Published in final edited form as:

Stat Med. 2018 November 30; 37(27): 4083–4095. doi:10.1002/sim.7898.

A pooling strategy to effectively use genotype data in quantitative traits genome-wide association studies

Wei Zhang¹, Aiyi Liu¹, Paul S. Albert², Robert D. Ashmead³, Enrique F. Schisterman¹, and James L. Mills¹

¹Division of Intramural Population Health Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Maryland

²Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland ³Center for Statistical Research and Methodology, US Census Bureau, Washington, District of Columbia

Abstract

The goal of quantitative traits genome-wide association studies is to identify associations between a phenotypic variable, such as a vitamin level and genetic variants, often single-nucleotide polymorphisms. When funding limits the number of assays that can be performed to measure the level of the phenotypic variable, a subgroup of subjects is often randomly selected from the genotype database and the level of the phenotypic variable is then measured for each subject. Because only a proportion of the genotype data can be used, such a simple random sampling method may suffer from substantial loss of efficiency, especially when the number of assays is relative small and the frequency of the less common variant (minor allele frequency) is low. We propose a pooling strategy in which subjects in a randomly selected reference subgroup are aligned with randomly selected subjects from the remaining study subjects to form independent pools; blood samples from subjects in each pool are mixed; and the level of the phenotypic variable is measured for each pool. We demonstrate that the proposed pooling approach produces considerable gains in efficiency over the simple random sampling method for inference concerning the phenotype-genotype association, resulting in higher precision and power. The methods are illustrated using genotypic and phenotypic data from the Trinity Students Study, a quantitative genome-wide association study.

Keywords

biallelic model; efficiency; estimation; group testing; homozygous minor allele; multiple comparison; phenotypes; pooling; power; random sampling; single nucleotide polymorphism

Correspondence: Aiyi Liu, Biostatistics and Bioinformatics Branch, Eunice Kennedy Shriver National Institute of Child Health and Human Development, Bethesda, MD 20817. liua@mail.nih.gov.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

1 | INTRODUCTION

The goal of a quantitative traits genome-wide association study (GWAS) is to identify genetic variants, often single-nucleotide polymorphisms (SNPs) that influence levels of a biological compound of interest, such as a vitamin, called the phenotypic variable. The phenotypic variable is generally measured on a continuous scale. The SNP is generally biallelic, that is, there are two possible versions, and each parent contributes one or the other to the study subjects. Thus, the possible combinations are *AA* (two copies of the major allele), *Aa* (one copy of the major allele and one copy of the minor allele), and *aa* (two copies of the minor allele).

Current technology can test an extraordinary number of SNPs, ie, 2.5 million, at a cost of only about \$600 per person. In contrast, measuring the phenotypic variable or variables can be very expensive, depending on the compound, or compounds, selected. For example, measuring one phenotypic variate, homocysteine, could cost about \$200 per person. Such cost constraints sometime necessitate measuring phenotypes in randomly selected samples rather than samples from all genotyped subjects. A random sampling strategy is simple in concept and yields phenotypic levels from each selected individual subject that make analysis by the standard methods such as two-sample t-test, ANOVA, and regression feasible.

However, disadvantages of simple random sampling are also apparent, a major one being the absence of phenotypic data from subjects not selected by the random sampling. Given the low genotype frequencies of the minor alleles in many SNPs, the number of subjects with the homozygous minor allele genotype can be very small or even zero in the subgroup randomly selected for phenotype testing. Thus, there may be insufficient data to detect an effect of the genotype on the phenotypic variables. By chance, the selected subgroup is likely to contain most subjects with the minor allele for some SNPs, but very few subjects with the minor allele for other SNPs, because there is little or no connection between the presence of rare alleles from SNP to SNP. The small number of subjects who have the uncommon genotypes of interest when random sampling is performed may seriously hinder the ability to find important genotypes in a GWAS, especially because of the need to adjust for multiple comparisons.

In the present paper, we propose a pooling strategy to improve the efficiency of an existing data sample obtained from the simple random sampling approach. Suppose that a subgroup, referred to as the reference group, of subjects has been randomly selected from the existing genotypic database. For each subject in the reference subgroup, we randomly select a number of subjects from among the remaining subjects in the genotype database to align with the subject to form a pool, regardless of their genotypes. For subjects within each pool, serum samples are mixed and the level of the phenotypic variable is measured. The result is the average of the unobserved individual phenotypic levels of the subjects in the pool. The genotypic-phenotypic association for each SNP is then estimated based on these averages of phenotypic levels across the different pools of samples. A special case to be investigated both analytically and numerically is the so called *one-big-pool* approach, also defined as 1 and $n - 1$ approach by Schisterman et al,¹ where individual phenotypic levels are measured

for all but one subject selected by simple random sampling, whereas the remaining subjects are all put into one pool to yield an average measure. A unique feature of the proposed pooling strategy is that pooling is carried out so that each pool contains only one subject from the reference group (the subjects already selected by random sampling).

Group testing was first proposed by Dorfman² to improve screening for a rare disease by pooling of biospecimens. It has drawn considerable attention in the statistical and medical literature for screening and estimation of a binary outcome; see, among others, related works.^{3–7} Note that, in a group testing framework, the outcome of each pool is the maximum of individual binary outcomes. The group testing method has been extended to measuring continuous variables where the outcome of a pool is the average of the individual observations in the pool; see, among others, other works.^{8–13}

The proposed pooling strategy has some similarities to both Dorfman's² group testing concerning a binary outcome and the pooling strategy concerning the assessment of an exposure measured on a continuous scale (eg, the work of Schisterman et al¹³). By aligning subjects, the chances are increased that a pool contains subjects with homozygous minor allele genotypes of interest, which, in turn, could yield more efficient inference on the association.

The pooling strategy proposed in the present paper, however, differs from conventional pooling methods in one aspect. In the conventional pooling of continuous variable data, pooling is *supervised* in that the subjects are randomly pooled according to certain characteristics being compared (eg, cases or controls); see, eg, the work of Schisterman et al.¹³ The assignment of subjects to each pool from the proposed strategy, on the other hand, is *unsupervised*, not involving genotypes or other characteristics. (Indeed, in a quantitative traits GWAS, it is impossible to assign subjects to a pool according to their genotypes because the number of SNPs is so large.)

The paper is arranged as follows. In Section 2, we describe the Trinity Students Study, a quantitative traits GWAS, which motivated the present research. In Section 3, we discuss estimation and hypothesis testing based on the random sampling strategy in the context of quantitative traits GWAS. In Section 4, we describe the general pooling approach and develop methods for estimation and hypothesis testing with data generated from such a strategy. In Section 5, we focus on the *one-big-pool* approach and derive least-squares estimates of the parameters of interest. In Section 6, we demonstrate the efficiency gains of the proposed pooling strategy over simple random sampling. Specifically, we present conditions under which pooling yields more efficient estimation than the simple random sampling method. We further develop methods to compute the probability that these conditions hold. The methods are illustrated in Section 7 using data from the Trinity Students GWAS to evaluate the genotype-phenotype association between phenotypic variables and genotypes of interest. Finally, some discussions are presented in Section 8.

2 | THE TRINITY STUDENTS GWAS: A MOTIVATING EXAMPLE

The Trinity Students (TSS) GWAS was conducted in 2003–2004 by investigators at The Epidemiology Branch, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, Trinity College, Dublin, National Human Genome Research Institute and the Health Research Board of Ireland. From a starting population of a total of 3569 students attending the University of Dublin, Trinity College, 2507 students who had Irish grandparents and no major medical problems, who completed the study questionnaire, and who provided the necessary blood samples constituted the final study population. Written informed consent and IRB approval were obtained.

DNA was collected for the GWAS and over 750 000 SNPs were assayed via the Illumina system. The GWAS analysis is ongoing. As expected, many SNPs have low minor allele frequencies, producing rates of homozygosity for the minor allele of 1%–10%.

Serum samples were collected, processed, and stored until they were shipped to laboratories that measured over 40 target phenotype measures including hematologic factors, liver function tests, B vitamins, and related metabolites.

The large number of subjects makes measuring phenotype measures costly. In fact, financial constraints have made it possible to measure additional phenotypic variables of interest in only half of the study participants. For this reason, the strategy of random sampling was discussed. Random sampling is not optimal because genotype data from subjects not selected by the random sampling approach are wasted. Therefore, pooling was discussed as an alternative, potentially more informative, approach.

Later, we will use the genotype data and phenotype data collected on all subjects from the Trinity Students GWAS to evaluate the efficiency of the proposed pooling strategy, assuming for illustration that the phenotypic data had not been available for all subjects.

3 | THE SIMPLE RANDOM SAMPLING APPROACH

Let Y denote the phenotypic level and G the genotype of a single biallelic SNP. For simplicity, we consider binary genotypes with $G = 1$ if the SNP has homozygous minor allele aa and $G = 0$ if otherwise. Define

$$\mu_1 = E(Y|G = 1), \sigma_1^2 = \text{var}(Y|G = 1), \mu_0 = E(Y|G = 0), \sigma_0^2 = \text{var}(Y|G = 0).$$

Our interest is to estimate $\delta = \mu_1 - \mu_0$ and to test the null hypothesis $H_0 : \delta = 0$.

Suppose that genotypic data are available from a total number N of subjects, denoted by $\mathcal{S} = \{1, 2, \dots, N\}$. Let $\mathcal{S}_1 = \{1 \leq i \leq N : G_i = 1\}$ and $\mathcal{S}_0 = \{1 \leq i \leq N : G_i = 0\}$ be the subset of subjects with $G = 1$ and $G = 0$, respectively. Furthermore, let N_1 and N_0 be the number of subjects in the two subsets, respectively, and $\rho = N_1/N$ be the proportion of subjects with $G = 1$ in the genotype data set. If the N subjects are randomly selected from the study

population, then ρ is an unbiased estimate of the genotype ($G = 1$) frequency in the population.

For subject $i \in \mathcal{S}$, denote by Y_i its phenotypic level and G_i its genotype. Note that genotypes are observed for all subjects. However, the phenotypic level Y_i can be observed for only $n < N$ subjects.

A simple random sampling selects a subset $\mathcal{S}_R$ of size n from $\mathcal{S}$. For subjects in this reference group, define

$$\mathcal{S}_{1R} = \{i \in \mathcal{S}_R : G_i = 1\}, \quad \mathcal{S}_{0R} = \{i \in \mathcal{S}_R : G_i = 0\}$$

to be the (reference) group of subjects having genotype 1 and 0, respectively, and write n_1 and n_0 as the number of subjects in the two groups, respectively. Let $\mathbf{Y}_{1R}$ and $\mathbf{Y}_{0R}$ be the vector of phenotypic levels for subjects in $\mathcal{S}_{1R}$ and $\mathcal{S}_{0R}$, respectively, where throughout vectors refer to column vectors. Then, we have, conditioning on the random number n_1 ,

$$E \left\{ \begin{pmatrix} \mathbf{Y}_{1R} \\ \mathbf{Y}_{0R} \end{pmatrix} \right\} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} \\ 0 & \mathbf{1}_{n_0} \end{pmatrix} \begin{pmatrix} \delta \\ \mu_0 \end{pmatrix}, \quad \text{var} \left\{ \begin{pmatrix} \mathbf{Y}_{1R} \\ \mathbf{Y}_{0R} \end{pmatrix} \right\} = \begin{pmatrix} \sigma_1^2 \mathbf{I}_{n_1} & 0 \\ 0 & \sigma_0^2 \mathbf{I}_{n_0} \end{pmatrix}, \quad (1)$$

where $\mathbf{1}_m$ is the vector of 1's with order m , and $\mathbf{I}_m$ is the identity matrix of order m . Without further distributional assumption, the parameters can be estimated using least-squares method.¹⁴ It follows that the least squares estimate of δ is simply the difference in the observed average of phenotypic levels between the two genotypes, that is,

$$\hat{\delta}_R = \bar{Y}_{1R} - \bar{Y}_{0R},$$

where $\bar{Y}_{1R} = \sum_{i \in \mathcal{S}_{1R}} Y_i / n_1$, and $\bar{Y}_{0R} = \sum_{i \in \mathcal{S}_{0R}} Y_i / n_0$. The variance of $\hat{\delta}_R$ is given by

$$\sigma_R^2 = \frac{1}{n_1} \sigma_1^2 + \frac{1}{n_0} \sigma_0^2 = \frac{1}{n_1} \sigma_1^2 + \frac{1}{n - n_1} \sigma_0^2. \quad (2)$$

The two variances σ_1^2 and σ_0^2 can be estimated by their sample variances

$$\hat{\sigma}_1^2 = \sum_{i \in \mathcal{S}_{1R}} (Y_i - \bar{Y}_{1R})^2 / (n_1 - 1), \quad \hat{\sigma}_0^2 = \sum_{i \in \mathcal{S}_{0R}} (Y_i - \bar{Y}_{0R})^2 / (n_0 - 1).$$

The null hypothesis H_0 can then be tested using the Student's t-test.

For given n , σ_1 , and σ_0 , the variance σ_R^2 is a function of n_1 and decreases as n_1 increases until n_1 becomes larger than $n\sigma_1/(\sigma_1 + \sigma_0)$. This fact can be easily verified by investigating the derivative of σ_R^2 with respect to n_1 . When the genotype frequency of the homozygous minor allele ($G = 1$) is relative low, the simple sampling strategy can suffer from substantial loss of efficiency because the number n_1 of subjects with genotype $G = 1$ can be very small. With low frequency of a genotype, it is possible that the random sample may not include any subjects with this genotype, in which case inference on δ becomes infeasible. The most efficient (minimum variance) estimator will be achieved if $n_1 = N_1$, that is, if all subjects who are homozygous for the minor allele ($G = 1$) are selected. However, because N_1 is much smaller than n , the chance that the simple random sampling selects all N_1 subjects with $G = 1$ is very low, noting that the probability that the simple random sampling picks n_1 subjects with genotype $G = 1$ from $\mathcal{S}$ is $C_{N_1}^{n_1} C_{N-N_1}^{n-n_1} / C_N^n$, where C_N^n is the binomial coefficient.

Moreover, even if the simple random sampling yields an efficient estimation for certain SNPs, it may still lose substantial efficiency for other SNPs in the GWAS analysis because of the variation in genotype frequency across SNPs.

Due to the low frequency of $G = 1$, the variance σ_1^2 is much more influential than σ_0^2 on the precision of $\hat{\delta}_R$. For $N = 2500$ and $N_1 = 250$, reflecting a typical SNP in the Trinity Students GWAS example with 10% frequency for $G = 1$, Figure 1 presents the variance σ_R as a function of $1 - n_1/N_1$, for σ_1^2 and 1.5, respectively, with $n = N/2 = 1250$. In the figure, σ_0 is set to be 1. The figure shows clearly that the precision of the estimator increased substantially when n_1 increases from a relatively small value. It is also worth noting that, for this specific example, the probability is about 0.0005 that the simple random sampling includes at least 150 subjects with $G = 1$; the probability approaches to zero ($\approx 10^{-25}$) if at least 200 such subjects are selected.

4 | A SAMPLE ENRICHMENT APPROACH BY POOLING

Given a reference group, a randomly selected subset $\mathcal{S}_R$ of n subjects, as described in the previous section, we are interested in whether the remaining $N - n$ subjects in the subset $\mathcal{S} \setminus \mathcal{S}_R$ can be used together with those in $\mathcal{S}_R$ to produce an estimate of δ with smaller variance or a test with higher power than the estimate or test based on subjects in $\mathcal{S}_R$. Here, we propose a sample enrichment strategy as follows. For each subject $i \in \mathcal{S}_R$, we randomly select $k_i - 1$ subjects from $\mathcal{S} \setminus \mathcal{S}_R$ to form a pool with subject i . Thus, k_i is the size of the pool (total number of subjects in the pool), with $k_i \geq 1$, $\sum_{i=1}^m k_i = N$. If $k_i = 1$, then the phenotypic level is measured for subject i . However, for subjects in a pool with size $k_i \geq 2$, individual phenotypic levels are not observed. Instead, we observe the average of individual phenotypic levels in the pool, that is,

$$\tilde{Y}_{iR} = \frac{(Y_i + \text{sum of the } Y \text{ s from subjects aligned to subject } i)}{k_i}.$$

Let k_{1i} be the number of subjects with $G=1$ in the i th pool. Write $r_i = k_{1i}/k_i$ as the proportion of subjects with $G=1$ in the pool, and $\theta_i = r_i\sigma_1^2 + (1-r_i)\sigma_0^2$ as the weighted variances. Then, we have

$$E(\tilde{Y}_{iR}) = \mu_0 + r_i\delta, \quad \text{var}(\tilde{Y}_{iR}) = \frac{\theta_i}{k_i}, \quad (3)$$

which reduces to Equation (1) for individual phenotypic observations if $k_i = 1$.

An estimate $\hat{\delta}_E$ of δ can be obtained using the least squares methods. Let $\hat{\sigma}_1^2$ and $\hat{\sigma}_0^2$ be consistent estimates of σ_1^2 and σ_0^2 , respectively, and define $\hat{\theta}_i = r_i\hat{\sigma}_1^2 + (1-r_i)\hat{\sigma}_0^2$. Then, the two-stage weighted least squares estimate of δ is given by

$$\hat{\delta}_E = \frac{\left(\sum \frac{k_i}{\hat{\theta}_i}\right)\left(\sum \frac{k_{1i}\tilde{Y}_{iR}}{\hat{\theta}_i}\right) - \left(\sum \frac{k_{1i}}{\hat{\theta}_i}\right)\left(\sum \frac{k_i\tilde{Y}_{iR}}{\hat{\theta}_i}\right)}{\left(\sum \frac{k_i}{\hat{\theta}_i}\right)\left(\sum \frac{k_{1i}r_i}{\hat{\theta}_i}\right) - \left(\sum \frac{k_{1i}}{\hat{\theta}_i}\right)^2}, \quad (4)$$

and its variance is asymptotically given by

$$\sigma_E^2 = \frac{\sum \frac{k_i}{\theta_i}}{\left(\sum \frac{k_i}{\theta_i}\right)\left(\sum \frac{k_{1i}r_i}{\theta_i}\right) - \left(\sum \frac{k_{1i}}{\theta_i}\right)^2}, \quad (5)$$

where, unless otherwise stated, the summation is over $\{i: i \in \mathcal{S}_R\}$.

With homoscedasticity, that is, $\sigma_1^2 = \sigma_0^2 = \sigma^2$, the least squares estimate and its variance reduce to

$$\hat{\delta}_E = \frac{\sum k_i(r_i - \rho)\tilde{Y}_{iR}}{\sum k_{1i}(r_i - \rho)}, \quad \sigma_E^2 = \frac{\sigma^2}{\sum k_{1i}(r_i - \rho)}.$$

For low frequency of $G = 1$, we expect that, for most cases, the enrichment approach will yield smaller variance, that is, $\text{var}(\tilde{\delta}_E) = \sigma_E^2 < \sigma_R^2$, and thus tests with higher power, because it fully utilizes the genotype data.

Consistent estimates of σ_1^2 and σ_0^2 are needed for interval estimation and hypothesis testing concerning δ . These estimates can be obtained as follows. Let $\mu_0^\star$ and $\delta^\star$ be the least squares estimate from (3), derived by assuming that $\sigma_0^2 = \sigma_1^2$. Define $Z_i = k_i(\tilde{Y}_{iR} - \mu_0^\star - r_i\delta^\star)^2$. Then, it follows from (3) that

$$E(Z_i) \approx r_i\sigma_1^2 + (1 - r_i)\sigma_0^2.$$

The variances can then be estimated by the ordinary least squares estimates from the aforementioned regression equation. This estimation process can also be iterated a few times till the estimates converge, with μ_0 and δ being estimated using least squares methods in further iterations.

Alternatively, we can obtain the maximum likelihood estimates if the phenotypic levels are assumed to be normally distributed. However, the least square methods are more robust against normal distributional assumptions.

5 | THE ONE-BIG-POOL APPROACH

The *one-big-pool* approach is a special case of the sample enrichment, with $k_i = N - n + 1$ for some i and 1 for all others. Recall that subjects in the reference group $\mathcal{S}_R$ are selected from the random sampling procedure, among which n_1 subjects have genotype $G = 1$ and n_0 subjects are of genotype $G = 0$. The *one-big-pool* strategy randomly selects a subject from $\mathcal{S}_R$ to be combined with all subjects in $\mathcal{S} \setminus \mathcal{S}_R$ to form a single pool, hence the name “one-big-pool.” We denote by $\mathcal{S}_R^-$ the subset by excluding from $\mathcal{S}_R$ the selected subject for alignment, and define

$$\tilde{n}_1 = \#\{i \in \mathcal{S}_R^- : G_i = 1\}, \quad \tilde{n}_0 = \#\{i \in \mathcal{S}_R^- : G_i = 0\},$$

to be the number of subjects with $G = 1$ and $G = 0$ in $\mathcal{S}_R^-$, respectively, where G_i is the genotype of the i th subject. Thus, $\tilde{n}_1 = n_1$ if the subject selected for alignment is of genotype $G = 0$ and $\tilde{n}_1 = n_1 - 1$ if the subject selected for alignment is of genotype $G = 1$.

Individual phenotypic levels are measured for each subject in $\mathcal{S}_R^-$. For the *one-big-pool*, its outcome is denoted by $\tilde{Y}$, which is the average of the unobserved individual phenotypic levels in the pool. Note that, in this pool, there are $N_1 - \tilde{n}_1$ subjects with genotype $G = 1$ and $N_0 - \tilde{n}_0$ subjects with genotype $G = 0$. Note that, for $i \in \mathcal{S}_R^-$, we have $k_i = 1$, $k_{1i} = G_i = r_i$, and

$\theta_i = G_i\sigma_1^2 + (1 - G_i)\sigma_0^2$. Define $r = (N_1 - \tilde{n}_1)/(N - n + 1)$, the proportion of subjects in the pool that have genotype $G = 1$, and

$$\bar{Y}_1 = \frac{1}{\tilde{n}_1} \sum_{i \in \mathcal{S}_R^-, G_i = 1} \tilde{Y}_{iR}, \quad \bar{Y}_0 = \frac{1}{\tilde{n}_0} \sum_{i \in \mathcal{S}_R^-, G_i = 0} \tilde{Y}_{iR}.$$

Then, following tedious yet straightforward algebraic manipulation, we can show that Equation (4) becomes

$$\hat{\delta}_E = \frac{\frac{\tilde{n}_1}{\sigma_1^2} \left(\frac{\tilde{n}_0}{\sigma_0^2} + \frac{N_0 - \tilde{n}_0}{\hat{\theta}} \right) \bar{Y}_1 - \frac{\tilde{n}_0}{\sigma_0^2} \left(\frac{\tilde{n}_1}{\sigma_1^2} + \frac{N_1 - \tilde{n}_1}{\hat{\theta}} \right) \bar{Y}_0 + \frac{N - n + 1}{\hat{\theta}} \left(\frac{r\tilde{n}_0}{\sigma_0^2} - \frac{(1-r)\tilde{n}_1}{\sigma_1^2} \right) \bar{Y}}{\frac{\tilde{n}_0}{\sigma_0^2} \left(\frac{\tilde{n}_1}{\sigma_1^2} + \frac{r(N_1 - \tilde{n}_1)}{\theta} \right) + \frac{\tilde{n}_1(1-r)^2(N - n + 1)}{\theta\sigma_1^2}},$$

and Equation (5) reduces to

$$\sigma_E^2 = \frac{\frac{\tilde{n}_1}{\sigma_1^2} + \frac{\tilde{n}_0}{\sigma_0^2} + \frac{N - n + 1}{\theta}}{\frac{\tilde{n}_0}{\sigma_0^2} \left(\frac{\tilde{n}_1}{\sigma_1^2} + \frac{r(N_1 - \tilde{n}_1)}{\theta} \right) + \frac{\tilde{n}_1(1-r)^2(N - n + 1)}{\theta\sigma_1^2}}, \quad (6)$$

where $\theta = r\sigma_1^2 + (1-r)\sigma_0^2$.

Among all pooling strategies, the *one-big-pool* approach achieves the maximum amount of individual phenotypic data. This is appealing because the sample variances based on these individual phenotypic data provide convenient estimates of the variances.

With homoscedasticity, the least squares estimate further reduces to

$$\hat{\delta}_E = \frac{(1-\rho)\tilde{n}_1\bar{Y}_1 - \rho\tilde{n}_0\bar{Y}_0 + (N - n + 1)(r - \rho)\bar{Y}}{(1-r)\tilde{n}_1 + (1-\rho)N_1},$$

with variance

$$\sigma_E^2 = \frac{\sigma^2}{(1-r)\tilde{n}_1 + (r-\rho)N_1}. \quad (7)$$

6 | GAINS IN EFFICIENCY

The sample enrichment yields more efficient estimation of δ and subsequently more powerful test if $\sigma_E^2 < \sigma_R^2$. Given the genotypic data $\mathcal{S}$ and the number n , a specific reference group $\mathcal{S}_R$ of n subjects is selected by random sampling from $\mathcal{S}$. We are interested in whether gains in efficiency can be achieved by enriching the reference group $\mathcal{S}_R$ via pooling of samples with subjects in $\mathcal{S}_R$. Specifically, we will investigate conditions under which $\sigma_E^2 < \sigma_R^2$, and the probability that such variance inequality holds. Due to the large number of combinations of aligning the subjects, these issues can be addressed analytically for simple cases such as the *one-big-pool* approach, whereas, for others, they can be assessed via simulations.

For the *one-big-pool* case, the variances are given in (6) for heteroscedastic error variances and (7) for homoscedastic error variances, respectively. Fixing other components in the variance functions, the random sampling variance $\sigma_R^2 = \sigma_R^2(n_1)$ depends on n_1 , the number of subjects with $G = 1$ in $\mathcal{S}_R$, and the enrichment variance $\sigma_E^2 = \sigma_E^2(\tilde{n}_1)$ depends on $\tilde{n}_1$, the number of subjects with $G = 1$, not in the pool. Recall that $\tilde{n}_1$ is either n_1 or $n_1 - 1$. The probability of having the sampling outcome $(n_1, \tilde{n}_1)$ is

$$p(n_1, \tilde{n}_1) = \begin{cases} \frac{C_{N_1}^{n_1} C_{N-N_1}^{n-n_1}}{C_N^n} \times \frac{n-n_1}{n}, & \text{if } \tilde{n}_1 = n_1; \\ \frac{C_{N_1}^{n_1} C_{N-N_1}^{n-n_1}}{C_N^n} \times \frac{n_1}{n}, & \text{if } \tilde{n}_1 = n_1 - 1. \end{cases}$$

Therefore, the probability that the *one-big-pool* enrichment results in more efficient estimation is given by

$$P\{\sigma_E^2(\tilde{n}_1) < \sigma_R^2(n_1)\} = \sum_{(n_1, \tilde{n}_1) \in \Omega} p(n_1, \tilde{n}_1),$$

where $\Omega = \{(n_1, \tilde{n}_1): \sigma_E^2(\tilde{n}_1) < \sigma_R^2(n_1)\}$.

Explicit conditions in terms of n_1 can be derived under which $\sigma_E^2(\tilde{n}_1) < \sigma_R^2(n_1)$. Note that the equation $\sigma_E^2(\tilde{n}_1) < \sigma_R^2(n_1)$ reduces to a quadratic equation for either $\tilde{n}_1 = n_1$ or $\tilde{n}_1 = n_1 - 1$. Let $L < U$ be the two roots of the quadratic equation, then $\sigma_E^2(\tilde{n}_1) < \sigma_R^2(n_1)$ if and only if $n_1 < L$ or $n_1 > U$. For the homoscedastic error variances in particular, the two roots are given by

$\rho\{nN \pm \sqrt{nN(N-n+1)}\}/(N+1)$ if $\tilde{n}_1 = n_1$, which is obtained by solving the quadratic equation $\sigma_E^2(\tilde{n}_1) = \sigma_R^2(n_1)$ with $\sigma_R^2(n_1)$ defined in (2) and $\sigma_E^2(\tilde{n}_1)$ in (7).

Assume that the variances are homoscedastic, we compute the relative efficiency of the *one-big-pool* enrichment to the random sampling with N and n , as given in Figure 2, but with $N_1 = 250$ and 750 , respectively, representing the case of relatively low and high frequency of $G = 1$.

With $N_1 = 250$, if the one subject to be aligned with the remaining subject has genotype $G = 0$, then the enrichment approach is more efficient if $n_1 \leq 122$ or ≥ 128 . Otherwise, if the subject has genotype $G = 1$, then the enrichment is more efficient if $n_1 \leq 102$ or ≥ 148 . The probability is found to be 0.665 that the enrichment is more efficient than the random sampling.

With $N_1 = 750$, if the one subject to be aligned with the remaining subject has genotype $G = 0$, then the enrichment approach is more efficient if $n_1 \leq 367$ or ≥ 383 . Otherwise, if the subject has genotype $G = 1$, then the enrichment is more efficient if $n_1 \leq 357$ or ≥ 393 . Notably, the probability that the enrichment approach is more efficient than the random sampling approach is found to be almost one.

Figure 2 presents the variance ratios of $\sigma_E^2(n_1)/\sigma_R^2(n_1)$ and $\sigma_E^2(n_1-1)/\sigma_R^2(n_1)$ for $1 \leq n_1 \leq N_1$. For both cases, the enrichment yields more efficient estimation for most values of n_1 . The enrichment gains slightly more efficiency if the one subject to be aligned with the remaining subject has genotype $G = 0$. However, the genotype of this subject becomes less influential when genotype frequency is relatively high. Substantial gains in efficiency occur when n_1 is small, that is, when the random sampling selects fewer subjects with $G = 1$. The figure also indicates that the loss of efficiency due to enrichment, if occurring, is minimal.

We wanted to see how the enrichment performs compared to the random sample for different values of N_1 , the number of homozygous minor alleles, especially for small values. In order to make this comparison, we calculated the power of each approach as a function of N_1 , ranging from 100 to 550.

For a given number of subjects with each genotype out of a total of 2500, data were simulated from a normal distribution. To be specific, we generated the phenotype data with $G = 0$ from the normal distribution $N(0, 11)$, and $N(2, 11)$ for those with $G = 1$. That suggests that the two populations of genotypes had an effect size of $2/11 = 0.1818$, which is similar to the largest effect sizes we saw in the TSS data. Let $n = 1250$. The significance level was set to be 0.05/500 000 to adjust for multiple comparisons in a GWAS.

Figure 3 shows that, for SNPs with low percentages of homozygous minor alleles, the enrichment gains a fairly large advantage in power over the random sample approach. While the difference in power decreases as the percentage of homozygous minor alleles increases, the *one-big-pool* approach appears to always yield larger power.

7 | ILLUSTRATION WITH DATA FROM THE TRINITY STUDENTS STUDY

Since we had individual phenotype value on all subjects, we averaged the individual phenotypic values from groups of subjects to create the value that would be obtained by random sampling and pooling their serum samples before testing. We then examined how the results obtained by analyzing data on these pools compared with the complete result calculated by including all the individual data.

For illustrations, we used a biochemical analyte as the phenotypic variable along with 120 SNPs to test the different approaches. The genotype data were selected from among the genes related to this factor biologically, eg, enzyme genes involved in the metabolism of the biochemical analyte. These genotypic data were supplemented with data on genes that were considered potentially important in other phenotypic variables but were not thought to be related to the biochemical analyte we studied.

A small number of values were missing from the data set, with 0.6% of the phenotypic variable's entries missing and between 1.6% and 8.4% of entries missing for the different SNPs. Missing values of the phenotypic variable was replaced by the overall average of the observed values, and missing SNPs were randomly assigned a genotype according to the proportion in the sample for that particular SNP.

The two-stage least-squares analysis showed that, in general, the *one-big-pool* enrichment yielded smaller p-values than the random sample approach for SNPs deemed significant in the complete approach. Depending on the sample drawn, the methods varied in their ability to identify the same significant SNPs as the complete approach. For the comparison of methods, we averaged the p-values over 1000 iterations of random samples from the data. The results are presented in Figure 4.

We found that, for the SNPs that were at least marginally significant, the average p-value for the *one-big-pool* approach was considerably smaller than for the random sample. While this was not always the case for an individual sample, on average, the *one-big-pool* approach gave results closer to the complete.

In this illustration, many of the SNPs had a very low homozygous minor allele percentage. Among the SNPs with the 17 lowest p-values for the complete approach, the percentage of homozygous minor alleles ranged from 0.8% to 5.9%. We saw in Figure 3 that the gains in efficiency were greatest for samples with small percentages of homozygous minor alleles, which corresponds to the relatively large gains in efficiency for the *one-big-pool* approach that we see in this example.

8 | DISCUSSION

In this paper, we proposed a sample enrichment approach to use existing genotypic data more effectively to estimate the genotype-phenotype association in a quantitative GWAS. Through technical development, simulation, and illustration with data from the TSS GWAS, we demonstrated that the *one-big-pool* approach could gain substantial efficiency as compared to the simple random sampling approach. Pooling DNA samples for genotyping

has been discussed in the GWAS literature. To the best of our knowledge, this is the first paper on pooling phenotypic measures in the context of quantitative GWAS. We discuss next a few issues and some potential research directions arising from the proposed enrichment design.

8.1 | Other alignment approaches

The performance of other enrichment strategies deserves further investigation, both analytically and numerically. Unlike the *one-big-pool* approach, explicit conditions under which other enrichment strategies are more efficient than the random sampling approach are extremely difficult to derive because of the number of possible combinations of aligning subjects. The challenge here is to figure out all possible configurations of the number of subjects in each pool having the homozygous minor allele and the probability that each configuration occurs.

Equation (5) and its special case under homoscedasticity provide the basis for evaluating other designs. Given the reference group, it is possible to produce Figure 2 to compare between random sample design and the one-big-pool design because the variance depends on whether or not the reference subject in the big pool has the homozygous minor allele ($\tilde{n}_1 = n_1$ or $\tilde{n}_1 = n_1 - 1$). For other designs, however, such graphical presentation may not be feasible due to the many different configurations of $\{k_i: i \in S_R\}$. For example, consider the 1–1 design where each subject in the reference group of size $n < N/2$ is aligned with one subject from the remaining subjects. The variance involves the number of pools with both subjects having the allele, only one subject having the allele and neither subject having the allele. Understandably, there are many configurations of these three numbers. Instead of comparing variances for each configuration, we can compare the expected variance $E(\sigma_E^2)$ with the expectation taken over all possible configurations of $\{k_i: i \in S_R\}$.

To demonstrate, consider 1–1 alignment with $N = 2500$, $N_1 = 50$, and $n = 250$, reflecting a rare allele. For each $n_1 = \{1, 2, \dots, N_1\}$, we calculate the average variance of the sample enrichment approach based on 5000 simulations (ie, 5000 possible configurations). The ratio of the expected variance of the sample enrichment approach over that of the random sampling approach is 0.928, showing the efficiency of the 1–1 design.

8.2 | Relative cost efficiency

In the present paper, the relative statistical efficiency, defined as the ratio of variances of estimators, was used to compare different designs. This is a common practice in statistical literature. In some practical applications, both statistical precision and real cost are of concern, in which case, Thomas' (2007) relative cost efficiency¹⁵ can be used, which provides further informative understanding for the comparison of different study designs. According to Thomas (2007), the relative cost efficiency is the relative statistical efficiency multiplied by the ratio of costs.

The present paper deals with a secondary problem following the completion of genotyping, therefore no additional cost will incur for genotyping. There are, however, two potential sources of cost in the follow-up study. One is the laboratory assay cost for phenotypic

measures for each pool; the total cost of such is proportional to the number of pools. Another is the cost associated with each subject selected for follow-up, including possibly purchasing cost of the biospecimens and the cost for combining samples for pooling. If only the laboratory assay cost occurs, then the cost for assaying a random sample of size n is the same as for assaying n pools; the relative cost efficiency then reduces to the relative statistical efficiency.

Otherwise, let the cost per laboratory assay is C_1 and the cost per subject is C_2 . Then, the relative cost efficiency of a pooling design with n pools and pool size $\{k_i, i = 1, \dots, n\}$ to a random reference sample of size n is

$$\frac{nC_1 + C_2 \sum_{i=1}^n k_i}{n(C_1 + C_2)} = \frac{nc + \sum_{i=1}^n k_i}{n(c + 1)}$$

multiplied by the variance ratio, where $c = C_1/C_2$. In practice, the assay specific cost C_1 is usually much higher than the subject specific cost, that is, c is usually large. Then, the relative cost efficiency is close to the relative statistical efficiency.

Future research is needed to investigate the performance of various enrichment designs and to search for optimal designs based on the relative cost efficiency criterion.

8.3 | Other remarks

A difficult problem that remains unsolved is whether optimal strategies, in terms of sampling methods and pool sizes, exist. Even if they do exist, the optimal strategy may itself depend on the genotype frequency, making this characterization not practically useful when multiple SNPs are of interest. To this end, methods in the work of Schisterman et al¹³ for calculating the optimal pool and unpooled samples to maximize efficiency could be extended for this problem.

The two-group comparison we focused on in the present paper applies to recessive and dominant genetic model, which essentially divides subjects into two groups. If the underlying genetic model is additive, then the genotype will be coded as 0, 1, and 2, which in turn yields three groups of subjects. In this case, the ANOVA and Kruskal-Wallis test can be used instead to compare a phenotypic variable among the three groups. It is expected that the enrichment design will continue to outperform the random reference group, when the allele frequency is relatively low. Extensions of the enrichment approach to multiple-group comparison deserve further investigation.

When multiple correlated genotypes are of interest, the problem for association studies becomes one that tests the equality of phenotypic means in all possible genotype (haplotype) combinations. For example, if k genetic markers (loci) are considered in an association study, then there are 2^k possible haplotype pairs that are compatible with the observed genotypes and the null hypothesis is that there is no difference between the phenotypic means of these 2^k genotypic groups. One potential method is to directly extend the sample enrichment approach discussed in the present manuscript to the comparison of multiple

groups. The correlations between the markers are automatically embedded in the grouping of the genotypes. The effect of genotyping errors on the efficiency and means to correct for genotyping errors is also worth further investigating.

Multiple phenotypes are commonly encountered in practice including GWAS. Indeed, the TSS GWAS that motivated our interest measured concentration levels of a number of vitamins and biomarkers from each student, although in this study, investigators are interested in each individual phenotypic variable. When multiple phenotypes are being tested simultaneously, the problem becomes one that tests the equality of phenotypic mean vectors between independent groups. This can be accomplished by applying a proper multivariate test such as Hotelling's T^2 or multivariate analysis of variance.

In the present paper, the sampling approach is constructed by assuming a homogenous population, which is the case in the TSS GWAS (participants are students in a college in Ireland). If stratification exists in the population, then both genotypes and phenotypes need to be adjusted for population stratification; see, among others, the work of Price et al.¹⁶ When the sample enrichment design is employed, only the averages of the individual phenotypic levels are available. This would make it extremely difficult to adjust for stratification for the phenotypes. Further research is needed along these lines.

Finally, it is worth pointing out that, in the general context of pooling biospecimens to measure the levels of biomarkers on a continuous scale, a key assumption is that the value calculated from a pool of samples is the average of the individual levels of the samples in the pool. While in many cases, this assumption seems well justified, in other cases, caution needs to be exercised when making such an assumption.

ACKNOWLEDGEMENTS

The authors thank the editor and two anonymous referees for their insightful comments and suggestions that substantially improved the manuscript. Research of the authors is supported by the Intramural Research Program of the Eunice Kennedy Shriver National Institute of Child Health and Human Development. The design, recruitment, and all metabolite and genomic analyses of the TSS cohort were implemented through the collaboration of Prof. John Scott and Dr. Anne Molloy, Trinity College Dublin, Ireland, Dr. Peadar Kirke, Health Research Board, Dublin, Ireland, Dr. Lawrence Brody and Dr. Faith Pangilinan, National Human Genome Research Institute (NHGRI), and Dr. James Mills, Eunice Shriver National Institute for Child Health and Human Development (NICHD) with funding through NICHD contract NO1-HD-3-3348 and with additional financial support from NHGRI and the Health Research Board, Ireland. The authors thank Dr. Yaakov Malinovsky for helpful comments and suggestions, and Mary Conley for help with the TSS GWAS data.

Funding information

Intramural Research Program of the Eunice Kennedy Shriver National Institute of Child Health and Human Development; NICHD, Grant/Award Number: NO1-HD-3-3348; NHGRI; Health Research Board, Ireland

APPENDIX

TECHNICAL DERIVATIONS

Equations (4) and (5)

By Equation (3), we can use a regression model $E(Y) = X\beta$ to relate $\tilde{Y}_{iR}$ and r_i , $i = 1, \dots, n$, where

$$Y = \begin{pmatrix} \tilde{Y}_{1R} \\ \vdots \\ \tilde{Y}_{nR} \end{pmatrix}, X = \begin{pmatrix} 1 & r_1 \\ \vdots & \vdots \\ 1 & r_n \end{pmatrix}, \text{ and } \beta = \begin{pmatrix} \mu_0 \\ \delta \end{pmatrix},$$

with $\text{Cov}(Y) = \Sigma$, an $n \times n$ variance-covariance matrix with diagonal elements being $\text{Var}(\tilde{Y}_{iR}) = \theta_i/k_i$ and off-diagonal elements all being 0.

Assuming $\theta_i = r_i\sigma_1^2 + (1 - r_i)\sigma_2^2$ are known, then the weighted least squares estimate of β is given by

$$\tilde{\beta} = \begin{pmatrix} \tilde{\mu}_0 \\ \tilde{\delta}_E \end{pmatrix} = (X' \Sigma^{-1} X)^{-1} X' \Sigma^{-1} Y,$$

yielding

$$\begin{aligned} \tilde{\mu}_0 &= \frac{\left(\sum \frac{k_i r_i^2}{\theta_i} \right) \left(\sum \frac{k_i \tilde{Y}_{iR}}{\theta_i} \right) - \left(\sum \frac{k_i r_i}{\theta_i} \right) \left(\sum \frac{k_i \tilde{Y}_{iR}}{\theta_i} \right)}{\left(\sum \frac{k_i}{\theta_i} \right) \left(\sum \frac{k_i r_i}{\theta_i} \right) - \left(\sum \frac{k_i}{\theta_i} \right)^2}, \text{ and} \\ \tilde{\delta}_E &= \frac{\left(\sum \frac{k_i}{\theta_i} \right) \left(\sum \frac{k_i \tilde{Y}_{iR}}{\theta_i} \right) - \left(\sum \frac{k_i r_i}{\theta_i} \right) \left(\sum \frac{k_i \tilde{Y}_{iR}}{\theta_i} \right)}{\left(\sum \frac{k_i}{\theta_i} \right) \left(\sum \frac{k_i r_i}{\theta_i} \right) - \left(\sum \frac{k_i}{\theta_i} \right)^2}. \end{aligned}$$

Note that $\text{Cov}(\tilde{\beta}) = (X' \Sigma^{-1} X)^{-1}$. Thus, we have

$$\text{Var}(\tilde{\delta}_E) = \frac{\sum \frac{k_i}{\theta_i}}{\left(\sum \frac{k_i}{\theta_i} \right) \left(\sum \frac{k_i r_i}{\theta_i} \right) - \left(\sum \frac{k_i}{\theta_i} \right)^2}.$$

The estimator $\hat{\delta}_E$ in Equation (4) is then obtained by replacing θ_i in $\tilde{\delta}_E$ with its consistent estimator $\hat{\theta}_i$ s. The (asymptotic) variance of $\hat{\delta}_E$ is given by Equation (5), utilizing that fact that $\text{Var}(\hat{\delta}_E) = \text{Var}(\tilde{\delta}_E)$ (in large sample sense).

Equations (6) and (7)

Note that the *one-big-pool* approach is a special case of the sample enrichment with $k_i = N - n + 1$ for the selected subject and 1 for all others. For $i \in \mathcal{S}_R^-$, we observe that $k_i = 1$, $k_{1i} = G_i = r_i$, and $\theta_i = G_i\sigma_1^2 + (1 - G_i)\sigma_0^2$. With the estimators $\hat{\sigma}_1^2$ and $\hat{\sigma}_0^2$, θ_i can be estimated by

$\hat{\theta}_i = G_i \hat{\sigma}_1^2 + (1 - G_i) \hat{\sigma}_0^2$. As denoted earlier, there are $\tilde{n}_1$ subjects with $G = 1$ and $\tilde{n}_0$ subjects with $G = 0$ in $\mathcal{S}_R^-$, where $\tilde{n}_1 + \tilde{n}_0 = n - 1$; in the big pool, a total of $N_1 - \tilde{n}_1$ subjects have genotype $G = 1$ and $N_0 - \tilde{n}_0$ subjects have genotype $G = 0$. Hence, it follows that

$$\begin{aligned} \sum_{i \in \mathcal{S}_R} \frac{k_i}{\hat{\theta}_i} &= \frac{\tilde{n}_1}{\hat{\sigma}_1^2} + \frac{\tilde{n}_0}{\hat{\sigma}_0^2} + \frac{N - n + 1}{\hat{\theta}}, & \sum_{i \in \mathcal{S}_R} \frac{k_{1i}}{\hat{\theta}_i} &= \frac{\tilde{n}_1}{\hat{\sigma}_1^2} + \frac{N_1 - \tilde{n}_1}{\hat{\theta}}, \\ \sum_{i \in \mathcal{S}_R} \frac{k_i \bar{Y}_{iR}}{\hat{\theta}_i} &= \frac{\tilde{n}_1 \bar{Y}_1}{\hat{\sigma}_1^2} + \frac{\tilde{n}_0 \bar{Y}_0}{\hat{\sigma}_0^2} + \frac{(N - n + 1) \bar{Y}}{\hat{\theta}}, & \sum_{i \in \mathcal{S}_R} \frac{k_{1i} \bar{Y}_{iR}}{\hat{\theta}_i} &= \frac{\tilde{n}_1 \bar{Y}_1}{\hat{\sigma}_1^2} + \frac{(N_1 - \tilde{n}_1) \bar{Y}}{\hat{\theta}}, \end{aligned}$$

where $\hat{\theta} = [(N_1 - \tilde{n}_1) \hat{\sigma}_1^2 + (N_0 - \tilde{n}_0) \hat{\sigma}_0^2] / (N - n + 1)$. Plugging them into (4) yields

$$\hat{\sigma}_E^2 = \frac{\frac{\tilde{n}_1}{\hat{\sigma}_1^2} \left(\frac{\tilde{n}_0}{\hat{\sigma}_0^2} + \frac{N_0 - \tilde{n}_0}{\hat{\theta}} \right) \bar{Y}_1 - \frac{\tilde{n}_0}{\hat{\sigma}_0^2} \left(\frac{\tilde{n}_1}{\hat{\sigma}_1^2} + \frac{N_1 - \tilde{n}_1}{\hat{\theta}} \right) \bar{Y}_0 + \frac{N - n + 1}{\hat{\theta}} \left(\frac{r \tilde{n}_0}{\hat{\sigma}_0^2} - \frac{(1 - r) \tilde{n}_1}{\hat{\sigma}_1^2} \right) \bar{Y}}{\frac{\tilde{n}_0}{\hat{\sigma}_0^2} \left(\frac{\tilde{n}_1}{\hat{\sigma}_1^2} + \frac{r(N_1 - \tilde{n}_1)}{\hat{\theta}} \right) + \frac{\tilde{n}_1 (1 - r)^2 (N - n + 1)}{\hat{\theta} \hat{\sigma}_1^2}},$$

where $r = (N_1 - \tilde{n}_1) / (N - n + 1)$. Analogously, the Equation (5) becomes

$$\sigma_E^2 = \frac{\frac{\tilde{n}_1}{\sigma_1^2} + \frac{\tilde{n}_0}{\sigma_0^2} + \frac{N - n + 1}{\theta}}{\frac{\tilde{n}_0}{\sigma_0^2} \left(\frac{\tilde{n}_1}{\sigma_1^2} + \frac{r(N_1 - \tilde{n}_1)}{\theta} \right) + \frac{\tilde{n}_1 (1 - r)^2 (N - n + 1)}{\theta \sigma_1^2}},$$

where $\theta = r \sigma_1^2 + (1 - r) \sigma_0^2$. Hence, we obtain the expected results in (6) and (7).

REFERENCES

- Schisterman EF, Vexler A, Ye A, Perkins NJ. A combined efficient design for biomarker data subject to a limit of detection due to measuring instrument sensitivity. *Ann Appl Stat*. 2011;5:2651–2667.
- Dorfman D The detection of defective members of large populations. *Ann Math Statist*. 1943;14:436–440.
- Sobel M, Groll PA. Group testing to eliminate efficiently all defectives in a binomial sample. *Bell System Tech J*. 1959;38:1179–1252.
- Gastwirth J, Johnson W. Screening with cost-effective quality control: potential applications to HIV and drug testing. *J Am Stat Assoc*. 1994;89:972–981.
- Chen CL, Swallow WH. Using group testing to estimate a proportion, and to test the binomial model. *Biometrics*. 1990;46:1035–1046. [PubMed: 2085624]
- Hughes-Oliver JM, Swallow WH. A two-stage adaptive group-testing procedure for estimating small proportions. *J Am Stat Assoc*. 1994;89:982–993.
- Tu XM, Litvak E, Pagano M. On the informativeness and accuracy of pooled testing in estimating prevalence of a rare disease: application to HIV screening. *Biometrika*. 1995;82:287–289.

8. Weinberg CR, Umbach DM. Using pooled exposure assessment to improve efficiency in case-control studies. *Biometrics*. 1999;55:718–726. [PubMed: 11314998]
9. Faraggi D, Reiser B, Schisterman EF. ROC curve analysis for biomarkers based on pooled assessments. *Statist Med*. 2003;22:2515–2527.
10. Kendziora CM, Zhang Y, Lan H, Attie D. The efficiency of pooling mRNA in microarray experiments. *Biostatistics*. 2003;4:465–477. [PubMed: 12925512]
11. Liu A, Schisterman EF. Comparison of diagnostic accuracy of biomarkers with pooled assessments. *Biom J*. 2003;45:631–644.
12. Yuan M, Yang Y, Zheng G. Two-stage genome-wide association studies with DNA pooling and genetic model selection. *Stat Sin*. 2009;19:1769–1786.
13. Schisterman EF, Vexler A, Mumford SL, Perkins NJ. Hybrid pooled-unpooled design for cost-efficient measurement of biomarkers. *Stat Sin*. 2010;29:597–613.
14. Tebbs JM, Swallow WH. Estimating ordered binomial proportions with the use of group testing. *Biometrika*. 2003;90:471–477.
15. Thomas DC. Multistage sampling for latent variable models. *Lifetime Data Anal*. 2007;13:565–581. [PubMed: 17943440]
16. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genet*. 2006;38(8):904–909. [PubMed: 16862161]

FIGURE 1.

Variance of the estimator by simple random sampling, $\sigma_1^2 = 0.5$ (dash line), 1 (solid line), and 1.5 (dotted line)

FIGURE 2.

Relative efficiency of enrichment to random sampling (solid lines correspond to $\tilde{n}_1 = n_1 - 1$, and dash lines to $\tilde{n}_1 = n_1$)

FIGURE 3.
Power as a function of the number of homozygous minor alleles

FIGURE 4.

Log averaged p-values of the The Trinity Students data for each sampling method (the dashed line corresponds to the type I error correction threshold of $\log(0.05/120)$). Single-nucleotide polymorphisms were sorted by the ascending order of p-values for the complete analysis