Published in final edited form as:

Genet Epidemiol. 2015 July ; 39(5): 334–346. doi:10.1002/gepi.21890.

Deciphering Genome-Environment-Wide Interactions Using Exposed Subjects Only

Lue Ping Zhao^{1,2,*}, Wenhong Fan¹, Gary Goodman^{1,3}, Jerry Radich¹, and Paul Martin¹

¹Fred Hutchinson Cancer Research Center, Seattle, WA

²School of Public Health Sciences, University of Washington, Seattle, WA 98105

³Swedish Medical Center Cancer Institute, Seattle, WA

Abstract

The recent successes of genome-wide association studies (GWAS) have renewed interest in genome-environment-wide interaction studies (GEWIS) to discover genetic factors that modulate penetrance of environmental exposures to human diseases. Indeed, gene-environment interactions (GxE), which have not been emphasized in the GWAS era, could be a source contributing to the missing heritability, a major bottleneck limiting continuing GWAS successes. In this manuscript, we describe a design and analytic strategy to focus on GxE using only exposed subjects, dubbed as e-GEWIS. Operationally, an e-GEWIS analysis is equivalent to a GWAS analysis on exposed subjects only, and it has actually been used in some earlier GWAS without being explicitly identified as such. Through both analytics and simulations, e-GEWIS have been shown better efficiency than the usual cross-product-based analysis of GxE interaction with both cases and controls (cc-GEWIS), and they have comparable efficiency to case-only analysis of GxE (c-GEWIS), with potentially smaller sample sizes. The formalization of e-GEWIS here provides a theoretical basis to legitimize this framework for routine investigation of GxE, for more efficient GxE study designs, and for improvement of reproducibility in replicating GEWIS findings. As an illustration, we apply e-GEWIS to a lung cancer GWAS dataset to perform a GEWIS, focusing on gene and smoking interaction. The e-GEWIS analysis successfully uncovered positive genetic associations on chromosome 15 among current smokers, suggesting a gene-smoking interaction. While this signal was detected earlier, the current finding here serves as a positive control in support of this e-GEWIS strategy.

Keywords

Case-control; Exposed subjects; genome scan; GEWIS; GWAS; GxE

* Correspondence to: Lue Ping Zhao at Fred Hutchinson Cancer Research Center, Division of Public Health Sciences, 1100 Fairview Ave NE, Seattle, WA 98109. lzhao@fhcrc.org.

Authors declare no financial conflict of interest.

Introduction

Following decades of epidemiological and genetic research of complex diseases, it is clear that both nature and nurture play roles in human health [Bix 1999; Butcher and Plomin 2008; Flowers, et al. 2011; Hernandez and Blazer 2006; Jones 1999; Ordovas and Smith 2010; Spires and Hannan 2005]. Besides independent and separate disease associations with environment and host-genetics (typically measured by single nucleotide polymorphisms, or SNPs), their interactions, dubbed as gene-environment interactions (GxE), may also play important roles in many complex diseases. To investigate GxE interactions, Khoury and Wacholder (2009) proposed a framework for genome-environment-wide interaction studies (GEWIS), searching for SNPs involved in GxE interactions on a genome-wide scale, a natural extension to genome-wide association studies (GWAS) [Khoury and Wacholder 2009]. In last few years, the National Institutes of Health (NIH) stimulated the research community to pursue GEWIS by launching an NIH-wide initiative, known as the gene-environment initiative, organizing workshops focusing on GxE [Bookman, et al. 2011; Hutter, et al. 2013; Mechanic, et al. 2012] and injecting new funding supports into GxE studies (<http://grants.nih.gov/grants/guide/pa-files/PAR-13-382.html>).

While conceptually clear, interpretations of GxE interactions in biomedical research literature are discipline-specific. In population-based etiological studies, presence of GxE interaction means that the joint effect of genetic polymorphisms and environmental factors is greater (or less) than the sum of two separate effects [Khoury, et al. 1993]. In disease prevention studies, a GxE interaction means that environment factors, like diets and behaviors, may modify genetic associations with diseases [Song, et al. 2011]; identifications of such environment factors offer means to prevent diseases among high risk individuals. In pharmacogenetics or precision medicine, GxE interaction implies that the treatment (regarded as an environmental exposure) may have specific desirable/undesirable responses depending on patients' genetic polymorphisms [Superko, et al. 2012], and identification of such genetic polymorphisms would allow physicians to personalize treatment options.

Besides the conceptual variations of GxE interactions, there are two distinct scales used to quantify GxE interactions: additive or multiplicative scale [Clayton 2009; Rothman, et al. 1980]. Traditionally, the multiplicative scale is adopted in population-based studies, mostly because odds ratios (OR), typically used to measure disease association, are naturally parameterized in the logistic regression model and computation and interpretation are statistically tractable [Breslow and Day 1980]. To be succinct, let $OR(G,-)$, $OR(-,E)$ and $OR(G,E)$ represent odds ratios of G alone, E alone and their joint effect, respectively. The presence of a multiplicative GxE means that $OR(G,E)/[OR(G,*)OR(*,E)] > 1$ (or < 1) corresponds to a synergistic interaction (or antagonistic interaction). On the other hand, the recent development of causal models for human diseases suggests that the additive scale is probably more appropriate for characterizing GxE interactions [Rothman and Greenland 2005]. Using the same odds ratio representations, the presence of an additive GxE interaction implies that $[OR(G,E)-1]-[OR(G,-)-1]-[OR(-,E)-1] > 0$ (or < 0) corresponds to a synergistic interaction (or antagonistic interaction). Quantitatively, these two interactions are not equivalent, i.e., the absence of additive interaction does not imply the absence of multiplicative interaction, vice versa. Regardless of the scale used for quantifying GxE, one

can use the logistic regression model with a flexible parameterization to capture main effects and their interactions. The primary difference between two scales lies in their interpretation. Without complicating our presentation, we use the logistic regression to quantify these three ORs, and to interpret GxE on a multiplicative scale in general, noting the differences when we explore the additive GxE interpretation.

Since few studies are designed to optimize GEWIS at this time, a common practice is to examine GxE interactions using available GWAS that were designed to discover main disease associations with SNPs, using various analytic methods. Several analytic methods and related software tools are published and reviewed elsewhere [Bookman, et al. 2011; Hutter, et al. 2013; Mechanic, et al. 2012; Thomas 2010a; Thomas 2010b]. On the multiplicative scale, a default approach is to quantify GxE interaction as a cross-product of genetic polymorphism with environmental factor in the logistic regression [Thomas 2010b]. Unfortunately, such a GxE analysis has been found to have limited power in typical GWAS [Clayton 2012; Gauderman, et al. 2010; Thomas 2010a; Thomas 2010b]. A desire to improve analytic power has been a major impetus to develop new approaches for discovering GxE interactions. One class of approaches is to exploit the “likely” independence between genetic polymorphisms and environmental factors in general populations (or controls) [Albert, et al. 2001; Chatterjee and Carroll 2005; Chatterjee, et al. 2005; Cornelis, et al. 2012; Helbig, et al. 2012; Piegorsch, et al. 1994; Spinka, et al. 2005]. Under this “independence” assumption, retrospective likelihood methods have been proposed, and they have shown substantial analytic power gain over the usual interaction analysis [Chatterjee and Carroll 2005; Chatterjee, et al. 2005; Cheng and Lin 2005; Spinka, et al. 2005]. In fact, under this independence assumption, all one needs are cases for testing the presence of GxE interaction via examining dependence between G and E among cases [Albert, et al. 2001; Cornelis, et al. 2012; Helbig, et al. 2012; Piegorsch, et al. 1994]. Because of its prominent role in GEWIS, it is referred to as case-only GEWIS or c-GEWIS hereafter. The main concern with this general approach is that the independence assumption is not readily verifiable with empirical data, and any positive discovery from c-GEWIS could be attributed to either genuine GxE or to violation of the independence assumption. Another class of methods use a multistage strategy with combinations of tests at various stages [Breslow and Chatterjee 1999; Gauderman, et al. 2010; Gauderman, et al. 2013; Hsu, et al. 2012; Kooperberg and LeBlanc 2008; Kraft, et al. 2007; Murcray, et al. 2011; Murcray, et al. 2009]. For example, Kraft et al (2007) described several joint test statistics, and proposed to use them according to the analytic expectations. Kooperberg and LeBlanc (2008) advocated detecting main effects first and then testing for interactions at the second stage [Kooperberg and LeBlanc 2008]. Hsu et al (2012) described a cocktail procedure with multiple tests [Hsu, et al. 2012]. Gauderman et al (2013) utilizes both marginal associations and correlation between G and E to improve the analytic power of discovering GxE [Gauderman, et al. 2013].

Sharing the same motivation, this manuscript describes a seemingly intuitive approach that maximizes the analytic power of deciphering GxE interaction, without relying on the unverifiable G by E independence assumption. Specifically, assuming a binary exposure under consideration, an efficient GEWIS scans the genetic association genome-wide, using only “exposed subjects”, i.e., performing a GWAS analysis on exposed subjects. This

approach is called e-GEWIS for its exclusive use of exposed subjects. For simplicity, consider a simplest case-control study for GxE where data can be organized as Table 1a. The e-GEWIS uses c/C and d/D to assess GxE. In contrast, c-GEWIS uses only (a, b, c, d). The traditional case-control analysis uses all of the data and is thus referred to as cc-GEWIS hereafter. In the remainder of this manuscript, we provide biological and statistical motivations for e-GEWIS, describe an analytic procedure, and show the power comparison with the c-GEWIS and also cc-GEWIS through a simulation study. Finally, we apply e-GEWIS to discover genetic polymorphisms that may interact with status of current smoking as an illustration.

Methods

The GxE Problem

Suppose that a case-control GWAS with n_1 cases and n_0 controls is available for GEWIS. Let d_i take value 1 or 0 for the i th subject to be a case or control, respectively. On the i th subject, the GWAS scans his/her genome with J SNPs ($j=1,2,\dots,J$), in which the genotype for the j th SNP is denoted as g_{ij} ($=a_{ij1}:a_{ij2}$), where each allele takes value of either 0 or 1 for major or minor allele, respectively. In addition, from the i th subject, the study gathers an array of covariates and designates one of them as a primary exposure variable E_i and remaining covariates as possible confounders denoted as x_i . The analytic objective of GEWIS is to assess which SNP interacts with E_i in their joint penetrance to the phenotype d_i . Without losing generality, we consider here a binary exposure ($E_i=0$ and 1 for a non-exposed and exposed subject, respectively), and a binary genotypic variation ($g_{ij}=0$ or 1 for the absence or presence of the minor allele, respectively). Assume that the disease probability given g_{ij} and E_i follows the logistic penetrance model:

$$\text{logit } \Pr(d_i=1|g_{ij}, E_i, x_i) = \alpha'_j x_i + \beta_j E_i + \gamma_j g_{ij} + \delta_j E_i g_{ij}, \quad (1)$$

where regression coefficient β_j quantifies the disease association with the exposure, γ_j with the j th SNP, δ_j with the multiplicative GxE, and the term $\alpha'_j x_i$ is a linear combination of all covariates plus intercept, the last of which is determined by baseline incidence, sampling fractions of cases and controls, and all confounding effects. The logistic regression is commonly used, primarily because of its desirable statistical properties [Breslow and Day 1980]. Also, exponentiations of regression coefficients correspond to odds ratios, which approximate relative risks for rare diseases [Prentice and Pyke 1979]. As noted earlier, the cross-product term in the above logistic regression (1) leads to a natural interpretation of the multiplicative GxE with the single parameter δ_j . Specifically, the absence of the multiplicative GxE implies the null hypothesis $H_0: \delta_j=0$, i.e., $\exp(\beta_j+\gamma_j) = \exp(\beta_j)\exp(\gamma_j)$, or $\text{OR}(G,E) = \text{OR}(G,-)\text{OR}(-,E)$ [Breslow and Day 1980].

It is also important to note that the parameterization with the cross-product in the above logistic regression (1) is sufficiently flexible with three parameters ($\beta_j, \gamma_j, \delta_j$). One can transform these coefficients to odds ratios $\text{OR}(-,E) = \exp(\beta_j)$, $\text{OR}(G,-) = \exp(\gamma_j)$ and $\text{OR}(G,E) = \exp(\beta_j + \gamma_j + \delta_j)$. To test the presence of the additive GxE, one can examine the null hypothesis: $H_0: [\exp(\beta_j + \gamma_j + \delta_j) - 1] = [\exp(\beta_j) - 1] + [\exp(\gamma_j) - 1]$, and interpret result as an

additive GxE interaction [Rothman and Greenland 2005]. Hence, the choice of the above logistic regression does not negate the statistical and biological validity for quantifying additive GxE.

Motivations

Genetic polymorphisms and environmental factors in typical GEWIS are inherently different from both biological and statistical perspectives, even though their symbolic representations by GxE or quantitative representations (E_i and g_{ij}) in the logistic regression model (1) seem to be symmetric. An environmental factor chosen in typical GEWIS is likely to have a strong association with the disease phenotype, for example, exposures to asbestos, cigarettes, fatty diets, known chemical toxins, or therapeutic drugs. As expected, such exposure E may act upon multiple biological networks, each of which involves multiple genes. Their etiological associations with phenotypes likely have been observed repeatedly across multiple populations. In contrast, genetic factors in typical GEWIS are a collection of SNPs (or even structural polymorphisms) with unknown associations with the phenotype since they are randomly chosen to cover genome variations and are only known to be polymorphic in general population [Wang, et al. 1998]. As expected, most SNPs do not associate with the phenotype, and discovering those exceptions is the focus of GWAS, as demonstrated by the identification of disease-associated nucleotide polymorphisms in the GWAS catalog database [Hindorff, et al. 2009; Li, et al. 2012; Welter, et al. 2014]. The number of relatively fewer SNPs that associate with the phenotype through GxE interactions may be even smaller. In the context of the logistic model (1), a realistic expectation is that the environment factor E has a substantial association with the phenotype ($\beta_j \neq 0$), while most SNPs do not associate with the phenotype ($\gamma_j = 0$). The primary interest of GEWIS is to discover those few SNPs that are involved in GxE interactions. In other words, the null hypothesis of GEWIS is to test the null hypothesis $H_0: \delta_j = 0$ on a multiplicative scale.

Statistically speaking, genetic polymorphism G and environmental exposure E should be dealt differently, despite the fact that the logistic model (1) treats G and E in a symmetrical fashion. Let us consider a simple scenario useful for studying GxE with binary SNP variation (such as presence of minor allele under the dominant penetrance mode) and binary exposure in a case-control study, data from which can be organized in a 2x2x2 frequency table (Table 1a). Under the logistic model (1), one can write down four odds to quantify GxE association pattern (Table 1b). Among all possible association patterns, a likely GxE pattern is that the main genetic effect, defined via the logistic regression model (1), is absent ($\gamma_j = 0$), resulting in a probable GxE association pattern (Table 1c). In such a case, those in the unexposed group do not contribute information to the GxE interaction. Intuitively, therefore GEWIS should center the estimation and inference of the parameter (δ_j) on exposed subjects only (contrasting the ratio c/C with d/D in Table 1a), which leads to the proposal of the e-GEWIS analysis. When the assumption of zero main genetic association ($\gamma_j = 0$) is violated, the proposed e-GEWIS analysis actually tests the combined main genetic effect and GxE interaction. In other words, the null hypothesis becomes $H_0: \gamma_j + \delta_j = 0$. Collectively, both biological and statistical insights suggest that an effective analytic strategy of e-GEWIS is to use only exposed subjects for assessing GxE interaction, at the expense of reducing analytic power of discovering main genetic associations. Note that the

main effect of the SNP (γ_j) can be confused with the marginal association of SNP, the latter of which is resulted from integrating all known and unknown etiologic and confounding factors. Based upon the biological motivation, e-GEWIS is to test the absence of GxE with the null hypothesis $H_0: \delta_j=0$, if the main effect of SNP is absent ($\gamma_j=0$). This assumption could be violated (see the next section for more discussion on consequences due to violating this assumption).

An e-GEWIS Approach

Under the assumed logistic regression model (1), an e-GEWIS analysis assesses genetic association with the disease phenotype among exposed subjects only. Specifically, given an exposure variable within a GWAS, one selects a sub-population of subjects who are exposed to the designated variable. On this sub-population, e-GEWIS performs a SNP-by-SNP association scan throughout the entire genome, or equivalently, GWAS analysis on exposed subjects. In light of earlier discussions, the genetic associations detected by e-GEWIS are composite of the main effect and interaction ($\gamma_j + \delta_j$) (bottom row of Table 1b). Under the assumption that the main genetic associations are absent for most SNPs, the detected genetic association by e-GEWIS likely resulted from pure GxE interactions. The e-GEWIS may experience inflated false-positive discoveries, to test the null hypothesis $H_0: \delta_j=0$, if the main genetic association is present ($\gamma_j \neq 0$). Of course, “falsely discovered SNPs in GxE” is actually the main genetic effect, and such “false discoveries” would be serendipity of e-GEWIS!

Computationally, e-GEWIS uses the same likelihood procedure as the GWAS [Cox 1986]. Briefly, assuming the independence across n subjects, one derives a log-likelihood function $L(\alpha_j, \beta_j, \gamma_j, \delta_j)$ with the summation of logarithmic Bernoulli probability $\log[\Pr(d_i/g_{ij}, E_i=1)]$ over n exposed subjects, in which each probability function is defined by the logistic regression above (1). A maximum likelihood estimate is obtained via maximizing the log-likelihood function. To obtain this estimate, one typically uses the Newton-Raphson method, iteratively obtaining the estimate [Zhao 1989]. The estimate is known to have an asymptotic normal distribution, which can be used to construct test-statistics for a specific hypothesis [Breslow and Day 1980]. Under the null hypothesis $H_0: (\gamma_j + \delta_j=0)$, one can construct test statistics $T_{\gamma+\delta}$ based on log likelihood function, score function or point estimates [Cox 1986]. Estimated test statistics can be converted to p-values, which can be used to account for multiple comparisons in the genome-wide scan [Bookman, et al. 2011; Hutter, et al. 2013; Liu, et al. 2012; Mechanic, et al. 2012; Thomas 2010a].

Simulation-Based Comparisons

In recent years, c-GEWIS has been promoted as a promising approach for GEWIS, so a direct comparison between e-GEWIS and c-GEWIS is warranted. At a conceptual level, e-GEWIS and c-GEWIS share the same analytic objective, i.e., to discover GxE interaction. In addition, both e-GEWIS and c-GEWIS use only subsets of case-control data, which should lead to cost-saving on genotyping without substantially sacrificing analytic power. However, these two approaches have several noticeable differences, which we will detail in the Discussion section. Here we focus on comparing their analytic powers. To place this comparison in a familiar context, we include the cross-product-based GxE test with

complete case-control data, i.e., cc-GEWIS. The power comparison of e-GEWIS, c-GEWIS and cc-GEWIS is made through simulation studies.

This simulation study first considers a simplest case-control setup on GxE (Table 1) and then a typical case-control GEWIS, with 2,000 cases and 2,000 controls. In the simple case-control setup, the study simulates a SNP (G) with a given minor allelic frequency (MAF) and an exposure (E) with certain prevalence. All binary random variables are simulated by Bernoulli process. Using the logistic regression model (1) as the penetrance, we compute the disease probability $\text{Pr}(d_i = 1 | g_{ij}, E_i)$, with pre-specified regression parameters $\alpha = -5$, $\gamma = 0$, β and δ , in which last two parameters are directly associated with powers and vary from $\log(1)$ to $\log(2)$. Based on the computed disease probabilities, we simulate binary disease status for every subject in the population of 500,000 subjects, which yields approximately 3000 cases and remainders controls. From the simulated population, we randomly sample 2000 cases and 2000 controls, with G_i , E_i and d_i , $i = 1, 2, \dots, 4000$. On the simulated case-control data, we perform e-GEWIS, c-GEWIS and cc-GEWIS analyses, as a replicate. For each set of chosen parameters, we repeat the above simulation 1000 times. Despite its simplicity, this simulation scenario captures the essence of GxE analysis, but it has several major limitations, such as ignoring variations of MAFs, presence of linkage-disequilibrium (LD), indirect causal associations or presence of confounding variables. All of these factors could impact on performances of GEWIS analysis.

To augment the above simple setup, we align the second simulation scenario with the illustrative example on lung cancer GWAS (to be described below), with the focus on an investigation of gene by smoking interaction. For the target study population, the gender distribution is balanced with 50% male and 50% female. The male smoking prevalence is set at 25%, and the female smoking prevalence rate is set at 10%, based on which smoking status, stratifying over gender, is simulated by Bernoulli process. Participants tend to be elder subjects, with age centering around 60 with the minimum age truncated at 18, and they are simulated with truncated Gaussian process. In human genome, distributions of SNP genotypes are complicated, including Hardy-Weinberg disequilibrium (HWD) within SNP loci, or LD between SNP loci, or other structural variations. To retain the integrity of human genome, we selected consecutive SNPs from chromosome 6 from 11,508 subjects involved in an NCI lung cancer study (dbGaP: http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000336.v1.p1) as the population basis, and randomly chose their chromosome-specific SNP genotypes, with replacement, as simulated genotypes. In this simulation, the number of SNPs was limited to 100 in order to facilitate computational feasibility. While power estimates in the simulation studies with more SNPs are diluted by correction for multiple comparisons, the relative comparison of these three analytic strategies remains valid and informative. On each simulated subject with covariates and genotypes, we compute the corresponding disease probability by the logistic regression model (1), with designated intercept (α) and log odds ratios for gender, age, smoking, SNP, and smoking-SNP interaction. Based on the computed disease probability, we simulate the disease phenotype using the Bernoulli process.

In the second step of the simulation setup, we assume that smoking is a major risk factor, with an odds ratio of 2.0. The disease incidence increases with age linearly on the logit

scale. Gender, after adjusting for smoking and age, does not associate with disease incidence. We assume that one SNP, with the MAF exceeding 10%, is randomly chosen to be a causal SNP with a strong interaction with smoking; the corresponding GxE interaction parameter δ_i ranges from $\log(1)$ to $\log(4)$ for each scenario. For each individual scenario, we used 1,000 replicates. The simulation plan is largely consistent to the above description. Briefly, on each replicate, we first generated a targeted study population with simulated covariates (gender, age, and smoking), randomly drawn 100 SNP genotypes, a randomly chosen SNP locus to interact with smoking, and simulated disease phenotype. The resulting study population typically had approximately 3000 disease subjects, with the remainder representing non-diseased subjects. By randomly sampling the study population without replacement, we obtained 2,000 diseased subjects as cases and 2,000 non-diseased subjects as controls. Once a case-control study data set was simulated, we performed cc-GEWIS, c-GEWIS and e-GEWIS to scan the short segment of the genome. To accommodate some associations through LD, we assume that a significant association with any SNP, within the proximity of 5,000 base pairs to the true SNP locus, is a true-positive discovery.

Considering realistic situations where true actual casual SNPs may or may not be directly genotyped on the chip, we conducted simulations of genotype data with and without the true “causal SNP” included. The former analysis is known as the direct association analysis, while the latter is the indirect analysis. The power of detecting GxE signals by indirect analysis is expected to be much lower, since signals are captured only through adjacent SNP markers that are in LD with the “causal SNP.”

Given the known GxE SNP locus within the simulation, we can determine if discovered SNP loci, if any, include the true GxE interaction. We define true discovery power (TDP) as the percentage of true GxE interactions that are discovered by GEWIS. This may be considered as a conditional probability being discovered given the truth. On the other hand, any GEWIS discovery can be false. Across replicates, we computed the total number of discoveries and number of false discoveries, and estimate the false discovery rate (FDR). We use TDP and FDR to compare three approaches cc-GEWIS, c-GEWIS and e-GEWIS, along with their corresponding sample sizes.

An Illustrative Example

To illustrate, we apply e-GEWIS, c-GEWIS and cc-GEWIS to a lung cancer GWAS. While several lung cancer GWAS have deposited their data to dbGAP, only two GWAS have deposited pertinent smoking history to dbGaP for public availability. After going through a formal application to our local IRB and the dbGaP steering committee, we were granted access to these two lung cancer GWAS (study id=phs000093.v2.p2). These two data sets were collected by Environment and Genetics in Lung Cancer Etiology (EAGLE) and Prostate, Lung, Colon, Ovary Screening Trial (PLCO)[Landi, et al. 2009]. The two studies have rather different designs, and such differences would negatively affect GxE analysis of the pooled GWAS data sets. For this example, we use EAGLE data, which is genotyped by Illumina 610QUAD. Table 2 lists the distribution of all included covariates: gender, age, smoking status (never, former and current), and pack years. Here we focus on interaction of genes with smoking status (current smoker or not).

Results

Simulation-Based Comparisons with a Single SNP in GxE

The first portion of this simulation study centers on power comparison of three methods for testing GxE with a single SNP. Figure 1a shows the power curves of e-GEWIS, c-GEWIS and cc-GEWIS, when fixing other parameters ($\alpha = -5$, $\gamma = 0$, $\beta = \log(1.5)$, MAF=20%, exposure prevalence=50%). It appears that e-GEWIS is slightly more powerful than c-GEWIS, and both have superior powers to cc-GEWIS. Note that the average sample sizes by e-GEWIS range from about 2200 to 2400, i.e., 10% larger sample size. On balance, e-GEWIS and c-GEWIS have comparable power in this scenario. Since sample size, used by e-GEWIS, is closely tied with the exposure prevalence, we then fix interaction odds ratio, $OR(GxE)=1.3$, and then evaluate powers across exposure prevalence from 10% to 90% (Figure 1b). Again, both e-GEWIS and c-GEWIS are uniformly more powerful than cc-GEWIS. It appears that under the current scenario, c-GEWIS with fixed sample size of 2000 cases is more powerful than e-GEWIS when exposure prevalence is below, e.g., 40%. On the other hand, e-GEWIS becomes more powerful than c-GEWIS when the exposure prevalence is larger than 60%. As expected, associated sample sizes used by e-GEWIS increase with the exposure prevalence. Balancing between power and sample changes, it is reasonable to assert that both e-GEWIS and c-GEWIS have comparable power.

With the interaction odds ratio fixed at $OR(GxE)=1.5$ and exposure prevalence fixed at 50%, we examined whether or not MAF of SNP may affect computed powers by profiling over MAF from 5% to 50% (Figure 1c). It appears that both e-GEWIS and c-GEWIS, with comparable powers, have greater power than cc-GEWIS, and their differences diminish with MAF approaching 50%. Another relevant feature in e-GEWIS is the magnitude of risk association with the exposure. To examine its impact on e-GEWIS and its comparison with other designs, we profile the main exposure effect with β ranging from $\log(1)$ to $\log(2)$, and compute associated powers attained by e-GEWIS, c-GEWIS and cc-GEWIS (Figure 1d). It is interesting to note that c-GEWIS appears to become less powerful, as the disease association with the exposure increases, and its power approaches that of cc-GEWIS. Similarly, if a genetic main effect association with disease deviates from the null, c-GEWIS appears to lose powers (not shown), probably because cases are caused by not only by the GxE interaction but also by the genetic factor alone. In contrast, the power of e-GEWIS increases, partly because the test statistic captures both the GxE interaction and the main genetic association.

Simulation-Based Comparisons with Multiple SNPs in GxE

Recognizing limitations mentioned above, we now imitate a typical GWAS per the simulation plan described above, and estimate true discovery power (TDP) and false discovery rate (FDR) associated with cc-GEWIS, c-GEWIS and e-GEWIS under each scenario with varying magnitudes of GxE interaction. Throughout all simulations, cc-GEWIS requires the full case-control study, hence using a total of 2000 case and 2000 control subjects. c-GEWIS uses cases only, and thus its sample size is 2000 cases. In contrast, numbers of exposed subjects used by e-GEWIS are influenced by exposure-related factors documented above. In the current simulation study, it is noted that sample sizes are

linearly increasing with the increasing odds ratios of GxE, approximately from about 830 subjects under the null interaction, to about 995 subjects with OR_{GxE} of 2, and to about 1,263 subjects when OR_{GxE} approaches 4. Given the general expectation on OR_{GxE} around 2, it is reasonable to conclude that e-GEWIS, in the current simulation configuration, utilizes about 25% of the total sample size 4000 used by cc-GEWIS and about 50% of the sample size required by c-GEWIS.

With nearly 50% of the total sample sizes required by c-GEWIS, a natural question is whether e-GEWIS loses any power of detecting GxE with fewer samples. Figure 2 shows estimated TPD when the smoking association is relatively strong with $OR_E=2$. TPD curves are estimated for the direct association scan (left three curves) and for the indirect association scan (right three curves). In general, TPD curves for c-GEWIS and e-GEWIS are comparable and consistently higher than that for cc-GEWIS. Note that TPD estimates under the scenario with $OR_{GxE}=1$ are largely below 0.05, the pre-set false-positive error rate. Equally important are estimated FDRs across all different scenarios. Under the null hypothesis $OR_{GxE}=1$, the numbers of total discoveries are relatively small, and hence their FDR estimates are not reliable. Across alternative OR_{GxE} values, estimated FDR values vary from 0.30 to 0.55 (not shown). Collectively, simulation results under the null hypothesis or those from unlinked SNPs support the validity of the estimation and testing procedures. As expected, when the number of SNPs increases and type I error is controlled with a Bonferroni's correction, the corresponding power curves will shift towards the right, while relative differences of power curves among e-GEWIS, c-GEWIS and cc-GEWIS remain.

Interaction of Genes with Smoking Status in Lung Cancer

Smoking has long been established to be a key risk factor for lung cancer. In our example, we use a lung cancer GWAS, known as EAGEL, including a total of 1,992 controls and 1,820 cases (Table 2). EAGLE has a relatively small fraction of non-smokers (15%). Distributions of both cases and controls with respect to their gender, age, pack year of smoking and smoking status are listed in Table 2, and are skewed towards high-risk populations. In this GEWIS, our primary focus is to identify gene interactions with smoking. To minimize misclassification error on exposure status, we consider only current smokers as exposed subjects, leading to a total of 485 controls and 904 cases. Throughout all analyses by cc-GEWIS and e-GEWIS, gender, age, pack year, study sites and first ten principle components (based on GWAS data) are adjusted, while c-GEWIS is not amenable to covariate adjustment. This illustrative e-GEWIS analysis focuses on additive allelic associations with lung cancer SNP by SNP among current smokers, using PLINK, which has been widely used to scan genome-wide associations [Purcell, et al. 2007]. For c-GEWIS, we again use PLINK, to assess allelic associations with current smoking status (as outcome) among only lung cancer cases. Lastly, cc-GEWIS analysis utilizes all available cases and controls and assesses interactions as cross-products of smoking status and SNP genotypes (0, 1 and 2). The scan analysis with this interaction is performed again by PLINK. Figure 3 shows three Manhattan plots from e-GEWIS, c-GEWIS and cc-GEWIS. In the plot, we denote two threshold values, 10^{-6} (red line) and 10^{-5} (blue line). By e-GEWIS, the focus of this example, one can visually observe a major hit on chromosome 15, while two modest signals are observed on chromosomes 10 and 21. In contrast, c-GEWIS does not show any

signal on chromosome 15, but shows one consistent signal on chromosome 10 and another suggestive signal on chromosome 4. On the third panel, cc-GEWIS does not seem to indicate any major signal on chromosome 15, except for two sporadic signals on chromosome 16 and 20.

With a threshold value of 10^{-7} , we have identified three SNPs on chromosome 15, and further, with a more liberal threshold value of 10^{-5} , we select the 4 additional SNPs on chromosome 15, in addition to a few SNPs on chromosomes 10 and 21 that will not be discussed further. Table 3 lists estimated log odds ratios, standard errors, Z-scores and p-values, in addition to basic SNP annotations (rs number, alleles, MAFs). Seven SNPs on chromosome 15 have comparable MAFs with similar association profiles, mostly because these SNPs are in high LD (not shown). On the multiplicative OR scale, the signal on chromosome 15 is synergistic with the multiplicative log odds ratios around 0.39 to 0.54, with corresponding p-values less than 10^{-5} .

In light of available GWAS data, we use the full case-control data to characterize this GxE hit, in which the current smoker is coded as 1 and the never or former smoker as 0. Admittedly, such grouping of never and former smokers may lead to very specific interpretation on the smoking-associated effect, which warrants careful interpretation. For illustration, we control the same set of covariates, and estimate main effects (smoking, SNP) and their interactions as log odds ratios in the column 2-4 of Table 4. Based on estimated log odds ratios, all multiplicative GxE interactions remain to have significant associations with varying magnitudes of log odds ratios and associated p-values.

To facilitate a scale-independent interpretation of GxE interaction, we transform log odds ratios to odds ratios for main smoking effect, main SNP effect and their joint effect (columns 5-7 of Table 4). Specifically, the signal on chromosome 15 indicates that the smoking alone, in the absence of the minor SNP allele, appear to have modest risk from 1.08 to 1.25. For non-current smoker, the minor allele confirms odds ratios from 1.17 to 1.30. For those who carry the minor allele and are still smoking, the risk of lung cancer jumps to 1.74 to 1.90. Under either additive or multiplicative assumptions, current smoking status and minor alleles of these SNPs have synergistic interactions.

The analysis of GxE interaction by both cases and controls listed above is essentially equivalent to cc-GEWIS, and corresponding estimates can thus be used to explain why cc-GEWIS fails to achieve the genomewide significance. Actually, the corresponding highest p-value to the second SNP (rs1051730) is computed as 0.0014 (from Z-score of $2.91=0.32$ (coef) / 0.11 (SE), or $-\log_{10}(p\text{-value})=2.44$). Even though such a signal is “significant” at 1% level, its significance is overwhelmed by “statistical noises” with multiple comparisons. From this specific comparison, one can appreciate why use of powerful analytic design and analytic strategies is important to detect GxE. Now reviewing other SNPs, it is clear that estimated standard errors by e-GEWIS (Table 3) and cc-GEWIS (Table 4) are comparable, but their estimated coefficients are rather different, resulting in substantially different p-values in two analyses (Figure 3). Similarly, we also examined p-values for these selected SNPs computed from c-GEWIS (not shown). Two best p-values from c-GEWIS,

corresponding to rs8034191 and rs1051730, are 0.0192 and 0.0327, respectively. These marginally significant signals are thus buried by multiple comparison correction.

Discussion

Attention to GEWIS is increasing for a variety of purposes: discovering genes that mediate disease associations with environment for personalized health risk assessment, discovering modifiable environment factors in GxE interaction that may help the development of personalized prevention strategies, discovering a new generation of drugs or treatment modalities through patient-specific genetic polymorphisms for precision medicine, and investigating drug-induced toxicity among a subset of patients with certain genetic characteristics to improve efficiency of clinical trials. Despite these laudable goals, the GEWIS field has gone through its peaks and valleys. One major barrier to the success of GEWIS is that the analytic power of GEWIS is much lower than that of GWAS by as much as 75% [Thomas 2010a]. While increasing sample sizes of GEWIS is one plausible solution, the expense, especially in the current funding environment, becomes a major barrier. Another avenue is to come up with new analytic methods, some of which are reviewed above. A complementary avenue is to develop more efficient study methodology, i.e., designing studies that have maximum powers to detect GxE as the primary objective, a complementary but different goal from GWAS. Towards this objective, this manuscript describes an e-GEWIS approach to decipher GxE interactions genome-wide. Conceptually, e-GEWIS may be thought of as a GWAS among only exposed subjects. The key assumption, motivated by both epidemiological and genetic considerations, is that environmental exposure, under consideration for GxE, likely has a strong disease association; most SNPs, on the other hand, do not associate with the disease outcome of interest. Hence, by focusing on exposed subjects, e-GEWIS is expected to maximize the analytic power for deciphering GxE interactions.

e-GEWIS and c-GEWIS share the same scientific goal. c-GEWIS has been proposed as a promising approach for GEWIS. A typical c-GEWIS uses only cases, hence halving the required sample size required by cc-GEWIS, if 50% of subjects are cases. It has also been shown, and is supported by our simulation studies here, that c-GEWIS is in general more powerful than cc-GEWIS. Furthermore, c-GEWIS captures the presence of GxE via testing the departure from independence between SNP and exposure without requiring any specific penetrance model for GxE, such as the logistic regression model (1). Consequently, test statistics for c-GEWIS may be considered as omnibus test statistics, and hence are more robust than parametric tests implied by any regression models. The major weakness of c-GEWIS is that it requires the independence assumption between SNPs and exposure in the general population (or in controls, for uncommon disease researches), for which most studies are not powered to definitively verify this assumption [Piegorsch, et al. 1994]. Any violation to this assumption could inflate false-positive errors, and inflation errors are exaggerated in GEWIS with testing many SNPs [Albert, et al. 2001]. Another weakness is that there is no obvious way for c-GEWIS to adjust for confounders, including potential epidemiological covariates or population stratification factors.

To alleviate the concern with c-GEWIS, we describe e-GEWIS as an alternative approach for GEWIS, which may be suitable in some situations. Through simulation studies initially with a single SNP, we compared both approaches, with cc-GEWIS as a background comparator. In general, the power of e-GEWIS is comparable to that of c-GEWIS, after considering variable sample sizes by e-GEWIS. One noticeable exception is that the power of c-GEWIS declines with the increasing disease association with exposure, implying that e-GEWIS may be preferred when the exposure is known to have a strong main effect.

To complement the simulation study with a single SNP, we also conduct a relatively more realistic simulation, mimicking a lung cancer case-control study. In this simulation study, we showed that e-GEWIS, with less than 50% sample sizes required by c-GEWIS, has a power comparable to c-GEWIS, which is more powerful than cc-GEWIS (Figure 2). In the situations where SNPs and exposure are not independent, e-GEWIS maintains its false-positive error rate, while c-GEWIS would experience an inflated false-positive error rate (not shown). On the other hand, if a SNP of interest has an independent association with the disease, c-GEWIS would maintain its false-positive error rate for testing GxE, while e-GEWIS would suffer from an inflated “false discovery rate” with respect to GxE (not shown). In other words, those main SNP associations are falsely identified as GxE by e-GEWIS. Even so, such “false discoveries” would still be of interest to researchers.

Besides the power consideration, one advantage in favor of e-GEWIS over c-GEWIS is that one can adjust covariates through the logistic regression (1), as is the case for the usual cc-GEWIS. This is important, since typical GEWIS has to control multiple confounding variables. Further, it is essential to control population stratification via adjustment for principle components that reflect hidden ethnicities. Lastly, e-GEWIS is particularly suitable for meta-analysis of GEWIS from multiple sites where site-specific heterogeneity can be adjusted through the regression analysis.

While arguing in favor of e-GEWIS, we also need to recognize situations where e-GEWIS may be less desirable than c-GEWIS. For example, in some GEWIS studies, one may have an array of exposure variables that are equally important. Naively assembling “exposed subjects” by a single exposure variable could diminish analytic powers of discovering GxE with other exposure variables. Further, this conditional sampling would restrict interpretations of other analytic results that apply to those exposed subjects. Another scenario in favor of c-GEWIS, rather than e-GEWIS, is that the disease penetrance is complex and may not be easily modeled by, say, a logistic regression. In such a case, an omnibus test statistic may be useful as an initial screening for possible leads.

Before leaving the topic on simulation-based comparison of e-GEWIS and c-GEWIS, one legitimate question is whether choosing the simplest form of test for c-GEWIS leads to a fair comparison. In practice, most investigators are aware of this “problematic assumption” required by c-GEWIS, and likely adopt some form of two-stage methods, to minimize false-positive discoveries induced by violating the independence assumption. Regardless of how one chooses combinations of different methods at two or more stages, use of case-only methods at either the screening stage or the validating stage, will lead to excessive number of false-positive discoveries. False discoveries need to be controlled at an independent stage,

which may lead to reduction of power. Realistically, e-GEWIS, c-GEWIS and cc-GEWIS can be integrated for various analyses at multiple stages, to maximize analytic powers, while controlling false-positive discoveries at a pre-specified level. By focusing on the native form of e-GEWIS and c-GEWIS, we can address the fundamental issues regarding designs and methods, without engaging the complexities of hybrid multi-stage approaches.

Another concern with simulation and analytic comparisons here is if the comparison of e-GEWIS and c-GEWIS may seem unfair, since hypotheses under testing are different and so are study designs and analytic methods. However, both share the same scientific objective, and same design goal of maximizing powers. Hence, the comparison is appropriate and informative for practitioners to choose appropriate designs for GEWIS, without worrying detailed differences between their operational characteristics of e-GEWIS and c-GEWIS.

Applying e-GEWIS to a lung cancer case-control study, we set out to discover which SNPs/genes interact with smoking status in lung cancer (Figure 3). The strongest signal from e-GEWIS is on chromosome 15q25 with p-values around 3.27×10^{-9} , which reaches the typical significance threshold value of 10^{-8} . Actually, this genetic locus has been discovered by several earlier GWAS [Amos, et al. 2008; Hung, et al. 2008; Liu, et al. 2008; Thorgeirsson, et al. 2008]. In particular, it is worth noting that Amos et al carried out their GWAS, using current and former smokers for their cases and controls [Amos, et al. 2010; Amos, et al. 2008], employing e-GEWIS by the terminology here. As noted earlier, odds ratios associated with current smoking, SNP and their joint effect are approximately 1.08, 1.24 and 1.85 (Table 4), implying that these estimated odds ratios provide statistical evidence for possible “GxE” on either multiplicative scale ($OR_{G \times E} = 1.85 / (1.08 \times 1.24) = 1.37 > 1$) or additive scale ($OR_{G \times E} = (1.85 - 1) - (1.08 - 1) - (1.24 - 1) = 0.68 > 0$). For this discovery, it appears that the genetic associations with all identified SNPs are substantial, implying that both a genetic factor and its interaction with smoking contribute to lung cancer risk. This dual contribution may be one of reasons for this signal to have a high significance. In literature, this locus flanks nicotinic acetylcholine receptor subunits (CHRNA5, CHRNA3 and CHRNB4). While it is tempting to speculate the potential scientific significance, the presentation of results and interpretations of the discoveries on chromosome 15 are beyond the scope of this illustration and will be considered in a separate paper, with additional independent replication.

Upon completing three separate scans by e-GEWIS, c-GEWIS and cc-GEWIS, we were somewhat surprised with the discovery results (Figure 3). With modest sample size (485 controls and 904 cases, all current smoker smokers), e-GEWIS reveals the same signal on chromosome 15 discovered earlier (Table 3 and 4). In contrast, c-GEWIS, with 1,820 cases, has not yielded the same signal. Upon further investigation on these SNPs, their p-values range from 0.019 to 0.69 (not shown). This negative result is puzzling, partly because our simulation study suggests a comparable power between e-GEWIS and c-GEWIS. Several plausible explanations could account for this negative result. One possibility is that the smoker status is treated differently by c-GEWIS and e-GEWIS, in which c-GEWIS compares the current smokers with former and never smokers into unexposed category, given our emphasis on the current smoker. Secondly, e-GEWIS is able to adjust all available risk factors, while c-GEWIS is not amenable to any meaningful adjustment for these risk

factors. Lastly, it is entirely possible that smoking status and these SNPs are not independent in general population, since these SNPs flank nicotinic acetylcholine receptor subunits that associate with smoking behavior. Such departure could reduce the analytic power [Mukherjee, et al. 2012]. Note that lacking strong signals from cc-GEWIS is expected, since this approach must estimate two main effects associated with smoking and with SNP and quantify the interaction beyond their additive effects on the logit scale.

While appreciating the intuitive nature of e-GEWIS, we do realize the partial reliance of our conclusions on simulation studies (Figure 1 and 2), which have several limitations: either unrealistic scenario with a single SNP or relatively small number of SNPs, a single associated disease-causing locus, close resemblance to smoking exposure in lung cancer GWAS, and an assumed logistic regression model. With respect to genotype limitations, we have performed several simulation studies with many more SNPs from a large genome segment, SNPs from entire chromosomes, or genome-wide. The general trends remain, regardless of SNP configurations. Our choice of relatively few SNPs was intended to enable computational feasibility while retaining realism. We have performed several additional simulations with varying distributions of exposure variables, and simulation results are largely consistent with those reported here. Finally, mis-specified penetrance remains as an issue for further exploration in the future.

Closely related to model misspecification is the uncertainty of the scale on which GxE interaction may be quantified, namely, multiplicative or additive scale. Under the current logistic model, GxE interpretation is naturally quantified by a single parameter on a multiplicative scale, i.e., in the logistic regression model, the zero δ_j implies the null interaction on the multiplicative scale. However, the zero δ_j does not necessarily imply the absence of the GxE interaction on the additive scale, since $\exp(\beta_j + \gamma_j) - \exp(\beta_j) - \exp(\gamma_j) + 1$ may not equal zero. Now when performing e-GEWIS, we are actually testing the null hypothesis $H_0: (\gamma_j + \delta_j) = 0$. When both γ_j and δ_j values are positive or negative, this null hypothesis implies that $\gamma_j = 0$ and $\delta_j = 0$. With this constraint, the null hypothesis of zero δ_j actually implies that the additive GxE interaction equals zero, i.e., $\exp(\beta_j + \gamma_j + \delta_j) - \exp(\beta_j) - \exp(\gamma_j) + 1 = 0$. Further, if the main genetic effect is absent, which is true for most of SNPs, the null hypothesis of e-GEWIS is equivalent between multiplicative and additive scale, since $\exp(\beta_j + \gamma_j + \delta_j) - \exp(\beta_j) - \exp(\gamma_j) + 1 = \exp(\beta_j + \delta_j) - \exp(\beta_j) = \exp(\beta_j)[\exp(\delta_j) - 1]$. In less frequent situations, γ_j and δ_j values may be in opposite directions, so the interpretation of multiplicative GxE is no longer applicable to the additive GxE. In such cases, it is necessary to use the full case-control data to estimate all three regression coefficients (β_j , γ_j , δ_j) (as we have done in the illustrative example). With estimated regression coefficients from the logistic regression model, one can now estimate the additive GxE effect as $\exp(\beta_j + \gamma_j + \delta_j) - \exp(\beta_j) - \exp(\gamma_j) + 1$. Several test statistics have been proposed to test this additive GxE interaction [Knol, et al. 2011; Rothman 1986].

Following a successful e-GEWIS with one or more positive discoveries, one may be interested in learning if the positive discoveries resulted from GxE interactions or from main genetic associations, utilizing available GWAS data. Further, only by estimating all three parameters (β_j , γ_j , δ_j) one can then establish either additive or multiplicative GxE interaction. In pursuit of the analysis at this stage, one uses the logistic regression model (1) to estimate

$(\beta_j, \gamma_j, \delta_j)$ from available exposed and unexposed subjects, as the post e-GEWIS analysis. Additional characterization analyses may include further adjustments of possible confounders, assessment of dose-response effect in GxE, and characterization of haplotype effects of multiple adjacent SNPs.

In advocating for e-GEWIS, we recognize the risk of providing yet another statistical tool for dredging data. In the GWAS era, multiple comparisons have been a thorny problem, haunting the entire biomedical community that uses genomic technologies. Besides testing many SNPs, other multiple comparisons involve exploring multiple phenotypes, multiple covariates, and many sub-groups. Using e-GEWIS, one could conceptually identify many “exposure strata”, and look for indications of GxE by performing GWAS analysis on “exposed subjects”. Multiple comparisons would diminish the power of detecting GxE with excessive careless e-GEWIS analyses.

What we have learned from e-GEWIS also sheds insight into the lack of replications of GWAS discoveries. In recent years, many GWAS discoveries have not been replicated across different study populations; failure of replication is often taken as a sign of false-positive discoveries [Kraft, et al. 2009]. While consistent replications assure the true genetic associations, lacking replications of a GWAS discovery should not immediately negate the original discovery. It is important to examine if other environmental exposures are comparable between the original and replication populations. For example, if an original discovery from GWAS of lung cancer susceptibility genes in United States were not replicated in appropriately powered GWAS in United Kingdom or in China, one should not dismiss the original discovery, because gene by smoking interaction could distort replication results with drastically different smoking patterns and other lifestyle variables. To improve the validity in assessing reproducibility across populations, it is important for investigators to examine distributions of key exposure variables before reaching any meaningful conclusions. Whenever possible, the study should consider an e-GEWIS design, conditioning on the key exposure status, to improve the reproducibility across different study populations.

Formalizing e-GEWIS lays a foundation for designing future GEWIS that maximizes the power of detecting GxE. For example, when simulating a GWAS of lung cancer with genes by smoking interaction, we have shown that e-GEWIS requires sample sizes only about 50% of c-GEWIS and 25% of cc-GEWIS, without sacrificing the analytic power of detecting GxE interactions. When designing a GEWIS, one could choose to genotype only those exposed subjects, which would save on genotyping cost. The cost saving would allow investigators to genotype many more exposed subjects. By the same rationale, one can more effectively design a large consortium of multiple e-GEWIS that have comparable exposure distributions. In advocating for e-GEWIS, we recognize that this sampling will constrain any future secondary use of the data, in similar ways as single-compound clinical trials.

One practical question, facing e-GEWIS, is how to design such a study. As noted above, e-GEWIS is operationally equivalent to a GWAS among exposed subjects. Hence, many design considerations would be equivalent to those for GWAS, with the constraint of the “exposure” on the study population [Sham and Purcell 2014]. The appropriate design of an

e-GEWIS emphasizes the definition of the exposure variable(s) and the individuals who are considered as “exposed subjects”; balancing between study hypotheses, availability of exposed subjects, corresponding powers, and constraints on future use of GEWIS in secondary data analyses. Purely from the perspective on exposure prevalence, the simulation study (Figure 1) suggests that e-GEWIS has desirable power, once the exposure prevalence exceeds 30%, in comparison with c-GEWIS.

While the above description and discussion of e-GEWIS concern mostly case-control study design, the design principle is readily generalizable to several other design situations useful for GEWIS, namely cohort studies and clinical trial designs. For example, in a genetic study using large databases of electronic health records, one may have a large number of subjects for investigating gene by drug interactions, in which drug exposures are typically well defined. Based on the principle of e-GEWIS, a preferred study design is to genotype only patients who are taking the drug of interest. Similarly, when designing a clinical trial focusing on gene by drug interactions, one should genotype only those patients who have been assigned to the treatment arm. Lastly but not least, e-GEWIS could be an effective design strategy to investigate GxE interaction and drug-associated toxicity in rescuing effective drugs that fail to receive FDA approval due to toxicity. In such studies, e-GEWIS would genotype only those patients who have received treatments, saving costs and accelerating study completion without sacrificing study powers.

Acknowledgment

Authors would like to thank Dr. Maria Teresa Landi for her guidance on use of lung cancer GWAS data sets. Equally, we would like to thank anonymous reviewers whose comments have improved the presentation of this manuscript. This work is supported in part by institutional development fund and NIH grants (R01 CA139633, R01 HL105914).

References

- Albert PS, Ratnasinghe D, Tangrea J, Wacholder S. Limitations of the case-only design for identifying gene-environment interactions. *Am J Epidemiol.* 2001; 154(8):687–93. [PubMed: 11590080]
- Amos CI, Gorlov IP, Dong Q, Wu X, Zhang H, Lu EY, Scheet P, Greisinger AJ, Mills GB, Spitz MR. Nicotinic acetylcholine receptor region on chromosome 15q25 and lung cancer risk among African Americans: a case-control study. *J Natl Cancer Inst.* 2010; 102(15):1199–205. [PubMed: 20554942]
- Amos CI, Wu X, Broderick P, Gorlov IP, Gu J, Eisen T, Dong Q, Zhang Q, Gu X, Vijayakrishnan J. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat Genet.* 2008; 40(5):616–22. others. [PubMed: 18385676]
- Bix AS. The politics of heredity: Essays on eugenics, biomedicine, and the nature-nurture debate. *Journal of the History of Biology.* 1999; 32(2):395–397.
- Bookman EB, McAllister K, Gillanders E, Wanke K, Balshaw D, Rutter J, Reedy J, Shaughnessy D, Agurs-Collins T, Paltoo D. Gene-environment interplay in common complex diseases: forging an integrative model-recommendations from an NIH workshop. *Genet Epidemiol.* 2011 others.
- Breslow NE, Chatterjee N. Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis. *Journal of the Royal Statistical Society Series C-Applied Statistics.* 1999; 48:457–468.
- Breslow, NE.; Day, NE. Statistical methods in cancer research. International Agency for Research on Cancer; Lyon: 1980.
- Butcher LM, Plomin R. The nature of nurture: a genomewide association scan for family chaos. *Behav Genet.* 2008; 38(4):361–71. [PubMed: 18360741]

- Chatterjee N, Carroll RJ. Semiparametric maximum likelihood estimation in case-control studies of gene-environment interactions. *Biometrika*. 2005; 92:399–418.
- Chatterjee N, Kalaylioglu Z, Carroll RJ. Exploiting gene-environment independence in family-based case-control studies: increased power for detecting associations, interactions and joint effects. *Genet Epidemiol*. 2005; 28(2):138–56. [PubMed: 15593088]
- Cheng KF, Lin WJ. Retrospective analysis of case-control studies when the population is in Hardy-Weinberg equilibrium. *Stat Med*. 2005; 24(21):3289–310. [PubMed: 16206241]
- Clayton D. Commentary: reporting and assessing evidence for interaction: why, when and how? *Int J Epidemiol*. 2012; 41(3):707–10. [PubMed: 22596933]
- Clayton DG. Prediction and interaction in complex disease genetics: experience in type 1 diabetes. *PLoS Genet*. 2009; 5(7):e1000540. [PubMed: 19584936]
- Cornelis MC, Tchetgen EJ, Liang L, Qi L, Chatterjee N, Hu FB, Kraft P. Gene-environment interactions in genome-wide association studies: a comparative study of tests applied to empirical studies of type 2 diabetes. *Am J Epidemiol*. 2012; 175(3):191–202. [PubMed: 22199026]
- Cox, DRHDV. *Theoretical Statistics*. Chapman and Hall; New York: 1986.
- Flowers E, Froelicher ES, Aouizerat BE. Gene-environment interactions in cardiovascular disease. *Eur J Cardiovasc Nurs*. 2011
- Gauderman WJ, Thomas DC, Murcray CE, Conti D, Li D, Lewinger JP. Efficient genome-wide association testing of gene-environment interaction in case-parent trios. *Am J Epidemiol*. 2010; 172(1):116–22. [PubMed: 20543031]
- Gauderman WJ, Zhang P, Morrison JL, Lewinger JP. Finding novel genes by testing G x E interactions in a genome-wide association study. *Genet Epidemiol*. 2013; 37(6):603–13. [PubMed: 23873611]
- Helbig KL, Nothnagel M, Hampe J, Balschun T, Nikolaus S, Schreiber S, Franke A, Nothlings U. A case-only study of gene-environment interaction between genetic susceptibility variants in NOD2 and cigarette smoking in Crohn's disease aetiology. *BMC Med Genet*. 2012; 13:14. [PubMed: 22416979]
- Hernandez, LM.; Blazer, DG. Introduction to “Genes, Behavior, and the Social Environment: Moving Beyond the Nature/Nurture Debate”.. In: Hernandez, LM.; Blazer, DG., editors. *Genes, Behavior, and the Social Environment: Moving Beyond the Nature/Nurture Debate*. Washington (DC): 2006.
- Hindorf LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A*. 2009; 106(23):9362–7. [PubMed: 19474294]
- Hsu L, Jiao S, Dai JY, Hutter C, Peters U, Kooperberg C. Powerful cocktail methods for detecting genome-wide gene-environment interaction. *Genet Epidemiol*. 2012; 36(3):183–94. [PubMed: 22714933]
- Hung RJ, McKay JD, Gaborieau V, Boffetta P, Hashibe M, Zaridze D, Mukeria A, Szeszenia-Dabrowska N, Lissowska J, Rudnai P. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature*. 2008; 452(7187):633–7. others. [PubMed: 18385738]
- Hutter CM, Mechanic LE, Chatterjee N, Kraft P, Gillanders EM, Tank NCIG-ET. Gene-environment interactions in cancer epidemiology: a National Cancer Institute Think Tank report. *Genet Epidemiol*. 2013; 37(7):643–57. [PubMed: 24123198]
- Jones G. *The politics of heredity: Essays on eugenics, biomedicine, and the nature-nurture debate*. Isis. 1999; 90(4):851–852.
- Khoury, MJ.; Beaty, TH.; Cohen, BH. *Fundamentals of genetic epidemiology*. Oxford University Press; New York: 1993.
- Khoury MJ, Wacholder S. Invited commentary: from genome-wide association studies to gene-environment-wide interaction studies--challenges and opportunities. *Am J Epidemiol*. 2009; 169(2):227–30. discussion 234–5. [PubMed: 19022826]
- Knol MJ, VanderWeele TJ, Groenwold RH, Klungel OH, Rovers MM, Grobbee DE. Estimating measures of interaction on an additive scale for preventive exposures. *Eur J Epidemiol*. 2011; 26(6):433–8. [PubMed: 21344323]

- Kooperberg C, Leblanc M. Increasing the power of identifying gene x gene interactions in genome-wide association studies. *Genet Epidemiol.* 2008; 32(3):255–63. [PubMed: 18200600]
- Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ. Exploiting gene-environment interaction to detect genetic associations. *Human Heredity.* 2007; 63(2):111–9. [PubMed: 17283440]
- Kraft P, Zeggini E, Ioannidis JP. Replication in genome-wide association studies. *Statistical Science.* 2009; 24(4):561–573. [PubMed: 20454541]
- Landi MT, Chatterjee N, Yu K, Goldin LR, Goldstein AM, Rotunno M, Mirabello L, Jacobs K, Wheeler W, Yeager M. A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. *Am J Hum Genet.* 2009; 85(5):679–91. others. [PubMed: 19836008]
- Li MJ, Wang P, Liu X, Lim EL, Wang Z, Yeager M, Wong MP, Sham PC, Chanock SJ, Wang J. GWASdb: a database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res.* 2012; 40(Database issue):D1047–54. [PubMed: 22139925]
- Liu C, Batliwalla F, Li W, Lee A, Roubenoff R, Beckman E, Khalili H, Damle A, Kern M, Furie R. Genome-wide association scan identifies candidate polymorphisms associated with differential response to anti-TNF treatment in rheumatoid arthritis. *Mol Med.* 2008; 14(9-10):575–81. others. [PubMed: 18615156]
- Liu CY, Maity A, Lin X, Wright RO, Christiani DC. Design and analysis issues in gene and environment studies. *Environ Health.* 2012; 11:93. [PubMed: 23253229]
- Mechanic LE, Chen HS, Amos CI, Chatterjee N, Cox NJ, Divi RL, Fan R, Harris EL, Jacobs K, Kraft P. Next generation analytic tools for large scale genetic epidemiology studies of complex diseases. *Genet Epidemiol.* 2012; 36(1):22–35. others. [PubMed: 22147673]
- Mukherjee B, Ahn J, Gruber SB, Chatterjee N. Testing gene-environment interaction in large-scale case-control association studies: possible choices and comparisons. *Am J Epidemiol.* 2012; 175(3):177–90. [PubMed: 22199027]
- Murcray CE, Lewinger JP, Conti DV, Thomas DC, Gauderman WJ. Sample size requirements to detect gene-environment interactions in genome-wide association studies. *Genet Epidemiol.* 2011; 35(3):201–10. [PubMed: 21308767]
- Murcray CE, Lewinger JP, Gauderman WJ. Gene-environment interaction in genome-wide association studies. *Am J Epidemiol.* 2009; 169(2):219–26. [PubMed: 19022827]
- Ordovas JM, Smith CE. Epigenetics and cardiovascular disease. *Nat Rev Cardiol.* 2010; 7(9):510–9. [PubMed: 20603647]
- Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Stat Med.* 1994; 13(2):153–62. [PubMed: 8122051]
- Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika.* 1979; 66(3):403–11.
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007; 81(3):559–75. others. [PubMed: 17701901]
- Rothman, KJ. *Modern epidemiology.* Little, Brown and company; Boston: 1986.
- Rothman KJ, Greenland S. Causation and causal inference in epidemiology. *Am J Public Health.* 2005; 95(Suppl 1):S144–50. [PubMed: 16030331]
- Rothman KJ, Greenland S, Walker AM. Concepts of interaction. *Am J Epidemiol.* 1980; 112(4):467–70. [PubMed: 7424895]
- Sham PC, Purcell SM. Statistical power and significance testing in large-scale genetic studies. *Nat Rev Genet.* 2014; 15(5):335–46. [PubMed: 24739678]
- Song M, Lee KM, Kang D. Breast cancer prevention based on gene-environment interaction. *Mol Carcinog.* 2011; 50(4):280–90. [PubMed: 21465576]
- Spinka C, Carroll RJ, Chatterjee N. Analysis of case-control studies of genetic and environmental factors with missing genetic information and haplotype-phase ambiguity. *Genet Epidemiol.* 2005; 29(2):108–27. [PubMed: 16080203]

- Spires TL, Hannan AJ. Nature, nurture and neurology: gene-environment interactions in neurodegenerative disease. FEBS Anniversary Prize Lecture delivered on 27 June 2004 at the 29th FEBS Congress in Warsaw. FEBS J. 2005; 272(10):2347–61. [PubMed: 15885086]
- Superko HR, Momary KM, Li Y. Statins personalized. Med Clin North Am. 2012; 96(1):123–39. [PubMed: 22391257]
- Thomas D. Gene--environment-wide association studies: emerging approaches. Nat Rev Genet. 2010a; 11(4):259–72. [PubMed: 20212493]
- Thomas D. Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies. Annu Rev Public Health. 2010b; 31:21–36. [PubMed: 20070199]
- Thorgeirsson TE, Geller F, Sulem P, Rafnar T, Wiste A, Magnusson KP, Manolescu A, Thorleifsson G, Stefansson H, Ingason A. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. Nature. 2008; 452(7187):638–42. others. [PubMed: 18385739]
- Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science. May 15th.1998 280:1077–1082. others. [PubMed: 9582121]
- Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Research. 2014; 42(Database issue):D1001–6. others. [PubMed: 24316577]
- Zhao, LP. Multivariate analysis of binary data, ph.d. thesis. University of Washington School of Public Health; Seattle, Washington: 1989.

Figure 1.

Estimated powers by e-GEWIS on exposed subjects only (dark-solid line), by c-GEWIS on cases only (red-dashed line), and by cc-GEWIS on all available cases and controls (green-dotted line), when a single SNP is included in GxE study: a) estimated powers as odds ratio of GxE varies from 1 to 2; b) exposure prevalence ranges from 0.1 to 0.9 and also estimated ratios of required sample size by e-GEWIS, $N(\text{exposure})$, over the sample size by c-GEWIS, $N(\text{case})$. The large black dots correspond to ratios of required sample sizes, centering around the black dashed line of 1; c) minor allelic frequency (MAF) varies from 0.05 to 0.50; d) odds ratio of the main effect of exposure variable varies from 1 to 2.

Figure 2.

Estimated true discovery powers by eGEWIS on exposed subjects only, by cGEWIS on cases only, and by ccGEWIS on all available cases and controls, when a causal SNP is included for the direct association analysis (left three curves) and the causal SNP is not included in the candidate SNP list, leading to an indirect association analysis (right three curves).

Figure 3. Manhattan plots, with $-\log(\text{p-values})$ over all SNP positions across 22 autosomes, are obtained by eGEWIS, cGEWIS and ccGEWIS, when they are applied to scan two lung cancer GWAS, assessing genetic interactions with current smoking status

Table 1

A simple data representation useful for assessing GxE: a) summary table of binary exposure, genotype, and phenotype, b) association pattern of odds for quantifying GxE, and c) a plausible association pattern of odds in the absence of the main genetic effect

	(a) Counts [*]		(b) Odds ⁺		(c) Plausible Pattern	
	G=0	G=1	G=0	G=1	G=0	G=1
E=0	a/A	b/B	e^{α}	$e^{\alpha+\gamma}$	e^{α}	e^{α}
E=1	c/C	d/D	$e^{\alpha+\beta}$	$e^{\alpha+\beta+\gamma+\delta}$	$e^{\alpha+\beta}$	$e^{\alpha+\beta+\delta}$

^{*} Number of diseased subjects / Number of non-diseased subjects

⁺ Odds=Pr(D=1|G=1,E)/Pr(D=0|G=0,E)

Table 2

Distributions of gender, age, pack year and site over cases and controls, across smoking status (never, former, and current smoker) among a large GWAS on lung cancer conducted by National Cancer Institute

Smoking Status		Never		Former		Current	
Covariates		Control	Case	Control	Case	Control	Case
Gender	Male	378	31	765	673	384	731
	Female	263	107	101	105	101	173
Age	<60	178	23	151	105	174	267
	60-64	97	21	159	119	98	179
	65-69	146	35	208	180	101	191
	70-	220	59	348	374	112	267
Pack Year	>0-15			380	95	95	29
	>15-30			228	160	130	119
	>30-40			96	152	99	154
	>40-50			83	132	67	191
	>50-60			32	88	49	163
	>60-70			13	44	16	78
	>70-80			18	37	13	47
	>80			16	70	16	123

Table 3

Estimated coefficients, standard errors, Z-scores and their p-values for 7 selected SNPs that have been identified from e-GEWIS analysis. Associated rs#, minor/major allele, and minor allelic frequency (MAF) have also been listed.

Chr	SNP rs#	Minor/Major Allele	MAF	Coef.	SE	Z-scores	P-values
15	rs8034191	C/T	0.45	0.47	0.09	5.15	2.61E-07
15	rs1051730	T/C	0.44	0.54	0.09	5.92	3.27E-09
15	rs12914385	T/C	0.47	0.47	0.09	5.24	1.63E-07
15	rs1996371	G/A	0.46	0.43	0.09	4.85	1.22E-06
15	rs6495314	C/A	0.46	0.43	0.09	4.86	1.17E-06
15	rs4887077	T/C	0.45	0.39	0.09	4.42	9.96E-06
15	rs11638372	T/C	0.45	0.39	0.09	4.45	8.56E-06

Table 4

Estimated log odds ratios and standard errors using a full logistic regression model for these 7 selected SNPs, and corresponding odds ratios for current smoker, SNP allele, and their joint effect, from a full case-control data analysis with the logistic regression model on gene by smoking interaction in lung cancer

Chr	SNP rs#	Estimated Coefficients (SE)			Odds Ratios		
		Smoking	SNP	Interaction	Smoking	SNP	Joint
15	rs8034191	0.10(0.13)	0.16(0.10)	0.29(0.11)	1.11	1.17	1.74
15	rs1051730	0.07(0.13)	0.21(0.07)	0.32(0.11)	1.08	1.24	1.85
15	rs12914385	0.17(0.13)	0.26(0.07)	0.20(0.11)	1.19	1.30	1.90
15	rs1996371	0.19(0.13)	0.25(0.07)	0.19(0.11)	1.21	1.23	1.80
15	rs6495314	0.18(0.13)	0.24(0.07)	0.20(0.11)	1.20	1.27	1.86
15	rs4887077	0.22(0.13)	0.24(0.07)	0.15(0.11)	1.25	1.27	1.86
15	rs11638372	0.22(0.13)	0.24(0.07)	0.16(0.11)	1.25	1.28	1.87