Published in final edited form as:

Stat Med. 2012 November 20; 31(26): 3241–3252. doi:10.1002/sim.5357.

Estimation of Gene-Environment Interaction by Pooling Biospecimens

M.R. Danaher^{1,2}, E.F. Schisterman¹, A. Roy², and P.S. Albert¹

¹Division of Epidemiology, Statistics and Prevention Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development

²University of Maryland, Baltimore County

Summary

Case-control studies are prone to low power for testing gene-environment interactions (GXE) given the need for a sufficient number of individuals on each strata of disease, gene, and environment. We propose a new study design to increase power by strategically pooling biospecimens. Pooling biospecimens allows us to increase the number of subjects significantly, thereby providing substantial increase in power. We focus on a special, though realistic case, where disease and environmental statuses are binary, and gene status is ordinal with each individual having 0, 1 or 2 minor alleles. Through pooling, we obtain an allele frequency for each level of disease and environmental status. Using the allele frequencies, we develop new methodology for estimating and testing GXE that is comparable to the situation when we have complete data on gene status for each individual. We also explore the measurement process and its effect on the GXE estimator. Using an illustration, we show the effectiveness of pooling with an epidemiologic study which tests an interaction for fiber and PON1 on anovulation. Through simulation, we show that taking 12 pooled measurements from 1000 individuals achieves more power than individually genotyping 500 individuals. Our findings suggest that strategic pooling should be considered when an investigator designs a pilot study to test for a GXE.

Keywords

Allele frequency measurements; case-control study; gene-environment interaction; pooling; power

1. Introduction

In studying complex diseases such as infertility, cancer, heart disease, and obesity, gene-environment interactions are critical hypotheses [1-3]. One important implication of unmasking gene-environment interactions is to identify highly susceptible populations such that modifiable exposures that cause disease can be prevented. For example, the Clean Air Act mandates that the U.S. Environmental Protection Agency aims to set standards to protect children, the elderly, those with chronic disease, and genetically susceptible

individuals [3]. However, genetically susceptible individuals cannot be identified unless there exist a better understanding of how genes interact with individuals' environments.

The case-control study design is traditionally used to estimate gene-environment interactions because it is convenient both in terms of cost and study duration, particularly in studying rare diseases [4]. However, the traditional case-control interaction estimator is known to require a large sample size to obtain reasonable statistical power [5, 6]. This is due to the large number of individuals necessary to obtain sufficient information on each level of disease, environment, and genotype. Unfortunately genotyping a large number of individuals can be expensive and time consuming. While in recent years the cost of measuring a candidate gene for individuals has become less of a burden, the power to detect an interaction will always benefit when the cost associated with measurement decreases and the recruitment costs increase. Moreover, in certain situations the investigator might not have access to enough biospecimens from each individual to measure the genotype with current technology. Hence, cost and biospecimen availability often leave many interesting gene-environment interaction hypotheses unexplored.

As a cost and biospecimen saving alternative, multiple authors have proposed pooling specimens to increase statistical power [7-10]. The first to attempt to pool DNA in a case-control study was researched by [11], and since then pooling has become popular in genome-wide studies for estimating allele frequencies within a pool [10, 12, 13].

If an investigator had access to a large number of stored biospecimen samples then it is easy to understand that pooling methods would be very useful in preliminary assessments of gene-environment interaction hypotheses. Additionally, with new genotype technology constantly emerging, and investigator might choose to pool biospecimen using an expensive and accurate method rather than the less expensive and inaccurate alternative.

In this paper we explore a novel pooling strategy, where we pool within disease and environmental status to increase statistical power to test a gene-environment interaction hypothesis. First we propose a strategic pooling method that provides a gene-environment interaction estimator. Next we discuss the pooling laboratory process and provide an example of how pooled measurements are performed. Then we discuss the effect of the pooling process on the gene-environment interaction estimator and its efficiency. Subsequently, we explore the consequences of pooling in the BioCycle study of the gene human serum paraoxonase polymorphism (PON1), and its interaction with dietary fiber consumption on the effect of anovulation. Then we perform a simulation study to evaluate the power and coverage probability of the pooled gene-environment interaction estimator under different measurement errors and model assumptions. Lastly, we provide discussions and a small simulation study for estimating the gene-environment interaction adjusted for categorical covariates.

2. Pooled Gene-Environment Interaction Estimator

While these methods can be extended, for illustration, we focus on a special, though realistic case-control study, where the environmental variable of interest is binary (e.g. smoking status) and the data for genotype is ordinal. Let D represent disease status, where $D = 1$

represents individuals with disease presence (case), and $D = 0$ represents without (control). Let E represent environmental exposure status, where $E = 1$ represents with an exposure, and $E = 0$ represents without. We assume that we are exploring a biallelic single nucleotide polymorphism (SNP) at a given SNP locus. Therefore the SNP genotype consists of two alleles, one from each chromosome at a known location, where one allele is of interest. Therefore, for each individual, $G = 0, 1$, or 2 represents none, one, or two copies of the allele of interest respectively. Let the number of individuals with disease and environmental status D and E be r_{DE} , and the number of individuals with disease, environmental, and genotype status D , E and G be n_{DEG} .

Defining the genotype status as ordinal implies that we are assuming the multiplicative or additive model as the underlying genetic model as opposed to a dominant or recessive model. When the genetic model is assumed to be a dominant or recessive model, little methodological development is necessary, as will be discussed later. Moreover, we define the genotype status of an individual using a biallelic SNP. Therefore, we assume that the genotype comprises a pair of alleles at a single known locus, and for each individual there are two possible copies of the SNP location, one per chromosome. We assume that the pair of alleles are independent, and the probability of receiving a SNP copy from the father is identical to the probability of receiving it from the mother.

Assuming that the disease and environmental statuses are known before genotyping, which is typically the situation in a case control study, we propose pooling the individuals' biospecimens within all four levels of D and E and then performing R replicate measurements of the frequency within each of the four levels. Therefore we obtain R replicate measurements at each level of D and E ($4R$ overall), rather than obtaining r_{DE} individual genotypes. The pooled measurement will yield an estimate of the minor allele frequency of interest which we denote by f_{DE} . By pooling, we no longer observe n_{DEG} , rather we observe an estimate of f_{DE} , and we know r_{DE} (the pool size). Using estimates of the pooled information, f_{DE} and r_{DE} , we are able to estimate the gene-environment interaction as if we had observed n_{DEG} individual genotypes.

Let the allele status (A) be binary where $A = 1$ represents the allele of interest, and $A = 0$ otherwise. Assume that the disease is related to the allele and environmental statuses by

$$\text{logit}(P(D|A, E)) = \alpha + \beta_E E + \beta_A A + \beta_{A \times E} AE. \quad (1)$$

Therefore $\beta_{A \times E}$ is the allele-environment interaction parameter. While it is immediately clear that the maximum likelihood estimator for $\beta_{A \times E}$ has a closed form expression which only depends on the values of f_{DE} , it is not immediately clear how an allele-environment interaction relates back to a gene-environment interaction. For example an allele-environment interaction might be interpreted as having a specified allele and environmental status which increases the risk for a disease; however since an individual has a pair of alleles, one on each chromosome, the allele-environment interaction does not directly relate back to the individual. We need to formulate how we can relate an allele-environment interaction to a gene-environment interaction.

Define A_i ($i = 1, 2$) as the i^{th} allele for an individual where $A_i = 1$ if the i^{th} allele is the allele of interest, and $A_i = 0$ otherwise. Under the assumed additive model, for each individual G is the sum of two alleles $G = A_1 + A_2$. If A_1 and A_2 are independent within disease and environmental status, and the probability of having the disease with A_1 is equivalent to having it with A_2 , then we can use the set based logistic model of [7], where each individual is a set of two alleles. The assumption of independence between A_1 and A_2 , and the assumption that they are equally probable are common assumptions in genetic epidemiology. The assumptions imply that the parents' genotypes are independent, and the child is equally likely to inherit the gene from the mother as they are from the father. Using the set based logistic model (derived in section A of the supplementary material), we relate disease to genotype and environmental status by

$$\text{logit}(P(D|G, E)) = \tilde{\alpha} + \tilde{\beta}_E E + \beta_A G + \beta_{A \times E} G E. \quad (2)$$

The parameters $\tilde{\alpha}$ and $\tilde{\beta}_E$ are defined in section A of the supplementary material. The gene-environment interaction parameter of interest, is equivalent to the allele-environment interaction ($\beta_{G \times E} = \beta_{A \times E}$). Therefore, estimators of $\beta_{A \times E}$ using allele information will be equivalent to estimators of $\beta_{G \times E}$ obtained using full information on individual genotypes. Hence, inference of the allele-environment interaction ($\beta_{A \times E}$) is exactly the same as the inference of the gene-environment interaction ($\beta_{G \times E}$).

It is important to note that using (1) to estimate the gene-environment interaction is not equivalent to using the set based model as proposed by [7]. In their paper they developed a method which assumes that the pooled exposure measurement is the average of the individuals' exposure. This method can be adapted to handle allele frequencies (f_{DE}), which are half of the average genotyped status (G) within a pool. However, if one adapted this method, the estimator for the gene-environment interaction would not be in closed form as it is in (1), and more pools would be required for estimation. If the dominate or recessive genotype models were assumed, the genotype status (G), would be binary and thus the interaction estimator in (2) would be in closed form and only depend on estimates of $\Pr(G = 1|D, E)$. If one were to assume Hardy-Weinberg equilibrium so that $\Pr(A = 1|D, E) = f_{DE}$, then $\Pr(G=1|D, E) = f_{DE}^2 + 2f_{DE}(1 - f_{DE})$ for the dominate model, and $\Pr(G=1|D, E) = f_{DE}^2$ for the recessive model. Misspecification of the genotype model will bias our pooled gene-environment interaction estimator in the same way that misspecification would bias the traditional gene-environment interaction estimator.

Through the pooling process described in the next section, we have data on the alleles and therefore we can use (1), and obtain the closed form maximum likelihood estimate of $\beta_{A \times E}$,

$$\tilde{\beta}_{G \times E(MLE)} = \tilde{\gamma}_{11} - \tilde{\gamma}_{01} - \tilde{\gamma}_{10} + \tilde{\gamma}_{00} \quad (3)$$

where $\tilde{\gamma}_{DE} = \log\left(\frac{\tilde{f}_{DE}}{1 - \tilde{f}_{DE}}\right)$, $\tilde{f}_{DE}$ is the allele frequency in a sample of individuals with disease status D and environmental status E . However, due to measurement error inherent in

the pooling process, we do not observe $\tilde{f}_{DE}$, rather we observe $\hat{f}_{DE(k)}$ which is an estimate of the allele frequency in the pool with disease status D and environmental status E measured on replicate k , where $k = 1, 2, 3, \dots, R$, and R is the number of replicate measurements for each pool. We construct an estimate of the gene-environment interaction by averaging replicate allele frequencies on the logit scale. The intuition for taking the average over replication is based on our understanding that a single estimate of each allele frequency within strata of disease and environment can be used for asymptotically unbiased estimation for (3); however, replication is necessary for estimation of the variance of the estimator.

Therefore, we define an estimate of $\tilde{\gamma}_{DE}$ to be $\hat{\gamma}_{DE} = \frac{1}{R} \sum_{k=1}^R \log \left(\frac{\hat{f}_{DE(k)}}{1 - \hat{f}_{DE(k)}} \right)$. Using $\hat{\gamma}_{DE}$ we define the estimator by,

$$\hat{\beta}_{G \times E} = \hat{\gamma}_{11} - \hat{\gamma}_{01} - \hat{\gamma}_{10} + \hat{\gamma}_{00}. \quad (4)$$

Further explanation of the estimation process for the allele frequencies of the four pools (one pool on each level of disease and environmental status) is provided in the following sections.

3. The Pooling Process

A number of methods have been proposed for measuring SNP allele frequencies of pools [12-18]. There are three main stages in obtaining a pooled measurement of the gene frequency. The first stage is called pool formation. In this stage, equal quantities of individuals' DNA must be combined to form the pool. The second stage is called PCR (polymerase chain reaction). In this stage, a targeted sequence is amplified. Lastly, there is the estimation stage. In this stage, using a methodology of choice, the minor allele frequency is estimated [17].

There are a number of measurement procedures available for estimating the frequency of a SNP [19-23]. For illustration, we describe the method used in [12]. It is straightforward to extend our estimation procedure to other methods, including those in [10, 18]. In the method described by [12], the allele frequency is determined by comparing the number of cycles of PCR amplifications (each PCR amplification cycle doubles the frequency) required before a reaction crosses a predetermined frequency threshold. For two alleles, two PCR amplifications are required, and the number of PCR amplification cycles are compared (Figure 1 illustrates this process). If one allele has a one cycle delay to the other, then the minor allele ratio is 1:2. If one allele has a two cycle delay to the other, then the minor allele ratio is 1:4. If we define $\widetilde{\Delta C}_{tDE}$ as the cycle delay for the pool on the level DE , then in general the allele ratio is $1:2^{\widetilde{\Delta C}_{tDE}}$. Therefore, the minor allele frequency in the population of individuals with disease status D and environmental status E can be estimated as

$$\tilde{f}_{DE} = \frac{1}{1 + 2^{\widetilde{\Delta C}_{tDE}}}, \quad (5)$$

where the number of allele, and hence $\widetilde{\Delta C}_{tDE}$, cycles can be fractional.

4. The Pooled Gene-Environment Interaction Estimator with Measurement Error

After forming the four pools on each level of D and E , we obtain an estimate of the cycle difference, $\widetilde{\Delta C}_{tDE}$, for each replicate measurement. Note that we are defining a replicate as repeating the measurement processes, including forming the pools with equal quantities of specimen, PCR amplification and cycle delay measurement. Let $\hat{\Delta C}_{tDE(k)}$ be the estimated cycle difference for the k^{th} replicate of the pool with disease status D , and environmental status E . Define,

$$\hat{\Delta C}_{tDE(k)} = \widetilde{\Delta C}_{tDE} + \varepsilon_{DE(k)} \quad (6)$$

where $\varepsilon_{DE(k)}$ is the measurement error for the k^{th} replicate. We assume

$\varepsilon_{DE(k)} \sim \text{iid} \sim N(0, \sigma_{DE}^2)$, for $k = 1, 2, 3, \dots, R$. Using the pooled measurements, $\hat{\Delta C}_{tDE(k)}$, we can estimate the interaction,

$$\hat{\beta}_{G \times E} = \frac{-\log(2)}{R} \sum_{k=1}^R \left(\hat{\Delta C}_{t11(k)} - \hat{\Delta C}_{t01(k)} - \hat{\Delta C}_{t10(k)} + \hat{\Delta C}_{t00(k)} \right) \quad (7)$$

and its variance by,

$$\text{var}(\hat{\beta}_{G \times E}) = \sum_{D=0}^1 \sum_{E=0}^1 \left\{ \frac{1}{R} \sum_{k=1}^R \left[\frac{1}{r_{DE}(1 - \hat{f}_{DE(k)})(\hat{f}_{DE(k)})} \right] + \frac{1}{R} [\log(2)]^2 s_{DE}^2 \right\} \quad (8)$$

where s_{DE}^2 is the sample variance of the pooled measurements $\hat{\Delta C}_{tDE(k)}$, and

$\hat{f}_{DE(k)} = \frac{1}{1 + 2\hat{\Delta C}_{tDE(k)}}$. The derivation of this asymptotically unbiased estimator of the gene-environment interaction and the estimator of its variance are given in section B and C of the supplementary material. The expression for the variance depends on the frequency of each of the pools and the variance due to pooling. The variance estimator is shown to be positively biased in section C of the supplementary material.

5. Example

The BioCycle study is a prospective cohort study of 259 regularly menstruating women ages 18 to 44, who were followed for up to two menstrual cycles. Details of the study are described elsewhere [24]. Within the BioCycle study, Gaskins et al. found that a higher consumption of dietary fiber was significantly associated with lower concentrations of reproductive hormones and an increased risk of incident anovulation [25]. Also, the human serum paraoxonase (PON1) has been identified as a prime example of gene/polymorphism that plays an important role in ‘environmental susceptibility’ to multifactorial diseases and disorders [26]. Therefore a natural hypothesis might be to investigate whether there is an

interaction effect of fiber and PON1 on anovulation. Table 2 shows the data where all individuals in the study are genotyped, and dichotomized fiber intake as above and below the mean (there are two missing values).

Using the data in Table 2, the gene-environment interaction estimate is -0.4459 (S.E. = 0.6082), with a p-value of 0.4634. Thus, the gene-environment interaction estimate is not significant. However, this study has a very low power to detect an interaction if one were to exist. For illustration, we took a random sample of 1000 individuals with replacement from the BioCycle data. While the investigators were not able to observe 1000 individuals due to cost constraints, we took a random sample with replacement to show that the power of the interaction estimator could be improved if fewer costs were allocated for genotyping, and more funding were spent on collecting a large sample size. Table 3 shows an example of one random sample of 1000 individuals with replacement from the BioCycle data.

The data in Table 3 yield a gene-environment interaction estimate of -0.7236 (S.E. = 0.296), with a significant p-value of 0.014. This data is an illustration, and due to the combined cost of genotyping and collecting the larger sample, the investigator might not be able to collect a sample with 1000 individuals with individual measurements of genotype (G). On the other hand, the investigator might have enough funding to sample 1000 individuals if the cost due to genotyping was reduced significantly to 12 measurements, 3 repeated measurements of the allele frequency on each level of D and E , rather than 1000 measurements of G .

Using the random sample given in Table 3, we observe that $\tilde{f}_{11} = 0.0275$, $\tilde{f}_{10} = 0.4877$, $\tilde{f}_{01} = 0.3207$, $\tilde{f}_{00} = 0.3689$. Using (5) we can solve $\widetilde{\Delta C}_{tDE}$ for each level of D and E . Recall that $\hat{\Delta C}_{tDE(k)} = \widetilde{\Delta C}_{tDE} + \varepsilon_{DE(k)}$, where $\varepsilon_{DE(k)} \left(\varepsilon_{DE(k)} \sim \text{iid} \sim N(0, \sigma_{DE}^2) \right)$ is the measurement error for the k^{th} replicate, $k = 1, 2, 3$. For illustration, we set $\sigma_{DE}^2 = .05$ which is consistent with the measurement error seen in literature [17]. By simulating a value of $\varepsilon_{DE(k)} \sim \text{iid} \sim N(0, .05)$ for each level of D , E , and k , we obtain $\hat{\Delta C}_{tDE(k)}$ values presented in Table 4.

Using the data in Table 4, we can solve (7) and (8) to estimate the gene-environment interaction and its standard error. We find that the gene-environment interaction estimate is -0.7097 (S.E. 0.343), with a significant p-value of 0.0387.

This illustration shows that in a large sample from a population similar to the one the generated the data in the BioCycle Study, pooling can be implemented to increase the power of an estimator without increasing the cost of the study. We emphasize that this approach is only used to illustrate the performance of our design, and cannot be used to make statistical inference about the interaction of PONS and fiber on ovulation. However, our illustration shows that if an investigator is interested in a particular gene-environment interaction, they could consider increasing their funding for collecting individuals by decreasing their cost due to genotyping. If one adapts a pooling strategy such that the cost due to genotyping is reduced, often the investigator is capable of allocating more funding to collect a larger sample size, and thus can substantially increase the power to test the gene-environment interaction hypothesis.

6. Simulations

We performed simulations to compare the statistical properties of the pooled gene-environment interaction estimator to the estimator we would obtain if we had data on individual genotypes. In the BioCycle anovulation example explored in the previous section, the allele frequency of the sample is 0.37, and the proportion of individuals with the environmental exposure is 0.25. Using these parameters, we simulate the alleles with $\Pr(A = 1) = 0.37$, and then randomly pair the alleles for each genotype observation. The number of minor alleles ($A=1$) for each observation is defined as G (therefore G is 0, 1, or 2). Subsequently, we generate the environmental exposure with $\Pr(E = 1) = 0.25$. Lastly, we set the following logistic regression equation to generate D with a dependence on E and G ,

$$\text{logit}(\Pr(D=1|E, G)) = \alpha + \beta_E E + \beta_G G + \beta_{G \times E} GE$$

where $\alpha = -2.26$, $\beta_E = 1.06$, and $\beta_G = 0.18$. These beta coefficients are obtained by fitting the above model to the BioCycle anovulation example. To explore the properties of our estimate of the gene-environment interaction, we vary $\beta_{G \times E}$ between -1 and 1.

Once the cohort (disease, gene, and environmental status) are generated, we took a case-control sample (500 cases and 500 controls). Next, we pooled our sample on disease and environmental status to find the minor allele frequency ($\tilde{f}_{DE}$) for each of the four pools (four levels of disease and environmental status). Once we had the simulated frequencies for our case-control sample, using (5) we found the value of $\widetilde{\Delta C}_{tDE}$ for each level of D and E .

Lastly we add measurement error using $\hat{\Delta C}_{tDE(k)} = \widetilde{\Delta C}_{tDE} + \varepsilon_{DE(k)}$, where $\varepsilon_{DE(k)}$ is the measurement error of the k^{th} replicate ($\varepsilon_{DE(k)} \sim \text{iid} \sim N(0, \sigma_{DE}^2)$, for $k = 1, 2, 3, \dots, R$). In our simulation we defined R to be 3 and $\sigma_{DE}^2 = 0.05$ for all four levels of disease and environmental status. This variance is consistent with the measurement error seen in literature [17].

Under this simulation scenario, Table 5 shows that the gene-environment interaction estimator is unbiased for the true gene-environment interaction. The variance estimators are very close to the empirical variances of the gene-environment interaction estimators in the Monte Carlo simulation. However there is a trend for the variance estimators to be slightly positively biased. This slight positive bias is in accordance with our theoretical results, and yields coverage probabilities which are slightly smaller than 0.95.

Once we have estimated the gene-environment interaction and the variance of the estimator, we were able to compute the power function using a two-sided Wald test (i.e. a test based on

standard normal assumption for $\hat{\beta}_{G \times E} / \sqrt{\text{var}(\hat{\beta}_{G \times E})}$ under the null hypothesis). We have shown that the pooled estimator of the gene-environment interaction is unbiased. Therefore, compared with the traditional estimator, which is also unbiased, the power of the pooled estimator will depend on the distribution of the estimated variance of the estimator, $\text{var}(\hat{\beta}_{G \times E})$. As the number of replicates measurements, R , increases, the extra term in the

variance estimator decreases towards zero. Therefore, for a large number of replicates, the power of the pooled estimator is nearly identical to the power of the traditional estimator. However, the purpose of pooling is to allow the number of measurements, R , to be as small as possible such that we achieve reasonable power for an alpha level test. At minimum, we require two replicate measurements to estimate the measurement error, s_{DE}^2 . However, if the number of replicate measurements does not allow accurate estimation of the measurement error, then we risk underestimating the variance of the estimator, and thus inflating our type one error. We explore the relationship between the power of the estimator and the number of replicate measurements in Figure 2.

To compare our pooling method to traditional methods, Figure 2 shows the power functions of five different scenarios. The first scenario is the same scenario described for Table 5. In this scenario the investigator uses our pooling strategy to estimate the gene-environment interaction with a case-control sample of 1000 individuals. The investigator has 3 replicate measurements on each level of disease and environment, yielding 12 total measurements, and we assume that the measurement error variance (σ_{DE}^2) is 0.05 for all levels of D and E . The second scenario decreases the number of replicate measurements to 2, yielding 6 total measurements. To show that as the number of replicate measurements increases, the power approaches the power of the traditional estimator, the third scenario increases the number of replicate measurements to 100, yielding 400 total measurements. The fourth and fifth scenarios use the traditional estimator of the gene-environment interaction using a case-control sample of 1000 and 500 individuals, respectively. These scenarios require that the investigator genotypes 1000 or 500 individuals, respectively.

In Figure 2 we see that the gene-environment interaction estimators using our pooling strategy, involving 3 replicate measurements (12 total measurements), provides a test and achieves slightly more power than the test with the traditional gene-environment interaction estimator, involving 500 measurements. For higher number of replicates the power of the pooled test becomes nearly identical to the power of the traditional test. If the number of replicates is too small the type one error is generally inflated.

Figure 2 shows a dramatic decrease in the number of measurements required (12 pooled measurements compared to 500 individual measurements) to obtain approximately the same amount of power. However, as evident from Figure 2, care needs to be taken to accurately estimate the measurement error variance with replicates to obtain an alpha level test. Equation 5 suggests that measurement error near extreme true frequencies has less of an impact than measurement error near a true allele frequency of 0.5. Changing the population allele frequency, $\Pr(A = 1)$, will affect the allele frequencies within strata of disease and environmental statuses which are measured in pools. Therefore, we explored different population allele frequencies, with different magnitudes of measurement error to find the appropriate number of replicates to obtain an alpha level test. Table 6 shows the type I error under the same simulation scenario described for Table 5 except with different population allele frequencies, measurement error variances, and number of replicate measurements. Extreme frequencies are less impacted by measurement error, and thus require less replicate

measurements to obtain a test at alpha level. When the measurement error variance is 0.05 [17] more than 3 replicates are required to obtain a test at alpha level.

7. Incorporating Covariates

Until now we have only discussed a simple gene-environment interaction estimator which is not adjusted for covariates. However, incorporating a reasonable number of categorical covariates is straightforward. In this situation we suggest incorporating categorical covariates by stratifying the pools by disease, environment, and covariate statuses. Once pools are formed for each stratum, frequencies can be measured in replicate and gene-environment interactions can be estimated using the methods we have already discussed. Assuming that the gene-environment interactions are the same within each stratum, one can estimate the gene-environment interaction of interest using the weighted average,

$$\hat{\beta}_{G \times E} = \frac{1}{\sum_{c=1}^C w_c} \sum_{c=1}^C w_c (\hat{\beta}_{G \times E(c)}) \quad (7)$$

and its variance by,

$$\text{var}(\hat{\beta}_{G \times E}) = \left(\frac{1}{\sum_{c=1}^C w_c} \right)^2 \sum_{c=1}^C w_c^2 (\text{var}(\hat{\beta}_{G \times E(c)})) \quad (8)$$

where, $\hat{\beta}_{G \times E(c)}$ is the estimated gene-environment interaction within covariate strata, c . Also,

we define the weights to be the inverse variance, $w_c = \frac{1}{\text{var}(\hat{\beta}_{G \times E(c)})}$.

To explore the power of our pooled estimator adjusted for covariates, we considered a simple simulation scenario where we adjusted for a covariate with three levels, 0, 1, or 2, with the following probabilities, $\Pr(C=0) = 0.25$ and $\Pr(C=1) = 0.5$. We used the following logistic regression equation to generate E , with dependence on the covariate C ,

$$\text{logit}(\Pr(E=1|C)) = \gamma_0 + \gamma_c C$$

where $\gamma_0 = 0.2$. We varied γ_c to explore the effect of the relationship between E and C on our pooled estimator. We used the following logistic regression equation to generate D with a dependence on E , G , and C ,

$$\text{logit}(\Pr(D=1|E, G)) = \alpha + \beta_E E + \beta_G G + \beta_{G \times E} GE + \beta_C C$$

where we set $\alpha = -2.26$, $\beta_E = 1.06$, $\beta_G = 0.18$, $\beta_C = 5$, and varied $\beta_{G \times E}$ between -1 and 1. The number of replicate measurements, R , was set to 3, and the measurement error variance, σ_{DE}^2 , was 0.05. For our pooled estimates we took a case-control sample of 1000 individuals

to be measured in 12 pools (4 levels of D and E and 3 levels of C), and we compared the power to that of an individually measured case-control sample of size 500. Figure 3 shows that the test based on the adjusted pooled estimator, which requires only 36 measurements (12 pools each measured with 3 replicates), maintains the size at a nominal level while having more power than the test based on the traditional estimator with 500 individual genotype measurements.

8. Discussion

In this paper we have proposed an effective method for obtaining a gene-environment interaction estimator using pooled biospecimens. Pooling can greatly reduce the number of measurements, which in turn greatly reduces the cost due to genotyping. Pooling can also save valuable biospecimens because a pooled measurement requires less biospecimens from each individual. These savings provide a feasible method for investigating previously unexplored gene-environment interaction hypotheses with sufficient statistical power.

We have shown that when disease and environmental statuses are binary, the pooled frequency at each level of disease and environmental status is a sufficient statistic for the gene-environment interaction under some assumptions. Through simulation, we have demonstrated that with 3 replicate measurements on each level of disease and environment, and with a reasonable measurement error of 0.05, our pooled gene-environment interaction estimator and the estimate of the variance of the estimator are nearly unbiased. The simulation results indicate that a test based on 1000 individuals that uses our pooling strategy (12 genotyping measurements) has approximately the same power of a test with 500 individuals and genotypes of all 500 individuals.

While we have focused on the case where disease and environmental statuses are binary, these methods can be expanded to environmental statuses which are categorical. However, the requirement that the pools are formed within disease and environmental status suggest that for a large number of environmental categories a large number of pools will be necessary. In this situation, pooling might not dramatically decrease the number of assay measurements, and thus might not be recommended. Our pooling methods require that the disease and environment statuses, and possible confounders, must be known before the pooling can be performed. This is a slightly stronger assumption than those in case control studies. The requirement that the environmental status be known before pooling implies that there is no straightforward extension of our method to estimation of gene-gene interaction, where both genes are estimated using pooling. Further work is necessary to explore cheap, powerful estimation methodology for gene-gene interactions. The set based logistic model assumes that A_1 and A_2 (i.e. the mother and father) are independent within disease and environmental status, and the probability of receiving the allele from the mother is the same as that from the father. These assumptions may not be realistic for certain genes, and methods need to be developed to deal with departures from these assumption. Lastly, while these methods can adjust for categorical covariates, as the number of covariates increases the number of pools will also increase. Because the purpose of this method is to dramatically decrease the number of measurements required this method might not be appropriate when

an investigator would like to adjust for a large number of covariates. Adjusting for a large number of covariates might be done using the pooled method described by [7].

Although our method has limitations in its assumptions, it may be very useful in a pilot study. For example, if an investigator had access to a large number of stored biospecimen samples then our pooling method would be very useful in preliminary assessments of gene-environment interaction hypotheses. Additionally, our method might be very useful in high-dimensional SNP studies, as a direct extension to [10], where they related SNP allele frequencies based on disease statuses, and we related SNP allele frequencies based on disease and environmental statuses.

In conclusion, we proposed a powerful method to estimate a gene-environment interaction by pooling biospecimens. We have derived a pooled gene-environment interaction estimator, and the estimated variance of that estimator. Our findings show that strategic pooling should be considered when an investigator designs a pilot study to test for a gene-environment interaction.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

grant information: This research was supported partially by the Intramural Research Program of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, National Institutes of Health and by the Long-Range Research Initiative of the American Chemistry Council.

References

1. Amato R, Pinelli M, D'Andrea D, Miele G, Nicodemi M, Raiconi G, Coccozza S. A novel approach to simulate gene-environment interactions in complex diseases. *BMC Bioinformatics*. 2010; 11:8.10.1186/1471-2105-11-8 [PubMed: 20051127]
2. Marchand LL, Wilkens LR. Design considerations for genomic association studies: importance of gene-environment interactions. *Cancer Epidemiology Biomarkers and Prevention*. 2008; 17:263–267.10.1158/1055-9965.EPI-07-0402
3. Thomas D. Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies. *Annual Review of Public Health*. 2010; 31:21–36.10.1146/annurev.publhealth.012809.103619
4. Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika*. 1979; 66:403–411.10.1093/biomet/66.3.403
5. Foppa I, Spiegelman D. Power and sample size calculations for case-control studies of gene-environment interactions with a polytomous exposure variable. *American Journal of Epidemiology*. 1997; 146:596–604. [PubMed: 9326439]
6. Hwang SJ, Beaty TH, Liang KY, Coresh J, Khoury MJ. Minimum sample size estimation to detect gene-environment interaction in case-control designs. *American Journal of Epidemiology*. 1994; 140(11):1029–1037. [PubMed: 7985651]
7. Weinberg CR, Umbach DM. Using pooled exposure assessment to improve efficiency in case-control studies. *Biometrics*. 1999; 55:718–726. [PubMed: 11314998]
8. Faraggi D, Reiser B, Schisterman EF. ROC curve analysis for biomarkers based on pooled assessments. *Statistics in Medicine*. 2003; 22:2515–2527.10.1002/sim.1418 [PubMed: 12872306]

9. Schisterman EF, Vexler A, Mumford SL, Perkins NJ. Hybrid pooled-unpooled design for cost-efficient measurement of biomarkers. *Statistics in Medicine*. 2010; 29:597–613.10.1002/sim.382 [PubMed: 20049693]
10. Prentice RL, Qi L. Aspects of the design and analysis of high-dimensional SNP studies for disease risk estimation. *Biostatistics*. 2006; 7:339–354. DOI: 0.1093/biostatistics/kxj020. [PubMed: 16443924]
11. Arnheim N, Strange C, Erlich H. Use of pooled DNA samples to detect linkage disequilibrium of polymorphic restriction fragments and human disease: studies of the HLA class II loci. *Proceedings of the National Academy of Sciences of the United States of America*. 1985; 82:6970–6974. [PubMed: 2995996]
12. Germer S, Holland MJ, Higuchi R. High-throughput SNP allele-frequency determination in pooled DNA samples by kinetic PCR. *Genome Research*. 2000; 10:258–266. [PubMed: 10673283]
13. Werner M, Synch M, Herbon N, Illig T, König IR, Wjst M. Large-scale determination of SNP allele frequencies in DNA pools using MALDI-TOF mass spectrometry. *Human Mutation*. 2002; 20:57–64.10.1002/humu.10094 [PubMed: 12112658]
14. Pacek P, Sajantila A, Syvanen AC. Determination of allele frequencies at loci with length polymorphism by quantitative analysis of DNA amplified from pooled samples. *PCR Methods Application*. 1993; 2:313–317.10.1101/gr.2.4.313
15. Sham P, Bader JS, Craig I, O'Donovan M, Owen M. DNA Pooling: a tool for large-scale association studies. *Nature Review Genetics*. 2002; 3:862–871.
16. Shaw SH, Carrasquillo MM, Kashuk C, Puffenberger EG, Chakravarti A. Allele frequency distributions in pooled DNA samples: applications to mapping complex disease genes. *Genome Research*. 1998; 8:111–123.10.1101/gr.8.2.111 [PubMed: 9477339]
17. Barratt BJ, Payne F, Rance HE, Nutland S, Todd JA, Clayton DG. Identification of the sources of error in allele frequency estimations from pooled DNA indicates an optimal experimental design. *Annals of Human Genetics*. 2002; 66:393–405.10.1046/j.1469-1809.2002.00125.x [PubMed: 12485472]
18. Hoogendoorn B, Norton N, Kirov G, Williams N, Hamshire ML, Spurlock G, Austin J, Stephens MK, Buckland PR, Owen MJ, O'Donovan MC. Cheap, accurate and rapid allele frequency estimation of single nucleotide polymorphisms by primer extension and DHPLC in DNA pools. *Human Genetics*. 2000; 107:488–493.10.1007/s004390000397 [PubMed: 11140947]
19. Breen G, Harold D, Ralston S, Shaw D, St Clair D. Determining SNP allele frequencies in DNA pools. *Biotechniques*. 2000; 28:464–466. [PubMed: 10723558]
20. Giordano M, Mellai M, Hoogendoorn B, Momigliano-Richiardi P. Determination of SNP allele frequencies in pooled DNAs by primer extension genotyping and denaturing high-performance liquid chromatography. *Journal of Biochemical and Biophysical Methods*. 2001; 47:101–110.10.1016/S0165-022X(00)00156-1 [PubMed: 11179766]
21. Kosaki K, Yoshihashi H, Ohashi Y, Kosaki R, Suzuki T, Matsuo N. Fluorescence-based DHPLC for allelic quantification by single-nucleotide primer extension. *Journal of Biochemical and Biophysical Methods*. 2001; 47:111–119.10.1016/S0165-022X(00)00157-3 [PubMed: 11179767]
22. Ross P, Hall L, Haff LA. Quantitative approach to single-nucleotide polymorphism analysis using MALDI-TOF mass spectrometry. *Biotechniques*. 2000; 29:620–629. [PubMed: 10997276]
23. Sasaki T, Tahira T, Suzuki A, Higasa K, Kukita Y, Baba S, Hayashi K. Precise estimation of allele frequencies of single-nucleotide polymorphisms by a quantitative SSCP analysis of pooled DNA. *American Journal of Human Genetics*. 2001; 68:214–218.10.1086/316928 [PubMed: 11083945]
24. Wactawski-Wende J, Schisterman EF, Hovey KM, Howards PP, Browne RW, Hediger M, Liu A, Trevisan M. BioCycle Study Group. BioCycle study: design of the longitudinal study of the oxidative stress and hormone variation during the menstrual cycle. *Paediatric and Perinatal Epidemiology*. 2009; 23:171–184.10.1111/j.1365-3016.2008.00985.x [PubMed: 19159403]
25. Gaskins AJ, Mumford SL, Zhang C, Wactawski-Wende J, Hovey KM, Whitcomb BW, Howards PP, Perkins NJ, Yeung E, Schisterman EF. BioCycle Study Group. Effect of daily fiber intake on reproductive function: the BioCycle Study. *American Journal of Clinical Nutrition*. 2009; 90:1061–1069.10.3945/ajcn.2009.27990 [PubMed: 19692496]

26. Costa LG, Richter RJ, Li WF, Cole T, Guizzetti M, Furlong CE. Paraoxonase (PON 1) as a biomarker of susceptibility for organophosphate toxicity. *Biomarkers*. 2003; 8:1–12. [10.1080/1354750021014831](https://doi.org/10.1080/1354750021014831) [PubMed: 12519632]

Figure 1.
Illustration of obtaining minor allele frequencies using PCR.

Figure 2.

Power functions of five different scenarios. All solid lines are for pooled estimators with 1000 individuals.

The solid line with the least power uses 2 repeated measurements of pools on each level of disease and environments. The solid line above that uses 3 repeated measurements, and the solid line with the most power uses 1000 repeated measurements. The dotted line with the least power observes 500 individual measurements, and the dotted line with the most power uses 1000 individual measurements

Figure 3. Power functions of four different scenarios. All solid lines are for pooled estimators, with 1000 individuals and dotted lines are for the traditional estimators with 500 individual measurements. The solid and dotted lines with the most power uses `_c D_1`, and the solid and dotted lines with the least power uses `_c D 1`.

Table 1

The Setup for the Original Non Pooled Data

D = 0				D = 1			
E = 0		E = 1		E = 0		E = 1	
r_{00}		r_{01}		r_{10}		r_{11}	
G = 0	G = 1	G = 2	G = 0	G = 1	G = 2	G = 0	G = 1
n_{000}	n_{001}	n_{002}	n_{010}	n_{011}	n_{012}	n_{100}	n_{101}
						n_{102}	n_{110}
							n_{111}
							n_{112}

Table 2

The BioCycle Data

D = 0 (No anovulatory cycles)					
E = 0 (A fiber intake below the mean)			E = 1 (A fiber intake above the mean)		
$r_{00} = 173$			$r_{01} = 50$		
G = 0	G = 1	G = 2	G = 0	G = 1	G = 2
$n_{000} = 66$	$n_{001} = 81$	$n_{002} = 26$	$n_{010} = 23$	$n_{011} = 22$	$n_{012} = 5$
D = 1 (At least one anovulatory cycle)					
E = 0 (A fiber intake below the mean)			E = 1 (A fiber intake above the mean)		
$r_{10} = 21$			$r_{11} = 13$		
G = 0	G = 1	G = 2	G = 0	G = 1	G = 2
$n_{100} = 7$	$n_{101} = 10$	$n_{102} = 4$	$n_{110} = 6$	$n_{111} = 7$	$n_{112} = 0$

Table 3

A Random Sample with Replacement of the BioCycle Data

D = 0 (No anovulatory cycles)					
E = 0 (A fiber intake below the mean)		E = 1 (A fiber intake above the mean)			
$r_{00}= 675$		$r_{01}= 184$			
G = 0	G = 0	G = 0	G = 0	G = 0	G = 0
$n_{000}= 270$	$n_{001}= 312$	$n_{002}= 93$	$n_{010}= 85$	$n_{011}= 80$	$n_{012}= 19$
D = 1 (At least one anovulatory cycle)					
E = 0 (A fiber intake below the mean)		E = 1 (A fiber intake above the mean)			
$r_{10}= 81$		$r_{11}= 60$			
G = 0	G = 1	G = 2	G = 0	G = 1	G = 2
$n_{100}= 21$	$n_{101}= 41$	$n_{102}= 19$	$n_{110}= 27$	$n_{111}= 33$	$n_{112}= 0$

Table 4
Simulated Values of $\hat{\Delta}C_{tDE(k)}$ for the Random Sample with Replacement

Replicate	Pool			
	D = 1, E = 1	D = 1, E = 0	D = 0, E = 1	D = 0, E = 0
1	1.2105	0.0869	0.9776	0.8056
2	1.7081	0.4539	0.9411	1.0492
3	1.1178	-0.0636	1.0192	0.5954

Table 5
The Mean of the Sample Variances

$\beta_{G \times E}$	Mean $\hat{\beta}_{G \times E}$ (95% CI)	Empirical Variance of $\hat{\beta}_{G \times E}$	Mean $\left(\widehat{\text{var}}(\hat{\beta}_{G \times E}) \right)$ (95% CI)	Coverage Probability (level = .05) (95% CI)
-.5	-0.5017 (-0.594, -0.4940)	0.0764	0.0759 (0.0764, 0.0755)	0.9416 (0.9351, 0.9481)
-.3	-0.2996 (-0.3072, -0.2920)	0.0757	0.0757 (0.0752, 0.0761)	0.9432 (0.9367, 0.9496)
0	0.0023 (-0.0052, 0.0098)	0.0732	0.0756 (0.0751, 0.0760)	0.9478 (0.9413, 0.9540)
.3	0.2926 (0.2851, 0.3002)	0.0744	0.0768 (0.0764, 0.0773)	0.9508 (0.9443, 0.9568)
.5	0.4915 (0.4838, 0.4992)	0.0773	0.0785 (0.0781, 0.0790)	0.9470 (0.9405, 0.9532)

From left to right the columns are: The true simulated gene-environment interaction. The mean of the Monte Carlo simulated gene-environment interaction estimators. The empirical variance of the gene-environment interaction estimators in the Monte Carlo simulation. The mean of the Monte Carlo simulated variance estimators. The empirical coverage probability at level .05. The number of Monte Carlo replicates is 5000.

Table 6

Type 1 error

σ^2_{DE}	R	$\Pr(A = 1)$				
		0.1	0.3	0.5	0.7	0.9
0.05	2	0.0492*	0.0668*	0.0686*	0.0642*	0.0556*
	3	0.0482*	0.0510*	0.0550*	0.0518*	0.0498*
	5	0.0550*	0.0552*	0.0520*	0.0464*	0.0478*
	10	0.0508*	0.0460*	0.0494*	0.0500*	0.0480*
0.1	2	0.0576*	0.0816*	0.0834*	0.0738*	0.0596*
	3	0.0476*	0.0572*	0.0664*	0.0648*	0.0506*
	5	0.0468*	0.0498*	0.0510*	0.0558*	0.0460*
	10	0.0502*	0.0446*	0.0496*	0.0578*	0.0476*
0.5	2	0.0956*	0.1080*	0.1068*	0.1082*	0.0826*
	3	0.0586*	0.0782*	0.0824*	0.0748*	0.0648*
	5	0.0506*	0.0528*	0.0650*	0.0570*	0.0504*
	10	0.0444*	0.0492*	0.0516*	0.0442*	0.0480*

* 0.05 is less than the upper 95% confidence interval with 5000 Monte Carlo replicates