'}^{Q_3} \} = \{ -0.207, -0.119, 0.306 \}$. The average sample distribution of the resulting phase 2 designs across the 500 replicates is portrayed via mosaic plots (Figures S4 and S5). GA, RDS, and TZL designs show rather similar distributions across optimality criteria especially when $n = 2500$. LM, on the other hand, selects only from the extremes of the distribution for common heterozygous ($Z = 0$) for $n = 1250$. In addition, the intersection of the subsamples taken across each design/optimality criterion combination is presented via upset plots for a single replicate (Figures S3 and 3). These plots show that almost a third of the phase 2 subsamples are common among GA, RDS, and TZL designs and optimality criteria when $n = 1250$. Notably, the number of common subsamples jumps to about a half among GA, RDS, and TZL and optimality criteria when $n = 2500$.
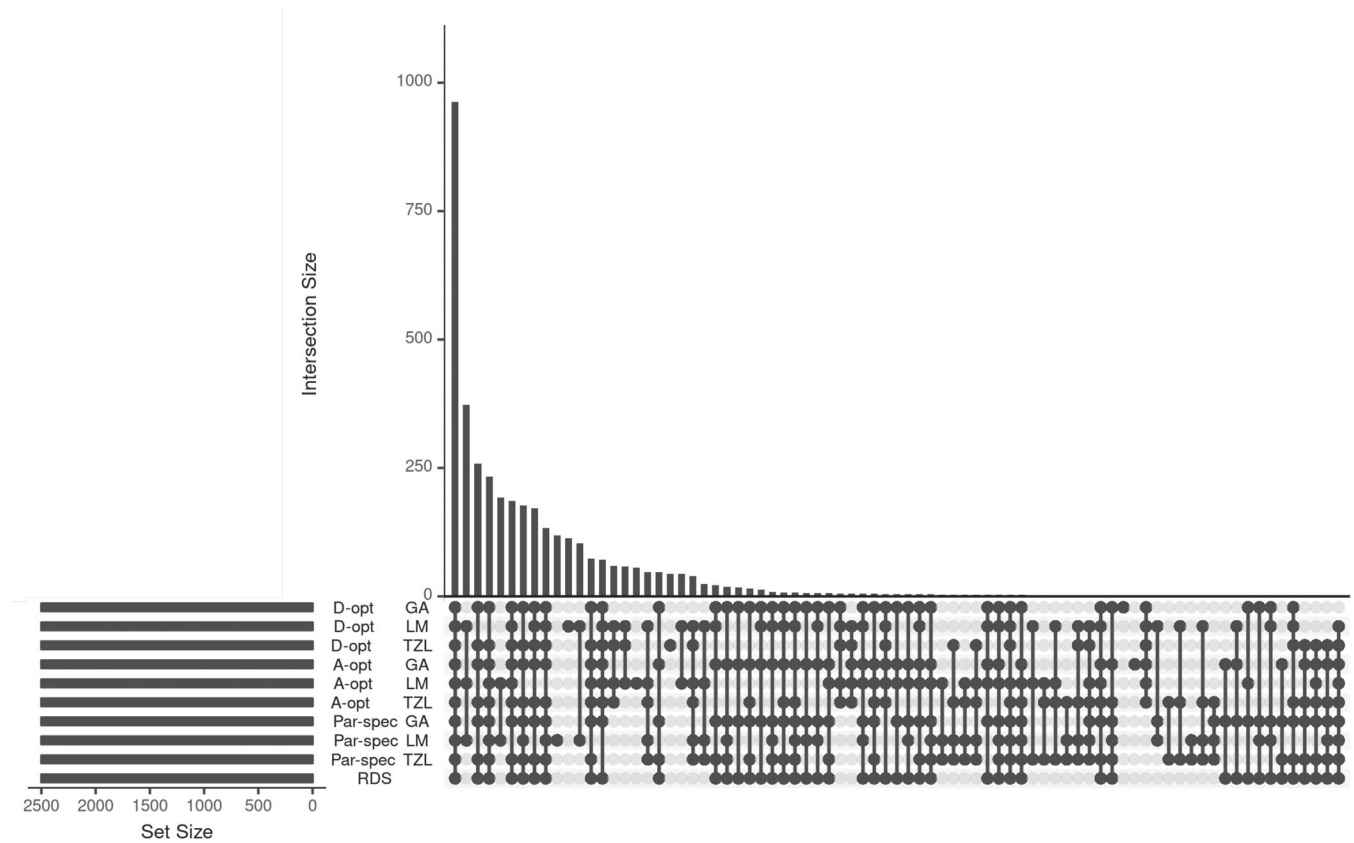
FIGURE 3 Upset plot for a single replicate in the realistic simulation to quantify the intersection sizes across studied designs and optimality criteria when $n = 2500$. Each bar denotes the size of a given intersection highlighted in the x-axis, that is, the number of subjects common among designs. The matrix in the x-axis corresponds to each optimality criterion (parameter-specific/A-/D-optimality) and design (LM/GA/TZL) combination as well as RDS

4.4 | Single-variant fine-mapping analysis

Once the phase 2 sample is selected in a given replicate, we perform a region scan using the 29 variants, that is, we test for association one variant at a time, across each phase 2 sample size ($n = 1250, 2500$). To decrease the collinearity between G and Z in the model, we treat the GWAS SNP, Z (rs247617), as a (three-level) categorical covariate at design and analysis stages. We summarize the point estimates, asymptotic standard errors, and empirical power rates (under the LR test) for the region scans across replicates for each design and optimality criteria (Tables S4 and 4).

GA, RDS, and TZL show similar results in terms of estimation and power across different values of n and optimality criteria (for GA and TZL) whereas LM exhibits considerably lower power (Tables S6 and 4). We observe similar distributions of the LR test p-value (in $-\log_{10}$ scale) across replicates for the studied designs with the exception of some outliers; LM shows the smallest ($-\log_{10}$) p-values compared with the other designs (Figures S6 and 4). Lastly, no optimality criterion shows consistently best estimation nor power, suggesting that no specific criteria substantially improves overall performance when the design values, θ_{des}, are misspecified.

It is obvious that in most cases the mean of the estimate for causal sequence variants does not correspond with its true value being both over- and underestimated (Tables S4 and S5). In fact, this discrepancy occurs even for the complete data case, which is unsurprising considering the unaccounted variation resulting from the single-variant analysis.

The power to detect association (at $\alpha = 0.05/29$) is above 80% in the complete data for three out of four causal variants: 89830, 94990, 95236. The power for the remaining causal variant (93324) is almost zero (Table S6). This decrease in power is likely due to the high LD between this variant and the GWAS SNP, Z ($r = 0.94$ and $D' = 0.95$) which is already included in the regression, thus, diluting its signal. Moreover, there are additional noncausal variants that display power above

TABLE 4 Empirical power rates at significance level $\alpha = 0.05/29$ for causal variants only and $n = 2500$ across 500 replicates in realistic fine-mapping simulation single-variant analysis, that is, misspecified $Var(G|Z)$

G pos.	β_{1G}	Complete	RDS	Par-spec			A-opt			D-opt		
				LM	GA	TZL	LM	GA	TZL	LM	GA	TZL
85805	0.00	8.4	8.6	6.6	8.4	8.0	6.0	8.6	8.2	6.2	8.6	8.8
86045[a]	0.00	76.4	72.4	60.4	72.0	73.0	56.4	72.2	72.6	56.4	72.4	72.2
86762[a]	0.00	100.0	100.0	99.4	100.0	100.0	99.2	100.0	100.0	99.0	100.0	100.0
86914	0.00	8.6	7.2	5.6	7.0	8.0	5.0	7.6	7.6	5.8	7.0	7.8
87015	0.00	0.2	0.2	0.2	0.2	0.2	0.4	0.2	0.0	0.4	0.2	0.0
87765	0.00	0.4	0.4	0.0	0.4	0.4	0.0	0.2	0.2	0.2	0.4	0.2
88044	0.00	0.4	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
88958[a]	0.00	83.6	78.8	62.0	79.2	81.4	59.4	78.6	78.4	60.2	78.8	78.0
89015	0.00	5.4	4.8	2.4	4.8	5.6	4.2	4.8	5.8	3.8	5.0	5.0
89830[b]	-0.20	100.0	100.0	99.6	100.0	100.0	99.8	100.0	100.0	99.6	100.0	100.0
90803[a]	0.00	77.2	74.2	61.8	74.0	74.6	58.2	73.6	74.8	57.4	73.4	74.6
91143[a]	0.00	67.4	63.6	49.2	63.4	64.8	48.0	63.0	64.4	44.8	63.4	62.6
91524	0.00	6.0	5.2	4.4	5.0	6.0	4.6	5.0	5.8	3.6	5.2	5.4
92017	0.00	7.0	5.6	4.8	5.8	6.0	4.6	6.0	6.2	4.2	5.4	5.8
93161	0.00	0.8	0.4	0.2	0.4	0.4	0.0	0.4	0.6	1.0	0.4	0.2
93211	0.00	7.0	6.8	2.4	6.6	7.0	3.2	7.0	6.4	3.4	6.8	6.4
93324[b]	0.12	3.0	2.4	1.0	2.4	2.2	1.2	2.4	2.2	1.4	2.4	2.0
93886	0.00	0.2	0.2	0.0	0.2	0.0	0.0	0.2	0.2	0.2	0.2	0.2
93897	0.00	20.0	17.6	9.2	17.8	16.2	10.2	17.6	16.6	9.6	17.4	17.2
93901	0.00	18.0	16.0	10.4	16.0	15.4	9.8	16.2	16.4	8.2	16.0	17.4
93935[a]	0.00	80.6	76.0	62.6	75.6	75.6	61.4	75.8	76.4	60.2	76.0	75.4
94192[a]	0.00	80.6	76.4	63.2	76.0	75.8	61.8	75.8	75.8	61.6	76.2	75.8
94212	0.00	3.8	2.8	1.8	2.8	2.8	1.0	2.6	2.8	1.4	3.0	2.8
94244	0.00	0.2	0.2	0.0	0.2	0.0	0.0	0.2	0.2	0.2	0.2	0.2
94528	0.00	0.2	0.2	0.0	0.2	0.0	0.0	0.2	0.2	0.2	0.2	0.2
94990[b]	0.25	84.4	82.0	65.8	82.0	81.8	66.2	81.4	81.6	65.8	81.0	80.4
95038[a]	0.00	79.6	74.6	62.2	74.6	75.0	60.4	74.2	74.8	60.2	74.6	74.0
95234	0.00	0.2	0.6	0.2	0.6	0.2	0.0	0.6	0.4	0.8	0.8	0.4
95236[b]	-0.15	95.2	93.0	77.8	92.8	92.4	79.0	93.0	92.0	78.4	93.0	92.4

Note: Base pair positions (pos.) marked with "b" denote causal variants whereas "a" denote hitchhikers. The remaining are noncausal. Positions are truncated to the last five digits.

80% in the complete data analysis (Table S6). These so-called "hitchhiker variants" achieve significant association as a consequence of their LD with causal variants. The performance of the studied designs for the hitchhiker variants resembles the complete data analysis and its ranking is similar to the one shown with the causal variants. These results indicate that single-variant analysis does not distinguish well between causal and hitchhiker SNPs in complete nor two-phase analysis.

A common strategy to identify potential causal variants from hitchhikers consists of adjusting for the most significant variant (or variants) in the region and performing a new—conditional—scan (ie, one variant at a time) fitting the following model: $Y = \beta_0 + \beta_1 G + \beta_2 G_{top} + \beta_z' Z$, where G denotes a variant in the region, G_{top} is the most significant locus from
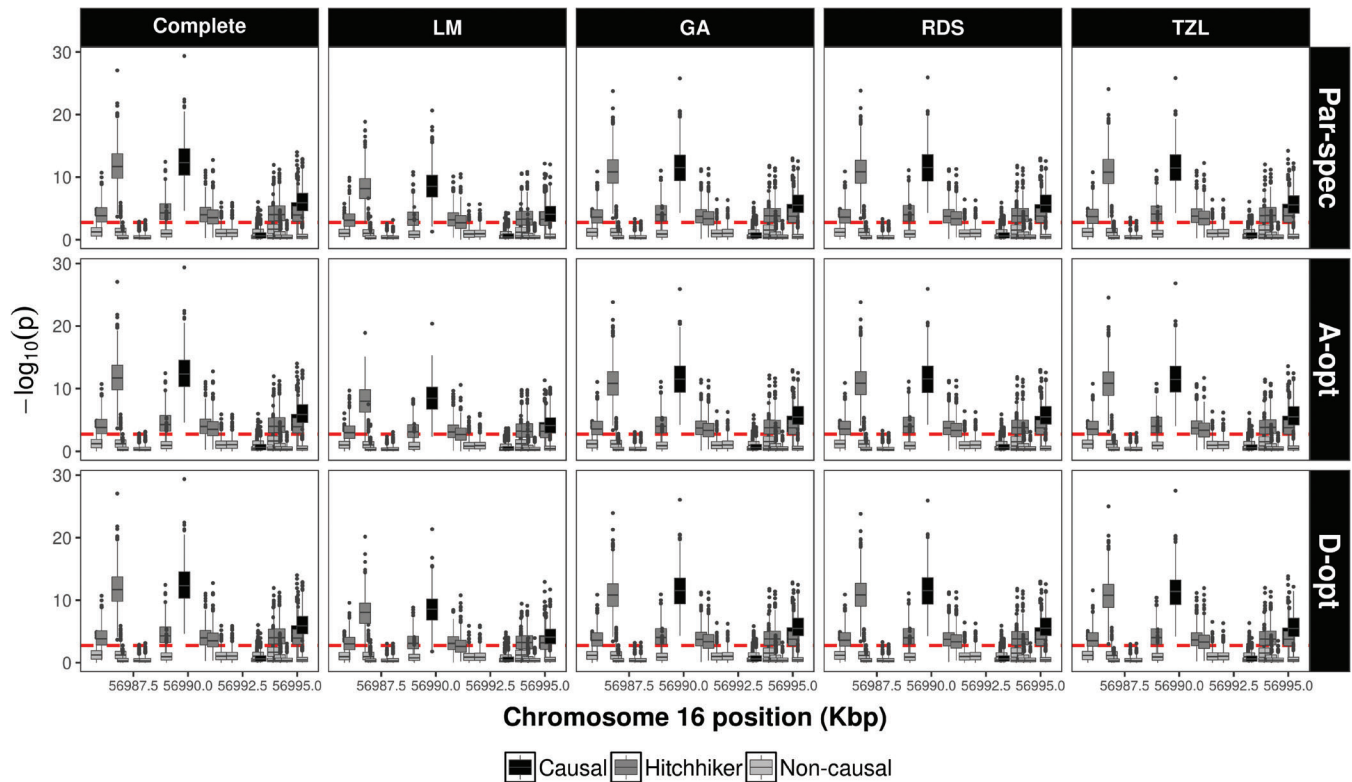
FIGURE 4 Boxplots of the $(-\log_{10})$ p-values across 500 replicates in the fine-mapping simulation single-variant analyses for a phase 2 sample size of $n = 2500$ across optimality criteria (for LM, GA, and TZL only): parameter-specific, A- and D-optimality in each row facet. Each column facet corresponds to the complete data analysis and studied designs respectively. The dashed line corresponds to a Bonferroni-corrected significance threshold of $\alpha = 0.05/29$ [Colour figure can be viewed at wileyonlinelibrary.com]

the single-variant analysis, and Z is the GWAS SNP treated as a (three-level) categorical variable. This approach aims to discover independent signals in the region. Results for this conditional analysis can be found on Section S6, Online Supplementary Material.

5 | APPLICATION IN THE NORTHERN FINLAND BIRTH COHORT OF 1966

We illustrate the methods outlined in Sections 2.4 and 4 using the Northern Finland Birth Cohort of 1966 (NFBC1966), which is a longitudinal, prospective birth cohort consisting of women and their offspring from the two northernmost provinces in Finland: Oulu and Lapland. Comprehensive phenotypic, lifestyle, and demographic data were collected after birth via questionnaires and clinical evaluations on the offspring at years 1, 7, 14-16, and 31. The NFBC1966 aims to study genetic, biological, social, or behavioral risk factors associated with the onset of different diseases as well as morbidity and mortality derived from adverse events such as preterm birth and intrauterine growth retardation.[32,33] In particular, as part of an NHLBI-sponsored project designed to characterize the genetic determinants of metabolic and cardiovascular diseases, special attention was paid to a selected list of heritable quantitative traits related to cardiovascular diseases or type 2 diabetes. These traits are body mass index (BMI), high density lipoproteins (HDL), low density lipoproteins (LDL), triglycerides (TG), glucose (GLU), insulin (INS), C-reactive protein (CRP), systolic blood pressure (SBP), and diastolic blood pressure (DBP).

We focus on 5402 subjects for which genotype information was collected using the Illumina Infinium platform, which is comprised by 346 590 SNPs (after standard quality control). In addition to the genotype information, custom targeted sequencing (CTS) was collected for 4511 of them (83.5%) as part of a series of resequencing studies to deepen the understanding of genotypic variation on metabolic traits.[34] The CTS data contain the coding sequence and 5′ and 3′ untranslated regions of 78 genes, which were selected based on previous GWAS meta-analyses of cardiovascular diseases. Details of these regions can be found in Service et al.[34] The purpose of this illustration is to optimally select a subsample of subjects

for targeted sequencing study (phase 2) and fine-mapping to locate potential causal variants using the methods outlined in the previous sections.

5.1 | Phase 2 subsample selection

We first identify GWAS-SNPs by performing genome-wide associations on the available quantitative traits. Although genome-wide scans on these very same metabolic traits for the NFBC1966 have been previously carried out in Sabatti et al, our analyses differ in a couple of aspects: (1) the sample size we utilized is slightly larger because additional subjects were genotyped at a later time and (2) we perform multiple linear regression adjusting by the SexOCPG covariate described in Sabatti et al, which is a composed categorical variable determined by sex, oral contraceptive use, and pregnancy status.[33]

We center attention on one trait, log-transformed TG (Y), as its GWAS has few peaks that identify only two genetic regions for further study: GCKR in chromosome 2 and LPL in chromosome 8 (Figure S7). We comment on the challenges of more complex GWAS scenarios in the discussion. We locate one SNPs in each of these regions that meet the usual genome-wide significance threshold (5×10^{-8}): rs1260326 (chr2:27730940, $\hat{\gamma}_1 = 0.0614$, s.e.$(\hat{\gamma}_1) = 0.0093$, $p = 5.67 \times 10^{-11}$, MAF= 35.7%), and rs10096633 (chr8:19830921, $\hat{\gamma}_1 = -0.0897$, s.e.$(\hat{\gamma}_1) = 0.015$, $p = 3.24 \times 10^{-9}$, MAF= 9.7%). Due to missing data on the TG values, the available subjects for the genome scan was 5300. Of these, the number of subjects with both GWAS and CTS data is $N = 4493$, which is the phase 1 sample size considered for phase 2 analyses. In addition to the GCKR and LPL regions used in the phase 2 selection, we consider another region for analysis: APOA5, to illustrate the correspondence of the two-phase design and analysis with the complete data approach pursued in Service et al.[34]

Using the two identified GWAS-SNPs, we select phase 2 subsamples under three of the previously described designs: GA, RDS, and TZL. We drop LM as it showed the worst performance in simulations. The phase 2 sample size is specified to be approximately 25%, or 50% of the phase 1 sample size ($n = 1123, 2246$). To define Z, we use all allele combinations of the GWAS-SNPs rs1260326 and rs10096633, which results in a nine-category variable. Considering that no optimality criterion performed best in Section 4, we deem it appropriate to assume β_1 is a scalar and use a parameter-specific criterion, which greatly simplifies the specification of the design quantities, particularly $\boldsymbol{p}_{des}$. Phase 2 subsample selection is performed separately per each phase 2 sample size. In each case, a set of design quantities is defined as follows: First, $\boldsymbol{\beta}_{des} = (\hat{\gamma}_0, \beta_1 = 0, \hat{\boldsymbol{\gamma}}')'$, where $\hat{\gamma}_0$ and $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_1, \hat{\gamma}_2, \hat{\boldsymbol{\gamma}}'_x)'$ correspond to the MLEs from the following regression model based on phase 1 data: $Y = \gamma_0 + \gamma_1 \text{rs1260326} + \gamma_2 \text{rs10096633} + \boldsymbol{\gamma}'_x \boldsymbol{X}$, where $\boldsymbol{X}$ is a vector of additional covariates including SexOCPG and the first four genetic principal components (PC1-4). Second, for $\boldsymbol{p}_{des}$ and given that MAF and LD values are unavailable a priori, we postulate the following ranges for these design quantities: $q_G \in \langle 0.05 + 0.05j | j = 0, 1, \ldots, 6 \rangle$ and $\boldsymbol{r} = \{0.0, \pm 0.17, \pm 0.33, \pm 0.50\}$. We use a categorical coding to define Z in the ML analysis because it reduces the potential collinearity with G, whereas there is no such issue at subsample selection stage, where the typical GWAS analysis involves additive coding for Z (rs1260326 and rs10096633 in this case).

For visualization purposes, we categorize TG (Y_{st}) into three groups corresponding to commonly used blood test ranges, that is, normal (<150 mg/dL), borderline high (150-199 mg/dL), and high (≥200 mg/dL). Notably, the groups in Y_{st} are asymmetrical with respect the middle Y_{st} stratum. On the other hand, the distribution of the nine-category variable determined by the two GWAS-SNPs (rs1260326 and rs10096633) has a small number of subjects for some categories due to the relatively low MAFs of SNPs rs10096633, which differs largely from the simulations (Figure S8). Consequently, the phase 2 subsample distributions tend to not select subjects from those categories. Notably, GA, RDS, and TZL show similar category distribution across phase 2 sample sizes (Figure S9). The proportion of subjects that are common among designs in the phase 2 subsamples is above 95% between pairs of GA, RDS, and TZL across phase 2 sample sizes (Figure S10).

5.2 | Fine-mapping analysis

CTS data for genes GCKR, LPL, and APOA5 were downloaded from the NCBI's dbGAP repository according to their GRCh37.p13 location ±5kbps. Since aligned reads were available, we performed variant calling using the Got-Cloud pipeline developed by the Center for Statistical Genetics at the University of Michigan.[35] Sequence data were
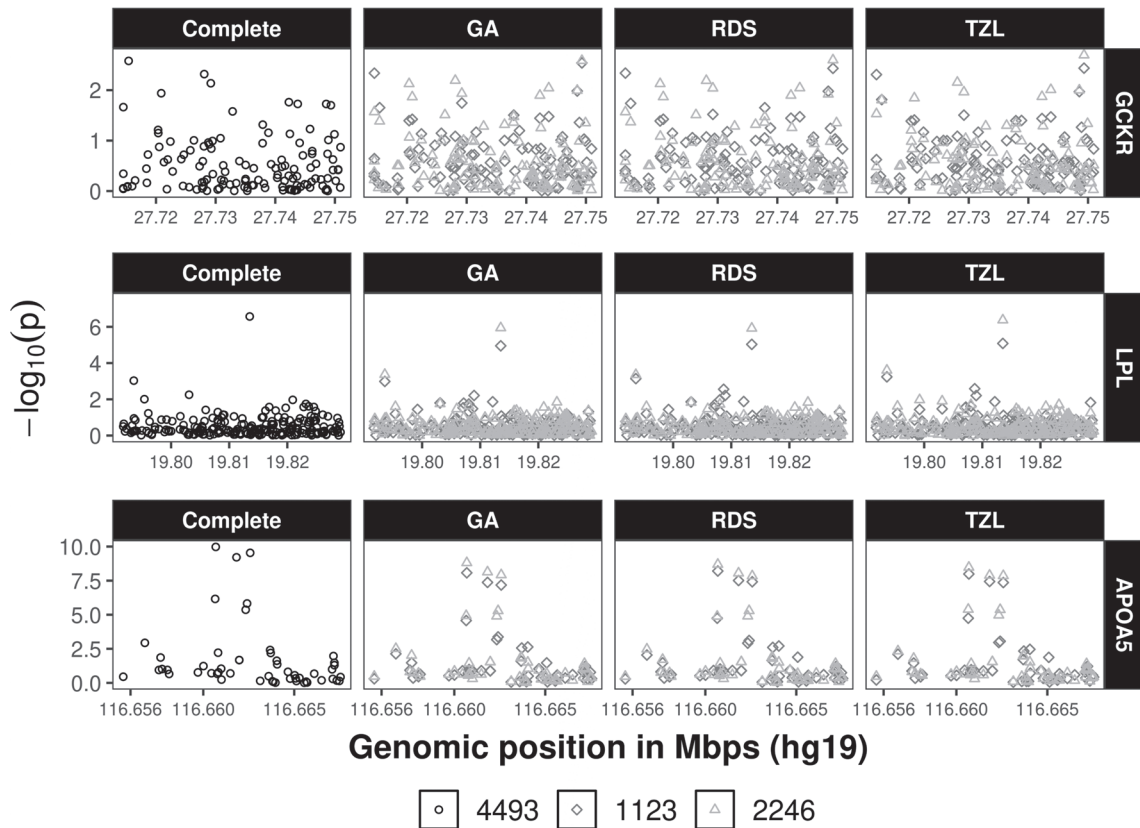
FIGURE 5 Region plots of the NFBC1966 CTS data for the ML analyses under the studied designs compared with the complete data analysis across phase 2 sample sizes ($n = 1123, 2246$). Column facets denote each of the studied designs plus the complete data analysis whereas row facets show each loci of interest

analyzed in two ways. First, we used linear regression with complete data, that is, subjects with both genotyping and CTS data ($N = 4493$), and second via the ML approach described above for each studied design: GA, RDS, and TZL ($n = 1123, 2246$). For the ML analysis, the nine-category variable defined by the GWAS SNPs was used as auxiliary variable Z. All analyses were adjusted by the GWAS-SNPs (rs1260326 and rs10096633), SexOCPG and PC1-PC4 as covariates.

Our main interest lies in gauging the performance of the two-phase designs with respect to the complete data analysis, evaluating both estimation and hypothesis testing. For estimation, we focus on three sequence variants reported in table 2 of Service et al:[34] rs268, rs2266788, and rs3135506 (Table 5). For these variants, GA, RDS, and TZL show similar results with improving performance as n increases. Comparisons of association estimates in Beta-Beta plots show similar spread estimates across designs (Figure S11). Similarly, region plots of association signals for complete data and ML analyses across the studied designs indicate that all designs tend to display results closer to those of the complete data case as n increases with no overwhelmingly better design (Figure 5). In addition to the region scans performed for analyzing common variants, we demonstrate that rare variants can be investigated under a two-phase design via burden tests in Section S7, Online Supplementary Material.

6 | DISCUSSION

In this report, we propose and evaluate state-of-the-art sample selection strategies for two-phase designs in the context of post-GWAS fine-mapping studies. We pay special attention in the comparison against a recently proposed optimal design, TZL. Our first set of simulations, considering a parameter-specific criterion, shows a clear advantage of TZL under the strong assumption of correctly specified design values ($\boldsymbol{p}_{des}$ or $\text{Var}(G|Z)$). On the other hand, TZL demonstrates biased estimation when the phase 2 sample size/nonmissing fraction is small ($n/N = 0.10, 0.25$) and the effect sizes are farther from the null $\beta_1 \geq 0.5$, with improvements noticed when n increases. These results are aligned with those obtained under LM, which reinforces our initial belief that LM is a crude approximation to TZL. Moreover, LM is only competitive

TABLE 5 Estimation and testing results for analyzing (log-transformed) triglyceride levels across three sequence SNPs from the fine-mapping analysis in the NFBC1966

n	Gene	Chr.	pos. (hg19)	Variant	$\hat{\beta}_1$ (S.E.) p-value			
					Complete	**GA**	**RDS**	**TZL**
1123	LPL	8	19813529	rs268	0.186 (0.04) p = 2.69e−07	0.221 (0.06) p = 1.12e−05	0.223 (0.06) p = 9.28e−06	0.221 (0.06) p = 8.24e−06
	APOA5	11	116660686	rs2266788	0.108 (0.02) p = 1.06e−10	0.115 (0.02) p = 8.41e−09	0.116 (0.02) p = 6.23e−09	0.114 (0.02) p = 1.03e−08
			116662407	rs3135506	0.100 (0.02) p = 1.50e−06	0.083 (0.02) p = 4.23e−04	0.081 (0.02) p = 7.78e−04	0.079 (0.02) p = 8.87e−04
2246	LPL	8	19813529	rs268	0.186 (0.04) p = 2.69e−07	0.186 (0.04) p = 1.14e−06	0.185 (0.04) p = 1.20e−06	0.198 (0.04) p = 4.21e−07
	APOA5	11	116660686	rs2266788	0.108 (0.02) p = 1.06e−10	0.106 (0.02) p = 1.53e−09	0.105 (0.02) p = 2.02e−09	0.103 (0.02) p = 3.32e−09
			116662407	rs3135506	0.100 (0.02) p = 1.50e−06	0.096 (0.02) p = 5.10e−06	0.096 (0.02) p = 5.07e−06	0.097 (0.02) p = 4.22e−06

when $n = 2500$, suggesting that the sampling variability introduced by this method can only be overcome under larger nonmissing fractions. In contrast, GA demonstrates unbiased estimation and competitive power (often larger than RDS) across all simulation settings. The appeal of GA lies in its generality as it can be extended to other settings beyond linear, logistic and Cox models while also avoiding uncertainty associated with sampling. Thus, GA provides an alternative approach to obtain efficient and robust two-phase designs across a wider range of settings, including large effect sizes.

Additionally, we investigate the use of different variances considered in the optimization: $\mathbb{J}^{-1}$, and $\mathbb{V}^{-1}$, which are, respectively, the inverse of the Fisher information matrix and the variance-covariance matrix of the most powerful test under the null hypothesis (Sections 3.2 and S3, Online Supplementary Material). The results support the use of $\mathbb{V}^{-1}$ for the parameter-specific criterion when the effect size is close to the null, as expected.

Correct design values specification is never attainable in practice. Hence, our second set of simulations evaluates a more realistic setting for which we propose a grid search approach to select a unique phase 2 design assuming a set of (misspecified) design values is available. In this scenario, GA, RDS, and TZL have comparable performance with LM falling behind. GA is advantageous as it can be easily applied to general $\Lambda(\cdot)$ functions. Thus, apart from the parameter-specific criterion, we examine two additional criteria to select a phase 2 design: A- and D-optimality. Notably, based on our simulations we found no evidence in favor of any particular optimality criterion. However, the parameter-specific criterion may be preferred as designating its design values requires fewer assumptions. A- and D-optimality criteria were chosen because they have been amply explored in the literature, possess solid roots in experimental designs, and have natural connection with hypothesis testing. Nonetheless, other criteria may be better suited for optimizing power, for instance, maximizing the noncentrality parameter of the likelihood ratio or score test χ^2 statistic may improve the power performance of the phase 2 designs.

An important observation from the simulations is that although the best performing designs have higher relative efficiency compared with less favorable designs (eg, combined or simple random sampling) across all phase 2 sample sizes, this improvement does not automatically translate to a closer agreement with the complete data analysis. That is, power performance is contingent upon nonmissing fraction and not necessarily phase 2 sample size itself. This finding is consistent throughout our investigations, where both parameter estimates and p-values achieve similar values as in the complete data analysis only when the phase 2 sample size is half of the phase 1 sample size ($n/N = 0.5$). Thus, careful evaluation of the statistical power of the phase 2 design needs to be considered in advance.

The competitive performance of GA notwithstanding, there are other considerations in implementing this algorithm. For instance, we provide in all simulations an ad-hoc approach to initialize the population of possible solutions to accelerate convergence. In general, giving a particular initialization is not necessary, however, a completely random initial population may need a larger number of generations ($\mathcal{H}$) to achieve good performance. In addition, given the stochastic nature of the search, there are few guarantees that the final solution has indeed reached a global as opposed to a local optimum. It is also worth mentioning that the tuning parameter settings for the proposed GA are intended as a guideline only and do not replace a more careful evaluation in specific problems. Walters suggests iterative calls,[29] which consist of running the GA multiple times, so that the final population in each run serves as the initial population in subsequent runs.

The budgetary constraint implicitly assumes that the cost of sequencing samples is the same for all study samples. This assumption may be relaxed as sequencing costs may vary due to location, tissue availability or number of samples. Thus, extensions to consider differential costs are yet to be considered; one example of such approach under tracing study designs can be found in Moon et al.[36] Another important issue that deserves further investigation in terms of budget constraints (or otherwise) is use of differential sequencing depths across samples.

Further investigation regarding selection of optimal phase 2 subsample under a set of loosely defined design values, $\theta = (\beta', p')' \in \Theta$, is warranted. Indeed, beyond the proposed grid search approach, alternative means to select a phase 2 subsample across ranges of θ are possible, for example, via min-max approaches.[37] However, this selection problem may also be addressed under a Bayesian framework for which a prior (joint) distribution for β and p needs to be specified.[3,38-40] The appeal of this approach is that it may better incorporate the uncertainty in the design values for selecting an phase 2 subsample although at the expense of computational complexity.

Beyond the feasibility and applicability of the proposed methodologies in practice, the illustration on the amply studied NFBC1966 raises some additional questions on considerations posed in the design and analysis of two-phase post-GWAS fine-mapping studies. First, our rare-variant analysis shows no association in either of the studied regions. This result is not that surprising for a couple of reasons: the limited sample size in the NFBC1966 and the low correlation between the GWAS SNPs and the computed genetic score. Additionally, the burden test assumes the same direction

across all variants, which is a limitation in various settings. Thus, further investigations are required for variance component tests under the proposed post-GWAS scenario, potentially by extending ideas from extreme-phenotype sampling designs for cross-sectional association studies.[41] Second, current practice in the field involves the imputation of GWAS data using high-quality reference panels such as the TOPMed Imputation Panel.[42] In principle, one can use the imputed variants to construct a suitable auxiliary variable, Z, to select a phase 2 subsample using the proposed framework. Alternative methods that accommodate differences between genotyped and imputed data for subjects not selected for phase 2 sequencing have been discussed.[43,44] Hence, comparisons between these methods and the approach undertaken in this article can be also evaluated. Lastly, in a similar vein, methodological extensions for situations when multiple loci are pinpointed by GWAS and/or multiple traits of equal interest are collected remain as topics of future work. A starting point in this direction may involve the calculation and application of polygenic risk scores to inform the phase 2 subsampling.

Another issue deserving further investigation is the influence of the phase 2 sample selection in the variant calling pipeline. For simplicity, in this application, variant calling was performed per each locus on all available CTS samples. However, even though genotype likelihoods are typically inferred by sample, population-specific filters such as MAF may change with the design. Thus, sensitivity analyses of these filters can be additionally explored.

We emphasize that although this report focuses on a normally distributed continuous trait, all the derivations apply in the context of generalized linear models within the exponential family. For instance, we include an example on how to obtain the GA design when the response variable is a count and can be modeled using Poisson regression as part of the accompanying Github repository. Furthermore, we are engaged in the development of an R package for this general case. Results of this research and the accompanying software aim to support investigators decision-making pertaining to study design, evaluation, and analysis of two-phase studies. These tools can serve to make more efficient use of limited budgetary resources for data acquisition and analysis.

Two-phase study designs can be sought in other contexts. In particular, their use in a variety of 'omics problems is broadly relevant as new and more costly technologies continue to arise. Beyond the case of fine-mapping where causal variants from GWAS-identified regions can be pinpointed at a fraction of the cost, this approach can be extended for phase 2 variables that are not categorical, for example, methylation, gene expression, or other 'omics measurements. Additionally, methods that introduce functional knowledge to further inform the inference (possibly through Bayesian methods) deserve further investigation.

CONFLICT OF INTEREST
The authors declare no potential conflict of interests.

AUTHOR CONTRIBUTIONS
Osvaldo Espin-Garcia wrote a first draft of this manuscript and subsequent revisions, performed simulations, and analyzed the NFBC1966 data. Radu V. Craiu and Shelley B. Bull provided statistical guidance and methodological support. All authors contributed in the design of the numerical experiments, overall analysis plan, as well as reviewing and approving the final version of the manuscript.

fellow trainee of OICR Biostatistics Training Initiative and CIHR STAGE (Strategic Training for Advanced Genetic Epidemiology) - CIHR Training Grant in Genetic Epidemiology and Statistical Genetics.

ORCID
Osvaldo Espin-Garcia 🄳 https://orcid.org/0000-0003-2052-2626

REFERENCES
1. Walker AM. Anamorphic analysis: sampling and estimation for covariate effects when both exposure and disease are known. *Biometrics*. 1982;38(4):1025-1032.
2. White JE. A two stage design for the study of the relationship between a rare exposure and a rare disease. *Am J Epidemiol*. 1982;115(1):119-128.
3. Karvanen J, Kulathinal S, Gasbarra D. Optimal designs to select individuals for genotyping conditional on observed binary or survival outcomes and non-genetic covariates. *Comput Stat Data Anal*. 2009;53(5):1782-1793. https://doi.org/10.1016/J.CSDA.2008.02.010
4. Schaid DJ, Jenkins GD, Ingle JN, Weinshilboum RM. Two-phase designs to follow-up genome-wide association signals with DNA resequencing studies. *Genet Epidemiol*. 2013;37(3):229-238. https://doi.org/10.1002/gepi.21708
5. Breslow NE, Chatterjee N. Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis. *Appl Stat*. 1999;48(4):457-468. https://doi.org/10.1111/1467-9876.00165
6. Reilly M. Optimal sampling strategies for two-stage studies. *Am J Epidemiol*. 1996;143(1):92-100.
7. McNamee R. Optimal design and efficiency of two-phase case-control studies with error-prone and error-free exposure measures. *Biostatistics*. 2005;6(4):590-603. https://doi.org/10.1093/biostatistics/kxi029
8. McIsaac MA, Cook RJ. Response-dependent two-phase sampling designs for biomarker studies. *Can J Stat*. 2014;42(4):268-284. https://doi.org/10.1002/cjs.11228
9. Zhao Y, Lawless JF, McLeish DL. Design and relative efficiency in two-phase studies. *J Stat Plan Inference*. 2012;142(11):2953-2964. https://doi.org/10.1016/j.jspi.2012.04.013
10. Tao R, Zeng D, Lin DY. Optimal designs of two-phase studies. *J Am Stat Assoc*. 2020;115(532):1946-1959. https://doi.org/10.1080/01621459.2019.1671200
11. 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*. 2018;42(1):104-116. https://doi.org/10.1002/gepi.22099
12. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. 2nd ed. Hoboken, NJ: Wiley; 2002.
13. Robins JM, Hsieh F, Newey W. Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates. *J R Stat Soc B*. 1995;57(2):409-424.
14. Lawless JF, Kalbfleisch JD, Wild CJ. Semiparametric methods for response-selective and missing data problems in regression. *J R Stat Soc B*. 1999;61(2):413-438. https://doi.org/10.1111/1467-9868.00185
15. Zhao Y, Lawless JF, McLeish DL. Likelihood methods for regression models with expensive variables missing by design. *Biom J*. 2009;51(1):123-136. https://doi.org/10.1002/bimj.200810487
16. Zhang Z, Rockette HE. Semiparametric maximum likelihood for missing covariates in parametric regression. *Ann Inst Stat Math*. 2006;58(4):687-706. https://doi.org/10.1007/s10463-006-0047-7
17. Lin DY, Zeng D, Tang ZZ. Quantitative trait analysis in sequencing studies under trait-dependent sampling. *Proc Natl Acad Sci*. 2013;110(30):12247-52. https://doi.org/10.1073/pnas.1221713110
18. Ibrahim JG. Incomplete data in generalized linear models. *J Am Stat Assoc*. 1990;85(411):765-769. https://doi.org/10.1080/01621459.1990.10474938
19. Ibrahim JG, Chen MH, Lipsitz SR. Monte Carlo EM for missing covariates in parametric regression models. *Biometrics*. 1999;55(2):591-596. https://doi.org/10.1111/j.0006-341X.1999.00591.x
20. Zhang Z, Rockette HE. An EM algorithm for regression analysis with incomplete covariate information. *J Stat Comput Simul*. 2007;77(2):163-173.
21. Louis TA. Finding the observed information matrix when using the EM algorithm. *J R Stat Soc*. 1982;44:226-233.
22. Schaid DJ, Chen W, Larson NB. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet*. 2018;19(8):491-504. https://doi.org/10.1038/s41576-018-0016-z
23. Reilly M, Pepe MS. A mean score method for missing and auxiliary covariate data in regression models. *Biometrika*. 1995;82(2):299-314.
24. Whittemore A, Halpern J. Multi-stage sampling in genetic epidemiology. *Stat Med*. 1997;16(1-3):153-167.
25. Holland JH. Genetic algorithms. *Sci Am*. 1992;267(1):66-73.
26. Whitley D. A genetic algorithm tutorial. *Stat Comput*. 1994;4(2):65-85. https://doi.org/10.1007/BF00175354
27. Ruiz-Torrubiano R, García-Moratilla S, Suárez A. optimization problems with cardinality constraints. *Computational Intelligence in Optimization*. Berlin/Heidelberg, Germany: Springer; 2010:105-130.

28. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2016.

29. Wolters MA. A genetic algorithm for selection of fixed-size subsets with application to design problems. *J Stat Softw*. 2015;68(Code Snippet 1):1455. https://doi.org/10.18637/jss.v068.c01

30. Emery AF, Nenarokomov AV. Optimal experiment design. *Meas Sci Technol*. 1998;9:864-876.

31. McIsaac MA, Cook RJ. Adaptive sampling in two-phase designs: a biomarker study for progression in arthritis. *Stat Med*. 2015;34(21):2899-2912. https://doi.org/10.1002/sim.6523

32. Järvelin MR, Sovio U, King V, et al. Early life factors and blood pressure at age 31 years in the 1966 Northern Finland birth cohort. *Hypertension*. 2004;44(6):838-846. https://doi.org/10.1161/01.HYP.0000148304.33869.ee

33. Sabatti C, Service SK, Hartikainen AL, et al. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat Genet*. 2009;41(1):35-46. https://doi.org/10.1038/ng.271

34. Service SK, Teslovich TM, Fuchsberger C, et al. Re-sequencing expands our understanding of the phenotypic impact of variants at GWAS loci. *PLoS Genet*. 2014;10(1):e1004147. https://doi.org/10.1371/journal.pgen.1004147

35. Jun G, Wing MK, Abecasis GR, Kang HM. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Res*. 2015;25(6):918-925. https://doi.org/10.1101/gr.176552.114

36. Moon NC, Zeng L, Cook RJ. Tracing studies in cohorts with attrition: selection models for efficient sampling. *Stat Med*. 2018;37(15):2354-2366. https://doi.org/10.1002/sim.7646

37. Sitter RR. Robust designs for binary data. *Biometrics*. 1992;48(4):1145-1155.

38. Chaloner K, Verdinelli I. Bayesian experimental design: a review. *Stat Sci*. 1995;10(3):273-304.

39. Zhou X, Joseph L, Wolfson DB, Bélisle P. A Bayesian A-optimal and model robust design criterion. *Biometrics*. 2003;59(4):1082-1088.

40. Karvanen J, Vanhatalo J, Auranen K, Kulathinal S, Mäntyniemi S. Optimal design of observational studies: overview and synthesis; 2016. arxiv.org/abs/1609.08347

41. Bjørnland T, Bye A, Ryeng E, Wisløff U, Langaas M. Powerful extreme phenotype sampling designs and score tests for genetic association studies. *Stat Med*. 2018;37(28):4234-4251. https://doi.org/10.1002/sim.7914

42. Taliun D, Harris DN, Kessler MD, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed program. *Nature*. 2021;590(7845):290-299. https://doi.org/10.1038/s41586-021-03205-y

43. Hu YJ, Li Y, Auer PL, Lin DY. Integrative analysis of sequencing and array genotype data for discovering disease associations with rare mutations. *Proc Natl Acad Sci*. 2015;112(4):201406143. https://doi.org/10.1073/pnas.1406143112

44. Wong KY, Zeng D, Lin DY. Robust score tests with missing data in genomics studies. *J Am Stat Assoc*. 2019;114(528):1778-1786. https://doi.org/10.1080/01621459.2018.1514304

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

APPENDIX. DERIVATION OF THE INFORMATION MATRICES

This section derives the observed and expected information matrices, which allow us to compute the optimality criterion. The expected information matrix in particular is called repeatedly during the optimization performed in both LM and GA approaches.

A.1 Log-likelihood

Considering the observed-data likelihood in (1), the log-likelihood takes the form

$$
\begin{aligned}
\ell(\boldsymbol{\beta}, \boldsymbol{p}) &= \sum_{i=1}^{N} \left[R_i \times \left(\log \left\{ f_{\boldsymbol{\beta}}(y_i|g_i, z_i) \right\} + \log \left\{ p_{g_i, z_i} \right\} \right) + (1 - R_i) \times \log \left\{ \sum_{g \in \mathcal{G}} f_{\boldsymbol{\beta}}(y_i|g, z_i) p_{g, z_i} \right\} \right] + c \\
&= \sum_{i=1}^{N} \left[R_i \times \left(\log \left\{ f_{\boldsymbol{\beta}}(y_i|g_i, z_i) \right\} + \log \left\{ \sum_{j=1}^{J} \mathbb{1}\left\{ (g_i, z_i) = j \right\} p_j \right\} \right) \right. \\
&\quad \left. + (1 - R_i) \times \log \left\{ \sum_{j=1}^{J} f_{\boldsymbol{\beta}}(y_i|g^j, z_i) p_j \right\} \right] + c,
\end{aligned}
\tag{A1}
$$

where $\boldsymbol{p} = (p_1, \dots, p_J)$ is the vector of probabilities corresponding to the J unique pairs (g, z) in $\mathcal{G} \times \mathcal{Z}$ using similar notation as in Reference 17 and g^j is the value $g' \in \mathcal{G}$ indexed by $(g', z_i) = j$.

A.2 Score equations

Taking the first derivative of (A1) with respect to $\boldsymbol{\beta}$ and $\boldsymbol{p}$ for a single observation, we have

$$\frac{\partial \ell_i(\boldsymbol{\beta}, \boldsymbol{p})}{\partial \boldsymbol{\beta}} = R_i \frac{\partial \log \{f_{\boldsymbol{\beta}}(y_i|g_i, z_i)\}}{\partial \boldsymbol{\beta}} + (1 - R_i) \frac{\partial \log \left\{\sum_{j=1}^J f_{\boldsymbol{\beta}}(y_i|g^j, z_i)p_j\right\}}{\partial \boldsymbol{\beta}}$$

$$= R_i S_{\boldsymbol{\beta}}(y_i|g_i, z_i) + (1 - R_i) \sum_{j=1}^J \omega_{i,g^j} S_{\boldsymbol{\beta}}(y_i|g^j, z_i),$$

$$\frac{\partial \ell_i(\boldsymbol{\beta}, \boldsymbol{p})}{\partial p_j} = R_i \frac{\partial \log \left\{\sum_{j=1}^J \mathbb{1}\{(g_i, z_i) = j\}p_j\right\}}{\partial p_j} + (1 - R_i) \frac{\partial \log \left\{\sum_{j=1}^J f_{\boldsymbol{\beta}}(y_i|g^j, z_i)p_j\right\}}{\partial p_j}$$

$$= R_i \frac{\mathbb{1}\{(g_i, z_i) = j\}}{\sum_{m=1}^J \mathbb{1}\{(g_i, z_i) = m\}p_m} + (1 - R_i) \frac{f_{\boldsymbol{\beta}}(y_i|g^j, z_i)}{\sum_{m=1}^J f_{\boldsymbol{\beta}}(y_i|g^m, z_i)p_m}$$

$$= R_i \frac{\mathbb{1}\{(g_i, z_i) = j\}}{p_j} + (1 - R_i) \frac{\omega_{i,g^j}}{p_j} \quad j = 1, \dots, J,$$

for the exponential family case, $S_{\boldsymbol{\beta}}(y_i|g_i, z_i) = (y_i - \mu_i) \frac{1}{V(\mu_i)a(\phi)} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}}$, where $V(\mu_i)$ is the variance function, $a(\phi)$ is the dispersion parameter and $\omega_{i,g^j} = \frac{f_{\boldsymbol{\beta}}(y_i|g^j, z_i)p_j}{\sum_{j'=1}^J f_{\boldsymbol{\beta}}(y_i|g^{j'}, z_i)p_{j'}} = \Pr(G = g^j|y_i, z_i)$ is the profile weight.

A.3 Observed information matrix

The second derivatives with respect to $\boldsymbol{\beta}$ and $\boldsymbol{p}$ for a single observation are as follows

$$\frac{\partial \ell_i^2(\boldsymbol{\beta}, \boldsymbol{p})}{\partial \beta_q \partial \beta_l} = R_i \frac{\partial S_{\beta_l}(y_i|g_i, z_i)}{\partial \beta_q \partial \beta_q} + (1 - R_i) \sum_{j=1}^J \left\{ \omega_{i,g^j} \frac{\partial S_{\beta_l}(y_i|g^j, z_i)}{\partial \beta_q} + S_{\beta_l}(y_i|g^j, z_i) \frac{\partial \omega_{i,g^j}}{\partial \beta_q} \right\}$$

$$= R_i \frac{\partial S_{\beta_l}(y_i|g_i, z_i)}{\partial \beta_q} + (1 - R_i) \left\{ \sum_{j=1}^J \omega_{i,g^j} \left[\frac{\partial S_{\beta_l}(y_i|g^j, z_i)}{\partial \beta_q} + S_{\beta_l}(y_i|g^j, z_i) S_{\beta_q}(y_i|g^j, z_i) \right] \right.$$

$$\left. - \left(\sum_{j=1}^J \omega_{i,g^j} S_{\beta_l}(y_i|g^j, z_i) \right) \left(\sum_{j=1}^J \omega_{i,g^j} S_{\beta_q}(y_i|g^j, z_i) \right) \right\},$$

$$= R_i \frac{\partial S_{\beta_l}(y_i|g_i, z_i)}{\partial \beta_q} + (1 - R_i) \left\{ \sum_{j=1}^J \frac{\omega_{i,g^j}}{f_{\boldsymbol{\beta}}(y_i|g^j, z_i)} \frac{\partial^2 f_{\boldsymbol{\beta}}(y_i|g^j, z_i)}{\partial \beta_q \partial \beta_l} \right.$$

$$\left. - \left(\sum_{j=1}^J \omega_{i,g^j} S_{\beta_l}(y_i|g^j, z_i) \right) \left(\sum_{j=1}^J \omega_{i,g^j} S_{\beta_q}(y_i|g^j, z_i) \right) \right\},$$

$$\frac{\partial \ell_i^2(\boldsymbol{\beta}, \boldsymbol{p})}{\partial p_m \partial p_j} = -R_i \frac{\mathbb{1}\{(g_i, z_i) = j\}\mathbb{1}\{j = m\}}{p_j^2} - (1 - R_i) \frac{\omega_{i,g^j} \omega_{i,g^m}}{p_j p_m},$$

$$\frac{\partial \ell_i^2(\boldsymbol{\beta}, \boldsymbol{p})}{\partial \boldsymbol{\beta} \partial p_j} = \frac{\partial \ell_i^2(\boldsymbol{\beta}, \boldsymbol{p})}{\partial p_j \partial \boldsymbol{\beta}} = (1 - R_i) \left[\frac{1}{p_j} \frac{\partial \omega_{i,g^j}}{\partial \boldsymbol{\beta}} \right] = (1 - R_i) \sum_{m=1}^J S_{\boldsymbol{\beta}}(y_i|g^m, z_i) \frac{\partial \omega_{i,g^m}}{\partial p_j}$$

$$= (1 - R_i) \left[\frac{\omega_{i,g^j}}{p_j} \left(S_{\boldsymbol{\beta}}(y_i|g^j, z_i) - \sum_{m=1}^J \omega_{i,g^m} S_{\boldsymbol{\beta}}(y_i|g^m, z_i) \right) \right].$$

For the generalized linear model with covariates $x_i = (x_{i0}, \dots, x_{ip})$, $\frac{\partial S_{\beta_l}(y_i|x_i)}{\partial \beta_q} = (y_i - \mu_i) \frac{\partial}{\partial \beta_q} \left(\frac{1}{V(\mu_i)a(\phi)} \frac{\partial \mu_i}{\partial \beta_l} \right) - x_{iq} W_i x_{il}$, where $W_i = \frac{(\partial \mu_i/\partial \eta_i)^2}{V(\mu_i)a(\phi)}$; in this case $x_i = (1, g_i, z_i)$.

A.4 Expected information matrix

The expected information matrix is obtained by taking the expectation of the expressions above with respect to (R, Y, G, Z).

Note that

$$E_{R,Y,G,Z}\left[R \times h(Y, G, Z)\right] = \int \sum_{g,z} h(y, g, z) \Pr(R = 1, Y = y, G = g, Z = z) dy$$

$$= \int \sum_{g,z} h(y, g, z) \Pr(R = 1|y, g, z) \Pr(Y = y, G = g, Z = z) dy$$

$$= \int \sum_{g,z} h(y, g, z) \Pr(R = 1|y, z) \Pr(Y = y, G = g, Z = z) dy \quad \text{MAR assumption}$$

$$= \int \sum_{g,z} h(y, g, z) \pi(y, z; \psi) \Pr(Y = y, G = g, Z = z) dy,$$

where the sums are taken with respect to all observed values of (g, z) and $\Pr(Y = y, G = g, Z = z) = \Pr(y, g, z) = f_\beta(y|g, z) p_{g,z}$.

Analogously,

$$E_{R,Y,G,Z}\left[(1 - R) \times h(Y, G, Z)\right] = \int \sum_{g,z} h(y, g, z) \Pr(R = 0, Y = y, G = g, Z = z) dy$$

$$= \int \sum_{g,z} h(y, g, z) \Pr(R = 0|y, g, z) \Pr(Y = y, G = g, Z = z) dy$$

$$= \int \sum_{g,z} h(y, g, z) \Pr(R = 0|y, z) \Pr(Y = y, G = g, Z = z) dy$$

$$= \int \sum_{g,z} h(y, g, z)[1 - \pi(y, z; \psi)] \Pr(Y = y, G = g, Z = z) dy.$$

Thus,

$$E_{R,Y,G,Z}\left[\frac{\partial \ell_i^2(\beta, p)}{\partial \beta_q \partial \beta_l}\right] = \int \sum_{g,z} \pi(y, z; \psi) \frac{\partial S_{\beta_l}(y|g, z)}{\partial \beta_q} \Pr(y, g, z) dy$$

$$+ \int \sum_{g,z} [1 - \pi(y, z; \psi)] \sum_{j=1}^{J} p_j \frac{\partial^2 f_\beta(y|g^j, z)}{\partial \beta_q \partial \beta_l} dy$$

$$- \int \sum_{g,z} [1 - \pi(y, z; \psi)] \left(\sum_{j=1}^{J} \omega_{g^j} S_{\beta_l}(y|g^j, z)\right) \left(\sum_{j=1}^{J} \omega_{g^j} S_{\beta_q}(y|g^j, z)\right) \Pr(y, z) dy,$$

$$E_{R,Y,G,Z}\left[\frac{\partial \ell_i^2(\beta, p)}{\partial p_m \partial p_j}\right] = -\int \sum_{g,z} \pi(y, z; \psi) \frac{\mathbb{1}\{(g, z) = j\} \mathbb{1}\{j = m\}}{p_j^2} \Pr(y, g, z) dy$$

$$- \int \sum_z [1 - \pi(y, z; \psi)] \frac{\omega_{g^j} \omega_{g^m}}{p_j p_m} \Pr(y, z) dy,$$

$$E_{R,Y,G,Z}\left[\frac{\partial \ell_i^2(\beta, p)}{\partial p_j \partial \beta}\right] = \int \sum_z [1 - \pi(y, z; \psi)] \frac{\partial f_\beta(y|g^j, z)}{\partial \beta} dy$$

$$- \int \sum_z [1 - \pi(y, z; \psi)] f_\beta(y|g^j, z) \sum_{m=1}^{J} \omega_{g^m} S_\beta(y|g^m, z) dy,$$

where $\Pr(y, z) = \sum_g \Pr(y, g, z)$ and $\omega_{g^j} = \frac{f_\beta(y|g^j, z) p_j}{\sum_{j'=1}^{J} f_\beta(y|g^{j'}, z) p_{j'}}$.

However, the expressions above can be computationally challenging especially when called multiple times to find the optimal design. Instead, considering that Y, Z are observed for all subjects in phase 1, we propose the following approximation by noting that $E_{R,Y,G,Z}[\cdot] = E_{Y,Z}[E_{R,G|Y,Z}[\cdot]]$, which showed adequate results in practice.

First, observe that

$$
\mathrm{E}_{R,G|Y,Z}\left[\frac{\partial \ell_i^2(\boldsymbol{\beta},\boldsymbol{p})}{\partial \beta_q \partial \beta_l}\right] = \pi(y_i,z_i;\psi)\sum_g \frac{\partial S_{\beta_l}(y_i|g,z_i)}{\partial \beta_q}\Pr(g|y_i,z_i)
$$

$$
+ [1-\pi(y_i,z_i;\psi)]\sum_{j=1}^{J}\frac{\omega_{i,g^j}}{f_\beta(y_i|g^j,z_i)}\frac{\partial^2 f_\beta(y_i|g^j,z_i)}{\partial \beta_q \partial \beta_l}
$$

$$
- [1-\pi(y_i,z_i;\psi)]\left(\sum_{j=1}^{J}\omega_{i,g^j}S_{\beta_l}(y_i|g^j,z_i)\right)\left(\sum_{j=1}^{J}\omega_{i,g^j}S_{\beta_q}(y_i|g^j,z_i)\right),
$$

$$
\mathrm{E}_{R,G|Y,Z}\left[\frac{\partial \ell_i^2(\boldsymbol{\beta},\boldsymbol{p})}{\partial p_m \partial p_j}\right] = -\pi(y_i,z_i;\psi)\sum_g \frac{\mathbb{1}\{(g,z_i)=j\}\mathbb{1}\{j=m\}}{p_j^2}\Pr(g|y_i,z_i)
$$

$$
- [1-\pi(y_i,z_i;\psi)]\frac{\omega_{i,g^j}\omega_{i,g^m}}{p_j p_m},
$$

$$
\mathrm{E}_{R,G|Y,Z}\left[\frac{\partial \ell_i^2(\boldsymbol{\beta},\boldsymbol{p})}{\partial p_j \partial \boldsymbol{\beta}}\right] = [1-\pi(y_i,z_i;\psi)]\frac{\omega_{i,g^j}}{p_j}S_\beta(y_i|g^j,z_i)
$$

$$
- [1-\pi(y_i,z_i;\psi)]\frac{\omega_{i,g^j}}{p_j}\sum_{m=1}^{J}\omega_{i,g^m}S_\beta(y_i|g^m,z_i).
$$

Second, we approximate the expected information matrix, $\mathcal{I}$, by averaging across all observed values of (y_i,z_i) as follows

$$
\mathcal{I}_{[\beta_l,\beta_q]} \doteq \sum_{i=1}^{N}\Pr(y_i,z_i)\left[\pi(y_i,z_i;\psi)\sum_g \frac{\partial S_{\beta_l}(y_i|g,z_i)}{\partial \beta_q}\Pr(g|y_i,z_i)\right.
$$

$$
+ [1-\pi(y_i,z_i;\psi)]\sum_{j=1}^{J}\frac{\omega_{i,g^j}}{f_\beta(y_i|g^j,z_i)}\frac{\partial^2 f_\beta(y_i|g^j,z_i)}{\partial \beta_q \partial \beta_l}
$$

$$
\left. - [1-\pi(y_i,z_i;\psi)]\left(\sum_{j=1}^{J}\omega_{i,g^j}S_{\beta_l}(y_i|g^j,z_i)\right)\left(\sum_{j=1}^{J}\omega_{i,g^j}S_{\beta_q}(y_i|g^j,z_i)\right)\right],
$$

$$
\mathcal{I}_{[p_m,p_j]} \doteq \sum_{i=1}^{N}\Pr(y_i,z_i)\left[-\pi(y_i,z_i;\psi)\sum_g \frac{\mathbb{1}\{(g,z_i)=j\}\mathbb{1}\{j=m\}}{p_j^2}\Pr(g|y_i,z_i)\right.
$$

$$
\left. - [1-\pi(y_i,z_i;\psi)]\frac{\omega_{i,g^j}\omega_{i,g^m}}{p_j p_m}\right],
$$

$$
\mathcal{I}_{[p_j,\beta]} \doteq \sum_{i=1}^{N}\Pr(y_i,z_i)\left[[1-\pi(y_i,z_i;\psi)]\frac{\omega_{i,g^j}}{p_j}S_\beta(y_i|g^j,z_i)\right.
$$

$$
\left. - [1-\pi(y_i,z_i;\psi)]\frac{\omega_{i,g^j}}{p_j}\sum_{m=1}^{J}\omega_{i,g^m}S_\beta(y_i|g^m,z_i)\right].
$$

Moreover, since $\pi(y_i,z_i;\psi) = \pi_i(\psi) = \pi_k(\psi)$ for all (y_i,z_i) in the kth stratum, $k=1,\dots,K$. The expressions above can be calculated separately per each group, simplifying the computations.