



Review of Statistical Methods for Gene-Environment Interaction Analysis

Summer S. Han^{1,2} · Nilanjan Chatterjee³

Published online: 12 February 2018
© Springer International Publishing AG, part of Springer Nature 2018

Abstract

Purpose of Reviews Complex diseases are caused by a combination of genetic and environmental factors, creating a challenge for understanding the disease mechanisms. Understanding the interplay between genes and environmental factors is important, as genes do not operate in isolation but rather in complex networks and pathways influenced by environmental factors. The advent of new technologies has made a massive amount of genetic data available, and various statistical methods have been developed to analyze genetic data and to identify interactions between genes and the environment, i.e., gene-environment (G-E) interactions.

Recent Findings In this review article, we introduce various statistical methods for identifying G-E interactions using case-control designs. We review a range of disease risk models for modeling the joint effects of genetic and environmental factors such as multiplicative and additive models. We then introduce various inference methods under these disease risk models, which include a standard prospective likelihood, case-only designs, a retrospective likelihood that exploits a gene-environment independence assumption to boost power, and an empirical Bayes type approach that uses the independence assumption in a data-adaptive way. Several tests for detecting genetic associations in the presence of G-E interactions are also introduced, which include a joint test and a maximum score test that provides a unified approach by integrating a class of disease risk models to maximize over a class of score tests.

Summary There are several challenges of G-E interaction analysis that include replication issues. While more powerful statistical methods for detecting interactions are helpful, ultimately studies with larger sample sizes are needed to identify interactions through consortium-based studies to achieve adequate power for G-E analysis.

Keywords Gene-environment interaction · GxE interaction · Complex diseases · Gene-environment independence · Retrospective likelihood · Empirical Bayes type estimator

This article is part of the Topical Collection on *Genetic Epidemiology*

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s40471-018-0135-2>) contains supplementary material, which is available to authorized users.

✉ Nilanjan Chatterjee
nchatte2@jhu.edu

- ¹ Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA
- ² Department of Neurosurgery, Stanford University School of Medicine, Stanford, CA 94305, USA
- ³ Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Johns Hopkins University, 615 N. Wolfe Street Room E3612, Baltimore, MD 21205, USA

Introduction

Complex diseases are caused by a combination of genetic and environmental factors, creating a challenge for understanding the disease mechanisms. Understanding the interplay between genes and environmental factors is important, as genes do not operate in isolation but rather in complex networks and pathways influenced by environmental factors. In addition to providing insights into disease etiology, exploiting gene-environment (G-E) interaction can help discover novel susceptibility loci for complex diseases, where genetic effects are modified and masked by the effects of environmental factors. Therefore, evaluating the main effects of a gene without considering its interaction with environmental factors can miss true association signals [1–3]. From a public health perspective, G-E interaction is useful because findings based on interactions can help develop strategies for targeted

intervention; conducting an intervention focusing on a subset of the population identified by G-E interactions can provide efficiency in disease prevention [4•, 5•].

Although G-E interaction has various meanings in epidemiology, it can be generally defined as a joint effect of genetic and environmental risk factors that cannot be explained by their separate marginal effects [6]. The recent advent of new technologies has made a massive amount of genetic data available, and various statistical methods have been developed to analyze genetic data and to identify G-E interactions. These methods include approaches that exploit additional assumptions such as G-E independence to improve power such as case-only analysis, retrospective likelihood-based analysis as well as empirical based estimators, methods that incorporate alternative disease risk models such as additive models, and tests for identifying interactions between rare variants and exposures based on exome or whole genome sequencing data. Various software packages have been also developed, which can be used to apply newly developed statistical methods for detecting G-E interactions.

The purpose of this article is to introduce recently developed statistical methods for evaluating G-E interactions across various complex diseases. While several study designs are available for examining G-E interactions such as prospective cohort studies, case-control studies, and family studies designs, we will focus on case-control studies that are mostly commonly used for genome-wide association studies (GWAS). In this review, we will first discuss the statistical models for joint effects of genetic and environmental factors and then introduce various statistical inferences methods under these models such as methods based on prospective and retrospective likelihoods as well as empirical base type approaches. We will then introduce statistical approaches that test for genetic associations in the presence of G-E interactions, various methods for two-stage analyses for GWAS, and methods for identifying interactions between rare variants and environmental exposures. We finalize this article with current challenges and future directions for analyzing G-E interactions.

Statistical Models for GxE and Interpretations of Interactions

There are several disease risk models for the joint effects of G and E, and interpretations of G-E interactions depend on the underlying disease risk models. A multiplicative model is one of the most commonly used models via logistic regression: $\text{logit}(\Pr(D = 1|G, E)) = \beta_0 + \beta_G G + \beta_E E$, where G is a genotype of a single nucleotide polymorphism (SNP), E is an environmental risk factor, and D is the disease status. Depending on the assumed genetic model, G can be coded for an additive genetic model (i.e., the number of the variant allele), dominant model (i.e., 1 for variant allele carriers and 0 for non-carriers),

or recessive model (i.e., 1 if one carries two copies of the variant allele and 0 otherwise). A departure from this model is called a multiplicative interaction, which can be tested by $H_0 : \beta_{GE} = 0$ in the following saturated model:

$$\text{logit}(\Pr(D = 1|G, E)) = \beta_0 + \beta_G G + \beta_E E + \beta_{GE} GE \quad (1)$$

Assuming binary factors for both G and E, a 2×2 table for a disease risk for each combination of G and E values can be constructed based on this model (see Table 1). Assuming a rare disease (i.e., relative risks can be approximated by odds ratios), “no multiplicative interaction” implies that the genetic effects measured by the ratios of the risks (e.g., $\frac{R_{10}}{R_{00}}$ for $E = 0$) are the same across different exposure levels with the null hypothesis of $H_0 : \frac{R_{10}}{R_{00}} = \frac{R_{11}}{R_{01}}$. On the other hand, an additive model is shown as $\text{logit}(\Pr(D = 1|G, E)) = b_0 + b_G G + b_E E$, where the effects of G and E are additive on the disease risk scale, but not on the logit scale. An additive interaction is defined by the departure from this model, which implies that the genetic effects measured as the differences of absolute risks (e.g., $R_{10} - R_{00}$ for $E = 0$) vary by exposure levels with the corresponding null hypothesis of $H_0 : R_{10} - R_{00} = R_{11} - R_{01}$. A number of researchers have shown that conceptual models for biologic interactions translate to the presence of interaction on the additive scale and not necessarily on the multiplicative scale [7]. In public health, evaluation of risk differences and additive interactions is directly relevant to problems such as whether it is beneficial to target individuals for intervention for an exposure based on genetic susceptibility [2, 8]. In addition to these multiplicative and additive models, there are some other non-standard models discussed in the literature [9, 10], including a liability threshold model, where the effects of G and E are additive on the probit scale [9].

Inferences: Methods for Testing G-E Interactions

Standard Prospective Likelihood-Based Approaches

Based on the models introduced in the previous section, several inference methods have been developed to test for G-E interactions. Standard analyses of case-control studies are typically based on a prospective likelihood of case-control data. While this approach does not take into account the retrospective nature of the sampling design, it is shown that such prospective

Table 1 Disease risk for binary factors G and E, where $R_{ij} = \Pr(D = 1|G = i, E = j)$ for $i, j = 0, 1$

	E = 0	E = 1
G = 0	R_{00}	R_{01}
G = 1	R_{10}	R_{11}

treatment of case-control data is valid when there is no assumption made about the joint distribution of covariates, including genetic and environmental factors and other confounders [11]. This likelihood can be used for both additive and multiplicative models, and several studies have used this approach for evaluating G-E interactions for various complex diseases [12•, 13, 14]. Figueroa et al. conducted a genome-wide interaction study of smoking for bladder cancer risk by applying both multiplicative and additive interactions based on a prospective likelihood and a retrospective likelihood [12•]. They identified 10 significant SNPs that interact with smoking status (ever versus never smokers) for bladder cancer; these included rs1711973 that had an increased risk (OR = 1.34; 95% confidence interval (CI): 1.2–1.5) among never smokers (multiplicative interaction $P = 6.38E-06$) and rs12216499 that had a reduced risk (OR = 0.75; CI: 0.67–0.84) for ever-smokers (additive interaction $P = 1.41E-06$). Multiplicative interactions based on a prospective likelihood can be tested using any statistical software package (e.g., SAS, R, or Stata). For example, in R, the `glm()` function can be used for testing a multiplicative interaction using a logistic regression based on a prospective likelihood. An R package, CGEN (<https://bioconductor.org/packages/release/bioc/html/CGEN.html>) implements the methods for both additive and multiplicative interaction based on a prospective likelihood; the `additive.test` function can be used for performing an additive interaction test and `snp.logistic` and `snp.score` for conducting a multiplicative interaction test (see Supplemental Fig. 1).

Case-Only Design

In evaluating G-E interactions, there have been several approaches that assume that G and E are independent in the underlying population. This assumption is plausible because the genetic variation an individual receives from a parent is determined during meiosis, and hence is not affected by subsequent environmental exposures after birth. Genetic susceptibility is unlikely to influence various exogenous exposures such as environmental pollutants or occupation exposures with some exceptions, whereas this assumption can become questionable for endogenous exposures, such as biomarkers. The case-only design is one of the non-traditional methods that depend on an assumption of G–E independence in the underlying population, which can be used to test for multiplicative interactions [15]. In brief, under the assumption of G-E independence in the underlying population (i.e., controls), a multiplicative interaction test statistic becomes equivalent to testing the association between G and E among cases. This method has been applied to the analyses of G-E interaction for various complex diseases [16, 17]. Freedman et al. used a case-only interaction test to evaluate the interaction between two independent genes, *FRMD3* and *MYH9* for end-stage renal disease risk. Any standard statistical software can be

used to conduct this test. For example, in R, a linear regression function (`lm` function) or generalized linear regression functions (`glm` function) can be used to evaluate an association between an environmental exposure and a genotype based on the data for controls. One major limitation of the case-only design is that while the case-only method has improved power over the traditional methods when G and E are independent in the underlying population, this method has an increased type I error if the independence assumption is violated [18]. In addition, the regression parameters for the main effects of G and E cannot be estimated using this method because the case-only test is only for evaluating a multiplicative interaction.

Retrospective Likelihood Approach

To address the limitations of case-only approaches that can only test for multiplicative interactions (not for the main effects of G and E), Umbach and Weinberg (1997) generalized the case-only design idea to use a log-linear model based on case-control data. They showed the maximum-likelihood estimates for all parameters of a logistic regression model can be obtained using a log-linear model [19]. Along the same line, Chatterjee and Carroll developed a general method using a retrospective likelihood that exploits the G-E independence assumption to test for multiplicative interaction, but can use both cases and controls to estimate all of the parameters in a general logistic regression model [20]. Basically, this method employs a retrospective likelihood that explicitly models the conditional probability of G given E mediated by an association parameter θ that can be constrained to be zero when the G-E independence assumption holds. This likelihood can be used for testing both multiplicative and additive interactions; recently, Han et al. developed a likelihood ratio test that exploits the G-E independence assumption using a retrospective likelihood [21•]. Their numerical investigation of power suggests that the incorporation of the independence assumption can enhance the efficiency of the test for additive interaction by 2- to 2.5-fold. The multiplicative and additive interaction tests based on a retrospective likelihood are implemented in the CGEN R package. The function `snp.score` can be used for testing a multiplicative interaction and the `additive.interaction` function (with an argument `indep = T`) can be used for testing an additive interaction using the G-E independence assumption.

Empirical Bayes Type Approaches

Despite the power gain using methods that rely on the G-E independence assumption—such as the case-only, log-linear, and retrospective methods—they can cause a large type I error when the underlying assumption is violated [18]. To address this issue, an empirical Bayes type method was

developed that uses a weighted average of the case-control and case-only estimators of the multiplicative interaction, which yields an acceptable trade-off between bias and efficiency [22•]. A stochastic framework is used to allow for uncertainty around the G-E independence assumption, which estimates the uncertainty parameter using data. The empirical-Bayes type estimator is provided as follows: $\hat{\beta}_{EB} = \frac{\hat{\sigma}_{CC}^2}{(\hat{\sigma}_{GC}^2 + \hat{\sigma}_{CC}^2)} \hat{\beta}_{CO} + \frac{\hat{\sigma}_{CC}^2}{(\hat{\sigma}_{GC}^2 + \hat{\sigma}_{CC}^2)} \hat{\beta}_{CC}$, where $\hat{\beta}_{CO}$ is a case-only estimator and $\hat{\beta}_{CC}$ is a case-control estimator for a multiplicative interaction, respectively. Here, $\hat{\theta}_{GE}$ is the measure of the G-E association among controls and $\hat{\sigma}_{CC}^2$ is the estimated variance of the case-control estimator. The intuition is that $\hat{\theta}_{GE}$ is a measure of the bias of the case-only method, and the empirical Bayes method provides more weight to the case-control method when this bias is large. How much weight will be given is calibrated by $\hat{\sigma}_{CC}^2$, which is the variance of the less efficient case-control estimator. If the G-E independence assumption is violated, i.e., true $\hat{\theta}_{GE} = 0$, then the empirical Bayes estimator will asymptotically behave the same as the case-control estimator. However, when G-E independence holds, the asymptotic weight for the empirical Bayes estimator will be non-zero for both case-control and case-only estimators and thus will have efficiency in between. A general approach for deriving empirical Bayes-type shrinkage estimators was also proposed for all of the parameters of a general logistic regression model [22•, 23], which is implemented in the CGEN R package. The empirical Bayes type estimator for an additive interaction was also developed in the general regression setting [24].

Testing for Genetic Association in the Presence of G-E Interaction

When identifying susceptibility loci for complex diseases, allowing for interactions to test for association could increase power when such interactions exist. It has been shown that a joint test of genetic association and interaction has robust performance over a wide range of underlying models [25•], although it could be less powerful than a marginal association test when there is no evidence of G-E interaction. Using the equation in (Eq. 1), the null hypothesis of the joint test is given as $H_0: \beta_G = \beta_{GE} = 0$, which has increased degrees of freedom compared to a marginal association test (i.e. $H_0: \beta_G = 0$) that can lead to a decrease in power when there is no interaction effect, i.e., $\beta_{GE} = 0$. Various likelihoods with or without the assumption of G-E independence can be used for joint tests. Recently, Hamza et al. conducted a genome-wide joint test for gene x coffee interaction for Parkinson's disease and identified a novel susceptibility locus in the *GRIN2A*

gene. In the gene, the T allele of the SNP rs4998386 is associated with a reduced risk among heavy coffee drinkers, whereas this variant has a minimal effect among light coffee drinkers [26•]. While a joint test can be powerful when the assumed interaction exists, the increased degrees of freedom of this test (versus a marginal association test) can lead to a reduced power when such interaction effects are relatively small or when these effects do not exist. A maximum score test was developed to overcome the potential loss of power of a joint test due to increased degrees of freedom [10]. This method provides a unified approach that integrates a class of disease risk models by maximizing over a class of score tests, each of which involves modified standard tests of genetic association through a weight function. This weight function reflects the potential heterogeneity of the genetic effects by levels of environmental exposures. Both joint test and maximum score test are implemented in the CGEN R package.

Two-Stage Analysis or G-E Interactions for Rare Variants

Several approaches have been proposed to conduct a two-stage analysis to improve the efficiency of detecting G-E interactions on a genome-wide scale [27–29]. In general, these methods suggest selecting a subset of SNPs based on the marginal effects of SNPs or G-E correlation tests in the first stage and conducting standard G-E interaction tests in the second stage, where the independence between the test statistics used in the two stages is required to provide a valid screening procedure. Applications of such methods are shown in a recent G-E analysis for colorectal cancer [30] that involves (i) a screening step based on marginal associations and gene-diet correlations and (ii) a testing step for multiplicative interactions. They identified a significant interaction between rs4143094 and processed meat consumption (OR = 1.17; $p = 8.7E-09$), which was consistently observed across studies. With the advent of high-throughput technologies, various statistical methods have been developed for identifying G-E interactions based on data for rare variants, generated by whole genome sequencing and exome sequencing [31–35]. A standard approach for this problem is a set-based G-E interaction framework that tests for an interaction between a set of rare variants and an environmental risk factor. Burden type tests [33, 36] and variance component tests [31] are available for analyzing G-E interaction in this framework. Some of these methods are implemented in the R packages *rareGE* (<https://www.hsph.harvard.edu/han-chen/software/>) and SIMreg (http://www4.stat.ncsu.edu/~jytzeng/software_simreg.php).

Software Available for G-E Analysis

There are several software packages that provide tools for conducting G-E interactions using the methods described in this review. The CGEN R package provides various functions that can conduct tests for multiplicative and additive interactions, joint tests, as well as maximum score tests under both prospective and retrospective likelihoods assuming the G-E independence assumption (<https://bioconductor.org/packages/release/bioc/html/CGEN.html>). The empirical Bayes type method for multiplicative interaction is also implemented in CGEN. The rareGE R package (<https://www.hsph.harvard.edu/han-chen/software/>) provides various functions for detecting G-E interaction as well as for testing the joint effect of a gene and G-E interaction under a set-based framework. The SIMreg R package (http://www4.stat.ncsu.edu/~jytzeng/software_simreg.php) offers functions for testing a set-based G-E interaction by using genetic similarity to aggregate information across SNPs, and incorporating adaptive weights depending on allele frequencies to accommodate rare and common variants. For calculating power for G-E interactions, the powerGWASinteraction R package is available (<https://cran.r-project.org/web/packages/powerGWASinteraction/index.html>), which includes a power calculation tool for four two-stage screening and testing procedures. Several studies compared the power of various G-E interaction tests including standard prospective likelihood approaches, case-only designs, retrospective likelihood methods, empirical Bayes-type estimators, and two-stage analyses [37, 38].

Challenges for G-E Analysis and Future Directions

There are several challenges of G-E interaction analysis. One main challenge is replication issues. While various GWAS findings of the main effects of SNPs have been replicated by independent studies for many complex diseases (<http://www.ebi.ac.uk/gwas/>), relatively few interactions have been reproduced. It is likely that the sample sizes of GWAS that have required measurements on environmental exposures are not yet adequate to reliably identify G-E interactions of modest magnitude. In addition, differences in the underlying distribution of environmental exposures across various studies as well as difficulties in accurately measuring environmental exposures can also lead to reduced power of detecting G-E interactions. While more powerful statistical methods for detecting interactions are helpful, ultimately studies with larger sample sizes are needed to identify interactions (e.g., through consortium-based studies) to achieve adequate power for G-E analysis. A reasonable goal for the future will be to at least identify parsimonious models that adequately describe the

risks of diseases associated with a combination of genetic and environmental risk factors. The lack of reporting of interaction in current studies so far indicates that linear logistic models, i.e., multiplicative models, in general may be a good starting point for building models for evaluating the joint effects of genetic and environmental factors [39].

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

Human and Animal Rights and Informed Consent This article does not contain any studies with human or animal subjects performed by any of the authors.

References

Papers of particular interest, published recently, have been highlighted as:

- Of importance
- Of major importance

1. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet.* 2010;11(6):446–50. <https://doi.org/10.1038/nrg2809>.
2. Thompson WD. Effect modification and the limits of biological inference from epidemiologic data. *J Clin Epidemiol.* 1991;44(3):221–32. [https://doi.org/10.1016/0895-4356\(91\)90033-6](https://doi.org/10.1016/0895-4356(91)90033-6).
3. Ottman R. Gene–environment interaction: definitions and study design. *Prev Med.* 1996;25(6):764–70. <https://doi.org/10.1006/pmed.1996.0117>.
4. Hutter CM, Mechanic LE, Chatterjee N, Kraft P, Gillanders EM. Gene-environment interactions in cancer epidemiology: a National Cancer Institute Think Tank Report. *Genet Epidemiol.* 2013;37(7):643–57. **Summarizes contemporary analytic methods for $G \times E$ interactions, provides an overview of motivation for performing $G \times E$ analysis, and discusses key considerations for analysis in case-control or nested case-control studies, and comments on interpretation of $G \times E$ interactions.** <https://doi.org/10.1002/gepi.21756>.
5. Maas P, Barrdahl M, Joshi AD, Auer PL, Gaudet MM, Milne RL, et al. Breast cancer risk from modifiable and nonmodifiable risk factors among white women in the United States. *JAMA Oncol.* 2016;2(10):1295–302. **Evaluates combined risk stratification utility of common low penetrant single nucleotide polymorphisms (SNPs) and epidemiologic risk factors for breast cancer. Their model for absolute risk of breast cancer including SNPs can provide stratification for the population of white women in the United States and also can identify subsets of the population at an elevated risk that would benefit most from risk-reduction strategies based on altering modifiable factors.** <https://doi.org/10.1001/jamaoncol.2016.1025>.
6. Thomas D. Gene-environment-wide association studies: emerging approaches. *Nat Rev Genet.* 2010;11(4):259–72. <https://doi.org/10.1038/nrg2764>.
7. Cordell HJ. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet.* 2002;11(20):2463–8. <https://doi.org/10.1093/hmg/11.20.2463>.

8. Lund E. Comparison of additive and multiplicative models for reproductive risk factors and post menopausal breast cancer. *Stat Med.* 1995;14(3):267–74. <https://doi.org/10.1002/sim.4780140305>.
9. Clayton D. Link functions in multi-locus genetic models: implications for testing, prediction, and interpretation. *Genet Epidemiol.* 2012;36(4):409–18. <https://doi.org/10.1002/gepi.21635>.
10. Han SS, Rosenberg PS, Ghosh A, Landi MT, Caporaso NE, Chatterjee N. An exposure-weighted score test for genetic associations integrating environmental risk factors. *Biometrics.* 2015.
11. Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika.* 1979;66(3):403–11. <https://doi.org/10.1093/biomet/66.3.403>.
12. Figueroa JD, Han SS, Garcia-Closas M, Baris D, Jacobs EJ, Kogevinas M, et al. Genome-wide interaction study of smoking and bladder cancer risk. *Carcinogenesis.* 2014;bgu064. **Conducted a genome-wide interaction study of smoking for bladder cancer risk by applying both multiplicative and additive interactions based on a prospective likelihood and a retrospective likelihood. They identified 10 significant SNPs that interact with smoking status (ever versus never smokers) for bladder cancer; these included rs1711973 that had an increased risk (OR=1.34; 95% confidence interval (CI): 1.2–1.5) among never smokers (multiplicative interaction P= 6.38E-06) and rs12216499 that had a reduced risk (OR=0.75; CI: 0.67–0.84) for ever-smokers (additive interaction P= 1.41E-06).**
13. Joshi AD, Lindström S, Hüsing A, Barrdahl M, VanderWeele TJ, Campa D, et al. Additive interactions between susceptibility single-nucleotide polymorphisms identified in genome-wide association studies and breast cancer risk factors in the breast and prostate cancer cohort consortium. *Am J Epidemiol.* 2014;180(10):1018–27. <https://doi.org/10.1093/aje/kwu214>.
14. Garcia-Closas M, Rothman N, Figueroa JD, Prokunina-Olsson L, Han SS, Baris D, et al. Common genetic polymorphisms modify the effect of smoking on absolute risk of bladder cancer. *Cancer Res.* 2013;73(7):2211–20. <https://doi.org/10.1158/0008-5472.CAN-12-2388>.
15. 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. <https://doi.org/10.1002/sim.4780130206>.
16. Freedman BI, Langefeld CD, Lu L, Divers J, Comeau ME, Kopp JB, et al. Differential effects of MYH9 and APOL1 risk variants on FRMD3 association with diabetic ESRD in African Americans. *PLoS Genet.* 2011;7(6):e1002150. <https://doi.org/10.1371/journal.pgen.1002150>.
17. Neslund-Dudas C, Levin AM, Rundle A, Beebe-Dimmer J, Bock CH, Nock NL, et al. Case-only gene–environment interaction between ALAD tagSNPs and occupational lead exposure in prostate cancer. *Prostate.* 2014;74(6):637–46. <https://doi.org/10.1002/pros.22781>.
18. Albert PS, Ratnasিংhe D, Tangrea J, Wacholder S. Limitations of the case-only design for identifying gene-environment interactions. *Am J Epidemiol.* 2001;154(8):687–93. <https://doi.org/10.1093/aje/154.8.687>.
19. Umbach DM, Weinberg CR. Designing and analysing case control studies to exploit independence of genotype and exposure. *Stat Med.* 1997;16(15):1731–43. [https://doi.org/10.1002/\(SICI\)1097-0258\(19970815\)16:15<1731::AID-SIM595>3.0.CO;2-S](https://doi.org/10.1002/(SICI)1097-0258(19970815)16:15<1731::AID-SIM595>3.0.CO;2-S).
20. Chatterjee N, Carroll RJ. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika.* 2005;92(2):399–418. <https://doi.org/10.1093/biomet/92.2.399>.
21. Han SS, Rosenberg PS, Garcia-Closas M, Figueroa JD, Silverman D, Chanock SJ, et al. Likelihood ratio test for detecting gene (G)-environment (E) interactions under an additive risk model exploiting GE independence for case-control data. *Am J Epidemiol.* 2012;176(11):1060–7. **Developed a likelihood ratio test for detecting additive interactions for case-control studies that incorporates the G-E independence assumption based on a retrospective likelihood. Numerical investigation of power suggests that incorporation of the independence assumption can enhance the efficiency of the test for additive interaction by 2- to 2.5-fold. The authors illustrate their method by applying it to data from a bladder cancer study.** <https://doi.org/10.1093/aje/kws166>.
22. 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):685–94. **Proposed a novel empirical Bayes-type shrinkage estimator to analyze case–control data that can relax the gene-environment independence assumption in a data-adaptive fashion. They also described a general approach for deriving the new shrinkage estimator and its variance within the retrospective maximum-likelihood framework developed by Chatterjee and Carroll (2005). Both simulated and real data examples suggested that the proposed estimator strikes a balance between bias and efficiency depending on the true nature of the gene-environment association and the sample size for a given study.** <https://doi.org/10.1111/j.1541-0420.2007.00953.x>.
23. Chen YH, Chatterjee N, Carroll RJ. Shrinkage estimators for robust and efficient inference in haplotype-based case-control studies. *J Am Stat Assoc.* 2009;104(485):220–33. <https://doi.org/10.1198/jasa.2009.0104>.
24. Liu G, Lee S, Lee AW, Wu AH, Bandera EV, Jensen A, et al. Robust tests for additive gene-environment interaction in case-control studies using gene-environment independence. *Am J Epidemiol.* 2017.
25. Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ. Exploiting gene-environment interaction to detect genetic associations. *Hum Hered.* 2007;63(2):111–9. **Present a joint test of marginal association and gene-environment interaction for case-control data. They compared the power and sample size requirements of this joint test to other analyses: the marginal test of genetic association, the standard test for gene-environment interaction based on logistic regression, and the case-only test for interaction that exploits gene-environment independence.** <https://doi.org/10.1159/000099183>.
26. Hamza TH, Chen H, Hill-Burns EM, Rhodes SL, Montimurro J, Kay DM, et al. Genome-wide gene-environment study identifies glutamate receptor gene GRIN2A as a Parkinson's disease modifier gene via interaction with coffee. *PLoS Genet.* 2011;7(8):e1002237. **Conducted a genome-wide joint test for gene x coffee interaction for Parkinson's disease and identified a novel susceptibility locus in the GRIN2A gene. In the gene, the T allele of the SNP rs4998386 is associated with a reduced risk among heavy coffee drinkers, whereas this variant has a minimal effect among light coffee drinkers.** <https://doi.org/10.1371/journal.pgen.1002237>.
27. 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. <https://doi.org/10.1002/gepi.21610>.
28. Kooperberg C, LeBlanc M. Increasing the power of identifying gene× gene interactions in genome-wide association studies. *Genet Epidemiol.* 2008;32(3):255–63. <https://doi.org/10.1002/gepi.20300>.
29. Murcray CE, Lewinger JP, Gauderman WJ. Gene-environment interaction in genome-wide association studies. *Am J Epidemiol.* 2008;169(2):219–26. <https://doi.org/10.1093/aje/kwn353>.
30. Figueiredo JC, Hsu L, Hutter CM, Lin Y, Campbell PT, Baron JA, et al. Genome-wide diet-gene interaction analyses for risk of colorectal cancer. *PLoS Genet.* 2014;10(4):e1004228. <https://doi.org/10.1371/journal.pgen.1004228>.

31. Lin X, Lee S, Wu MC, Wang C, Chen H, Li Z, et al. Test for rare variants by environment interactions in sequencing association studies. *Biometrics*. 2016;72(1):156–64. <https://doi.org/10.1111/biom.12368>.
32. Jiao S, Peters U, Berndt S, Bézieau S, Brenner H, Campbell PT, et al. Powerful set-based gene-environment interaction testing framework for complex diseases. *Genet Epidemiol*. 2015;39(8):609–18. <https://doi.org/10.1002/gepi.21908>.
33. Liu Q, Chen LS, Nicolae DL, Pierce BL. A unified set-based test with adaptive filtering for gene-environment interaction analyses. *Biometrics*. 2016;72(2):629–38. <https://doi.org/10.1111/biom.12428>.
34. Su Y-R, Di C-Z, Hsu L. A unified powerful set-based test for sequencing data analysis of GxE interactions. *Biostatistics*. 2017;18(1):119–31. <https://doi.org/10.1093/biostatistics/kxw034>.
35. Zhao G, Marceau R, Zhang D, Tzeng J-Y. Assessing gene-environment interactions for common and rare variants with binary traits using gene-trait similarity regression. *Genetics*. 2015;199(3):695–710. <https://doi.org/10.1534/genetics.114.171686>.
36. Jiao S, Hsu L, Bézieau S, Brenner H, Chan AT, Chang-Claude J, et al. SBERIA: set-based gene-environment interaction test for rare and common variants in complex diseases. *Genet Epidemiol*. 2013;37(5):452–64. <https://doi.org/10.1002/gepi.21735>.
37. 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*. 2011;175(3):177–90. <https://doi.org/10.1093/aje/kwr367>.
38. Mukherjee B, Ahn J, Gruber SB, Rennert G, Moreno V, Chatterjee N. Tests for gene-environment interaction from case-control data: a novel study of type I error, power and designs. *Genet Epidemiol*. 2008;32(7):615–26. <https://doi.org/10.1002/gepi.20337>.
39. Chatterjee N, Shi J, Garcia-Closas M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat Rev Genet*. 2016.