Ji-Hyung Shin*, Claire Infante-Rivard, Brad McNeney and Jinko Graham

A data-smoothing approach to explore and test gene-environment interaction in case-parent trios

Abstract: Complex traits result from an interplay between genes and environment. A better understanding of their joint effects can help refine understanding of the epidemiology of the trait. Various tests have been proposed to assess the statistical interaction between genes and the environment (*G*×*E*) in case-parent trio data. However, these tests can lose power when the form of *G*×*E* departs from that for which the test was developed. To address this limitation, we propose a data-smoothing approach to estimate and test *G*×*E* between a single nucleotide polymorphism and a continuous environmental covariate. For estimating *G*×*E*, we fit a generalized additive model using penalized likelihood. The resulting point- and interval-estimates of *G*×*E* lead to a graphical display, which can serve as a visualization tool for exploring the form of interaction. For testing *G*×*E*, we propose a permutation approach, which accounts for the extra uncertainty introduced by the smoothing process. We investigate the statistical properties of the proposed methods through simulation. We also illustrate the use of the approach with an example data set. We conclude that the approach is useful for exploring novel interactions in data-rich settings.

Keywords: gene-environment interaction; case-parent trio study; generalized additive model; penalized likelihood estimation; permutation test.

***Corresponding author: Ji-Hyung Shin,** Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, British Columbia, Canada, e-mail: shin@sfu.ca
Claire Infante-Rivard: Faculty of Medicine, Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, Canada
Brad McNeney and Jinko Graham: Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, British Columbia, Canada

1 Introduction

Complex traits result from an interplay between genes and environment. For example, obesity is more highly associated with dietary fat intake in carriers than in non-carriers of the Pro12Ala allele in the PPAR-γ gene (e.g., Garaulet et al., 2011). Therefore, a better understanding of the joint effects of genes and environment can help refine understanding about the epidemiology of the trait. For dichotomous traits, one way to summarize these joint effects is in terms of genetic relative risks that vary as a function of the environmental risk factor. Throughout, we refer to this statistical interaction as *G*×*E*.

In the case-parent trio design, genotype information is collected from unrelated affected children and their parents. Information on environmental or non-genetic covariates can also be collected from the affected children. For studying early-onset diseases such as childhood leukemia, for which controls can be difficult to obtain, this design is attractive because there is no need to collect controls. In effect, family-based controls, matched to the case for ancestry, are created by conditioning on parent genotypes in the analysis. The conditioning on parents provides inference of genetic main effects that is robust against population stratification (e.g., Schaid and Sommer, 1993). It also protects against population stratification for inference of *G*×*E*, provided that the genotype is measured on a causal marker (Shi et al., 2011).

Various conditional approaches to inference of *G*×*E* have been proposed. These approaches include conditional logistic regression (e.g., Schaid, 1999; Cordell et al., 2004), family-based association test of gene-environment interaction (FBATI; Lake and Laird, 2004), and a log-linear modeling approach (Umbach and Weinberg, 2000). However, all these methods are tuned for a particular form of $G \times E$. Conditional logistic regression fits the functional form of $G \times E$ that is specified in the regression model. FBAT-I is designed to detect the linear component of $G \times E$ because its test statistic is a measure of linear association between the genetic and environmental risk factors. The log-linear modeling approach parametrizes $G \times E$ in terms of a dichotomized non-genetic covariate. These approaches lose resolution and power when the form of $G \times E$ departs from that for which the approach was tuned. As the underlying form of $G \times E$ is generally unknown, we present a data-driven method to explore it. We also present a test for $G \times E$ which is appropriate when nothing is known about the form of interaction.

The remainder of the paper is structured as follows. In Section 2, we introduce the disease model for our problem. In Section 3, we propose a data-driven approach to estimating $G \times E$ and a permutation approach to testing $G \times E$. In Section 4, we present a simulation study that evaluates the performance of the proposed approach. In Section 5, we illustrate our methods on an example data set. In Section 6, we summarize and discuss our findings.

2 Model

Consider a single-nucleotide polymorphism (SNP) that is *causal* for a binary trait D (D=1 affected). Let G denote the number of copies of the SNP index allele carried by an individual and E, his/her continuously varying non-genetic covariate. We assume mating symmetry for the parental genotype pairs such that, for example, a mother having no copies of an allele and a father having one copy has the same probability of occurring as a mother having one copy and the father having no copies. We also assume the SNP segregates from the parents to the child according to Mendel's laws, with no mutation. Under these assumptions, only parental mating types with at least one heterozygous parent lead to variability in G, and hence are *informative*. Denote the informative mating types by

$$G_p = \begin{cases} 1 & \text{if one parent is heterozygous, and one parent is homozygous} \\ & \text{for the non-index allele,} \\ 2 & \text{if one parent is heterozygous, and one parent is homozygous} \\ & \text{for the index allele,} \\ 3 & \text{if both parents are heterozygous.} \end{cases} \tag{1}$$

We assume that, within families, G and E are independent; i.e., they are conditionally independent given G_p.

We use genotype relative risks as the basis for inference of $G \times E$. A genotype relative risk (GRR) is the ratio of disease risks between individuals with one genotype and those with another genotype (Schaid and Sommer, 1993). We define $G \times E$ as GRRs that vary with the levels of E. For a SNP, $G \times E$ can be assessed through two GRRs which, for a fixed value of E=e, we define as

$$\mathrm{GRR}_h(e) \equiv \frac{P(D=1|G=h, E=e)}{P(D=1|G=h-1, E=e)} = \exp\{\gamma_h + f_h(e)\}, \quad h=1,2. \tag{2}$$

The smooth functions $f_1(e)$ and $f_2(e)$ allow for $G \times E$ because GRRs can vary with E=e when $f_1(e) \neq 0$ or $f_2(e) \neq 0$. Under standard identifiability constraints on $f_1(e)$ and $f_2(e)$ discussed in §2.2, γ_h is the sample average of $\log(\mathrm{GRR}_h(e))$ in trios informative for GRR_h. In the absence of $G \times E$ (i.e., $f_1(e)=f_2(e)\equiv 0$), $\log(\mathrm{GRR}_h(e)) \equiv \gamma_h$ for all E=e. We discuss which trios are informative for GRR_h in §2.1.

The parameterizations in (2) follow from a log-linear model of disease risk (e.g., Shin et al., 2010) in which

$$P(D=1|G=g, E=e) = \exp\{k + z_1(g)\gamma_1 + z_2(g)\gamma_2 + \xi(e) + z_1(g)f_1(e) + z_2(g)f_2(e)\}. \tag{3}$$

In this risk model, the dummy variables $z_1(g)$ and $z_2(g)$ implement codominant genetic coding for the genotype $G=g$, with $z_1(g)$ indicating $g>0$ and $z_2(g)$ indicating $g=2$. However, dominant, recessive and log-additive penetrance modes may be specified by setting $\gamma_2=0$ and $f_2\equiv 0$; $\gamma_1=0$ and $f_1\equiv 0$; and $\gamma_1=\gamma_2$ and $f_1\equiv f_2$, respectively. In the absence of $G\times E$, the nuisance parameters $\xi(e)$ and k describe the disease risk of an individual with $G=0$.

3 Methods

3.1 Likelihood

We adopt a likelihood approach to inference of $G\times E$. The full likelihood can be written as

$$P(G=g,\ E=e,\ G_p=m|D=1)=P(G=g|E=e,\ G_p=m,\ D=1)P(E=e,\ G_p=m|D=1). \tag{4}$$

In general, conditioning results in a loss of information. However, conditional inference of linear $G\times E$, based on $P(G=g|E=e,\ G_p=m,\ D=1)$, is asymptotically efficient for data from case-parent trios (Moerkerke et al., 2010). Hence, we expect minimal loss of information from conditioning when working with the large numbers of trios required to characterize $G\times E$.

To write the likelihood, we introduce a binary variable Y_{mjg} indicating whether $G=g$ in the j^{th} trio from the m^{th} mating type; i.e.,

$$Y_{mjg}=\begin{cases}1 & \text{if the child has } G=g,\\ 0 & \text{otherwise}\end{cases}.$$

Let $\mu_{mg}(e)\equiv E(Y_{mjg}|E=e,\ G_p=m,\ D=1)=P(G=g|E=e,\ G_p=m,\ D=1)$. Expressions for $\mu_{mg}(e)$ are presented in Table 1 (Duke, 2007). In calculating these expressions, the nuisance parameters k and $\xi(e)$ of the disease risk model (3) cancel out and hence are not estimable.

Let $\mu(e)=(\mu_{10}(e), \mu_{11}(e), \mu_{21}(e), \mu_{22}(e), \mu_{30}(e), \mu_{31}(e), \mu_{32}(e))^\top$. Then the log-likelihood $l(\mu(e))$ can be written as a sum of contributions from trios in the informative mating types G_p=1,2 and 3:

$$l(\mu(e))=l_1(\mu_{10}(e), \mu_{11}(e))+l_2(\mu_{21}(e), \mu_{22}(e))+l_3(\mu_{30}(e), \mu_{31}(e), \mu_{32}(e)), \tag{5}$$

where the mating-type-specific log-likelihoods $l_m(\cdot)$ are defined below. From the definition (1) of mating types, trios with $G_p=1$ have one parent heterozygous and one parent homozygous for the non-index allele; hence $G=2$ is impossible. Similarly, trios with $G_p=2$ have one parent heterozygous and one homozygous for the index allele, and hence $G=0$ is impossible. All genotype values are possible for trios with $G_p=3$. Therefore, $l_1(\cdot)$, $l_2(\cdot)$ are binomial log-likelihoods, and $l_3(\cdot)$ is a trinomial log-likelihood; specifically,

Table 1 Expressions for $\mu_{mg}(e)\equiv P(G=g|E=e,\ G_p=m,\ D=1)$.

Mating type (*m*)	**Genotype (*g*)**		
	0	**1**	**2**
1	$\frac{1}{1+\exp(\gamma_1+f_1(e))}$	$\frac{\exp(\gamma_1+f_1(e))}{1+\exp(\gamma_1+f_1(e))}$	–
2	–	$\frac{1}{1+\exp(\gamma_2+f_2(e))}$	$\frac{\exp(\gamma_2+f_2(e))}{1+\exp(\gamma_2+f_2(e))}$
3[a]	$\frac{1}{d(\gamma_1,\gamma_2,f_1(e),f_2(e))}$	$\frac{2\exp(\gamma_1+f_1(e))}{d(\gamma_1,\gamma_2,f_1(e),f_2(e))}$	$\frac{\exp(\gamma_1+f_1(e)+\gamma_2+f_2(e))}{d(\gamma_1,\gamma_2,f_1(e),f_2(e))}$

[a] $d(\gamma_1, \gamma_2, f_1(e), f_2(e))\equiv 1+2\exp(\gamma_1+f_1(e))+\exp(\gamma_1+f_1(e)+\gamma_2+f_2(e))$.

$$l_m(\cdot)=\begin{cases}\sum_{j=1}^{n_1}[y_{1j0}\log(\mu_{10}(e_{1j}))+y_{1j1}\log(\mu_{11}(e_{1j}))]\equiv\sum_{j=1}^{n_1}l_{1j}(\gamma_1,f_1(e_{1j})) & \text{if } m=1,\\ \sum_{j=1}^{n_2}[y_{2j1}\log(\mu_{21}(e_{2j}))+y_{2j2}\log(\mu_{22}(e_{2j}))]\equiv\sum_{j=1}^{n_2}l_{2j}(\gamma_2,f_2(e_{2j})) & \text{if } m=2,\\ \sum_{j=1}^{n_3}[y_{3j0}\log(\mu_{30}(e_{3j}))+y_{3j1}\log(\mu_{31}(e_{3j}))+y_{3j2}\log(\mu_{32}(e_{3j}))] & \\ \qquad\qquad\equiv\sum_{j=1}^{n_3}l_{3j}(\gamma_1,\gamma_2,f_1(e_{3j}),f_2(e_{3j})) & \text{if } m=3,\end{cases}\tag{6}$$

where n_m is the number of trios from the m^{th} mating type, and e_{mj} is the value of the non-genetic covariate for j^{th} trio from the m^{th} mating type.

Substituting the expressions for $\mu_{mg}(e)$ in Table 1 into the mating-type-specific log-likelihoods of equation (6) enables the log-likelihood (5) to be re-written as

$$l(\gamma_1,\gamma_2,f_1(e),f_2(e))=\sum_{j=1}^{n_1}l_{1j}(\gamma_1,f_1(e_{1j}))+\sum_{j=1}^{n_2}l_{2j}(\gamma_2,f_2(e_{2j}))+\sum_{j=1}^{n_3}l_{3j}(\gamma_1,\gamma_2,f_1(e_{3j}),f_2(e_{3j})).\tag{7}$$

Thus, trios from G_p=1, 3 are used to estimate γ_1 and $f_1(e)$, while trios from G_p=2, 3 are used to estimate γ_2 and $f_2(e)$.

3.2 Flexible estimation of GRR curves

To estimate GRR curves, we take a penalized likelihood approach under a generalized additive modeling framework (Wood, 2006). The penalized log-likelihood $l_p(\cdot)$ is defined from the log-likelihood in (7), by subtracting terms that penalize models for $f_h(e)$ that are too wiggly:

$$l_p(\gamma_1,\gamma_2,f_1(e),f_2(e))=l(\gamma_1,\gamma_2,f_1(e),f_2(e))-\frac{1}{2}\sum_{h=1}^{2}\lambda_h\int\{f_h''(e)\}^2de,\tag{8}$$

where λ_h, for h=1, 2, are smoothing parameters that control the tradeoff between goodness-of-fit in the first (likelihood) term and smoothness in the second (penalty) term.

To represent $f_h(e)$ for h=1, 2, we use natural cubic spline functions with K_h knots (e.g., Wood, 2006). We aim to characterize interaction of reasonable complexity and set K_1=K_2=5, by default. We place knots at sample quantiles of the observed E in the trios from mating types G_p=1, 2 and G_p=2, 3, respectively. For example, with a total of five knots, we place three at the 25th, 50th and 75th quantiles and two at the endpoints of the data. Following Wood (2006), to identify the model parameters, we impose constraints on $f_1(e)$ and $f_2(e)$:

$$\sum_{m\in\{1,3\}}\sum_j f_1(e_{mj})=0;\ \text{and}\ \sum_{m\in\{2,3\}}\sum_j f_2(e_{mj})=0.\tag{9}$$

These constraints involve sums over all observed E=e_{mj} in trios from appropriate mating types. Estimating the GRR parameters γ_1, γ_2, $f_1(e)$ and $f_2(e)$ amounts to estimating $\boldsymbol{\beta}=(\gamma_1,\boldsymbol{c}_1^\top,\gamma_2,\boldsymbol{c}_2^\top)^\top$, where $\boldsymbol{c}_h=(c_{h1},c_{h2},\ldots,c_{hK_h-1})^\top$ is the spline coefficient vector for $f_h(e)$ that satisfies the constraint.

Given a fixed set of smoothing parameters (λ_1, λ_2), the penalized likelihood estimator $\hat{\boldsymbol{\beta}}$ can be obtained via penalized re-weighted least squares optimization (Shin, 2012, Section 3.3.2). The smoothing parameters are obtained by minimizing the generalized Akaike information criterion function (Shin, 2012, Appendix B.3). To obtain interval estimates of $G\times E$ curves, we adopt a Bayesian approach (e.g., Wood, 2006), which has been shown to yield good frequentist coverage probabilities when the bias is a relatively small fraction of the mean squared error (Marra and Wood, 2012). Details about the Bayesian credible bands may be found in Section

3.3.3 of Shin (2012). To explore the form of interaction, we plot the fitted curves $\hat{f}_1$ and $\hat{f}_2$ along with their 95% Bayesian credible bands.

3.3 Permutation test of *G*×*E*

We adopt a permutation approach to testing *G*×*E* that takes into account the extra uncertainty introduced by estimating the smoothing parameters. We define the test statistic *T* as

$$T=\hat{\boldsymbol{c}}^\top\{V_{\boldsymbol{c}}\}^{-1}\hat{\boldsymbol{c}},$$

where $\hat{\boldsymbol{c}}$ is the penalized likelihood estimate of $\boldsymbol{c}=(\boldsymbol{c}_1^\top,\boldsymbol{c}_2^\top)^\top$, and V_c is the Bayesian posterior variance-covariance matrix for $\hat{\boldsymbol{c}}$ (e.g., Wood, 2006). Details about the Bayesian posterior variance-covariance matrix may be found in Section 3.3.3 of Shin (2012). Under the null hypothesis of no *G*×*E*, independence between *G* and *E* within a random family implies independence between *G* and *E* within a random affected family (e.g., Umbach and Weinberg, 2000; Shi et al., 2010). Since the analysis is conditional on parental genotypes, we may therefore approximate the null distribution of *T* by shuffling *E* within parental mating types.

It should be noted that the approach to approximating the null distribution of *T* can be invalid when the test locus is not causal but in linkage disequilibrium (LD) with the causal locus, and the study population has multiple subpopulations. In a stratified population, a single mating type stratum may consist of families from multiple subpopulations. In that case, different haplotype frequency and *E* distributions in the subpopulations can induce correlation between *E* and the genotypes *G* at the test locus (e.g., Shi et al., 2011). Without *G-E* independence, one cannot guarantee exchangeability within mating types, and hence the validity of the test.

4 Simulation study

We conducted a simulation study to evaluate the proposed approach.

4.1 Performance measures

We investigated the performance of the smoothing estimators of the interaction curves f_1 and f_2 through the empirical integrated squared error for interaction (EISEI), defined as $\mathrm{EISEI}=\sum_{h=1}^{2}\mathrm{EISEI}_h$, where

$$\begin{aligned}\mathrm{EISEI}_h&=\int[f_h(e)-\hat{f}_h(e)]^2dP_{\tilde{n}_h}(e),\\&=\frac{1}{\tilde{n}_h}\left[\sum_{m\in\{h,3\}}\sum_j\{f_h(e_{mj})-\hat{f}_h(e_{mj})\}^2\right],\quad h=1,2\end{aligned}$$

with $\tilde{n}_h\equiv n_h+n_3$ denoting the number of informative trios used for estimating $\mathrm{GRR}_h(e)$ and $P_{\tilde{n}_h}$, the empirical measure of the values of *E* from the $\tilde{n}_h$ informative trios.

We assessed the performance of the proposed permutation test of *G*×*E* by evaluating its type 1 error rates and power. Tests of *G*×*E* that do not condition on parental mating types are subject to inflated type 1 error rates when there is population *G-E* dependence due to population stratification (Umbach and Weinberg, 2000). To verify the validity of our test, we evaluated its type 1 error rates under both *G-E* independence and *G-E* dependence induced by population stratification. We compared the power of the test to other popular tests, including the likelihood ratio test from conditional logistic regression (e.g., Schaid, 1999; Cordell et al., 2004), FBAT-I (Lake and Laird, 2004) and the likelihood ratio test from the log-linear modeling approach (Umbach and Weinberg, 2000).

Our test statistic T was calculated by fitting data with $K_1=K_2=5$ knots. The p-values were estimated from 1000 permutation replicates. For the other tests, we set $f_1(e)$ and $f_2(e)$ as follows. For conditional logistic regression, we specified linear $G\times E$ with $f_1(e)=\beta_{ge1}e$ and $f_2(e)=\beta_{ge2}e$. For FBAT-I, we specified a log-additive penetrance mode by setting $z_1(g)=z_2(g)=g$ and $f_1(e)=f_2(e)$. Lastly, for the log-linear approach, we dichotomized E based on its sample median $\tilde{\mu}$ such that $f_1(e)=\beta_{ge1}\mathrm{I}\{e>\tilde{\mu}\}$ and $f_2(e)=\beta_{ge2}\mathrm{I}\{e>\tilde{\mu}\}$.

4.2 Simulation settings

Under G-E independence, we considered a homogeneous population, having SNP index-allele frequency of $q=0.1$, and a normally distributed non-genetic attribute E with mean of 0 and variance of 1. To induce G-E dependence, we considered a stratified population composed of two equal-sized subpopulations $S=0$ and $S=1$ with different distributions of G and E. For G, the subpopulation-specific allele frequencies were $q_0=0.1$ and $q_1=0.9$. For E, the subpopulation-specific distributions were Normal with means $\mu_0=-0.8$ and $\mu_1=0.8$ and common variance 0.36. In the overall population, these settings induced a mean of 0 and a variance of 1 for E, with G-E correlation of 0.71. For details of the calculations, see Shin et al. (2010). In the homogeneous population or within each subpopulation, G was generated according to Mendel's laws, based on parental genotypes G_p, and E was generated independently of G and G_p. Parental genotypes were generated under Hardy-Weinberg proportions.

Disease status D was simulated according to the disease penetrance model (3). The baseline parameter k and the non-genetic main effect term $\xi(e)$ were set to be zero for convenience. Dominant, log-additive and recessive penetrance modes were considered. The penetrance modes were calibrated to have equal values of $\mathrm{GRR}_1(e)\times\mathrm{GRR}_2(e)$ for all $E=e$. This calibration ensures that the GRR for two versus zero copies of the index allele is the same across modes. For dominant, log-additive and recessive penetrance modes, the genetic main effect parameters (γ_1, γ_2) were, respectively: (log 3, 0), $\left(\log\sqrt{3},\log\sqrt{3}\right)$ and (0, log 3). The interaction functions $(f_1(e), f_2(e))$ were, respectively: $(f(e), 0)$, $\left(\frac{1}{2}f(e),\frac{1}{2}f(e)\right)$, and $(0, f(e))$.

To understand how the form of interaction affects performance, we considered no $G\times E$ (H_0), as well as linear (H_{1L}), piecewise-linear (H_{1P}) and quadratic (H_{1Q}) interaction functions. Under H_0, $f(e)\equiv 0$. Under H_{1L}, we let $f(e)$ be a linear function with slope β_{ge}. Under H_{1P}, we let $f(e)$ be a piecewise-linear function created by joining a straight line with slope $\beta_{ge}>0$ and a horizontal line together at a point $E=z_p$. Under H_{1Q}, we let $f(e)$ be a quadratic function with a quadratic coefficient β_{ge} and an axis of symmetry at z_p. We let z_p be the p^{th} quantile of the distribution of E to see the effect of the placement of the non-linear feature (i.e., joining point or axis of symmetry) relative to the mass of data. Under H_{1P} and H_{1Q}, for a fixed β_{ge}, varying p changes the strength of linear trend in $f(e)$. Under H_{1P}, the linear trend in $f(e)$ becomes weaker as p decreases. Under H_{1Q}, the linear trend becomes weaker as p gets closer to 0.5. The parameterizations and parameter values for $f(e)$ are summarized in Table 2.

Finally, to understand how data concentration around the non-linear interaction feature affects performance, we added a scenario with uniform E over the interval (−4, 4), in a homogeneous population. The limits of E were chosen to match the 99.99% prediction interval of a standard normal random variable.

Table 2 Parameterizations for $f(e)$ under different settings. The p^{th} quantile of the distribution of E is denoted by z_p.

$G\times E$	Setting	$f(e)$	β_{ge}	p
None	H_0	0	–	–
Linear	H_{1L}	$\beta_{ge}e$	(−0.10, −0.17, −0.24)	–
Piecewise-Linear	H_{1P}	$\beta_{ge}\mathrm{I}\{e<z_p\}\cdot(e-z_p)$	(0.2, 0.6, 0.10)	(0.1, 0.3, 0.5)
Quadratic	H_{1Q}	$\beta_{ge}(e-z_p)^2$	(−0.04, −0.12, −0.20)	(0.1, 0.3, 0.5)

4.3 Simulation results

To save space, we present only highlights of the simulation results. A more detailed summary of results can be found in Supplementary Tables A1 and A2.

As expected, EISEIs decreased as the number of informative trios increased (results not shown). When interaction was non-linear, EISEIs decreased as the concentration of data around the non-linear feature increased. For example, Figure 1 summarizes results under piecewise linear and quadratic $G\times E$ (settings H_{1P} and H_{1Q}) when the joining point or axis of symmetry of $f(e)$ is placed at the median of the distribution of E. In this case, the concentration of data around the non-linear feature is higher when E is normally distributed than when it is uniformly distributed. As a result, EISEIs are lower under the normal distribution.

As expected, our test maintained the nominal level of significance, within simulation error, whether G and E were independent or not, under all penetrance models (Table 3). Highlights of the results related to power are as follows. When $G\times E$ was non-linear with a stronger linear trend, or $G\times E$ was linear, the proposed test had lower power than conditional logistic regression, lower power than FBAT-I with correctly specified penetrance, comparable power to the log-linear modeling approach, and higher power than FBAT-I with incorrectly specified penetrance (e.g., Figure 2). When $G\times E$ was non-linear with a weaker linear trend, the proposed test had comparable power to that of conditional logistic regression and FBAT-I but higher power than the log-linear approach (e.g., Figure 3A and B, for $p=0.3$). However, when the linear trend in the interaction curve was negligible, the proposed test had the highest power among the four tests (e.g., Figure 3).

5 Illustration

We illustrate our methods with an example data set simulated to mimic real data from a study of childhood acute lymphoblastic leukemia (ALL).

ALL is the most common type of leukemia in children under 19 years old. As shown in Figure A1, ALL can occur at any age, but the age-adjusted incidence rates are highest between ages 2 and 6 years, decrease

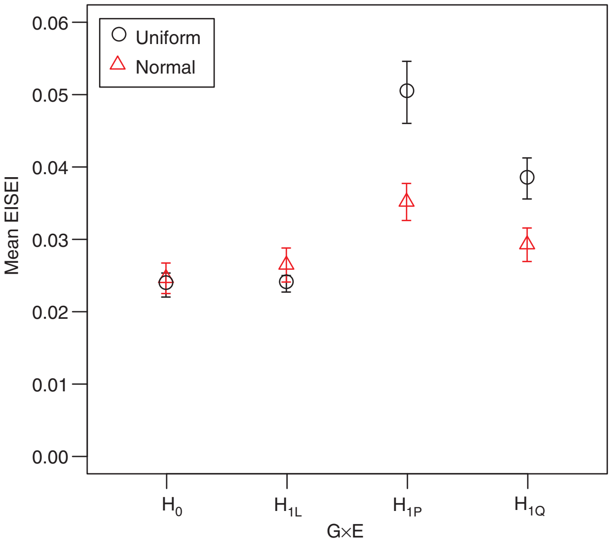

Figure 1 Average empirical integrated squared errors for interaction (EISEIs) for E simulated from a normal (red-triangles) or uniform (black-circles) distribution. In the simulations, G and E are independent and $f_1(e)=f(e)$, $f_2(e)=0$. Vertical lines connect the ±2 standard error bars. The horizontal axis represents different forms of $f(e)$; H_0, H_{1L}, H_{1P}, H_{1Q}, respectively, indicate $f(e)=0$; $f(e)=-0.24e$; $f(e)=\mathrm{I}\{e<0\}\cdot e$; and $f(e)=-0.2\cdot e^2$. Results are based on 1000 replicates of 3000 informative trios.

Table 3 Type 1 error rates (simulation error) for the proposed test of G×*E*. The nominal significance level of the test was 0.05.

Inheritance mode	*G-E* independence	
	Yes	**No**
Dominant	0.053 (0.007)	0.060 (0.008)
Log-additive	0.043 (0.006)	0.044 (0.006)
Recessive	0.049 (0.007)	0.061 (0.008)

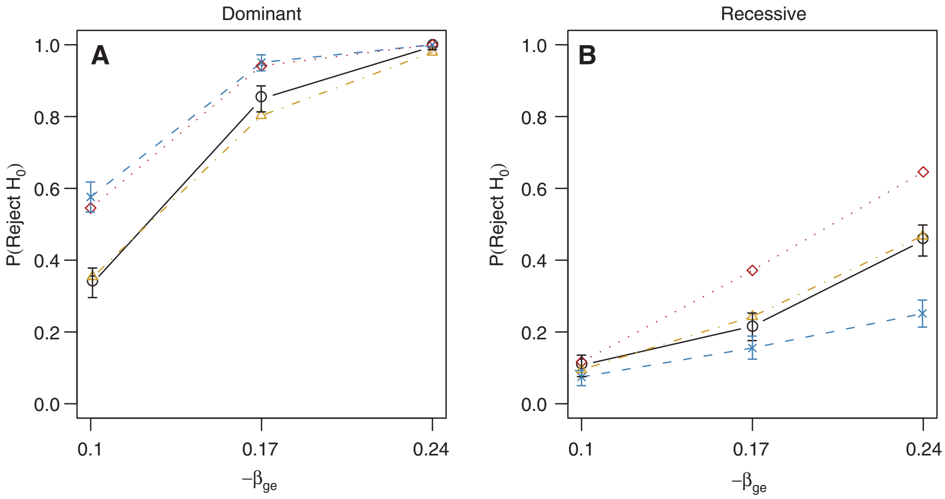

Figure 2 Empirical power under linear *G*×*E*. The data were generated under *G-E* independence and dominant (panel A) or recessive (panel B) penetrance modes. Proposed test, black circles with solid lines; FBAT-I, blue crosses with dashed lines; conditional logistic regression, red diamonds with dotted lines; log-linear modeling, yellow triangles with dot-dashed lines. To assist with comparison, we have included vertical bars representing ±2 simulation errors on results for the proposed test and FBAT-I. Results are based on 500 simulation replicates of 3000 informative trios.

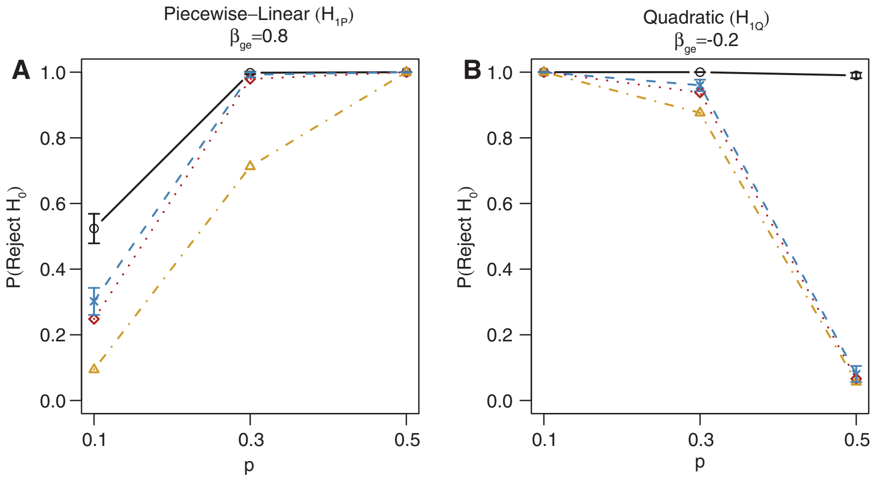

Figure 3 Empirical power under piecewise-linear (β_{ge}=0.8, panel A) and quadratic (β_{ge}=−0.2, panel B) *G*×*E*. The data were generated under *G-E* independence and a dominant penetrance mode. The quantile z_p of the distribution of *E* at which the non-linear feature is placed is controlled by *p* on the horizontal axis. In panel A, the linear trend increases with *p* while, in panel B, it decreases. Proposed test, black circles with solid lines; FBAT-I, blue crosses with dashed lines; conditional logistic regression, red diamonds with dotted lines; log-linear modeling, yellow triangles with dot-dashed lines. To assist with comparison, we have included vertical bars representing ±2 simulation errors on results for the proposed test and FBAT-I. Results are based on 500 simulation replicates of 3000 informative trios.

during young-adulthood and then start increasing again at ages >50 years (Ries et al., 1999). The bimodal distribution of incidence with age is consistent with different disease mechanisms for younger- and older-onset patients. For example, younger cases could have a genetic basis, whereas older cases could be sporadic. This motivates us to search for age-dependent genotype relative risks for ALL. The genetic risk factor we consider is the C609T polymorphism in the NAD(P)H:quinone oxidoreductase 1 (*NQO1*) gene. *NQO1* plays a role in detoxification of carcinogenic by-products (e.g., Ross and Siegel, 2004).

We generated a data set with 1000 informative case-parent trios mimicking those in a smaller data set. Details on how these example data were generated can be found in the Supplementary Materials. The mating-type-specific distribution of genotype frequencies and the histogram of age-at-diagnosis among cases in the example data set are shown in Table 4 and Figure 4, respectively.

According to Table 4, heterozygous parents transmit the variant allele to the cases 662/1190=56% of the time. The transmission disequilibrium test (TDT; Spielman et al., 1993) confirms that the variant allele was transmitted slightly more frequently than expected (p=0.0001). A similar trend was observed in the original data, for which the observed proportion was 171/322=53% (p=0.29). Figure 4 shows the age-at-onset distribution, which closely mimics the original data.

A co-dominant penetrance model was fitted to the simulated data. Three knots for $f_1(e)$ or ($f_2(e)$) were placed at the quartiles of the observed E in trios from G_p=1 or G_p=2 and 3; two knots were placed at the minimum and maximum values of E. Figure 5 shows the fitted interaction curves and the associated Bayesian 95% credible intervals for $f_1(e)$ and $f_2(e)$. The credible intervals are correctly suggestive of dominant non-linear $G \times E$ between *NQO1 C609T* and age (see Figure A2 for the true $G \times E$ curves).

We applied our permutation test of $G \times E$ to the example data, using a co-dominant penetrance model and 10,000 permutation replicates. We also applied our test using a dominant penetrance model (i.e., setting

Table 4 Mating-type-specific frequencies (%) of case genotypes in the example data.

Number of copies of *NQO1 C609T* variant	Informative mating type (m*)		
	1	2	3
0	328 (43)	–	36 (19)
1	440 (57)	19 (45)	109 (57)
2	–	23 (55)	45 (24)
n_m	768 (100)	42 (100)	190 (100)

*Mating types m=1, 2, 3 are defined in equation (1) of the text.

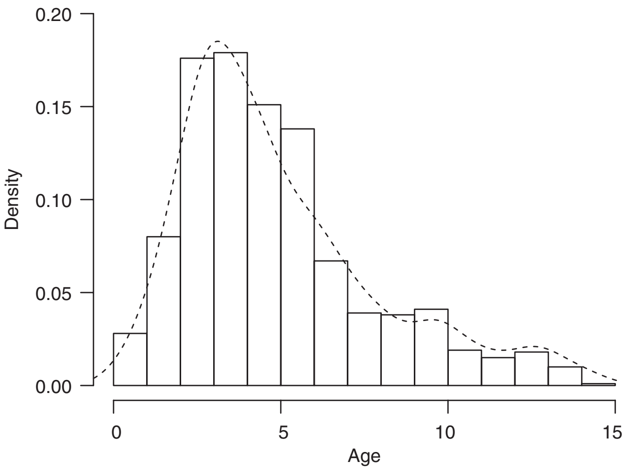

Figure 4 Histogram of age-at-diagnosis in informative trios of the example data. The dashed-line is the kernel density estimate obtained from the observed ages-at-onset of cases from informative trios in the original ALL data.

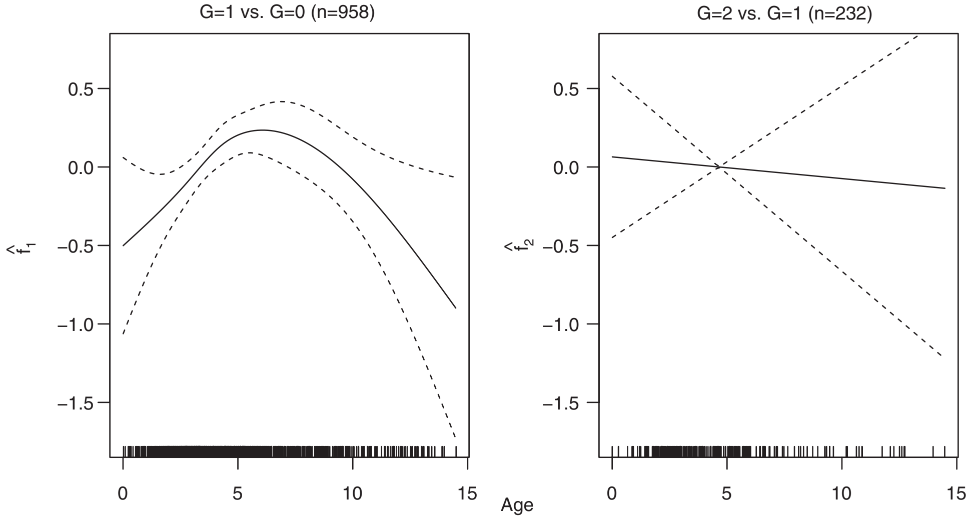

Figure 5 Fitted $G \times E$ for the example data set. The left and right panels, respectively, show the estimated interaction curves $\hat{f}_1$ and $\hat{f}_2$. The dashed lines indicate the 95% pointwise Bayesian credible limits.

$f_2(e) \equiv 0$) and 10,000 permutation replicates. For comparison, the other tests were applied using a dominant penetrance model. Referring to disease risk model (3), we specified $G \times E$ as linear such that $f_1(e)=\beta_{ge}e$ for conditional logistic regression and we dichotomized E based on its sample median $\tilde{\mu}$ such that $f_1(e)=\beta_{ge}\mathrm{I}\{e>\tilde{\mu}\}$ for the log-linear modeling approach. We used 10,000 permutation replications for FBAT-I. All tests were evaluated at significance level 5%. Our test indicated $G \times E$ under both codominant and dominant penetrance models (p=0.035 and p=0.008, respectively), while the other tests did not (p=0.803, 0.761 and 0.061, for conditional logistic regression, FBAT-I and log-linear modeling, respectively).

6 Discussion

In this work, we propose a data-smoothing approach to estimating and testing $G \times E$ using data from case-parent trios. The approach offers a flexible way to explore statistical interaction between a causal SNP and a continuous non-genetic covariate, using a generalized additive model (e.g., Wood, 2006). In particular, it allows for two separate genotype relative risks, depending on the number of copies of the variant allele and models interaction functions via regression splines. Consequently, instead of making assumptions about the penetrance mode and the parametric form for $G \times E$, the proposed approach lets the data determine them. We use this data-driven approach to obtain graphical displays for the point- and interval-estimates of the interaction curves (e.g., Figure 5). To our knowledge, this is the first tool for visualizing $G \times E$ for continuous E in case-parent trios. We also propose a permutation test of $G \times E$, which is appropriate when the form of interaction is unknown. This test accounts for the extra uncertainty introduced by the smoothing process.

We evaluated the proposed approach in a simulation study and found the following. The precision of the estimators depends on the number of informative case-parent trios, particularly the number concentrated around non-linear features of the interaction curve (e.g., Figure 1). To investigate the sample size required for accurate estimates of the $G \times E$ curves under our simulation settings, we varied the sample size in additional limited simulations. We found that precision was adequate when the number of informative trios was 2000, and did not greatly improve with the addition of more trios (results not shown). For a causal SNP, our permutation test is valid when there is G-E dependence due to population stratification. When $G \times E$ is non-linear with negligible linear trend, our test has more power than other tests of interaction that are tuned for linear $G \times E$ (e.g., Figure 3). When applied to an example data set, simulated to have non-linear $G \times E$, our graphical display correctly suggested $G \times E$, and our test provided stronger evidence for interaction than the other tests considered. These results suggest that our graphical display is a useful tool for exploring $G \times E$.

It is well known that detecting $G \times E$ requires large sample sizes (e.g., Smith and Day, 1984; Dempfle et al., 2008), and our method is no exception. In our context, the sample size is the number of informative case-parent trios. However, as the sample size increases, our permutation test becomes more computationally demanding. For example, on a Mac with 3.06 GHz Intel Core 2 Duo processor and 4 GB RAM, it took about 0.6 s for fitting the 1000 informative trios generated for illustration in Section 4. For data sets with 2000 and 3000 informative trios generated under the same scenario, fitting took about 2.4 and 4.5 s, respectively. Hence, to complete the proposed permutation test with 1000 replicates, about 10, 40 or 75 min would be required when the sample size is 1000, 2000 or 3000, respectively. To minimize the computational burden of the test, early termination (Besag and Clifford, 1991) can be deployed. On the other hand, the graphical display of $G \times E$ scales well with sample size and gives investigators a useful tool to characterize $G \times E$. In interpreting the graphical display, we caution that the Bayesian CIs are meant only as guide, since their coverage probabilities can be lower-than-nominal when the underlying interaction functions are non-linear but close to straight lines (Marra and Wood, 2012).

One of the key assumptions for all methods in this work is that the genotype is measured on a causal marker. However, in practice, we often have observations on nearby non-causal markers that are in linkage disequilibrium with the causal marker. Under such circumstances, approaches that condition on parental mating types, including our own, may be subject to biased assessment of gene-environment interaction when the study population has multiple hidden subpopulations with different distributions both for E and for haplotypes composed of the non-causal and causal loci. Such biased assessment can lead to either inflated type 1 error rates or reduced power, as alluded to by Shi et al. (2010) and later demonstrated by Shi et al. (2011) and Shin et al. (2012). Further research is underway to strengthen our approach against such population stratification, using ancestry informative markers or random markers measured on affected children.

Recently, many studies of $G \times E$ are genome-wide association studies. Methods for inference from such data (reviewed in Gauderman et al., 2013) focus on testing, rather than exploring the form of the interaction. In a genome-wide setting, application of the proposed permutation test to millions of markers will be computationally prohibitive. One way to reduce the computational burden is to adopt the two-step approach proposed by Gauderman et al. (2010): In step 1, screen variants, based on tests of association between G_p and E in cases (e.g., comparing the mean or median of E in different mating-type strata); and in step 2, apply the proposed permutation test only to those SNPs that pass the screening step, using appropriate methods to control for multiple testing.

Complex diseases result from both genetic and non-genetic risk factors interacting together. Characterizing statistical interaction between genes and the environment can thus provide important insights into the epidemiology of the disease. In this work, we show how a data-smoothing approach can be useful for exploring and testing statistical interaction between genetic and environmental covariates using data from case-parent trios. As with all tests of statistical interaction, our permutation test requires large sample sizes to ensure adequate power. Nevertheless, we anticipate that the proposed graphical display will be useful for understanding statistical interaction between genes and the environment. An advantage of our approach is that neither the parametric form of interaction or the mode of penetrance need to be specified. This feature is useful for investigating novel interactions, for which no prior information is available. The approach is implemented in an R package **trioGxE**, which is available on CRAN (R Core Team, 2012).

Although our method is designed to estimate and test $G \times E$ between one SNP and a continuous covariate, it can be easily extended to combinations of multiple SNPs by appropriately modifying the disease risk model (3). For example, with two SNPs, there will be eight genotype relative risks, and hence the risk model can be modified as $P(D=1|G_1=g_1, G_2=g_2, E=e)=\exp(k+\boldsymbol{z}(g)\boldsymbol{\gamma}+\boldsymbol{z}(g)\boldsymbol{f}(e))$, where $\boldsymbol{\gamma}=(\gamma_1, \gamma_2, \ldots, \gamma_8)^{\mathrm{T}}$ and $\boldsymbol{f}(e)=(f_1(e), f_2(e), \ldots, f_8(e))^{\mathrm{T}}$ represent the genetic effect and $G \times E$ interaction, respectively; and the genetic coding vector $\boldsymbol{z}(g_1, g_2)=(z_1(g_1, g_2), z_2(g_1, g_2), \ldots, z_8(g_1, g_2))$ is binary, as before. In such multi-variate SNP analyses, however, power to detect $G \times E$ may become an issue, considering that sample size requirements are large even for a single SNP analysis.

A common issue with case-parents designs is missing parental genotype data. Typically, in trios with missing genotype data, only the affected child and one parent contribute DNA. With parametric models, one can often include trios with a missing parental genotype in a likelihood-based data analysis by employing the EM algorithm or through multiple imputation. In the Supplementary Materials, we provide a sketch of a possible EM-based approach, whose implementation and operating characteristics are areas for future research.

Many other challenges remain in the study of gene-environment interaction, with the most daunting one being the measurement of environmental exposures (e.g., Thomas, 2010). Feasibility, cost and especially validity of such measures are challenging, particularly in large population studies; resulting measurement or misclassification error issues can affect the interpretation of gene-environment interaction results in ways that require more research in terms of identification and correction.

Acknowledgements: We thank two anonymous reviewers for their constructive and insightful comments that greatly improved the presentation of the manuscript. This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC); the Mathematics of Information Technology and Complex Systems, Canadian Networks of Centres of Excellence; the Canadian Institutes of Health Research; and the Brain Tumour Foundation of Canada. JS received PhD studentships from NSERC and the Michael Smith Foundation for Health Research.

References

Besag, J. and P. Clifford (1991): "Sequential Monte Carlo p-values," Biometrika, 78, 301–304.

Clavel, J., S. Bellec, S. Rebouissou, F. Ménégaux, J. Feunteun, C. Bonati-Pellié, A. Baruchel, K. Kebaili, A. Lambilliotte, G. Leverger, et al. (2005): "Childhood leukaemia, polymorphisms of metabolism enzyme genes, and interactions with maternal tobacco, coffee and alcohol consumption during pregnancy," Eur. J. Cancer Prev., 14, 531–540.

Cordell, H., B. Barratt and D. Clayton (2004): "Case/pseudocontrol analysis in genetic association studies: a unified framework for detection of genotype and haplotype associations, gene-gene and gene-environment interactions, and parent-of-origin effects," Genet. Epidemiol., 26, 167–185.

Database of Single Nucleotide Polymorphisms (dbSNP) (build 135): National Center for Biotechnology Information, National Library of Medicine dbSNP accession: rs1800566, Bethesda, MD, URL http://www.ncbi.nlm.nih.gov/SNP/, Available from: http://www.ncbi.nlm.nih.gov/SNP/.

Dempfle, A., A. Scherag, R. Hein, L. Beckmann, J. Chang-Claude and H. Schäfer (2008): "Gene-environment interactions for complex traits: definitions, methodological requirements and challenges," Eur. J. Hum. Genet., 16, 1164–1172.

Duke, L. (2007): A graphical tool for exploring SNP-by-environment interaction in case-parent trios., Master's thesis, Statistics and Actuarial Science: Simon Fraser University, URL http://www.stat.sfu.ca/content/dam/sfu/stat/alumnitheses/Duke-2007.pdf.

Duong, T. (2012): ks: Kernel smoothing, URL http://CRAN.R-project.org/package=ks, R package version 1.8.11.

Garaulet, M., C. Smith, T. Hernández-González, Y. Lee and J. Ordovás (2011): "PPARg Pro12Ala interacts with fat intake for obesity and weight loss in a behavioural treatment based on the Mediterranean diet," Mol. Nutr. Food Res., 55, 1771–1779.

Gauderman, W. J., D. C. Thomas, C. E. Murcray, D. Conti, D. Li and J. P. Lewinger (2010): "Efficient genome-wide association testing of gene-environment interaction in case-parent trios," Am. J. Epidemiol., 172, 116–22.

Gauderman, W. J., P. Zhang, J. L. Morrison and J. P. Lewinger (2013): "Finding novel genes by testing G×E interactions in a genome-wide association study," Genet. Epidemiol., 37, 603–613.

Green, P. J. (1990): "On use of the EM for penalized likelihood estimation," J. Roy. Stat. Soc. B Met., 52, 443–452.

Infante-Rivard, C. (2003): "Hospital or population controls for case-control studies of severe childhood diseases?" Am. J. Epidemiol., 157, 176–182.

Infante-Rivard, C., I. Fortier and E. Olson (2000): "Markers of infection, breast-feeding and childhood acute lymphoblastic leukaemia," Br. J. Cancer, 83, 1559–1564.

Infante-Rivard, C., J. Vermunt and C. Weinberg (2007): "Excess transmission of the NAD(P)H: Quinone Oxidoreductase 1 (NQO1) C609T polymorphism in families of children with acute lymphoblastic leukemia," Am. J. Epidemiol., 165, 1248–1254.

Kistner, E. and C. Weinberg (2004): "Method for using complete and incomplete trios to identify genes related to a quantitative trait," Genet. Epidemiol., 27, 33–42.

Lake, S. and N. Laird (2004): "Tests of gene-environment interaction for case-parent triads with general environmental exposures," Ann. Hum. Genet., 68, 55–64.

Marra, G. and S. Wood (2012): "Coverage properties of confidence intervals for generalized additive model components," Scand. J. Stat., 39, 53–74.

Moerkerke, B., S. Vansteelandt and C. Lange (2010): "A doubly robust test for gene-environment interaction in family-based studies of affected offspring," Biostatistics, 11, 213–225.

National Institute of Statistics and Economic Studies (2011): Pyramide des âges au 1er janvier 1999, URL http://www.insee.fr/fr/ppp/bases-de-donnees/donnees-detaillees/bilan-demo/pyramide/pyramide.htm?champ=fe&lang=fr&annee=1999, (accessed September 10, 2012).

Perrillat, F., J. Clavel, I. Jaussent, A. Baruchel, G. Leverger, B. Nelken, N. Philippe, G. Schaison, D. Sommelet, E. Vilmer, C. Bonaïti-Pellié and D. Hémon (2001): "Family cancer history and risk of childhood acute leukemia (France)," Cancer Causes Control, 12, 935–941.

R Core Team (2012): R: a language and environment for statistical computing, R Foundation for statistical computing, Vienna, Austria, URL http://www.R-project.org/, ISBN 3-900051-07-0.

Ries, L., M. Smith, J. Gurney, M. Linet, T. Tamra, J. Young and G. Bunin (Eds.) (1999): Cancer incidence and survival among children and adolescents: United States SEER program 1975-1995, National Cancer Institute, SEER Program, Bethesda, MD: NIH Pub. No. 99-4649.

Ross, D. and D. Siegel (2004): NAD(P)H:quinone oxidoreductase 1 (NQO1, DT-Diaphorase), functions and pharmacogenetics. In: Sies H., Packer L., (Eds.), Quinones and Quinone Enzymes, Part B, Methods in Enzymology, volume 382, Academic Press, 115–144, URL http://www.sciencedirect.com/science/article/pii/S0076687904820081.

Schaid, D. J. (1999): "Case-parents design for gene-environment interaction," Genet. Epidemiol., 16, 261–273.

Schaid, D. J. and S. S. Sommer (1993): "Genotype relative risks: methods for design and analysis of candidate-gene association studies," Am. J. Hum. Genet., 53, 1114–1126.

Shi, M., D. Umbach and C. Weinberg (2010): "Testing haplotype-environment interactions using case-parent triads," Hum. Hered., 70, 23–33.

Shi, M., D. Umbach and C. Weinberg (2011): "Family based gene-by-environment interaction studies: revelations and remedies," Epidemiology, 22, 400–407.

Shin, J.-H. (2012): Inferring gene-environment interaction from case-parent trio data: evaluation of and adjustment for spurious G×E and development of a data-smoothing method to uncover true G×E, Ph.D. thesis, Statistics and Actuarial Science: Simon Fraser University, URL https://theses.lib.sfu.ca/sites/all/files/public_copies/etd7214-j-shin-etd7214jshin.pdf.

Shin, J.-H., B. McNeney and J. Graham (2010): "On the use of allelic transmission rates for assessing gene-by-environment interaction in case-parent trios," Ann. Hum. Genet., 74, 439–451.

Shin, J.-H., C. Infante-Rivard, J. Graham and B. McNeney (2012): "Adjusting for spurious gene-by-environment interaction using case-parent triads," Stat. Appl. Genet. Mol. Biol., 11, Article 7, URL http://www.degruyter.com/view/j/sagmb.2012.11.issue-2/1544-6115.1714/1544-6115.1714.xml.

Smith, P. and N. Day (1984): "The design of case-control studies: the influence of confounding and interaction effects," Int. J. Epidemiol., 13, 356–365.

Spielman, R. S., R. E. McGinnis and W. J. Ewens (1993): "Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM)," Am. J. Hum. Genet., 52, 506–516.

Statistics Canada (2012): "Historical age pyramid," URL http://www12.statcan.gc.ca/census-recensement/2011/dp-pd/pyramid-pyramide/his/index-eng.cfm, (accessed September 10, 2012).

Thomas, D. (2010): "Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies," Annu. Rev. Publ. Health., 31, 21–36.

Umbach, D. and C. Weinberg (2000): "The use of case-parent triads to study joint effects of genotype and exposure," Am. J. Hum. Genet., 66, 251–261.

Wood, S. (2006): Generalized additive models: an introduction with R, Boca Raton, FL: Chapman & Hall/CRC.

Yee, T. W. (2010): "The VGAM package for categorical data analysis," J. Statist. Soft., 32, 1–34.

Yee, T. W. and C. Wild (1996): "Vector generalized additive models," J. Roy. Stat. Soc. B Met., 58, 481–493.