



Published in final edited form as:

Genet Epidemiol. 2013 September ; 37(6): 603–613. doi:10.1002/gepi.21748.

Finding Novel Genes by Testing G×E Interactions in a Genomewide Association Study

W. James Gauderman, Pingye Zhang, John L. Morrison, and Juan Pablo Lewinger

Department of Preventive Medicine, University of Southern California, Los Angeles

Abstract

In a genomewide association study (GWAS), investigators typically focus their primary analysis on the direct (marginal) associations of each SNP with the trait. Some SNPs that are truly associated with the trait may not be identified in this scan if they have a weak marginal effect and thus low power to be detected. However, these SNPs may be quite important in subgroups of the population defined by an environmental or personal factor, and may be detectable if such a factor is carefully considered in a gene-environment (G×E) interaction analysis. We address the question “Using a genome wide interaction scan (GWIS), can we find new genes that were not found in the primary GWAS scan?” We review commonly used approaches for conducting a GWIS in case-control studies, and propose a new 2-step screening and testing method (EDG×E) that is optimized to find genes with a weak marginal effect. We simulate several scenarios in which our 2-step method provides 70–80% power to detect a disease locus while a marginal scan provides less than 5% power. We also provide simulations demonstrating that the EDG×E method outperforms other GWIS approaches (including case only and previously proposed 2-step methods) for finding genes with a weak marginal effect. Application of this method to a G × Sex scan for childhood asthma reveals two potentially interesting SNPs that were not identified in the marginal-association scan. We distribute a new software program (G×Escan, available at <http://biostats.usc.edu/software>) that implements this new method as well as several other GWIS approaches.

Keywords

genomewide scan; environmental factor; power

Introduction

Many trait-related variants have been discovered through genomewide association scans of direct (marginal) effects [Hindorff et al., 2009]. However, after accounting for variants that have been identified, there remains a significant amount of heritability left unexplained for most traits. One reason we may not detect important SNPs is that the trait-related variant may only elevate risk in a subgroup of the population (e.g. only smokers), or there may be opposite genetic effects in different subgroups. Either situation is likely to produce a weak

marginal effect that is unlikely to be detected at a genome wide significance level. We will show that a genomewide interaction scan (GWIS) using an efficient testing method has the potential to identify such SNPs.

It is well known that a standard case-control (CC) analysis of G×E interaction using logistic regression generally has poor power. A case-only (CO) analysis [Piegorsch et al., 1994] can provide substantially greater power [Yang et al., 1997] but is only valid if G and E are independent in the source population. However, if G and E are not independent the CO analysis can have an unacceptably high false positive rate. A variety of approaches have been recently proposed in an attempt to provide greater power than a CC analysis without the potential type I error inflation of a CO analysis. These include empirical Bayes analysis [Mukherjee and Chatterjee, 2008], Bayes Model Averaging [Li and Conti, 2009], and various two-step approaches that include a screening and a testing step [Kooperberg and LeBlanc, 2008; Murcray et al., 2009; Murcray et al., 2011; Hsu et al., 2012]. Each of these two-step methods uses information in the case-control data to form a test statistic in the screening step that is independent of the test statistic in the testing step. In this paper, we describe a novel two-step approach that has greater power than all previously developed methods in many circumstances, particularly for a variant with a weak marginal effect that is likely to be missed in the primary scan.

Methods

Consider a case-control study consisting of N subjects, with N_1 cases and $N_0 = N - N_1$ controls, and let D_i , $i=1, \dots, N$ be indicators of disease status. We define E_i , $i=1, \dots, N$, to be an environmental factor, where “environment” is loosely defined to include an exogenous environmental variable (e.g., sunlight, air pollution), personal exposure (e.g., smoking, dietary fat), or other personal characteristic (e.g., sex, age). We assume for now that E is a binary indicator of ‘exposure’ with $P_E = \Pr(E=1)$ denoting the population exposure prevalence. We furthermore assume that M single nucleotide polymorphisms (SNPs) have been genotyped on each of the N study subjects. We let q_A denote the frequency of the minor (less common) allele “A” for a given SNP and let “a” denote the more common allele. For use in a statistical model, each SNP will be denoted G_i , $i=1, \dots, N$. In a GWAS, G is often coded according to an additive model, specifically $G = 0, 1$, or 2 for genotype aa, Aa, or AA, respectively. However, G could also be coded according to a dominant (G indicates AA or Aa genotype), recessive (G indicates AA genotype), or codominant (pair of indicators coding the 3 genotypes) model. For simplicity, we assume there is a single disease susceptibility locus (DSL), although the methods we develop can uncover multiple DSLs if they exist.

Marginal (G) association

In a case-control study, the marginal effect of a gene (G) on disease (D) is typically measured by the genetic odds ratio OR_G , which can be obtained as $\exp(\lambda_G)$ from a logistic regression model of the form:

$$\text{Logit}(\Pr(D_i=1|G)) = \lambda_0 + \lambda_G G_i. \quad (1)$$

Additional adjustment covariates can be included in this model if needed. A standard GWAS of marginal effects is conducted by testing the null hypothesis $\lambda_G=0$ for each of M SNPs in turn, for example using a likelihood ratio chi-squared test (S_{DG}) and significance level chosen to preserve the family-wise error rate (FWER). In the presence of an exposure factor (E) and a gene-environment ($G \times E$) interaction, OR_G is a weighted average of the corresponding genetic odds ratios in each exposure group ($OR_{G|E}$) if G is independent of E . If G is associated with E , then OR_G is also a weighted average if one includes E in Equation 1 as a covariate. The same magnitude of OR_G can result from quite different underlying patterns for the interactive effects of G and E on D .

Case-control (CC) test of $G \times E$

In follow-up to the primary scan, one could augment the model in Equation 1 to test each SNP in turn for a multiplicative $G \times E$ interaction using the model

$$\text{Logit}(\Pr(D_i=1|G_i)) = \beta_0 + \beta_G G_i + \beta_E E_i + \beta_{G \times E} G_i \times E_i \quad (2)$$

based on testing the null hypothesis $\beta_{G \times E}=0$ using test statistic $S_{G \times E}$. The quantity $OR_{G \times E} = \exp(\beta_{G \times E})$ is the interaction odds ratio, the genetic effect in exposed individuals relative to the genetic effect in unexposed (i.e. $OR_{G|E=1} / OR_{G|E=0}$). We denote analysis using this model as the standard case-control (CC) approach.

Case-only (CO) analysis

A more powerful test of $G \times E$ interaction can be obtained using a case-only (CO) analysis, in which association is tested between E and each SNP in affected individuals. Assuming a binary exposure factor, case-only analysis can be based on the model

$$\text{Logit}(\Pr(E_i=1|G_i, D_i=1)) = \gamma_0 + \gamma_{G \times E} G_i \quad (3)$$

The quantity $\exp(\gamma_{G \times E})$ is a consistent estimator of the $G \times E$ relative risk ratio [Piegorsch et al., 1994; Yang and Khoury, 1997] provided G and E are independent in the source population. A GWIS using the CO approach tests the null hypothesis that $\gamma_{G \times E} = 0$ for each of the M SNPs, with correction to preserve the FWER. A CO analysis can be substantially more powerful than a CC analysis [Yang et al., 1997], being equivalent to a comparable case control analysis with infinitely many controls, but it depends critically on the assumption of population-level G - E independence. Population-level G - E association can occur for SNPs that have a real effect on E , for example for gene variants that affect smoking behavior [Hodgson et al., 2012]. However, a factor (e.g. population sub-structure) that is associated with both E and SNP allele frequencies, can also induce spurious G - E associations. Either situation leads to invalid CO analysis and can produce an unacceptably high false positive rate [Mukherjee et al., 2012].

Empirical Bayes (EB)

Bayesian approaches, including Bayes model averaging [Li and Conti, 2009] and empirical Bayes (EB, [Mukherjee et al., 2010]), have been proposed for integrating direct information from a CC model with G - E correlation from a CO analysis. For example, in EB analysis a

Wald test statistic S_{EB} is formed based on a weighted average of $\beta_{G \times E}$ and $\gamma_{G \times E}$ with variance that is a function of the corresponding variances of these estimators. While both Bayesian approaches can provide greater power than a CC analysis, they can also have inflated Type I errors (though not as highly inflated as a CO analysis) in the presence of population-level G-E correlation [Mukherjee et al., 2010; Murcray et al., 2011]. For additional details see Mukherjee and Chatterjee [Mukherjee and Chatterjee, 2008].

Existing 2-step methods

Several two-step methods have been proposed to conduct a GWIS [Kooperberg and LeBlanc, 2008; Li and Conti, 2009; Murcray et al., 2009; Gauderman et al., 2010; Mukherjee et al., 2010; Murcray et al., 2011; Mukherjee et al., 2012], all of which generally provide greater power than a CC analysis while preserving the Type I error rate. A key requirement for any of the two-step methods is independence of the Step 1 screening and Step 2 testing statistics. All of the 2-step methods described below achieve this independence [Dai et al., 2012].

2-Step, DG | EB—Kooperberg and LeBlanc [Kooperberg and LeBlanc, 2008] proposed a 2-step procedure that uses the marginal DG association statistic S_{DG} to screen SNPs at Step-1 significance level α_1 . They proposed testing the subset $m \ll M$ SNPs that pass the Step-1 screen using Step-2 test statistic $S_{G \times E}$, with Bonferroni-corrected significance level α/m to preserve the FWER. This ‘DG|G×E’ approach was found to be less powerful than an alternative, DG|EB, in which screening is still based on S_{DG} but EB analysis is used for Step-2 testing [Hsu et al., 2012]. We implement the latter in our comparisons.

2-Step, EG | G×E—Murcray et al. [Murcray et al., 2009] demonstrated that in the presence of G×E interaction, there is an induced correlation between G and E in the combined case-control sample. In other words, based on the model

$$\text{Logit}(\Pr(E=1|G)) = \delta_0 + \delta_G G \quad (4)$$

applied to the full sample of cases and controls, one can expect $\delta_G \neq 0$ in the presence of G×E interaction. As a Step 1 screen, they proposed testing $H_0: \delta_G = 0$ at significance level α_1 using a likelihood ratio chi-squared test statistic (S_{EG}). As in the DG approach, they proposed testing the subset $m \ll M$ SNPs that pass the screen using Step-2 test statistic $S_{G \times E}$ at significance level α/m . The use of the model in Equation 4 rather than the case-only model in Equation 2 preserves the necessary independence between Steps 1 and 2 [Dai et al., 2012]. Murcray et al. also proposed a hybrid method (H2) that involved running both the DG and EG screening approaches in parallel, and adjusting the second step significance level to account for both sets of tests [Murcray et al., 2011].

2-Step, ‘Cocktail’—Hsu et al. [Hsu et al., 2012] proposed a different type of hybrid approach that mixes the different screening and testing statistics in an attempt to maximize efficient use of the data. In their ‘Cocktail I’ approach, they proposed a screening statistic $S_{CT} = S_{DG}$ if the p-value corresponding to S_{DG} is less than some threshold (they suggest 0.001), and $S_{CT} = S_{EG}$ otherwise. The Step-2 test is based on the test statistic (S_{EB}) from an EB analysis if $S_{CT} = S_{DG}$ and on $S_{G \times E}$ if $S_{CT} = S_{EG}$. The use of different statistics in Step 2 is

required to guarantee independence of the Step 1 and 2 tests. They furthermore suggested the use of weighted hypothesis testing [Ionita-Laza et al., 2007] (described below) in Step 2 rather than testing only a subset of SNPs that pass a Step-1 threshold. They showed that Cocktail I provides greater power than all of the approaches described above for most of the models they considered. They also proposed a variant on this approach (Cocktail II) based on defining S_{CT} as the maximum of S_{DG} and S_{EG} . In general, though, they found Cocktail I to have greater power than Cocktail II. We implement Cocktail I (called simply Cocktail) in our comparisons.

New 2-Step Method: EDG×E

The motivation for this new method comes from inspection of the standard retrospective likelihood for case-control data, which is based on the following conditional probability:

$$\Pr(G, E|D, Asc) \propto \Pr(D|G, E, Asc) \Pr(G, E|Asc) \quad (5)$$

Here ‘Asc’ denotes the ascertainment scheme used to obtain cases and controls. The first factor on the right hand gives rise to the model in Equation 2, and thus parameterizes both DG association and G×E interaction. The second factor can be expressed by the model in Equation 4 and captures EG association induced by the oversampling of cases from the source population. Previously proposed 2-step methods use different parts of the information contained in this likelihood to enhance power over a simple test of only G×E. For example, Kooperberg and LeBlanc [Kooperberg and LeBlanc, 2008] use DG information to screen and G×E to test, while Murcray et al. [Murcray et al., 2009] use EG to screen and G×E to test. The H2 hybrid method [Murcray et al., 2011] and Cocktail method [Hsu et al., 2012] consider both DG and EG information in screening, but ultimately use one or the other source of information to prioritize SNPs for testing in Step 2.

We propose a novel 2-step approach that uses all available surrogate information, i.e. both EG and DG association information combined, to screen SNPs. Specifically, for each of the M SNPs, we propose computing Step-1 screening statistic $S_{EG+DG} = S_{EG} + S_{DG}$, i.e. the sum of the EG and DG statistics described above. The two test statistics S_{EG} and S_{DG} are independent [Dai et al., 2012], and each follows a central chi-squared distribution with 1 degree of freedom (df) under their respective null hypotheses. Thus, S_{EG+DG} follows a central chi-squared distribution with 2 df under the joint null $H_0: \lambda_G = \delta_G = 0$. The Step-2 test is based on $S_{G \times E}$, which is independent of (S_{EG}, S_{DG}) [Dai et al., 2012] and thus is also independent of their sum S_{EG+DG} . The name EDG×E derives from Eg+Dg screening, with G×E testing. One can use either subset testing or weighted hypothesis testing in Step 2 (see below).

Hypothesis testing approaches in Step 2

As described above, some have proposed the use of subset testing and others weighted-hypothesis testing in Step 2. In the former, the analyst specifies α_1 , the significance threshold to pass Step 1. A larger value of α_1 will increase the chance of passing a truly associated SNP into Step 2, but at the cost of also increasing the number of unassociated

SNPs that pass into Step 2. A lower value of α_1 leads to lower m and thus greater power in Step 2, but at the potential cost of screening out a true SNP.

Rather than restrict Step-2 testing to a subset of the SNPs, one can test all M SNPs in Step 2 using a weighted significance level based on the ordered p -values from Step 1. The weighting scheme is designed to allocate a larger fraction of the genomewide significance level α to the most significant SNPs in Step 1. As proposed by Ionita-Laza et al. [Ionita-Laza et al., 2007], the B most significant (lowest p -value) SNPs based on Step 1 are evaluated in Step 2 at significance level $(\alpha/2)/B$, the next $2B$ SNPs are evaluated at $(\alpha/4)/(2B)$, the next $4B$ at $(\alpha/8)/(4B)$, etc. For example, when $B=5$ and $\alpha=0.05$, the top 5 SNPs from Step 1 are tested in Step 2 at significance level 0.005, the next 10 at 0.00125, etc. This weighting scheme guarantees that the overall significance level for the entire procedure does not exceed α . Under this weighting scheme, the top SNPs from Step 1 are tested at a more liberal significance threshold than the standard 5×10^{-8} level required in a standard exhaustive scan of all M SNPs (using CC, CO, or EB), and probably also a more liberal level than the threshold α/m required in subset testing. However, for the majority of SNPs not in the top bins, weighted testing will have a more stringent threshold than 5×10^{-8} . This indicates the importance of using an efficient Step-1 screening approach with strong likelihood of highly ranking any SNP with a true interaction.

Simulation Study

We use simulation to confirm the Type I error rates and to compare power of all of the above procedures. In all simulations we generated 2,000 replicate datasets, each consisting of equal numbers of cases and controls, and $M=1$ million SNPs. One SNP was designated as the DSL, assumed to have a $G \times E$ interaction effect on the trait. We considered two types of interaction models: 1) a modest interaction effect size ($OR_{G \times E}=1.5$), with marginal environmental effect size $OR_E=1.2$, common exposure ($p_E=0.4$), and common variant ($q_A=0.225$, yielding 40% carriers under a dominant model), and 2) a stronger interaction effect ($OR_{G \times E}=2.0$), with $OR_E=1.25$ and less common exposure ($p_E=0.10$), and variant ($q_A=0.134$, yielding 25% carriers). For each of these interaction models, we performed multiple simulations varying the magnitude of the marginal genetic effect (OR_G) from 1.0 to 1.35. These settings yielded a wide range of underlying disease risk models, encompassing both qualitative (effects of G in opposite directions depending on E) and quantitative (effects of G in the same direction but of differing magnitudes across levels of E) models of $G \times E$ interaction (Table 1). For each of the remaining $M - 1$ loci, we randomly sampled an allele frequency from a uniform distribution on the range 0.10 to 0.40. In our base model, none of these loci was associated with E or with disease. However, we considered alternative models in which 0.00001 or 0.00005 (corresponding to 10 or 50) of the 1 million SNPs were correlated with E in the population. We also considered an alternative model in which 10 loci had a marginal (but no $G \times E$) association with disease, with odds ratios for these 10 loci randomly sampled from a uniform distribution on the range 1.1 to 1.5. For all of the two-step methods (DG|EB, EG| $G \times E$, H2, Cocktail, and ED $G \times E$), we adopted weighted hypothesis testing in Step 2 and assumed an initial bin size of $B=5$ SNPs.

For each replicate data set, we performed genomewide analyses of G×E interaction using all of the methods described above. A dominant risk model was assumed in all analyses. The Type I error rate for each method was estimated as the proportion of replicates in which at least one of $M - 1$ non-DSL SNPs was declared statistically significant at a FWER of $\alpha = 0.05$. Power for each method was estimated as the proportion of replicates in which the DSL was identified as statistically significant. For each model, we also estimated power to detect the marginal effect of the DSL, based on Equation 1, to quantify the chance that the locus would have been identified in the primary G only scan. To explore the robustness of our power comparisons, we varied selected model and method settings around a base model with $OR_{G \times E} = 1.5$, $OR_G = OR_E = 1.2$, $q_A = 0.225$, and $p_E = 0.4$.

Asthma Analysis

Asthma is the most common chronic disease in children, with an estimated prevalence of 12.5% for diagnosis by a doctor before age 18[Merrick et al., 2005]. Prior GWAS scans have identified several loci that have a marginal association with asthma[Moffatt et al., 2010; Torgerson et al., 2011]. Asthma prevalence in children is known to vary by sex, particularly at young ages, with males exhibiting greater prevalence than females[Osman et al., 2007]. It is possible that this difference in prevalence is partly due to sex-specific effects of some genetic variants, for example if some sex-related personal characteristic such as hormone level has an effect on gene penetrance, or if there is a sex-specific difference in an environmental exposure that modifies gene penetrance. Thus, in an attempt to identify additional SNPs not found in the primary scans, we used the methods described in this paper to conduct a genomewide scan for G×Sex interaction.

We use data from the Children's Health Study (CHS) to conduct this analysis. The CHS is an ongoing cohort study spanning 16 southern California communities, investigating both genetic and environmental factors related to childhood asthma[McConnell et al., 2006] and lung function growth[Gauderman et al., 2007]. The CHS GWAS was based on a nested case-control sample selected from the Hispanic White (HW) and non-Hispanic White (NHW) children within the CHS cohorts. Based on questionnaire responses by parents, the presence or absence of doctor-diagnosed asthma, and for asthmatics the age of onset, were determined. For our analysis of G×Sex interaction, we focused on the subset of early-onset asthmatics, defined as reported asthma diagnosis prior to age 6. Controls were defined as subjects that were asthma free at age 6. A total of 2,382 HW or NHW subjects, including 631 cases and 1,751 controls were included in the analysis. Study samples were genotyped at the University of Southern California Genomic Center using the Illumina HumanHap550 or Human 610 Quad BeadChip microarrays. After quality control, a total of 536,857 SNPs were available for analysis. The CHS protocol was approved by the institutional review board for human studies at the University of Southern California, and written consent was provided by a parent or legal guardian for every study participant.

Results

Simulation Study

The CC and all 2-step methods achieved the nominal Type I error rate, whether or not some of the non-DSL SNPs had population level G-E correlation or marginal associations with disease (Table II). As expected, the CO analysis had unacceptably high Type I error rates when even a small fraction of the SNPs had G-E correlation in the population, but achieved nominal levels in the absence of such correlation. All 2-step methods also achieved the nominal Type I error rate using subset testing and for several alternative settings of model parameters (e.g. p_E , OR_E , data not shown).

Power to detect an interaction of magnitude $OR_{G \times E} = 1.5$ with 3,500 cases and 3,500 controls was quite low using the standard CC method (Figure 1a). As expected, a case-only analysis provided substantially higher power than a CC analysis, and EB provided power that was midway between CO and CC. Power for all three of these exhaustive approaches was nearly independent of the size of the marginal G effect (OR_G). On the other hand, power of the 2-step methods that utilize marginal G information in their screening step (DG|EB, H2, cocktail, and EDG×E) depended strongly on the size of OR_G . Of these 2-step methods, EDG×E was the most powerful when the magnitude of OR_G was small to moderate (in the range 1.10 to 1.25). For example, when $OR_G = 1.2$, power for EDG×E was 85% compared to 70% for Cocktail, 68% for H2, 63% for EG|G×E, and 66% for DG|EB. The EDG×E method was also more powerful than a case-only scan when $OR_G > \sim 1.12$. Power for DG|EB and Cocktail was about 5% higher than EDG×E when $OR_G > 1.35$, although at this and larger magnitudes of OR_G it is likely that the DSL would be identified in the marginal G-only scan. Similar trends were observed when the interaction effect size was larger ($OR_{G \times E} = 2.0$, Figure 1b), with EDG×E providing more power than other 2-step alternatives for a wider range of OR_G (1.10 to 1.30).

All 2-step methods provided greater power than the exhaustive CC or EB scans over a wide range of models (Table III). The H2, Cocktail and EDG×E methods also outperformed the case-only analysis in most scenarios. The improved power for EDG×E over other 2-step methods when $OR_G = 1.2$ (Base model) was robust to variations in DSL allele frequency (q_A), exposure frequency (p_E), exposure effect size (OR_E), bin size for weighted testing (B), and the presence of 10 additional loci with an effect on disease (G-D associations). Power for EDG×E was also greater than other 2-step methods if 10 SNPs were correlated with E in the population (G-E association), and greater than all but DG|EB if 50 SNPs were correlated with E. Power of all 2-step methods was higher if there was a positive correlation between the DSL and E in the population ($OR_{DSL-E} = 1.2$), but much lower if there was a negative correlation ($OR_{DSL-E} = 0.8$).

It is clear from Table III that the Cocktail and EDG×E methods are generally the most powerful methods across a range of scenarios. To further compare these two methods, we examined the ability of each to place the DSL among the highest ranked SNPs based on their respective Step 1 screens. For each of the models shown in Table III, Table IV shows the geometric mean rank of the Step-1 DSL p-values across replicate data sets, as well as the distribution across replicates of the DSL into Step-2 testing bins 1, 2, etc. As an example, for

the base model, the geometric mean Step-1 rank of the DSL was 2.3 (out of 1 million) using the EDG×E screen while it was 42.4 for the Cocktail screen. In addition, the DSL was among the top 5 SNPs (and thus in Bin 1 with most liberal Step 2 significance threshold) for 82% of the replicates using the EDG×E screen but for only 30% using the Cocktail screen. In general, the EDG×E screen more effectively ranked the DSL at the top of the Step 1 list than the Cocktail method across a wide range of models.

All of the above results were based on weighted hypothesis testing in Step 2. For the EDG×E method we also examined the power using subset testing, considering a range of possible settings of α_1 , the Step-1 significance threshold (Table V). Across most models, the highest power for subset testing occurred when a relatively small subset of markers (in the range of 10 to 100) was passed to Step 2. Note that power was significantly reduced with α_1 set to 0.05 or 0.01, thresholds that have been suggested in prior 2-step methods [Kooperberg and LeBlanc, 2008; Murcray et al., 2009]. However, except when there was a substantial number of markers with G-E correlation in the population, no choice of α_1 led to as much power as could be achieved using weighted testing.

Asthma Analysis

Exhaustive CC, CO, and EB scans for $G \times \text{Sex}$ interaction related to early-onset asthma did not produce any associations close to being genome-wide significant (Figure 2). The QQ plots for these analyses provide evidence that p-values were conservatively estimated for SNPs yielding more extreme configurations of G, E, and D (and thus the lowest p-values), likely due to our modest sample size. The QQ plot for the Step 1 screen of the EDG×E method (Figure 3) demonstrates that the p-values corresponding to the S_{DG+EG} statistic are consistent with the assumed 2-df chi-squared distribution. Indirectly, this QQ plot also shows that linkage disequilibrium among SNPs (present in these real data) does not affect the validity of the EDG×E screening test. Beginning at the upper end of this Step-1 QQ distribution, SNPs are placed by their rank-order into bins for Step-2 weighted hypothesis testing (we assume initial Bin size of 5). As shown in the Manhattan plot of Figure 3, one SNP in the first bin and one SNP in the second bin are close to their respective bin-specific significance thresholds. The first-bin SNP is rs6842542 (MAF=0.17), located on chromosome 4 near the GRIA2 locus, with Step-1 screening p-value 2.9×10^{-6} and Step-2 testing p-value 0.011 (Table VI). This locus exhibits a qualitative interaction, with $OR_{G|sex=F} = 1.13$ per allele for females and $OR_{G|sex=M} = 0.69$ for males. The second-bin SNP is rs7000310 (MAF=0.20) and has Step-1 screening p-value 1.1×10^{-5} and Step-2 testing p-value 0.0017. This SNP is located on chromosome 8q24 near the TNFRSF11B locus, a member of the TNF-receptor super-family. This locus exhibits a pure interaction, with no effect in females ($OR_{G|sex=F} = 0.99$) and a strong effect in males ($OR_{G|sex=M} = 1.78$). Neither of these SNPs or regions has been previously identified in marginal-effects scans of asthma, and although neither achieved 2-step genomewide significance, they are candidates for further investigation (e.g. in-silico replication analysis) in independent samples.

Discussion

We have presented a variety of scenarios for G×E interaction that produce a small marginal genetic effect. These kinds of loci are exactly the ones that are likely to be missed in our primary GWAS scans. While the potential importance of G×E interaction has been recognized for many diseases, poor power of a CC analysis and the potential biases in a CO analysis have likely reduced enthusiasm by investigators to conduct GWIS in their available samples. Prior investigations have shown that two-step methods can provide greater power than a CC analysis, and often greater power than a CO analysis. In this paper, we introduced a novel 2-step GWIS method that generally provides greater power than any other 2-step GWIS method when the marginal effect is small (OR_G is less than 1.3). This new approach may therefore provide the best opportunity to identify novel loci via GWIS analysis. We have developed a comprehensive and computationally efficient G×E analysis program that implements all of the methods described in this paper (G×Escan, available at <http://biostats.usc.edu/software>). For example, G×Escan required only 30 minutes on a single PC processor to conduct all analyses for the G×Sex scan in the CHS.

While we described the EDG×E method in the context of a binary environmental factor, we note that ‘E’ can be replaced by a quantitative exposure (e.g. air pollution), a personal factor (e.g. sex), or even a pre-specified candidate gene (e.g. GSTM1 genotype). An important concern in a GWIS is the availability of additional studies with comparable exposure data that can be used to replicate top G×E signals. Factors such as BMI or ever/never smoking are likely to be widely available, while variables such as dietary fat intake or pack-years of tobacco smoking may only have been measured in a limited number of studies. The investigator considering a GWIS needs to think carefully about the tradeoff between a high-quality exposure variable that may have little hope of being replicated compared versus a cruder exposure variable that may carry less information but be more widely available. In practice, it might be useful if results using the cruder but more widely available measure are routinely provided in an online supplement to enable meta-analyses of G×E interaction.

Instead of testing for G×E interaction, investigators may choose to simply repeat their marginal-effects scan focused on a specific exposure subgroup (e.g. a GWAS in smokers only). However, a genetic effect may be concentrated in either the exposed (e.g. smokers) or unexposed (e.g. non-smokers) subjects. Performing scans in both subgroups would ultimately reduce the power to detect G within either subgroup because of the additional multiple-testing correction required. On the other hand, power to detect a G×E interaction does not depend strongly on the direction of the effect.

For the models we considered, we have shown that adding Step-1 statistics ($S_{DG} + S_{EG}$) provides a more efficient screen than the Cocktail approach of using the maximum of the S_{DG} and S_{EG} statistics. When an interaction induces both a DG association and an EG correlation, our summation screen effectively uses both sources of information to prioritize SNPs for Step 2 testing. By taking only the maximum statistic to prioritize, the interaction information carried by the remaining statistic is not being utilized. There are alternative approaches one might adopt for combining Step-1 statistics, and we suggest this as a topic for future investigation.

The resources needed to conduct a GWIS are a proverbial drop in the bucket given the very large financial investment that has already been made in GWAS genotyping. We have highlighted the potential of a GWIS to identify novel SNPs. A fringe benefit of conducting GWIS is the potential to identify genes with modifiable effects and potentially susceptible subpopulations that would benefit from exposure modification. Successful identification of a G×E interaction may be the first step in learning about an important pathway, or may lead to additional studies, for example targeted sequencing, gene expression, or epigenetic studies focused on exposed or unexposed subjects. Although more efficient, the EDG×E approach still requires substantial sample sizes to achieve good power for detecting modest-sized interactions. For many traits, use of data from a consortium of genomewide association studies may provide the best opportunity for conducting a GWIS to uncover novel SNPs.

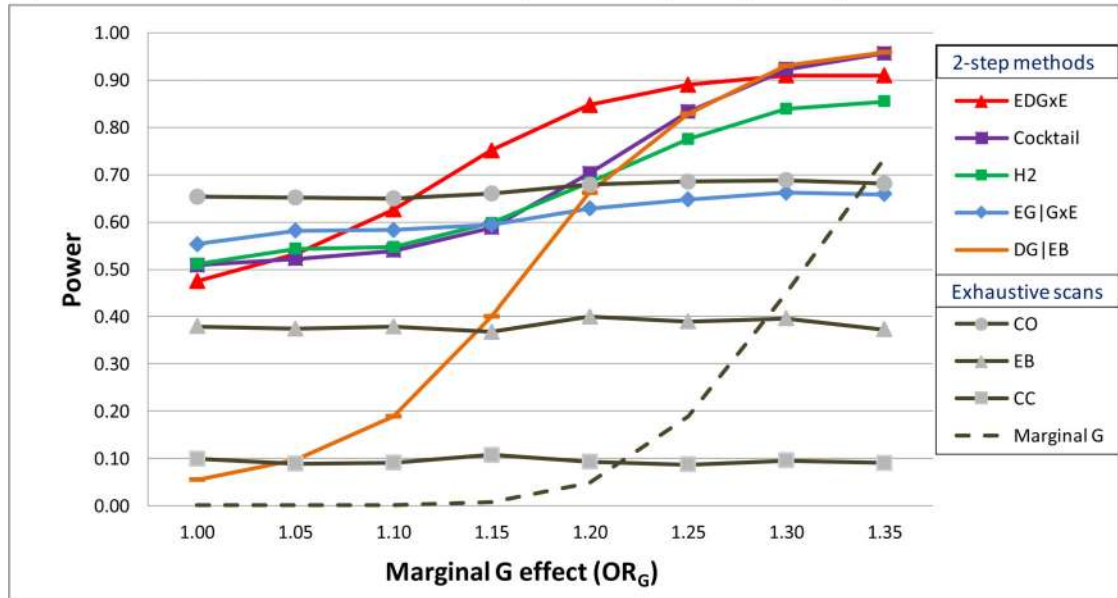
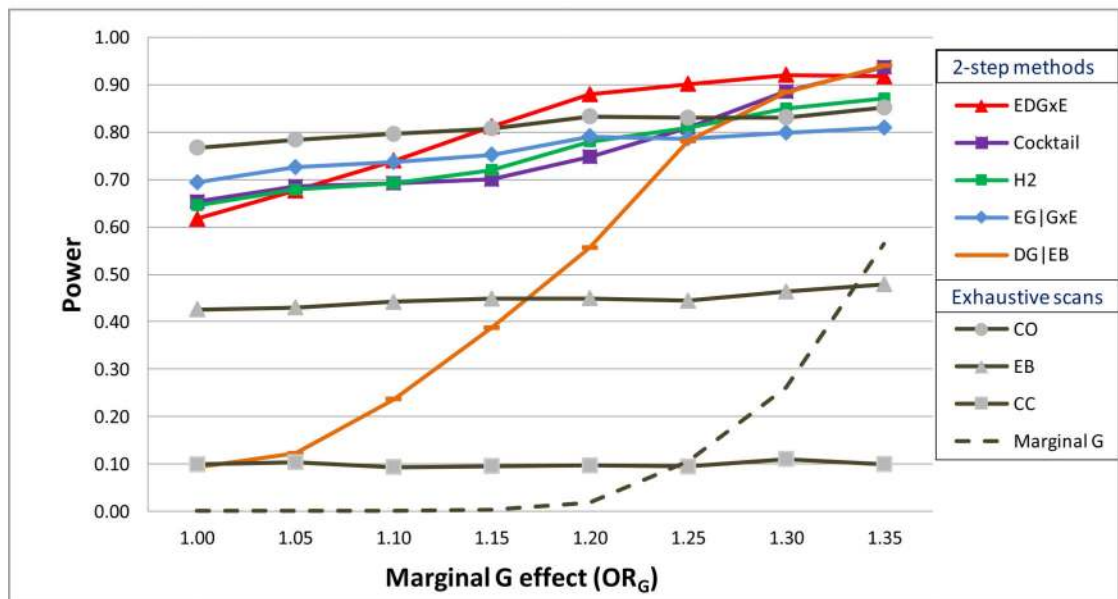
Acknowledgments

This work was supported by grants R21HL115606, R01HL087680, and 1RC2HL101651 from NHLBI; P30ES007048 from NIEHS, R41CA141852 from NCI, U01HG005927 from NHGRI, and U01HD061968 from NICHD.

References

- Dai J, Kooperberg C, LeBlanc M, Prentice R. Two-stage testing procedures with independent filtering for genome-wide gene-environment interaction. *Biometrika*. 2012; 99:929–944. [PubMed: 23843674]
- 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):116–122. [PubMed: 20543031]
- Gauderman WJ, Vora H, McConnell R, Berhane K, Gilliland F, Thomas D, et al. Effect of exposure to traffic on lung development from 10 to 18 years of age: a cohort study. *Lancet*. 2007; 369(9561): 571–577. [PubMed: 17307103]
- Hindorf LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. 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):9362–9367. [PubMed: 19474294]
- Hodgson ME, Olshan AF, North KE, Poole CL, Zeng D, Tse CK, et al. The case-only independence assumption: associations between genetic polymorphisms and smoking among controls in two population-based studies. *International journal of molecular epidemiology and genetics*. 2012; 3(4): 333–360. [PubMed: 23205185]
- 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):183–194. [PubMed: 22714933]
- Ionita-Laza I, McQueen MB, Laird NM, Lange C. Genomewide weighted hypothesis testing in family-based association studies, with an application to a 100K scan. *Am J Hum Genet*. 2007; 81(3):607–614. [PubMed: 17701906]
- Kooperberg C, LeBlanc M. Increasing the power of identifying gene × gene interactions in genome-wide association studies. *Genetic Epidemiology*. 2008; 32(3):255–263. [PubMed: 18200600]
- Li D, Conti DV. Detecting gene-environment interactions using a combined case-only and case-control approach. *Am J Epidemiol*. 2009; 169(4):497–504. [PubMed: 19074774]
- McConnell R, Berhane K, Yao L, Jerrett M, Lurmann F, Gilliland F, et al. Traffic, susceptibility, and childhood asthma. *Environ Health Perspect*. 2006; 114(5):766–772. [PubMed: 16675435]
- Merrick E, Hemmo-Lotem M, Merrick J. Recent trends in adolescent asthma. *Int J Adolesc Med Health*. 2005; 17(2):189–191. [PubMed: 15971739]

- Moffatt MF, Gut IG, Demenais F, Strachan DP, Bouzigon E, Heath S, et al. A large-scale, consortium-based genomewide association study of asthma. *N Engl J Med*. 2010; 363(13):1211–1221. [PubMed: 20860503]
- 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):177–190. [PubMed: 22199027]
- Mukherjee B, Ahn J, Gruber SB, Ghosh M, Chatterjee N. Case-control studies of gene-environment interaction: Bayesian design and analysis. *Biometrics*. 2010; 66(3):934–948. [PubMed: 19930190]
- Mukherjee B, Chatterjee N. Exploiting gene-environment independence for analysis of case-control studies: an empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics*. 2008; 64(3):685–694. [PubMed: 18162111]
- 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):201–210. [PubMed: 21308767]
- Murcray CE, Lewinger JP, Gauderman WJ. Gene-environment interaction in genome-wide association studies (with commentaries and rejoinder). *Am J Epidemiol*. 2009; 169(2):219–226. [PubMed: 19022827]
- Osman M, Tagiyeva N, Wassall HJ, Ninan TK, Devenny AM, McNeill G, et al. Changing trends in sex specific prevalence rates for childhood asthma, eczema, and hay fever. *Pediatr Pulmonol*. 2007; 42(1):60–65. [PubMed: 17133524]
- Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Statistics in Medicine*. 1994; 13(2):153–162. [PubMed: 8122051]
- Torgerson DG, Ampleford EJ, Chiu GY, Gauderman WJ, Gignoux CR, Graves PE, et al. Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations. *Nat Genet*. 2011; 43(9):887–892. [PubMed: 21804549]
- Yang Q, Khoury MJ. Evolving methods in genetic epidemiology. III. Gene-environment interaction in epidemiologic research. *Epidemiologic Reviews*. 1997; 19(1):33–43. [PubMed: 9360900]
- Yang Q, Khoury MJ, Flanders WD. Sample size requirements in case-only designs to detect gene-environment interaction. *American Journal of Epidemiology*. 1997; 146(9):713–720. [PubMed: 9366618]

A) Moderate interaction with common G and E ($OR_{G \times E}=1.5$, $q_A=0.23$, $p_E=0.40$)B) Strong interaction with less common G and E ($OR_{G \times E}=2.0$, $q_A=0.14$, $p_E=0.10$)**Figure 1.**

Power to detect GxE interaction across a range of magnitudes for the marginal G effect (OR_G) for several 2-step and exhaustive scan methods, with 3,500 cases and 3,500 controls (see Table 1 for additional details).

A) Moderate interaction with common G and E ($OR_{G \times E}=1.5$, $q_A=0.23$, $p_E=0.40$)

B) Strong interaction with less common G and E ($OR_{G \times E}=2.0$, $q_A=0.14$, $p_E=0.10$)

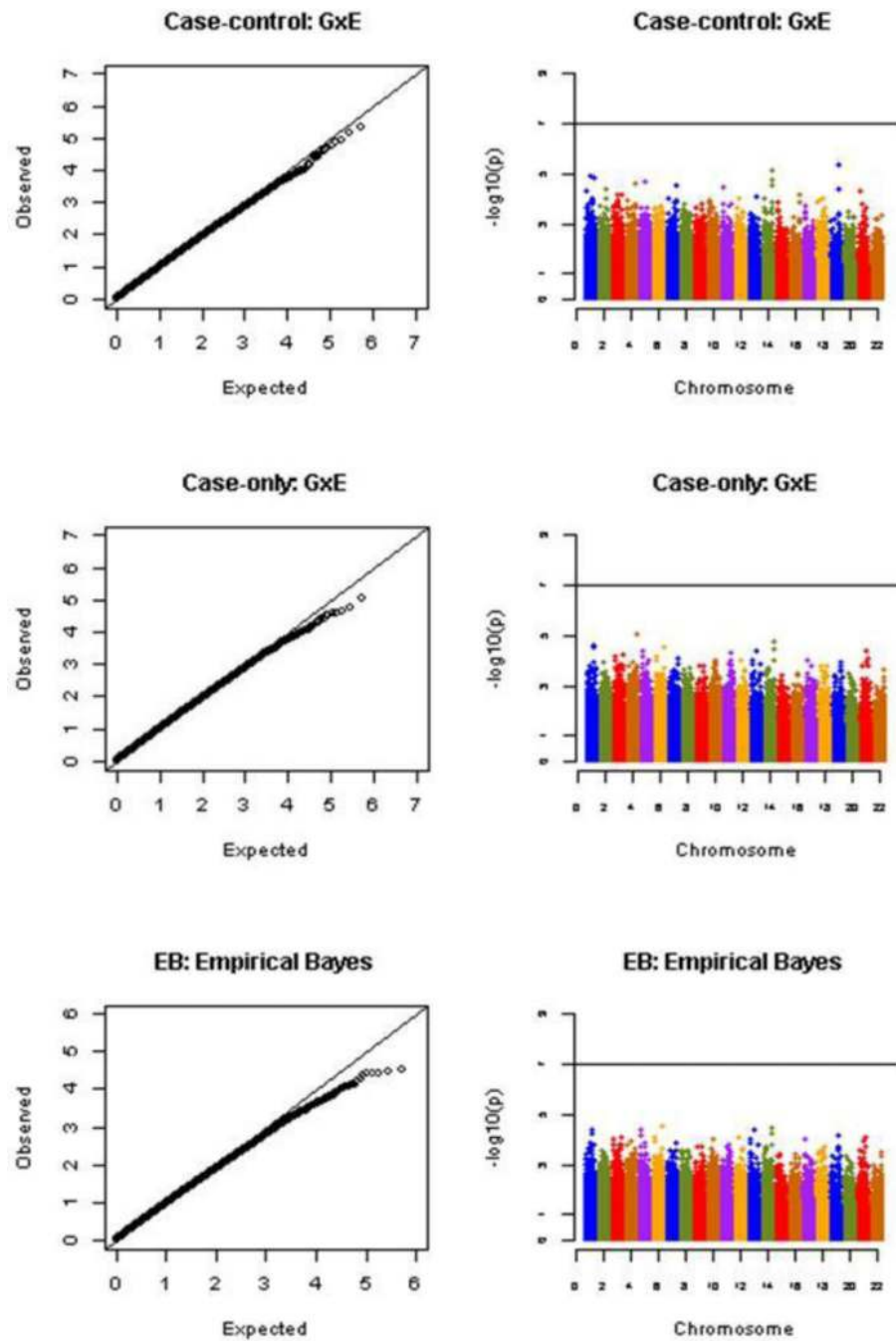


Figure 2. Quantile-quantile and Manhattan plots for case-control (CC), case-only (CO), and empirical-Bayes (EB) analysis of 536,857 SNPs for $G \times \text{Sex}$ interaction with young-onset childhood asthma.

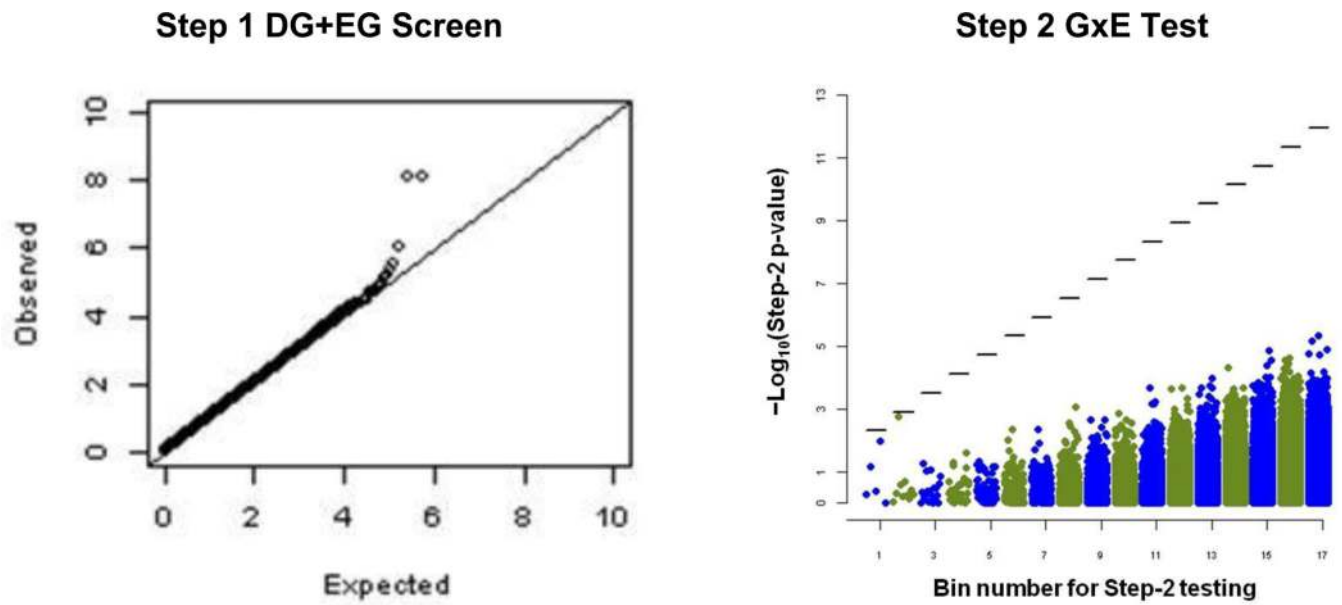


Figure 3.

Analysis of 536,857 SNPs for $G \times \text{Sex}$ interaction with young-onset childhood asthma using the EDGxE method. Shown are the QQ plot from Step 1 and the Manhattan plot from weighted Step-2 testing, with testing-bin assignment in Step 2 determined by Step-1 screening p-value. The initial bin size is $B=5$ SNPs, and each successive bin includes twice as many SNPs as the preceding bin.

















Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1
Models of GxE interaction that produce the indicated marginal genetic effect size (OR_G)

OR_G	$OR_{G \times E} = 1.5$			$OR_{G \times E} = 2.0$		
	$OR_{G E=0}$	$OR_{G E=1}$	Type	$OR_{G E=0}$	$OR_{G E=1}$	Type
1.00	0.83	1.27		0.91	1.84	
1.05	0.88	1.33		0.95	1.92	
1.10	0.92	1.39		1.00	1.99	
1.15	0.96	1.44		1.05	2.08	
1.20	1.00	1.50		1.09	2.15	
1.25	1.04	1.55		1.14	2.23	
1.30	1.08	1.61		1.18	2.30	
1.35	1.12	1.67		1.23	2.38	

Exposure-specific genetic odds ratios ($OR_{G|E}$) that produce the indicated marginal OR_G , for a moderate ($OR_{G \times E}=1.5$, with $q_A=0.23$ and $pE=0.4$) and strong ($OR_{G \times E}=2.0$, with $q_A=0.14$ and $pE=0.10$) interaction. The Type shows the pattern of risk for carriers ($G=1$, dashed line) relative to noncarriers (solid line) in subjects with $E=0$ (left side of each graph) or $E=1$ (right side).

Table II

Type 1 error rates for tests of GxE interaction across several methods

Model	Exhaustive Scans				2-Step Methods			
	CC	CO	EB	DG EB	EG GxE	H2	Cocktail	EDGxE
Base*	0.041	0.057	0.046	0.046	0.052	0.045	0.048	0.045
Population G-E Association&								
10 SNPs	0.042	1.000	0.042	0.048	0.048	0.045	0.050	0.046
50 SNPs	0.042	1.000	0.048	0.045	0.044	0.048	0.046	0.051
G-D Association&&								
10 SNPs	0.039	0.037	0.038	0.042	0.046	0.047	0.044	0.048
Both G-E and G-D Assoc								
10 G-E SNPs, 10 G-D SNPs	0.049	1.000	0.046	0.040	0.048	0.042	0.048	0.050
50 G-E SNPs, 10 G-D SNPs	0.049	1.000	0.046	0.043	0.046	0.048	0.044	0.049

Each estimate of Type I error is based on the proportion of 2,000 replicate datasets for which the indicated procedure identified at least one statistically significant result among 999,999 non-DSL SNPs. See Methods for description of each approach.

* The Base model is described in Methods and has no SNPs with G-E or G-D association

& Number of non-DSL SNPs simulated to have a population-level association with E

&& Number of non-DSL SNPs simulated to have an effect on disease risk (but no GxE interaction)

Table III

Power to detect the Disease Susceptibility Locus (DSL) across a range of models and methods

Model*	Exhaustive Scans				2-Step Methods			
	CC	CO	EB	DG EB	EG GxE	H2	Cocktail	EDGxE
Base**	0.093	0.679	0.400	0.662	0.629	0.683	0.703	0.847
q_A								
	0.15	0.051	0.523	0.259	0.518	0.473	0.521	0.727
	0.30	0.100	0.661	0.381	0.654	0.608	0.676	0.845
p_E								
	0.25	0.036	0.474	0.233	0.546	0.412	0.499	0.698
	0.50	0.098	0.675	0.388	0.668	0.621	0.686	0.839
Marginal OR_E								
	1.5	0.097	0.689	0.381	0.657	0.693	0.724	0.856
	1.8	0.106	0.666	0.382	0.626	0.710	0.714	0.852
Bin size (B)								
	10	0.096	0.666	0.380	0.669	0.613	0.673	0.811
	20	0.099	0.681	0.372	0.688	0.626	0.662	0.800
DSL-E Association***								
	OR _{DSL-E} =0.8	0.086	0.003	0.031	0.234	0.004	0.309	0.190
	OR _{DSL-E} =1.2	0.084	0.999	0.151	0.568	0.908	0.867	0.908
G-E Association[†]								
	10 SNPs	0.097	NA	0.365	0.654	0.558	0.630	0.747
	50 SNPs	0.093	NA	0.394	0.648	0.435	0.547	0.600
G-D Association[‡]								
	10 SNPs	0.088	0.683	0.400	0.638	0.626	0.681	0.831

Each estimate of power is based on the proportion of 2,000 replicate datasets for which the indicated procedure achieved statistical significance for GxE interaction at the DSL

* Each model varies the indicated parameter from the Base model setting. All results are based on a sample size of 3,500 cases and 3,500 controls and a total of 1 million SNPs

** The Base model has q_A=0.23, p_E=0.40, OR_{GxE}=1.5, OR_G=1.2, OR_E=1.2, Initial Bin size B=5 SNPs, no DSL-E association, and no additional SNPs with G-E or G-D association

*** Odds ratio between the DSL and E in the population

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

NA Not applicable since Type I error rate is inflated
Number of non-DSL SNPs simulated to have a population-level association with E
Number of non-DSL SNPs simulated to have an effect on disease risk (but no GxE interaction)

Table IV

Distribution of the DSL rank based on the Step-1 screening test from the EDGxE and cocktail methods

Model *	Method	Mean Rank **	Distribution of Step-2 Testing Bins ***					
			1	2	3	4	5	>=6
Base	EDGxE	2.3	82%	5%	4%	2%	2%	5%
	Cocktail	42.4	30%	9%	9%	8%	8%	36%
Marginal OR_G								
1.1	EDGxE	21.0	44%	10%	8%	6%	6%	26%
	Cocktail	62.2	30%	8%	6%	7%	7%	42%
1.3	EDGxE	1.1	98%	1%	1%	0%	0%	0%
	Cocktail	3.3	73%	8%	5%	3%	3%	8%
DSL-E Association								
OR _{DSL-E} =0.8	EDGxE	1,011	7%	5%	5%	5%	6%	72%
	Cocktail	2,593	9%	6%	5%	6%	6%	68%
OR _{DSL-E} =1.2	EDGxE	1.5	100%	0%	0%	0%	0%	0%
	Cocktail	17.4	47%	9%	7%	6%	7%	24%
G-E Association								
10 SNPs	EDGxE	17.4	0%	80%	9%	3%	2%	6%
	Cocktail	91.0	0%	28%	14%	10%	10%	38%
50 SNPs	EDGxE	63.9	0%	0%	0%	88%	5%	7%
	Cocktail	178.5	0%	0%	0%	43%	16%	41%
G-D Association								
10 SNPs	EDGxE	5.2	73%	12%	5%	4%	2%	4%
	Cocktail	62.5	18%	17%	10%	9%	8%	38%

* See Table 3 for a description of Models

** Geometric mean across replicate simulations of the DSL rank based on Step 1.
Total number of SNPs = 1 million

*** Bin 1 includes the 5 top ranked SNPs based on Step 1 p-value, Bin 2 includes ranks 6 – 15, Bin 3 16–35, Bin 4 36–75, Bin 5 76–155, and the last column ranks 156 - 1 million

Table VPower using subset testing for EDGxE method across a range of α_1 thresholds

Model*	α_1 : Expected # of markers passing to step 2**						Weighted testing ***
	0.05; 50,000	0.01; 10,000	0.001; 1,000	0.0001; 100	0.00001; 10		
Base	0.224	0.321	0.516	0.695	0.782		0.847
Marginal OR_G							
1.1	0.218	0.312	0.439	0.509	0.456		0.623
1.3	0.235	0.344	0.543	0.743	0.884		0.894
DSL-E Association							
OR _{DSL-E} = 0.8	0.191	0.227	0.235	0.182	0.098		0.258
OR _{DSL-E} = 1.2	0.234	0.332	0.517	0.743	0.899		0.908
G-E Association							
10 SNPs	0.224	0.328	0.511	0.666	0.727		0.747
50 SNPs	0.230	0.343	0.530	0.669	0.666		0.600
G-D Association							
10 SNPs	0.221	0.335	0.526	0.706	0.746		0.831

* See Table 3 for a description of models

**

*** α_1 is the significance threshold used for the Step 1 screen. The expected number of SNPs passing to Step 2 is based on screening 1 million SNPs in Step 1

Values in this column are the same as those shown in Table 3 for the EDGxE method

Table VI

Top 15 SNPs from DG+EG|GxE analysis of 536,857 SNPs for G \times Sex interaction with young-onset childhood asthma

Chr	SNP	Location	Reference Allele	Step 1			Step 2			Significance Threshold
				Chi-square	p-value	Bin	OR _{G×E}	t-test	p-value	
1	rs6697552	241192975	T	37.48	7.3E-09	1	1.13	0.64	0.52	0.005
1	rs1832719	200713649	C	37.23	8.2E-09	1	0.42	-1.81	0.07	0.005
7	rs1229492	81402058	T	27.88	8.8E-07	1	0.88	-0.82	0.41	0.005
4	rs6842542	158594467	C	25.50	2.9E-06	1	1.60	2.55	0.011	0.005
9	rs520613	109925873	G	24.79	4.1E-06	1	1.01	0.04	0.97	0.005
4	rs719525	76578274	C	24.18	5.6E-06	2	0.98	-0.11	0.91	0.0012
5	rs10069175	21923777	C	23.57	7.6E-06	2	1.15	0.70	0.48	0.0012
8	rs7000310	119837792	C	22.91	1.1E-05	2	0.57	-3.13	0.0017	0.0012
8	rs10505105	108376513	T	22.61	1.2E-05	2	0.82	-1.11	0.27	0.0012
9	rs630965	109925300	G	22.14	1.6E-05	2	1.07	0.53	0.59	0.0012
13	rs1988388	52944609	T	21.75	1.9E-05	2	1.19	1.26	0.21	0.0012
9	rs2767777	125998894	C	21.69	2.0E-05	2	0.91	-0.59	0.56	0.0012
9	rs865686	109928299	C	21.67	2.0E-05	2	1.05	0.34	0.74	0.0012
12	rs4765748	3724788	A	21.66	2.0E-05	2	0.84	-0.87	0.38	0.0012
15	rs1523526	59051579	C	21.31	2.4E-05	2	0.92	-0.62	0.53	0.0012