Using Genetic Algorithms in the Design of Two-phase Studies

Osvaldo Espin-Garcia, PhD[1,2]

[1] Department of Epidemiology and Biostatistics, Western University
[2] Dalla Lana School of Public Health and Department of Statistical Sciences, University of Toronto

Canadian Mathematical Society Winter Meeting
December 5th, 2022

Outline

Background

Given typically modest genetic effect sizes, large-scale population studies are necessary

Background

Given typically modest genetic effect sizes, large-scale population studies are necessary

Post-GWAS analysis often focuses on fine-mapping targeted genetic regions

Background

Given typically modest genetic effect sizes, large-scale population studies are necessary

Post-GWAS analysis often focuses on fine-mapping targeted genetic regions

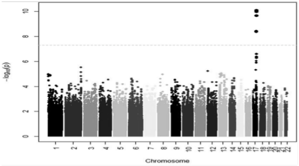
Costs of next-generation sequencing remain prohibitively expensive for large studies

Motivation: Two-phase Fine-mapping Studies

GWAS

$$Y = \gamma_0 + \gamma_Z Z$$

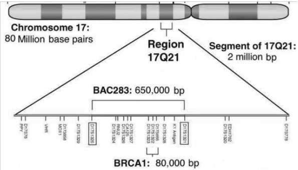
Phase 1 (N)



Fine-mapping

$$Y = \beta_0 + \beta_G G$$

Phase 2, S_2 (n)



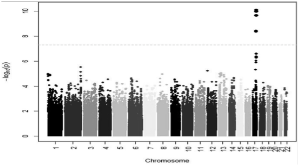
Y is an outcome of interest (quantitative trait, QT), Z is the lead SNP, and G is a (fine-mapped) sequence variant

Motivation: Two-phase Fine-mapping Studies

GWAS

$Y = \gamma_0 + \gamma_Z Z$

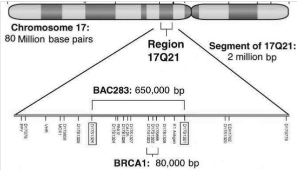
Phase 1 (N)



Fine-mapping

$Y = \beta_0 + \beta_G G$

Phase 2, S_2 (n)



Y is an outcome of interest (quantitative trait, QT), Z is the lead SNP, and G is a (fine-mapped) sequence variant

Aim: Identify designs that select informative individuals for S_2 data collection

Outline

Regression via Maximum Likelihood Methods

The observed data likelihood is defined as:

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

$$= \prod_{i \in S_2} f_\beta(y_i | g_i, z_i) p_{g_i, z_i} \prod_{i \in \bar{S}_2} \sum_{g \in \mathcal{G}} f_\beta(y_i | g, z_i) p_{g, z_i} \tag{2}$$

Regression via Maximum Likelihood Methods

The observed data likelihood is defined as:

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

$$= \prod_{i \in S_2} f_\beta(y_i|g_i, z_i) p_{g_i, z_i} \prod_{i \in \bar{S}_2} \sum_{g \in \mathcal{G}} f_\beta(y_i|g, z_i) p_{g, z_i} \tag{2}$$

where

- $R_i = \mathbb{1}\{i \in S_2\}$ with $\mathbb{P}(R_i = 1|y_i, z_i) = \pi(y_i, z_i; \psi)$, $\boldsymbol{R} = (R_1, \ldots, R_N)'$, and $\sum_{i=1}^{N} R_i = n$.
- $f_\beta(y|g, z)$ is a member of the exponential family, focus on $Y|G = g, Z = z \sim \mathcal{N}(\beta_0 + \beta_1 g + \beta_2 z, \sigma^2)$ with $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$.
- $p_{g,z}$ is the joint probability of $G = g$ and $Z = z$, $\boldsymbol{p} = \{p_{g,z}\}_{g \in \mathcal{G}, z \in \mathcal{Z}}$.
- $\mathcal{G}$, $\mathcal{Z}$ are sets of uniquely observed values of G (in S_2) and Z (in $S_2 \cup \bar{S}_2$).
- $\bar{S}_2$ is the set of $N - n$ unselected subjects.

Remarks on Statistical Analysis

- A missing (by design) data problem $\Rightarrow$ semiparametric maximum likelihood.

Remarks on Statistical Analysis

- A missing (by design) data problem $\Rightarrow$ semiparametric maximum likelihood.
    - Parametric part corresponds to a generalized linear model (β).

Remarks on Statistical Analysis

- A missing (by design) data problem $\Rightarrow$ semiparametric maximum likelihood.
  - Parametric part corresponds to a generalized linear model ($\boldsymbol{\beta}$).
  - Nonparametric part models the joint probability between Z and G ($\boldsymbol{p}$).

Remarks on Statistical Analysis

- A missing (by design) data problem $\Rightarrow$ semiparametric maximum likelihood.
  - Parametric part corresponds to a generalized linear model ($\boldsymbol{\beta}$).
  - Nonparametric part models the joint probability between Z and G ($\boldsymbol{p}$).
- Maximum likelihood estimates are obtained via the EM algorithm.

Remarks on Statistical Analysis

- A missing (by design) data problem $\Rightarrow$ semiparametric maximum likelihood.
    - Parametric part corresponds to a generalized linear model ($\boldsymbol{\beta}$).
    - Nonparametric part models the joint probability between Z and G ($\boldsymbol{p}$).
- Maximum likelihood estimates are obtained via the EM algorithm.
- The variance-covariance matrix (VCM) is computed via the Louis' method (1982).

Remarks on Statistical Analysis

- A missing (by design) data problem $\Rightarrow$ semiparametric maximum likelihood.
    - Parametric part corresponds to a generalized linear model ($\boldsymbol{\beta}$).
    - Nonparametric part models the joint probability between Z and G ($\boldsymbol{p}$).
- Maximum likelihood estimates are obtained via the EM algorithm.
- The variance-covariance matrix (VCM) is computed via the Louis' method (1982).
- Main interest lies in testing for $\mathrm{H}_0 : \beta_1 = 0$ vs $\mathrm{H}_A : \beta_1 \neq 0$.

Outline

Outcome/Covariate - Dependent Sampling

Let $R_i = \mathbb{1}\{i \in S_2\}$ and

$$\mathbb{P}(R_i = 1|Z_i, Y_i, G_i) = \mathbb{P}(R_i = 1|Z_i, Y_i) = \pi_i, \ i = 1, \ldots, N \quad \text{(Missing at random)}$$

Outcome/Covariate - Dependent Sampling

Let $R_i = \mathbb{1}\{i \in S_2\}$ and

$$\mathbb{P}(R_i = 1|Z_i, Y_i, G_i) = \mathbb{P}(R_i = 1|Z_i, Y_i) = \pi_i,\ i = 1, \ldots, N \ \text{(Missing at random)}$$

Marginal sampling:

$$\mathbb{P}(R_i = 1|Z_i, Y_i) = \begin{cases} \pi(Z_i) & \text{covariate-dependent} \\ \pi(Y_i) & \text{outcome-dependent} \end{cases}$$

Outcome/Covariate - Dependent Sampling

Let $R_i = \mathbb{1}\{i \in S_2\}$ and

$$\mathbb{P}(R_i = 1 | Z_i, Y_i, G_i) = \mathbb{P}(R_i = 1 | Z_i, Y_i) = \pi_i, \ i = 1, \dots, N \ \text{ (Missing at random)}$$

Marginal sampling:

$$\mathbb{P}(R_i = 1 | Z_i, Y_i) = \begin{cases} \pi(Z_i) & \text{covariate-dependent} \\ \pi(Y_i) & \text{outcome-dependent} \end{cases}$$

Joint sampling:

$$\mathbb{P}(R_i = 1 | Z_i, Y_i) = \pi(Z_i, Y_i)$$

For the lead SNP (Z) sampling, strata are naturally defined by the number of copies of the minor allele.

To operationalize the outcome sampling, we discretize Y into a K strata variable Y_{st}, ($T_1, \dots, T_K$) using pre-specified cut-offs $(C_1, \dots, C_{K-1})$.

A First Approach

Let N be the phase 1 sample size and n be the phase 2 sample size ($N > n$)

$Z \backslash Y_{st}$	**Phase 1 data** $\mathbf{T}_1$	$\ldots$	$\mathbf{T}_K$	$\mathbf{M}_Z$	**Phase 2 design** $\mathbf{T}_1$	$\ldots$	$\mathbf{T}_K$	$\mathbf{M}_Z$
0	N_{01}	$\ldots$	N_{0K}	$N_{0\cdot}$	n_{01}	$\ldots$	n_{0K}	$n_{0\cdot}$
1	N_{11}	$\ldots$	N_{1K}	$N_{1\cdot}$	n_{11}	$\ldots$	n_{1K}	$n_{1\cdot}$
2	N_{21}	$\ldots$	N_{2K}	$N_{2\cdot}$	n_{21}	$\ldots$	n_{2K}	$n_{2\cdot}$
$\mathbf{M}_Y$	$N_{\cdot 1}$	$\ldots$	$N_{\cdot K}$	N	$n_{\cdot 1}$	$\ldots$	$n_{\cdot K}$	n

where

$$n_{I_z,I_y} = \#\{i : Z_i \in I_z, Y_i \in I_y\}$$

is the number of subjects to be allocated in each stratum for the joint sampling design.

A First Approach

Let N be the phase 1 sample size and n be the phase 2 sample size ($N > n$)

$Z \backslash Y_{st}$
0
1
2
$\mathbf{M}_Y$

Phase 1 data			
$\mathbf{T}_1$	$\dots$	$\mathbf{T}_K$	$\mathbf{M}_Z$
N_{01}	$\dots$	N_{0K}	$N_{0\cdot}$
N_{11}	$\dots$	N_{1K}	$N_{1\cdot}$
N_{21}	$\dots$	N_{2K}	$N_{2\cdot}$
$N_{\cdot 1}$	$\dots$	$N_{\cdot K}$	N

Phase 2 design			
$\mathbf{T}_1$	$\dots$	$\mathbf{T}_K$	$\mathbf{M}_Z$
n_{01}	$\dots$	n_{0K}	$n_{0\cdot}$
n_{11}	$\dots$	n_{1K}	$n_{1\cdot}$
n_{21}	$\dots$	n_{2K}	$n_{2\cdot}$
$n_{\cdot 1}$	$\dots$	$n_{\cdot K}$	n

where

$$n_{I_z, I_y} = \# \{ i : Z_i \in I_z, Y_i \in I_y \}$$

is the number of subjects to be allocated in each stratum for the joint sampling design.

Q: How to allocate samples for the phase 2 design?

Heuristic designs

Strata are defined using Y and Z (observed at phase 1)

- Y, the QT, is discretized into a $K = 3$ strata variable Y_{st} (T_1, T_2, T_3)
- Z, the lead SNP, is defined by the number of copies of the minor allele $(0, 1, 2)$

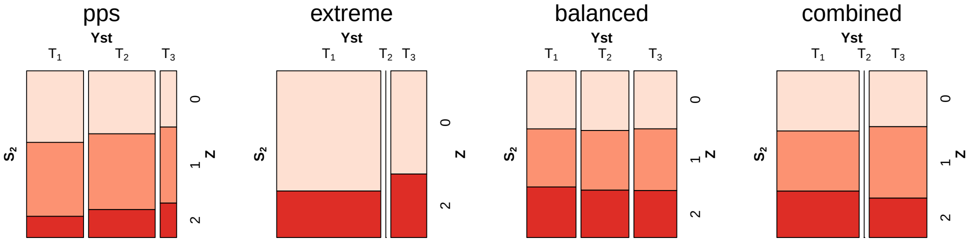


Figure: Heuristic designs when $n \approx N/2$, $Y \sim \mathcal{N}(2, 2.25)$, minor allele frequency (MAF) for Z, $q_Z = 0.3$

Heuristic designs

Strata are defined using Y and Z (observed at phase 1)

- Y, the QT, is discretized into a $K = 3$ strata variable Y_{st} (T_1, T_2, T_3)
- Z, the lead SNP, is defined by the number of copies of the minor allele $(0, 1, 2)$

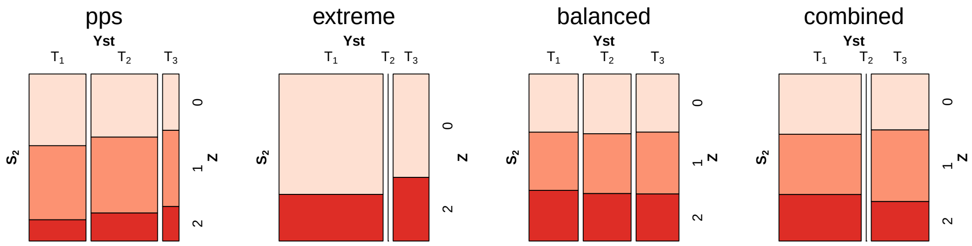


Figure: Heuristic designs when $n \approx N/2$, $Y \sim \mathcal{N}(2, 2.25)$, minor allele frequency (MAF) for Z, $q_Z = 0.3$

Q: Is there a better way to find phase 2 designs given a fixed n?

Optimality in Two-Phase Studies

Let $\theta = (\beta^t, \boldsymbol{p}^t)^t$, under regularity conditions, the limiting distribution of $\hat{\theta}$ follows asymptotically:

$$\sqrt{N}(\hat{\theta} - \theta) \sim \mathcal{N}(0, \mathbb{J}(\Phi)^{-1}),$$

where $\Phi = (\theta^t, \psi^t)^t$, and

- ψ is a parameter determining $\pi(y_i, z_i; \psi) = \mathbb{P}(R_i = 1 | y_i, z_i; \psi)$, and
- $\mathbb{J}(\Phi)$ is the expected information matrix.

Optimality in Two-Phase Studies

Let $\theta = (\beta^t, \boldsymbol{p}^t)^t$, under regularity conditions, the limiting distribution of $\hat{\theta}$ follows asymptotically:

$$\sqrt{N}(\hat{\theta} - \theta) \sim \mathcal{N}(0, \mathbb{J}(\Phi)^{-1}),$$

where $\Phi = (\theta^t, \psi^t)^t$, and

- ψ is a parameter determining $\pi(y_i, z_i; \psi) = \mathbb{P}(R_i = 1 | y_i, z_i; \psi)$, and
- $\mathbb{J}(\Phi)$ is the expected information matrix.

Following experimental design literature, different optimality criteria, $\Lambda(\cdot)$, can be investigated:

A-optimality: minimizes trace of VCM, i.e. $\Lambda\left[\mathbb{J}(\Phi)\right] = \sum \text{diag}(\mathbb{J}(\Phi)^{-1})$

D-optimality: minimizes determinant of VCM, i.e. $\Lambda\left[\mathbb{J}(\Phi)\right] = \det(\mathbb{J}(\Phi)^{-1})$

Parameter-specific: minimizes the entry of the VCM corresponding to β_1, i.e. $\Lambda\left[\mathbb{J}(\Phi)\right] = \left[\mathbb{J}(\Phi)^{-1}\right]_{\beta_1,\beta_1}$

The Optimization Problem

We want to identify a design that satisfies the following optimization problem:

$$\min \ \Lambda\left[\mathbb{J}(\Phi)\right] \ \text{ subject to } \ \sum_{i=1}^{N} R_i = n,$$

The Optimization Problem

We want to identify a design that satisfies the following optimization problem:

$$\min\ \Lambda\left[\mathbb{J}(\Phi)\right] \text{ subject to } \sum_{i=1}^{N} R_i = n,$$

or similarly

$$\min\ \Lambda\left[\mathbb{J}(\Phi)\right] \text{ subject to } \mathbb{P}(\boldsymbol{R} = 1|\boldsymbol{Y},\boldsymbol{Z}) = \frac{n}{N}.$$

Designs via Lagrange Multipliers

Aims to optimize the stratum-specific selection probabilities by minimizing the following expression

$$\Lambda\left[\mathbb{J}(\Phi)\right] + \lambda N^{-1}\left(\sum_{k,j \in Y_{st},Z} \pi_{kj}(\psi) N_{kj} - n\right)$$

where λ is a Lagrange multiplier (LM) that accounts for the budget constraint and N_{kj} is the phase 1 stratum size.

Designs via Lagrange Multipliers

Aims to optimize the stratum-specific selection probabilities by minimizing the following expression

$$\Lambda\left[\mathbb{J}(\Phi)\right] + \lambda N^{-1}\left(\sum_{k,j \in Y_{st},Z} \pi_{kj}(\psi) N_{kj} - n\right)$$

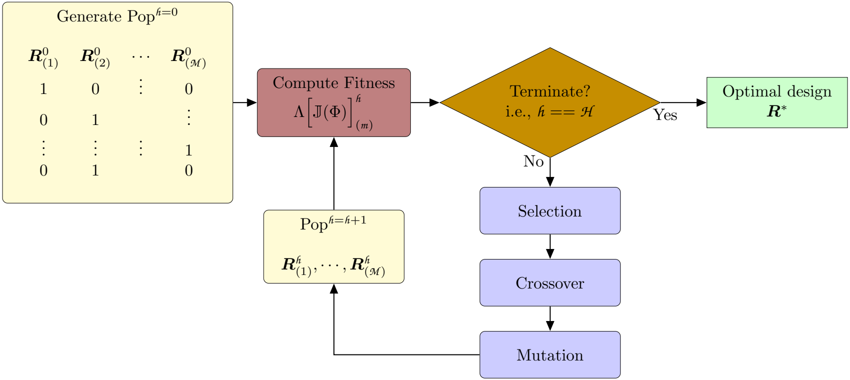
where λ is a Lagrange multiplier (LM) that accounts for the budget constraint and N_{kj} is the phase 1 stratum size.

Notes:

- $\theta = (\boldsymbol{\beta}^t, \boldsymbol{p}^t)^t$ are design quantities, i.e. need to be specified a priori
- Provides a set of optimal sampling fractions π^*_{kj} for each stratum
- A phase 2 subsample is drawn given $\boldsymbol{\pi}^* = (\pi^*_{10}, \ldots, \pi^*_{K2})$

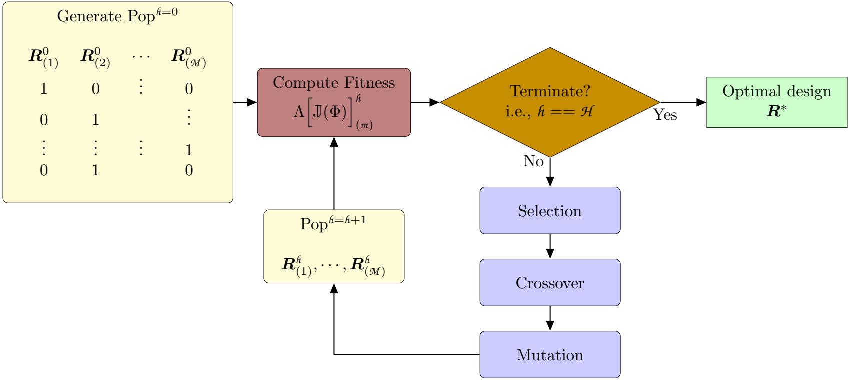
Designs via Genetic Algorithms

Aims to identify an "optimal" phase 2 subsample design ($\boldsymbol{R}^*$).



Designs via Genetic Algorithms

Aims to identify an "optimal" phase 2 subsample design ($\boldsymbol{R}^*$).



Notes:

- Suitable for discrete optimization in large search spaces (2^N).
- The budget constraint is accounted for via fixed-size subset selection.
- Hyper-parameter tuning is needed.

Rank-based Designs

Grounded in theoretical findings, these designs achieve a phase 2 sample size by selecting an equal number of subjects from each of the top and bottom rankings of the following quantities:

ODS: Outcome dependent sampling, i.e. $Y^{(1)}, \ldots, Y^{(N)}$

Rank-based Designs

Grounded in theoretical findings, these designs achieve a phase 2 sample size by selecting an equal number of subjects from each of the top and bottom rankings of the following quantities:

ODS: Outcome dependent sampling, i.e. $Y^{(1)}, \ldots, Y^{(N)}$

RDS: Residual dependent sampling, i.e. $\hat{e}^{(1)}, \ldots, \hat{e}^{(N)}$, where $\hat{e}_i = Y_i - \hat{Y}_i$, $i = 1, \ldots, N$

Rank-based Designs

Grounded in theoretical findings, these designs achieve a phase 2 sample size by selecting an equal number of subjects from each of the top and bottom rankings of the following quantities:

ODS: Outcome dependent sampling, i.e. $Y^{(1)}, \ldots, Y^{(N)}$

RDS: Residual dependent sampling, i.e. $\hat{e}^{(1)}, \ldots, \hat{e}^{(N)}$, where $\hat{e}_i = Y_i - \hat{Y}_i$, $i = 1, \ldots, N$

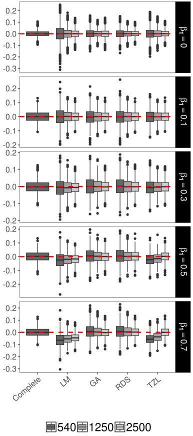
TZL: Tao-Zeng-Lin's scaled residuals design. Tao et al. (2020) showed that scaling the residuals by $\mathrm{Var}(G|Z)^{1/2}$ (unknown at design stage) yields a design with minimum variance under the score test, i.e. $\beta_1 = o(1)$.

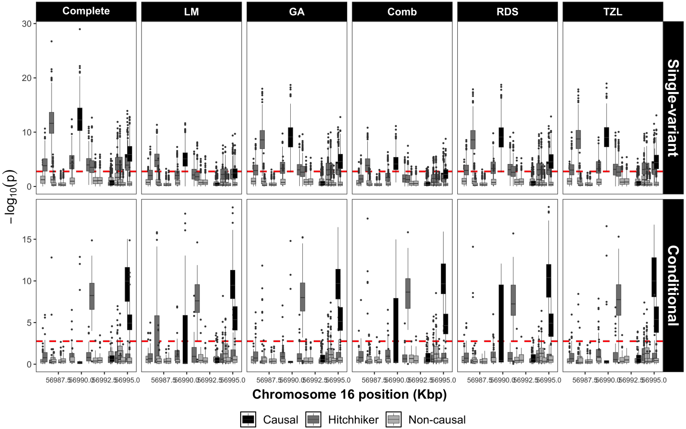
Characteristics of the Studied Designs

Feature	Design				
	Heuristic	LM	GA	ODS/RDS	TZL
Design quantities β, p need to be specified	$\times$	$\surd$	$\surd$	$\times$	$\surd$
Subject to sampling variability	$\surd$	$\surd$	$\times$	$\times$	$\times$
Computationally intensive	$\times$	$\times$	$\surd$	$\times$	$\times$
Depends on strata definition	$\surd$	$\surd$	$\times$	$\times$	$\times$

Simulation results in realistic scenario

- GA demonstrates less estimation bias compared to TZL.
- GA shows comparable power to TZL when $\mathrm{Var}(G|Z)$ is misspecified.



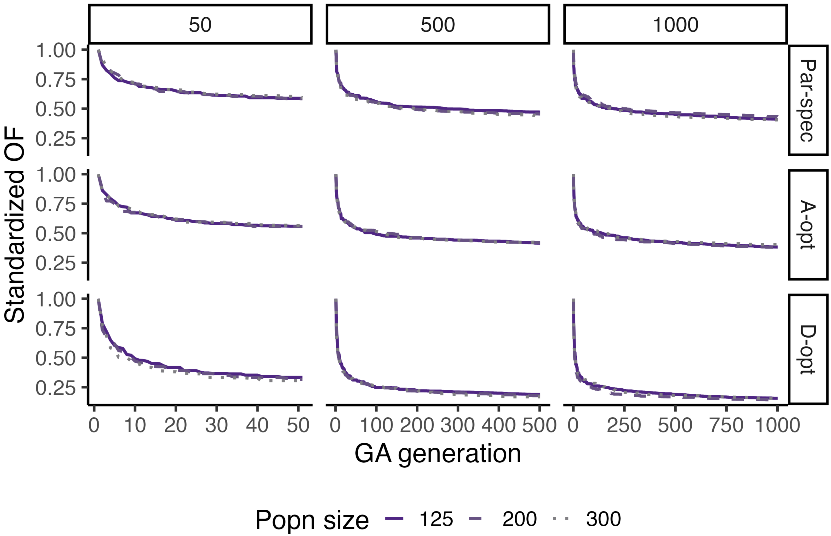


Summary

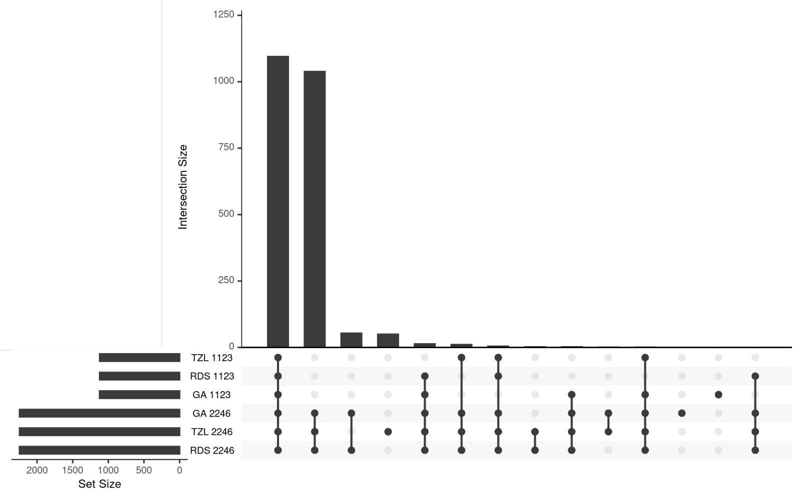
- The proposed GA designs can be more powerful than heuristic designs and comparable to rank-based designs.
- Similar results for GA over different MAFs, LDs and sample sizes (N,n).
- GA achieves power similar to complete data with half sequencing info.
- GA can be implemented across a variety of models.
- Implementation available at `github.com/egosv/twoPhaseGAS`.

Outline

Convergence



Solution Uniqueness



Outline

High-dimensional Covariate Settings

Let $\boldsymbol{Z} \in \mathbb{R}^p$ be a high-dimensional covariate vector, with observed likelihood:

$$L(\boldsymbol{\beta}, \gamma) = \prod_{i=1}^{N} \left\{ f_\beta(Y_i|G_i, \mathbf{Z}_i) h_\gamma(G_i, \mathbf{Z}_i) \right\}^{R_i} \times \left\{ \int f_\beta(Y_i|g, \mathbf{Z}_i) h_\gamma(g, \mathbf{Z}_i) \mathrm{d}g \right\}^{1-R_i},$$

High-dimensional Covariate Settings

Let $\boldsymbol{Z} \in \mathbb{R}^p$ be a high-dimensional covariate vector, with observed likelihood:

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \prod_{i=1}^{N} \left\{ f_\beta(Y_i|G_i, \boldsymbol{Z}_i) h_\gamma(G_i, \boldsymbol{Z}_i) \right\}^{R_i} \times \left\{ \int f_\beta(Y_i|g, \boldsymbol{Z}_i) h_\gamma(g, \boldsymbol{Z}_i) \mathrm{d}g \right\}^{1-R_i},$$

the optimization problem becomes:

$$\begin{aligned} \text{min} \quad & -\log L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{R}) + \Lambda\left[\mathbb{J}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{R})^{-1} \right] \\ \text{subject to} \quad & \sum_{i=1}^{N} R_i = n, \ \sum_{j=2}^{p+2} c_{1j} \mathbb{1}(\beta_j) \leq C_1, \ \sum_{j=2}^{p+2} c_{2j} \mathbb{1}(\gamma_j) \leq C_2, \end{aligned}$$

High-dimensional Covariate Settings

Let $\boldsymbol{Z} \in \mathbb{R}^p$ be a high-dimensional covariate vector, with observed likelihood:

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \prod_{i=1}^{N} \left\{ f_\beta(Y_i|G_i, \boldsymbol{Z}_i) h_\gamma(G_i, \boldsymbol{Z}_i) \right\}^{R_i} \times \left\{ \int f_\beta(Y_i|g, \boldsymbol{Z}_i) h_\gamma(g, \boldsymbol{Z}_i) \mathrm{d}g \right\}^{1-R_i},$$

the optimization problem becomes:

$$\begin{aligned} \text{min} \quad & -\log L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{R}) + \Lambda \left[\mathbb{J}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{R})^{-1} \right] \\ \text{subject to} \quad & \sum_{i=1}^{N} R_i = n, \ \sum_{j=2}^{p+2} c_{1j} \mathbb{1}(\beta_j) \leq C_1, \ \sum_{j=2}^{p+2} c_{2j} \mathbb{1}(\gamma_j) \leq C_2, \end{aligned}$$

here

- $\mathbb{1}(\cdot)$ is the indicator function denoting whether the jth parameter is to be selected,
- c_{kj}, $k = 1, 2; j = 2, \ldots, p+2$ represent covariate-specific costs, and
- C_k, $k = 1, 2$, are budget constraints for the parameters in f and h.

Alternative Optimization Methods

- Dynamic Programming (knapsack problem)
- Simulated Annealing
- Estimation of Distribution Algorithms

Acknowledgments

Shelley B. Bull (DLSPH, U of T/ LTRI)

Radu V. Craiu (Statistical Sciences, U of T)

Guan Wang (Master’s student, DLSPH, U of T)

OICR Ontario Institute for Cancer Research

CIHR IRSC Canadian Institutes of Health Research Instituts de recherche en santé du Canada

CANSSI INCASS ONTARIO

STAGE Strategic Training for Advanced Genetic Epidemiology

Selected References

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

[2] O. Espin-Garcia, R. V. Craiu, and S. B. Bull. Two-phase designs for joint quantitative-trait-dependent and genotype-dependent sampling in post-GWAS regional sequencing. *Genetic Epidemiology*, 42(1):104–116, 2018.

[3] O. Espin-Garcia, R. V. Craiu, and S. B. Bull. Two-phase sample selection strategies for design and analysis in post-genome wide association fine-mapping studies. *Statistics in Medicine*, 40(30):6792–6817, 2021.

[4] J. H. Holland. Genetic Algorithms. *Scientific American*, 267(1):66–73, 1992.

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

[6] D.-Y. Lin, D. Zeng, and Z. Z. Tang. Quantitative trait analysis in sequencing studies under trait-dependent sampling. *Proceedings of the National Academy of Sciences of the United States of America*, 110(30):12247–52, 2013.

[7] M. A. McIsaac and R. J. Cook. Response-dependent two-phase sampling designs for biomarker studies. *Canadian Journal of Statistics*, 42(4):268–284, 2014.

[8] M. Reilly. Optimal sampling strategies for two-stage studies. *American Journal of Epidemiology*, 143(1):92–100, 1996.

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

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

[11] R. Tao, D. Zeng, and D.-Y. Lin. Optimal Designs of Two-Phase Studies. *Journal of the American Statistical Association*, 115(532):1946–1959, 2020.

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

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

Thank you for listening!

Questions?