frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 04 January 2023
DOI 10.3389/fgene.2022.1065568

Check for updates

OPEN ACCESS

EDITED BY
Rongling Wu,
The Pennsylvania State University (PSU),
United States

REVIEWED BY
Tianyuan Lu,
McGill University, Canada
Zhaoxia Yu,
University of California, Irvine,
United States

*CORRESPONDENCE
Brad McNeney,
✉ mcneney@sfu.ca

SPECIALTY SECTION
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

RECEIVED 09 October 2022
ACCEPTED 29 November 2022
PUBLISHED 04 January 2023

CITATION
Ratnasekera P, Graham J and McNeney B (2023), Inference of gene-environment interaction from heterogeneous case-parent trios. *Front. Genet.* 13:1065568.
doi: 10.3389/fgene.2022.1065568

Inference of gene-environment interaction from heterogeneous case-parent trios

Pulindu Ratnasekera, Jinko Graham and Brad McNeney*

Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, Canada

Introduction: In genetic epidemiology, log-linear models of population risk may be used to study the effect of genotypes and exposures on the relative risk of a disease. Such models may also include gene-environment interaction terms that allow the genotypes to modify the effect of the exposure, or equivalently, the exposure to modify the effect of genotypes on the relative risk. When a measured test locus is in linkage disequilibrium with an unmeasured causal locus, exposure-related genetic structure in the population can lead to spurious gene-environment interaction; that is, to apparent gene-environment interaction at the test locus in the absence of true gene-environment interaction at the causal locus. Exposure-related genetic structure occurs when the distributions of exposures and of haplotypes at the test and causal locus both differ across population strata. A case-parent trio design can protect inference of genetic main effects from confounding bias due to genetic structure in the population. Unfortunately, when the genetic structure is exposure-related, the protection against confounding bias for the genetic main effect does not extend to the gene-environment interaction term.

Methods: We show that current methods to reduce the bias in estimated gene-environment interactions from case-parent trio data can only account for simple population structure involving two strata. To fill this gap, we propose to directly accommodate multiple population strata by adjusting for genetic principal components (PCs).

Results and Discussion: Through simulations, we show that our PC adjustment maintains the nominal type-1 error rate and has nearly identical power to detect gene-environment interaction as an oracle approach based directly on population strata. We also apply the PC-adjustment approach to data from a study of genetic modifiers of cleft palate comprised primarily of case-parent trios of European and East Asian ancestry. Consistent with earlier analyses, our results suggest that the gene-environment interaction signal in these data is due to the self-reported European trios.

KEYWORDS

gene-environment interaction, case-parent trios, population structure, genome-wide association study, cleft palate, principal components

1 Introduction

We start by considering a log-linear model of population disease risk that includes main effects for genotypes G, environmental exposures E, and a gene-environment interaction term $G \times E$. The $G \times E$ term allows genotypes to modify the effect of the exposure or, equivalently, the exposure to modify the effect of genotypes on the relative risk of developing the disease. Including a $G \times E$ term can improve model accuracy and provide a more detailed picture of disease etiology compared to models with just G and E main effects (Hunter, 2005). $G \times E$ is also useful for identifying environmental exposures with greater disease-association in individuals who carry particular alleles at susceptibility loci (Thomas, 2010). For example, dietary fat intake is more highly associated with obesity in carriers than in non-carriers of the Pro12Ala allele in the PPAR-γ gene (Garaulet et al., 2011).

We suppose throughout that G is an unmeasured causal locus in linkage disequilibrium with a measured non-causal test locus G', and that the distribution of GG' haplotypes differs across population strata (i.e. genetic structure). Stratum-specific differences in the GG' haplotype frequencies can lead to differences in G' risk across the population strata where none exist for G (Zaykin and Shibata, 2008). Exposure-related genetic structure occurs when the distribution of E also differs across the population strata (Weinberg et al., 2011). Without some adjustment for the population strata, E will tag the stratum-specific differences in G' risk (Figure 1), suggesting that E modifies G' risk, even in the absence of $G \times E$ (Shi et al., 2011; Weinberg et al., 2011); we refer to this as spurious $G' \times E$.

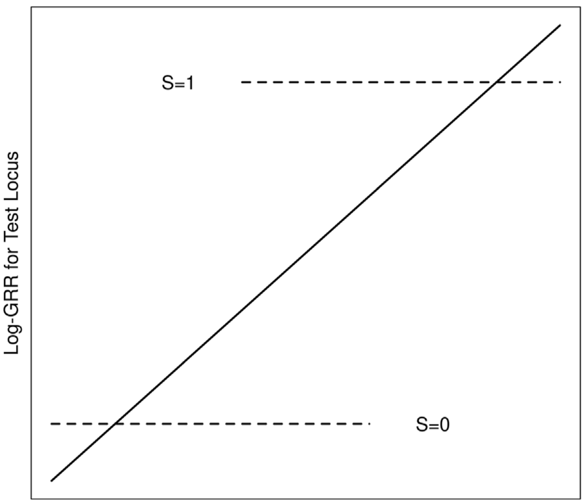

FIGURE 1
Schematic of log-GRRs for a non-causal test locus *versus* exposure in a structured population with two strata, S = 0 and S = 1. Dashed lines represent log-GRRs within each stratum. Horizontal positioning of these dashed lines indicates the support of the respective E distributions. High values of E are associated with S = 1, in which one of the alleles at the test locus is associated with increased disease risk. Low values of E are associated with S = 0 in which this same allele at the test locus is associated with low disease risk. Ignoring S yields the linear log-GRR curve indicated by the solid line, which erroneously suggests that E modifies the disease risk at the test locus.

A case-parent trio design can protect inference of genetic main effects from confounding bias due to genetic structure in the population (Weinberg, 1999). In this design, investigators collect information on G' and E in children affected with a disease of interest as well as the genotypes, G'_P, of their parents. To increase sample size, investigators may pool trios from multiple ancestral groups into one study; e.g., the GENEVA Oral Cleft Study (GENEVA, 2010) combined case-parent trios from recruitment sites in the United States, Europe and East Asia. Assuming G' and E are independent within families, a log-linear model of disease risk leads to a conditional likelihood for the G' and $G' \times E$ effects, based on the child's genotype given their exposure, affection status and parental genotypes (Shin et al., 2012). Unfortunately, when the genetic structure is exposure related, the protection against confounding bias for the genetic main effect does not extend to the gene-environment interaction term (Shi et al., 2011; Weinberg et al., 2011). Thus, spurious $G' \times E$ may be inferred from heterogeneous case-parent trio data in the absence of true $G \times E$.

Methods to mitigate this bias may be classified as design- or data-based. For a binary environmental exposure, the *design*-based tetrad approach of (Shi et al., 2011) augments the case-parent trio by adding the exposure of an unaffected sibling. These authors control the bias by including the sibship-averaged exposure in the log-linear model. They show that all information about the interaction in the tetrad design comes from the siblings, not the parents (Weinberg et al., 2011). Accordingly, they propose a sibling-augmented case-only design and analysis. By contrast, (Shin et al., 2012) takes a *data*-based approach, replacing the sibship-averaged exposure of (Shi et al., 2011) with the *predicted* exposure given ancestry. Predictions are obtained from a regression of exposure on principal components (PCs) computed from genetic markers that are unlinked to the test locus. This data-based approach may be applied to arbitrary exposures, including continuous exposures, and does not require siblings. However, its properties have not been evaluated in the case of more than two population strata.

We use the GENEVA Oral Cleft Study to motivate a new approach to unbiased inference of $G' \times E$ in case-parent trios. The analysis of (Beaty et al., 2011) found multiple single nucleotide polymorphisms (SNPs) that appeared to modify the effect of maternal smoking, maternal alcohol consumption or maternal multivitamin supplementation on the risk of cleft palate (CP). The self-reported ancestry of the study sample is primarily European or East Asian, and all three exposures are more common in self-reported Europeans than in self-reported East

TABLE 1 GG' **haplotype frequencies in four population strata.**

	Stratum			
GG'	$S = 0$	$S = 1$	$S = 2$	$S = 3$
R1	0.0	0.5	0.375	0.125
R0	0.5	0.0	0.125	0.375
N1	0.5	0.0	0.125	0.375
N0	0.0	0.5	0.375	0.125

Asians (Beaty et al., 2011, Table 2). If the frequencies of haplotypes spanning causal SNPs also vary by ancestral groups, exposure-related genetic structure may lead to spurious gene-environment interaction. (Ratnasekera and McNeney, 2021). focused on the self-reported Europeans and East Asians in the GENEVA Oral Cleft Study data. Applying the approach of (Shin et al., 2012), they confirmed the gene-environment interaction found by (Beaty et al., 2011), and concluded that the evidence for gene-environment interaction is predominantly from the data of self-reported Europeans. These authors also considered whether exposure-related genetic structure *within* self-reported Europeans could explain the apparent $G' \times E$. Their results were inconclusive, however, possibly owing to the methodology's limitation to just two ancestry groups. In modern datasets, the possibility of both inter- and intra-continental genetic structure necessitates methods that can more flexibly accommodate multiple ancestries. In this work we propose such an approach which relies on direct use of the genetic PCs to adjust for population structure.

The manuscript is structured as follows. In Section 2 we develop our direct PC-adjustment method and compare it to the indirect PC-based approach of (Shin et al., 2012). In Section 3 we present simulations to evaluate the statistical properties of both approaches. In Section 4 we re-analyze the GENEVA data. Section 5 includes a discussion and areas for future work.

2 Models and methods

2.1 Overview

We start with a log-linear model of disease risk parametrized in terms of genotype relative risks (GRRs) at a causal locus G. Under this model, G × E is equivalent to GRRs that depend on the exposure E. We then derive the GRRs at a non-causal test locus G′ in linkage disequilibrium with G and show that, in the absence of G × E, the G′-GRRs can depend on E when there is dependence between E and GG′ haplotypes in the population. Such dependence can lead to spurious inference of G′ × E in the absence of G × E. However, valid inference is obtained if we adjust the risk model for any variable X for which E and GG′ haplotypes are conditionally independent given X (Shin et al., 2012). We review the rationale for the adjustment used by (Shin et al., 2012) in this context, and propose an alternative adjustment based on inferred population structure. In particular, we use the method of (Gavish and Donoho, 2014) to select a parsimonious set of PCs with which to adjust the risk model. A key question is whether the PC-selection method yields a set of PCs that provide enough adjustment to maintain type 1 error in the absence of G × E, but not so much that we compromise power in the presence of G × E. The Models and Methods section concludes with a discussion of the simulation methods used to answer this question.

2.2 Risk model and likelihood

Let $G = 0$, 1 or 2 denote the number of copies of the variant allele at the causal locus and E denote the exposure variable. The disease-risk model of (Shin et al., 2012) can be obtained from a log-linear model of the GRRs

$$\begin{aligned}\log GRR_g(e) &= \log \frac{P(D=1|G=g,E=e)}{P(D=1|G=g-1,E=e)}\\ &= \beta_g + f_g(e) \quad \text{for } g = 1,2, \qquad (1)\end{aligned}$$

and the log-disease risk for carriers of the baseline genotype $G = 0$

$$\log P(D=1|G=0,E=e) \equiv \eta(e).$$

The parameters β_g and $f_g(\cdot)$ are, respectively, genotype-specific main effects and functions that allow for $G \times E$ interaction. We can also write disease risk in terms of the baseline risk $\eta(e)$ and the GRRs as follows. First define $GRR_0(e) \equiv 1$. Next, note that

$$\frac{P(D=1|G=1,E=e)}{P(D=1|G=0,E=e)} = GRR_1(e) = GRR_1(e)GRR_0(e)$$

and

$$\begin{aligned}\frac{P(D=1|G=2,E=e)}{P(D=1|G=0,E=e)} &= \frac{P(D=1|G=2,E=e)}{P(D=1|G=1,E=e)}\frac{P(D=1|G=1,E=e)}{P(D=1|G=0,E=e)}\\ &= GRR_2(e)GRR_1(e)GRR_0(e).\end{aligned}$$

it follows that

$$P(D=1|G=g,E=e) = \eta(e)\prod_{i=0}^{g} GRR_i(e) \quad \text{for } g = 0, 1 or 2. \qquad (2)$$

A likelihood for estimation of the GRR parameters β_g and $f_g(\cdot)$, $g = 1, 2$, from case-parent trio data can be derived under the assumption that G and E are conditionally independent given parental genotypes G_p. As shown in Supplementary Appendix S1, the likelihood is based on the conditional probability of the child's genotype given their exposure and parental genotypes. The function $\eta(\cdot)$ that parametrizes the environmental main effect drops out of the likelihood and cannot be estimated from case-parent trio data.

2.3 GRRs at a non-causal test locus

Let G' denote genotypes at a non-causal test locus in linkage disequilibrium with the causal locus G. We assume D and G' are conditionally independent given G and E, so that

$$P(D = 1|G = g, G' = g', E = e) = P(D = 1|G = g, E = e).$$

Therefore, the risk of disease given G' and E can be written as

$$\begin{aligned} &P(D = 1|G' = g', E = e) \\ &= \sum_{g=0}^{2} P(D = 1|G = g, E = e)P(G = g|G' = g', E = e). \end{aligned} \tag{3}$$

Eq. 3 is a latent-class model (Xu, 2017) with the unobserved causal locus G as the latent class having probabilities $P(G = g|G' = g', E = e)$. Eqs 2, 3 enable the log-GRRs at G' to be written in terms of the latent-class probabilities and the GRRs at G as follows:

$$\begin{aligned} \log GRR_{g'}(e) &\equiv \log\frac{P(D = 1|G' = g', E = e)}{P(D = 1|G' = g' - 1, E = e)} \\ &= \log\frac{\sum_{g=0}^{2} P(D = 1|G = g, E = e)P(G = g|G' = g', E = e)}{\sum_{g=0}^{2} P(D = 1|G = g, E = e)P(G = g|G' = g' - 1, E = e)} \\ &= \log\frac{\sum_{g=0}^{2}\left(\prod_{i=0}^{g} GRR_i(e)\right)P(G = g|G' = g', E = e)}{\sum_{g=0}^{2}\left(\prod_{i=0}^{g} GRR_i(e)\right)P(G = g|G' = g' - 1, E = e)}. \end{aligned} \tag{4}$$

Without $G \times E$, GRRs at G do not depend on E. Importantly, though, the log-GRRs at G' *can* depend on E through the latent-class probabilities $P(G = g|G' = g', E = e)$. In fact, as shown in Supplementary Appendix S2, these latent-class probabilities will depend on E whenever GG' haplotypes and E are associated, as happens when the population has exposure-related genetic structure. Since $G' \times E$ is equivalent to $GRR_{g'}$ varying with E, Eq. 4 gives insight into how exposure-related genetic structure creates spurious $G' \times E$.

2.4 Augmented risk model

The development so far has considered a disease-risk model that depends only on E and a causal locus G. We now consider an augmented disease-risk model that depends on E, G and a third variable X:

$$\begin{aligned} \log GRR_g(e, x) \equiv \log\frac{P(D = 1|G = g, E = e, X = x)}{P(D = 1|G = g - 1, E = e, X = x)} \\ = \beta_g + f_g(e, x) \quad \text{for } g = 1, 2, \end{aligned}$$

where β_g and $f_g(\cdot, x)$ are, respectively, genotype-specific main effects and functions that allow for $G \times E \times X$ interaction. Defining

$$GRR_0(e, x) \equiv 1,$$

an analogous development to Section 2.3 leads to the following X-adjusted log-GRRs at G':

$$\begin{aligned} \log GRR_{g'}(e, x) &\equiv \log\frac{P(D = 1|G' = g', E = e, X = x)}{P(D = 1|G' = g' - 1, E = e, X = x)} \\ &= \log\frac{\sum_{g=0}^{2}\left(\prod_{i=0}^{g} GRR_i(e, x)\right)P(G = g|G' = g', E = e, X = x)}{\sum_{g=0}^{2}\left(\prod_{i=0}^{g} GRR_i(e, x)\right)P(G = g|G' = g' - 1, E = e, X = x)}. \end{aligned} \tag{5}$$

In the next section we discuss choices for X that eliminate E from the latent-class probabilities for G, and hence eliminate spurious $G' \times E$ arising from exposure-related genetic structure.

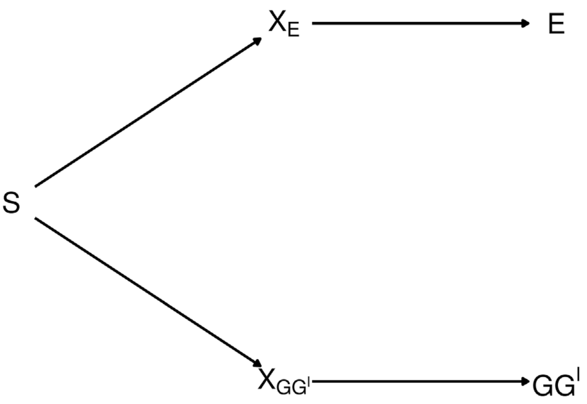

FIGURE 2
Diagram depicting exposure-related genetic structure. The latent population strata S induce dependence beetween E and GG'. Latent factors X_E and $X_{GG'}$ encode different distributions of E and GG', respectively. E and GG' are conditionally independent given any of the three variables that lie on the path between them.

2.5 Removing dependence of the latent-class probabilities on *E*

The diagram in Figure 2 depicts the dependence between GG' haplotypes and E from exposure-related genetic structure in the population. In the figure, S is a categorical variable that indicates population strata. The categorical variable X_E is a "coarsening" of S such that different levels of X_E correspond to different E distributions, and, similarly, $X_{GG'}$ is a coarsening of S such that different levels of $X_{GG'}$ correspond to different GG' haplotype distributions.

The path connecting E and GG' in Figure 2 is said to be *blocked* by each of the variables X_E, S and $X_{GG'}$ [(Pearl, 2009), Definition 1]. Therefore, E and GG' are conditionally independent given any of the blocking variables X_E, S or $X_{GG'}$ (Pearl, 1998). As shown in Supplementary Appendix S2, a consequence is that conditioning on any of these variables removes the dependence of the latent-class probabilities on E.

That is, letting X denote any of X_E, S or $X_{GG'}$, $P(G = g|G' = g', E = e, X = x) = P(G = g|G' = g', X = x)$. Consequently, from Eq. 5,

$$\log GRR_{g'}(e,x) \equiv \log \frac{P(D=1|G'=g', E=e, X=x)}{P(D=1|G'=g'-1, E=e, X=x)}$$
$$= \log \frac{\sum_{g=0}^{2}\left(\prod_{i=0}^{g} GRR_i(e,x)\right)P(G=g|G'=g', X=x)}{\sum_{g=0}^{2}\left(\prod_{i=0}^{g} GRR_i(e,x)\right)P(G=g|G'=g'-1, X=x)}. \quad (6)$$

GRRs at G' will thus depend on E if and only if GRRs at G do.

2.6 Linear model for the log GRRs

From Eq. 6 we see that, for fixed g' and x, log $GRR_{g'}(e, x)$ varies with e if and only if the $GRR_g(e, x)$ do. We can therefore test for $G \times E$ by fitting a model for log $GRR_{g'}(e, x)$ that allows separate curves in e for each combination of g' and x (Shin et al., 2014). We take these curves to be straight lines, and test whether any of them have non-zero slope. For a fixed value x of the adjustment variable X and a fixed value e of the environmental exposure E, the log-GRR is:

$$\log GRR_{g'}(e,x) = \beta_{g'} + \beta_{g'X}x + \beta_{g'E} \times e + \beta_{g'EX}x \times e;$$
$$g' = 1, 2. \quad (7)$$

The generalization of the above model to a vector X is to replace $\beta_{g'X}x$ with $\beta_{g'X}^T x$ and $\beta_{g'EX}$ with $\beta_{g'EX}^T x$ for coefficient vectors $\beta_{g'X}$ and $\beta_{g'EX}$. The intercepts of the log-GRR curves, $\beta_{g'}+\beta_{g'X}x$, are the genetic main effects in stratum x (i.e. when $e = 0$). The slopes, $\beta_{g'E}+\beta_{g'EX}x$, are the $G' \times E$ interaction terms in stratum x. We use a likelihood-ratio test of the null hypothesis that $\beta_{g'E} = \beta_{g'EX} = 0$ for $g' = 1, 2$, versus the alternative hypothesis that at least one of these slope parameters is non-zero to detect $G \times E$. We emphasize that the simplified log-GRR curves in e characterize $G \times E$ rather than environmental main effects, which are not estimable from case-parent trio data. Genetic main effects *are* estimable however and flexibly parametrized by the intercept terms of the log-GRR curves. The flexibility in the intercept terms avoids mis-specification of the genetic main effects which can lead to biased inference of interaction effects (Yu et al., 2015).

2.7 Choice of X

Following (Shi et al., 2011), (Shin et al., 2012) set X to be the categorical variable X_E that distinguishes E distributions among the genetic strata of the population. Since X_E is unobserved, (Shin et al., 2012), consider the expectation of E given genetic markers (EEGM) as a surrogate $\hat{X}_E$. The idea behind their EEGM approach is to distinguish exposure distributions by their mean, which may vary across genetic strata, S. Though S is not known, it is reflected in principal components (PCs), $\hat{S}$, computed from a set of genetic markers that are unlinked to G'. The expectation of E given $\hat{S}$ can be estimated by linear regression of E on $\hat{S}$ when E is continuous, or by logistic regression when E is binary. For EEGM adjustment, the expected exposure within genetic strata is estimated by $\hat{X}_E = \widehat{E(E|\hat{S})}$. (Shin et al., 2012). showed that EEGM adjustment works well where there are two population strata, but our simulation results (Section 3) indicate that it works poorly for more than two strata. We therefore propose to adjust for population strata directly; i.e., to take $X = S$. In particular, if the population has K+1 genetic strata, indexed 0, . . . , K, we let S denote a vector of K dummy variables that distinguish these strata such that the kth element $S_k = 1$ for trios in stratum $k > 0$ and 0 otherwise, for $k = 1, \ldots, K$.

2.8 Inferred population strata

The population stratum variable S reflects genetic ancestry and is not generally known. Since adjustment for self-reported ancestry can lead to bias (Wang et al., 2010) we use marker-based PCs, $\hat{S}$. An advantage of PC-adjustment is that it does not enforce discrete strata, and individuals whose PC values lie between those of clusters on the PC plot (e.g. admixed individuals) will have intermediate values of the slope and intercept of their log-GRR curve.

Standard PC adjustment in genetic association analyses relies on a relatively large set of PCs. For K PCs the degrees of freedom of the test for $G' \times E$ is equal to 2(K+1). Thus, using more PCs than are necessary reduces the power of the test for $G' \times E$. We seek methods to select a parsimonious set of PCs that provides enough adjustment to control type 1 error rate, without sacrificing power. We consider three PC-selection methods. The first (Zhu and Ghodsi, 2006) is an automated version of the graphical approach of looking for an "elbow" in the scree plot of variance explained by the PCs as a function of their number. The second (Gavish and Donoho, 2014) is an estimator of the rank of a matrix under a model in which the data matrix is a noisy version of a low-rank matrix. The third (Patterson et al., 2006) is to select PCs corresponding to eigenvalues that exceed a significance threshold determined from the distribution of the largest eigenvalue of an unstructured random matrix.

2.9 Simulation methods

2.9.1 Simulating G, G' and E on case-parent trios

To study the statistical properties of our proposed approach and compare it to the method of (Shin et al., 2012), we generated 5,000 data sets of 3,000 informative case-parent trios. Trios were sampled from one of four population strata labelled S = 0, 1, 2 or 3. We assumed random mating within and no mixing between strata. We performed some simulations using equal-sized strata and others using unequal-sized strata. In the case of unequal

TABLE 2 Estimated type 1 error rates (top entry) and corresponding 95% confidence intervals (bottom entry) when data are simulated from 2, 3 or 4 strata with equal (top three rows) or unequal (bottom three rows) stratum sizes.

Equal stratum sizes			
Number of strata			
Adjustment	2	3	4
S	0.0556	0.0524	0.0498
	(0.049, 0.062)	(0.046, 0.0586)	(0.044, 0.056)
EEGM	0.0538	1.0000	1.0000
	(0.048, 0.060)	NA	NA
PC	0.0546	0.0534	0.0496
	(0.048, 0.061)	(0.047, 0.060)	(0.044, 0.056)
Unequal stratum sizes			
	2	3	4
S	0.0524	0.0482	0.0536
	(0.046, 0.058)	(0.042 0.054)	(0.047,0.059)
EEGM	0.0536	1.0000	1.0000
	(0.047, 0.060)	NA	NA
PC	0.0540	0.0508	0.0527
	(0.048, 0.060)	(0.045, 0.057)	(0.046, 0.059)

stratum sizes, the split was 60%, 40% for two strata; 50%, 30% and 20% for three strata; and 40%, 30%, 20% and 10% for four strata.

For a given stratum, informative trios were simulated following the methods proposed by (Shin et al., 2013; Shin et al., 2014). Briefly, GG' haplotypes are first simulated on parents in a random-mating population according to the stratum-specific GG' haplotype distributions in Table 1. Child haplotypes are then simulated following Mendel's laws and assuming no recombination between G and G'. The child's exposure E is also simulated according to the stratum-specific distributions described below. Finally, the child's disease status is simulated based on the disease-risk model (1). Trios with an affected child and at least one heterozygous parent at the test locus are retained. The data recorded on each trio are G_p', G', and E, where G_p' is the pair of parental genotypes at the test locus.

Spurious $G' \times E$ is induced by specifying different distributions of E and GG' haplotypes in the four strata of Table 1. The GG' distributions for strata $S = 0$ and $S = 1$ are as in (Shin et al., 2012). Alleles at G are denoted R (risk) and N (non-risk), while alleles at G' are denoted 1 and 0. We summarize the haplotype distributions by the implied allelic correlations between the index alleles R and 1. Under the GG' haplotype frequencies given in Table 1, these correlations are $r_0 = -1$ in stratum $S = 0$, $r_1 = 1$ in stratum $S = 1$, $r_2 = 0.5$ in stratum $S = 2$ and $r_3 = -0.5$ in stratum $S = 3$.

The stratum-specific distributions of E are chosen to be normal with common variance $\sigma^2 = 0.36$, and means $\mu_0 = -0.8$, $\mu_1 = 0.8$, $\mu_2 = 2.4$ and $\mu_3 = 4.0$ in strata 0, 1, 2 and 3, respectively. The E distributions for strata $S = 0$ and $S = 1$ are as in (Shin et al., 2012).

The disease-risk model is specified as follows. The genetic main effect is $\beta_g = \log(3)/2$ for $g = 1, 2$, corresponding to a $\sqrt{3}$-fold increase in relative risk for each copy of the risk allele (R) in the absence of $G \times E$. To evaluate the type 1 error rate of the $G \times E$ test we set $f_g(e) = 0$ in our simulations. To investigate power we choose a linear interaction model for the $G \times E$ term, setting $f_g(e) = \beta_{gE} e$ with $\beta_{gE} = -0.10, -0.15, -0.20$ or -0.25.

2.9.2 Simulating markers for PC adjustment

A standard method of PC adjustment is to calculate PCs from a genomic region that is unlinked to the test locus. It is recommended that markers in this region be thinned, or LD pruned, to have pairwise correlations of $r^2 \leq 0.1$ (Grinde, 2019). We simulated such panels of markers based on data from the 1,000 genomes project (Clarke et al., 2016) using two East Asian (Chinese Dai in Xishuangbanna, China [CDX] and Han Chinese in Bejing China [CHB]) and two European (Iberian population in Spain [IBS] and Finnish in Finland [FIN]) populations. From the initial download of the genome-wide data, we retained 6,929,035 diallelic, autosomal markers with minor allele frequency (MAF) 0.05 or greater in all four of the population groups.

Our initial approach to simulating markers for a given population stratum was to fit a hidden Markov model (HMM) to the haplotypes in that stratum, chromosome by chromosome, using fastPHASE (Scheet and Stephens, 2006), and use this fitted model to simulate individual multilocus genotypes using SNPknock (Sesia et al., 2019). The simulated data are then LD pruned and principal components are computed from the thinned panel of markers. However, the computation involved in this approach proved to be prohibitive. For example, fitting the HMMs took up to 5 h per chromosome. We therefore considered two computationally cheaper alternatives. In the first alternative, we started from an LD-pruned set of markers in the original data and fit HMMs to this set. In the second alternative, we used the same panel of pruned markers, but simulated genotypes *independently* based on the MAFs in the population strata. In what follows we refer to the first and second alternatives as *LD-based* and *independent* marker simulation, respectively.

Independent markers could contain more information about the population strata than markers in LD. As a result, PC adjustment with independent markers might control type 1 error more effectively than adjustment with markers in LD. To assess this possibility, we completed 100 preliminary simulation replicates using LD-based marker simulation and 5,000 replicates using independent marker simulation. We simulated trios from four population strata under the null hypothesis of no $G \times E$, used the PC selection method of (Gavish and Donoho, 2014) to adjust the risk model and estimated the resulting type 1 error rates. Estimated type 1 error rates and their 95% confidence intervals under the LD-based and independent simulation methods were 0.04 (0.002, 0.078) and 0.0496 (0.044, 0.056), respectively, and consistent with similar type 1 error rates for the two approaches. We therefore used the faster simulation of independent markers for the simulation study.

In Section 3.2, Section 3.3 we present type I error and power results for two, three or four population strata. For two strata ($S = 0$ and $S = 1$), marker simulations were based on the CHB and IBS population groups. For three strata ($S = 0$, $S = 1$ and $S = 2$), simulations were based on the CHB, IBS and CDX population groups.

3 Results

3.1 Selection of principal components

All PC selection methods performed well when the sizes of the population strata were equal (results not shown), but not when the sizes were unequal. We illustrate with simulation results involving datasets of 3,000 trios sampled from four unequal-sized strata. For K+1 = 4 populations we require K = 3 PCs. In 5,000 simulation replicates, the method of (Gavish and Donoho, 2014) always selected three, the method of (Zhu and Ghodsi, 2006) always selected one, and the method of (Patterson et al., 2006) selected three PCs 4,942 times and four PCs 58 times. Other simulation results with unequal-sized strata (not shown) yielded similar results. Therefore, in what follows we use the method of (Gavish and Donoho, 2014) to select PCs.

3.2 Type I error rate

We compared the type I error rates of the test for $G' \times E$ using (i) adjustment with the true stratum membership S, (ii) the EEGM adjustment of (Shin et al., 2012), and (iii) PC adjustment. Results for simulated datasets with equal or unequal stratum sizes are shown in Table 2. For both equal and unequal stratum sizes, adjustment by S or direct PCs maintains the nominal 5% error rate regardless of the number of strata. By contrast, EEGM adjustment leads to an inflated type I error rate when there are more than two strata. In light of the inflated size of the test, we do not consider EEGM adjustment in the following section on power.

3.3 Power

Table 3 provides a comparison of estimated power when data are simulated from two, three or four strata. Results are shown for simulations using both equal and unequal stratum sizes and for different values of the $G \times E$ effect. From these results we see that power increases with effect size, decreases with number of strata and tends to be slightly larger for unequal strata than equal strata. Importantly, the estimated power under PC adjustment is always within simulation error of that under adjustment for true stratum membership.

4 The GENEVA Oral Cleft study

4.1 Data and objectives

The GENEVA Oral Cleft study (GENEVA, 2010) is comprised of 550 case-parent trios from 13 different sites across the United States, Europe, Southeast and East Asia. Data were obtained through dbGAP at https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000094.v1.p1 with accession number phs000094.v1.p1. Of the 550 trios, only 462 were available for analysis. Summaries of the trios by ancestry and gender of the affected child are shown in Table 4. From this table we see the ancestry of the sample is predominantly European (46%) and East Asian (51%).

The objective of the GENEVA study is to discover genetic contributions to orofacial clefts, the most common type of craniofacial birth defect in humans, and to assess whether these genes modify the effect of exposures known to be associated with cleft palate. Maternal exposure to multivitamins, alcohol and smoking were assessed through maternal interviews focused on the peri-conceptual period (3 months prior to conception through the first trimester), which includes the first 8–9 weeks of gestation when palatal development is completed. Exposure status is summarized in Table 5. From this table we see that the three dichotomous exposures are all more common in Europeans. In contrast to the continuous exposures of the simulation study, the exposures we consider in the GENEVA study are all dichotomous.

4.2 GENEVA data analysis

4.2.1 PC selection

LD pruning of the genome-wide panel of SNPs at an r^2 threshold of 0.1 yielded 63,694 markers. In a principal component analysis of these markers, the first PC explains 6.3% of the total variance and all others explain less than 0.4%. Not surprisingly, the method of (Gavish and Donoho, 2014) selects one PC. A plot of the projections of the data onto the first two PCs is shown in Figure 3, with points colored by self-reported ancestry. Each PC has been shifted by subtracting the minimum value and scaled by the range so that the values are between zero and one. The first PC distinguishes those with self-reported East Asian ancestry from those with self-

TABLE 3 Estimated power (top entry) and corresponding 95% confidence intervals (bottom entry) of different adjustment schemes for different $G \times E$ interaction effects β_{gE}, number of strata and stratum-size distributions.

Equal stratum sizes					
		β_{gE}			
Num. Strata	Adjustment	−0.10	−0.15	−0.20	−0.25
2	S	0.2602	0.5660	0.8420	0.9558
		(0.248, 0.272)	(0.552, 0.580)	(0.832, 0.852)	(0.950, 0.961)
	PC	0.2580	0.5660	0.8404	0.9564
		(0.246, 0.270)	(0.552, 0.580)	(0.830, 0.850)	(0.951, 0.962)
3	S	0.1742	0.3844	0.6498	0.8288
		(0.164, 0.185)	(0.371, 0.398)	(0.636, 0.663)	(0.818, 0.839)
	PC	0.1788	0.3920	0.6616	0.8316
		(0.168, 0.189)	(0.378, 0.406)	(0.648, 0.675)	(0.821, 0.842)
4	S	0.1306	0.2766	0.5010	0.6970
		(0.121, 0.140)	(0.264, 0.289)	(0.487, 0.515)	(0.684, 0.710)
	PC	0.1396	0.2936	0.5088	0.6918
		(0.130, 0.149)	(0.281, 0.306)	(0.495, 0.523)	(0.679, 0.704)
Unequal stratum sizes					
		β_{gE}			
		−0.10	−0.15	−0.20	−0.25
2	S	0.2636	0.5724	0.8328	0.9518
		(0.251, 0.276)	(0.559, 0.586)	(0.822, 0.843)	(0.946, 0.958)
	PC	0.2648	0.5722	0.8322	0.9514
		(0.252, 0.277)	(0.558, 0.586)	(0.822, 0.842)	(0.945, 0.957)
3	S	0.1950	0.4322	0.7082	0.8640
		(0.184, 0.206)	(0.418, 0.446)	(0.696, 0.721)	(0.854, 0.874)
	PC	0.1936	0.4334	0.7054	0.8632
		(0.183, 0.204)	(0.420, 0.447)	(0.693, 0.718)	(0.854, 0.873)
4	S	0.1614	0.3470	0.6028	0.7894
		(0.151, 0.172)	(0.334, 0.360)	(0.589, 0.616)	(0.778, 0.801)
	PC	0.1598	0.3380	0.5872	0.7820
		(0.150, 0.170)	(0.325, 0.351)	(0.574, 0.601)	(0.770, 0.794)

reported European ancestry; hence, a value near zero corresponds to a hypothetical East Asian and a value near one corresponds to a hypothetical European. The second PC separates the single self-reported African child from all others.

4.2.2 Inference of $G \times E$

The conditional-likelihood methods outlined in Supplementary Appendix S1 were applied to the data. We focused on inference of $G \times E$ between maternal alcohol consumption and the six SNPs in the *MLLT3* gene that had

TABLE 4 Gender of 462 affected children by self-reported ancestry.

Ancestry	Males	Females	Total	%
European	103	111	214	46%
Asian	93	141	234	51%
Other/Afr	3	11	14	3%
Total	199	263	462	100%

TABLE 5 Exposure rates for maternal alcohol consumption, maternal smoking and maternal vitamin supplementation by self-reported ancestry in affected trios.

Ancestry	Percent exposed to Maternal			Affected children
	Alcohol	Smoking	Vitamin Supp	
European	41%	28%	57%	214
East Asian	4%	3%	21%	234
Other/Afr	14%	7%	71%	14
Total	21%	14%	39%	462

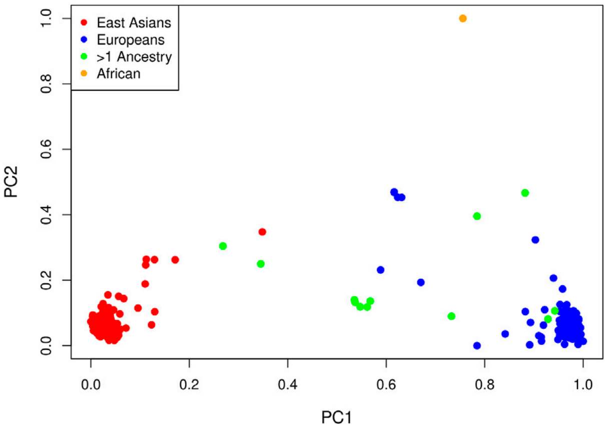

FIGURE 3
Projections of each affected child onto the first two PCs by self-reported ancestry: red = East Asian (234 trios), blue = European (214 trios), orange = African (one trio) and green = multiple ancestry/other (13 trios). Each PC has been shifted and scaled so that a PC1 value near zero corresponds to a hypothetical East Asian and a PC1 value near one corresponds to a hypothetical European.

significant $G \times E$ at the 5% level in the analysis of (Beaty et al., 2011). Displays of the LD between these SNPs and others nearby (Shin et al., 2006) are shown in Supplementary Figure S1, Supplementary Appendix S3, for self-reported European subjects and self-reported East Asian subjects. Table 6 shows the results of fitting three different log-linear models of $G' \times E$. Following (Beaty et al., 2011), each is based on an additive genetic model that specifies equal log-GRRs for genotypes $g' = 1$ or 2. Results based on fitting a more general co-dominant model (1) were similar (results not shown). The first model, as in (Beaty et al., 2011), makes no adjustment for exposure-related genetic structure in the population, the second uses EEGM adjustment and the third uses PC adjustment. From the table we see that, for each test SNP, p-values for the tests of $G' \times E$ are smallest when we make no adjustment. Comparing the EEGM and PC adjustment approaches we find that p-values from PC adjustment are similar to, but tend to be slightly smaller than, those from the EEGM adjustment. Of the six test SNPs show in the table, four retain significance at the 5% level after adjustment for exposure-related genetic structure.

The estimates shown in Table 6 are of the multiplicative factors by which maternal alcohol consumption modifies the GRRs at the six test SNPs. For a binary exposure such as maternal alcohol consumption, these modifying effects can be obtained by exponentiating the interaction term in the log-GRR model. With no adjustment for genetic structure there is a single interaction term and hence a single estimated modifying effect for all trios. For example, maternal alcohol consumption is estimated to increase the GRR at SNP rs4621895 by a factor of about 2.1 for all trios. By contrast, with EEGM or PC adjustment the interaction term depends on the value of the adjustment variable and we have reported estimates for hypothetical East Asian and European subjects in our sample. For example, maternal alcohol consumption is estimated to decrease the GRR at SNP rs4621895 by a factor of about 0.73 for East Asian trios and to increase the same GRR by a factor of about 2.4 for European trios. For these data, the adjustment variables used in the EEGM- and PC-adjustment approaches are highly correlated (Pearson correlation 0.996), and so the estimates for the two approaches are very similar. These estimates are also similar to those obtained from an analysis using self-reported ancestry (results not shown). The 95% confidence intervals for hypothetical East Asians cover one for each SNP but do not cover one for hypothetical Europeans, with the exception of SNP rs2780841. These results suggest that any $G \times E$ signal is from trios of European ancestry, where maternal alcohol consumption is more common.

5 Discussion

We consider a log-linear model of GRRs at a causal locus G. Under this model, $G \times E$ is equivalent to GRRs that vary with the exposure E. We show that exposure-related genetic structure in the population can lead to spurious $G' \times E$ at a non-causal test locus G' in LD with G. However, valid inference of $G' \times E$ can be obtained by augmenting the GRR model with a blocking variable X, such that GG' haplotypes and E are conditionally independent given X. We discuss the choice of X for inference of

TABLE 6 Estimated modifying effects of maternal alcohol consumption on GRRs, 95% confidence intervals and *p*-values from the analysis of the GENEVA data, at six SNPs in the MLLT3 gene (Chr 9) showing significant interaction with maternal alcohol consumption in (Beaty et al., 2011). Estimates, confidence intervals and tests are based on fitting an additive genetic model and use (i) no adjustment, (ii) EEGM adjustment or (iii) PC adjustment to control for exposure-related genetic structure in the population. The unadjusted analysis considers all trios without regard to genetic structure. The EEGM- and PC-adjusted analyses allow for genetic structure and we have reported estimates for hypothetical East Asian and European subjects.

	All			East Asian		European		
SNP	Adj	Est	95% CI	Est	95% CI	Est	95% CI	*p*-value
rs4621895	None	2.08	(1.36, 3.18)	–	–	–	–	0.0006
	EEGM	–	–	0.762	(0.214, 2.72)	2.44	(1.42, 4.20)	0.0047
	PC	–	–	0.701	(0.181, 2.72)	2.40	(1.42, 4.04)	0.0037
rs4977433	None	2.15	(1.40, 3.30)	–	–	–	–	0.0003
	EEGM	–	–	0.916	(0.244, 3.44)	2.47	(1.44, 4.25)	0.0036
	PC	–	–	0.854	(0.208, 3.45)	2.44	(1.45, 4.11)	0.0028
rs6475464	None	1.75	(1.13, 2.69)	–	–	–	–	0.0104
	EEGM	–	–	0.909	(0.271, 3.05)	2.25	(1.29, 3.95)	0.0158
	PC	–	–	0.840	(0.234, 3.02)	2.22	(1.29, 3.81)	0.0139
rs668703	None	2.02	(1.33, 3.07)	–	–	–	–	0.0008
	EEGM	–	–	0.588	(0.177, 1.95)	2.50	(1.45, 4.29)	0.0032
	PC	–	–	0.531	(0.148, 1.91)	2.43	(1.44, 4.09)	0.0025
rs623828	None	1.55	(1.00, 2.39)	–	–	–	–	0.0481
	EEGM	–	–	0.772	(0.239, 2.50)	1.77	(1.01, 3.11)	0.1368
	PC	–	–	0.757	(0.220, 2.60)	1.73	(1.00, 2.98)	0.1384
rs2780841	None	1.55	(1.01, 2.36)	–	–	–	–	0.0417
	EEGM	–	–	0.653	(0.217, 1.96)	1.71	(0.960, 3.04)	0.1613
	PC	–	–	0.620	(0.195, 1.97)	1.68	(0.965, 2.93)	0.1471

$G' \times E$ when data are collected from a study of case-parent trios. The population strata S would be an ideal choice for X but may not be known definitively. We propose to use principal components (PCs) instead. In particular, we calculate PCs from a genomic region unlinked to the test locus and select a parsimonious subset using the method of (Gavish and Donoho, 2014). We then specify a linear model for the log-GRRs whose intercept and slope depend on PC values. Slopes that vary with PC values allow the modifying effect of the exposure to vary with population strata, which can be important for maintaining power [20, Section 3.3]. Through simulations, we show that our PC adjustment maintains the nominal type-1 error rate and has nearly identical power to detect $G \times E$ as an oracle approach based directly on S. We illustrate our approach by applying it to an analysis of real data from case-parent trios in the GENEVA Oral Cleft Study. In our analysis of the GENEVA data, we focussed on SNPs and exposures identified by (Beaty et al., 2011). In a discussion of their results, these authors noted that the SNPs they identified are not in known cleft-palate susceptibility genes and are either intronic or are upstream/downstream of coding regions. This lack of compelling biological plausibility, coupled with the striking differences in exposure distributions between the self-reported European and East Asian strata, motivated our $G \times E$ analysis that adjusts for population structure. However, our results (Table 6) and those of (Ratnasekera and McNeney, 2021) do not contradict the hypothesis of $G \times E$, but rather suggest that any $G \times E$ signal is due to the self-reported European trios. Further data collection aimed at self-reported European trios may provide stronger conclusions regarding the presence of $G \times E$.

To reduce bias from exposure-related genetic structure, direct PC adjustment has advantages over the EEGM approach and design-based strategies such as the tetrad approach of (Shi et al., 2011) and the sibling-augmented case-only approach of (Weinberg et al., 2011). Unlike the EEGM approach, PC adjustment controls the type 1 error when there are more than two population strata. Unlike the design-based strategies, PC adjustment does not require siblings nor assume binary exposures.

Development of alternative approaches based on propensity scores is an area for future work. The EEGM approach is attractive in that it reduces the genetic principal components to a single score, $E(E|\hat{S})$. For binary exposures, such as those in the GENEVA study, the EEGM is a propensity score (Rosenbaum and Rubin, 1983). For continuous exposures, such as those in the simulation study, the analog to the EEGM is a continuous-treatment propensity score (Brown et al., 2021). With continuous exposures, we could predict E given the genetic markers and *then* convert the predictions to a Normal density score that takes low values for predictions far from their observed value. These density scores could be used either as predictors (Hirano and Imbens, 2004) or weights (Robins et al., 2000) in subsequent analyses. It would be interesting to explore the use of propensity-score methods in inference of $G' \times E$ from case-parent trios with continuous exposures, particularly when there are more than two population strata.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000094.v1.p1

Author contributions

PR developed the statistical methods, performed the simulations and data analyses, and wrote the initial draft of the manuscript. BM and JG conceptualized the study and revised the manuscript. All authors proofread and approved the final version of the manuscript.

Funding

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), grants RGPIN/04296-2018 (JG) and RGPIN/05595-2019 (BM). This research was enabled in part by support provided by the BC DRI Group and the Digital Research Alliance of Canada (http://alliancecan.ca). Funding support for the study entitled “International Consortium to Identify Genes and Interactions Controlling Oral Clefts” was provided by several previous grants from the National Institute of Dental and Craniofacial Research (NIDCR). Data and samples were drawn from several studies awarded to members of this consortium. Funding to support original data collection, previous genotyping and analysis came from several sources to individual investigators. Funding for individual investigators include: R21-DE-013707 and R01-DE-014581 (Beaty); R37-DE-08559 and P50-DE-016215 (Murray, Marazita) and the Iowa Comprehensive Program to Investigate Craniofacial and Dental Anomalies (Murray); R01-DE-09886 (Marazita), R01-DE-012472 (Marazita), R01-DE-014677 (Marazita), R01-DE-016148 (Marazita), R21-DE-016930 (Marazita); R01-DE-013939 (Scott). Parts of this research were supported in part by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (Wilcox, Lie). Additional recruitment was supported by the Smile Train Foundation for recruitment in China (Jabs, Beaty, Shi) and a grant from the Korean government (Jee). The genome-wide association study, also known as the Cleft Consortium, is part of the Gene Environment Association Studies (GENEVA) program of the trans-NIH Genes, Environment and Health Initiative (GEI) supported by U01-DE-018993. Genotyping services were provided by the Center for Inherited Disease Research (CIDR). CIDR is fully funded through a federal contract from the National Institutes of Health (NIH) to The Johns Hopkins University, contract number HHSN268200782096C. Funds for genotyping were provided by the NIDCR through CIDR’s NIH contract. Assistance with genotype cleaning, as well as with general study coordination, was provided by the GENEVA Coordinating Center (U01-HG-004446) and by the National Center for Biotechnology Information (NCBI).

Acknowledgments

We sincerely thank all of the patients and families at each recruitment site for participating in this study, and we gratefully acknowledge the invaluable assistance of clinical, field and laboratory staff who contributed to this effort over the years.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.1065568/full#supplementary-material

References

Beaty, T. H., Ruczinski, I., Murray, J. C., Marazita, M. L., Munger, R. G., Hetmanski, J. B., et al. (2011). Evidence for gene-environment interaction in a genome wide study of nonsyndromic cleft palate. *Genet. Epidemiol.* 35, 469–478. doi:10.1002/gepi.20595

Brown, D. W., Greene, T. J., Swartz, M. D., Wilkinson, A. V., and DeSantis, S. M. (2021). Propensity score stratification methods for continuous treatments. *Stat. Med.* 40, 1189–1203. doi:10.1002/sim.8835

Clarke, L., Fairley, S., Zheng-Bradley, X., Streeter, I., Perry, E., Lowy, E., et al. (2016). The international genome sample resource (igsr): A worldwide collection of genome variation incorporating the 1000 genomes project data. *Nucleic Acids Res.* 45, D854–D859. doi:10.1093/nar/gkw829

Garaulet, M., Smith, C. E., Hernández-González, T., Lee, Y. C., and Ordovás, J. M. (2011). PPARγ 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. doi:10.1002/mnfr.201100437

Gavish, M., and Donoho, D. L. (2014). The Optimal Hard Threshold for Singular Values is $4/\sqrt{3}$. *IEEE Trans. Inf. Theory* 60, 5040–5053. doi:10.1109/TIT.2014.2323359

GENEVA (2010). *GENEVA oral clefts project imputation report - HapMap III reference panel [pdf file]*. [Dataset]. Available at: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000094.v1.p1.

Grinde, K. (2019). *Statistical inference in admixed populations*. Ph.D. thesis. University of Washington.

Hirano, K., and Imbens, G. W. (2004). *The propensity score with continuous treatments*. John Wiley & Sons, 73–84. chap. 7. doi:10.1002/0470090456.ch7

Hunter, D. J. (2005). Gene-environment interactions in human diseases. *Nat. Rev. Genet.* 6, 287–298. doi:10.1038/nrg1578

Patterson, N., Prince, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genet.* 12, e190–e2093. doi:10.1371/journal.pgen.0020190

Pearl, J. (2009). Causal inference in statistics: An overview. *Stat. Surv.* 3, 96–146. doi:10.1214/09-SS057

Pearl, J. (1998). Graphs, causality, and structural equation models. *Sociol. Methods & Res.* 27, 226–284. doi:10.1177/0049124198027002004

Ratnasekera, P., and McNeney, B. (2021). Re-Analysis of a genome-wide gene-by-environment interaction study of case parent trios, adjusted for population stratification. *Front. Genet.* 11, 600232. doi:10.3389/fgene.2020.600232

Robins, J. M., Hernán, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* 11, 550–560. doi:10.1097/00001648-200009000-00011

Rosenbaum, P. R., and Rubin, D. B. (1983). "The central role of the propensity score in observational studies for causal effects," in *Matched sampling for causal effects* (Cambridge University Press), 170–184. doi:10.1017/cbo9780511810725.016

Scheet, P., and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* 78, 629–644. doi:10.1086/502802

Sesia, M., Sabatti, C., and Candès, E. (2019). Gene hunting with hidden markov model knockoffs. *Biometrika* 106, 1–18. doi:10.1093/biomet/asy033

Shi, M., Umbach, D. M., and Weinberg, C. R. (2011). Family-based gene-by-environment interaction studies: Revelations and remedies. *Epidemiology* 22, 400–407. doi:10.1097/ede.0b013e318212fec6

Shin, J.-H., Blay, S., McNeney, B., and Graham, J. (2006). Ldheatmap: An r function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. *J. Stat. Softw.* 16. Code Snippet 3. doi:10.18637/jss.v016.c03

Shin, J.-H., Infante-Rivard, C., Graham, J., and McNeney, B. (2012). Adjusting for spurious gene-by-environment interaction using case-parent triads. *Stat. Appl. Genet. Mol. Biol.* 11, 1714. doi:10.2202/1544-6115.1714

Shin, J.-H., McNeney, B., and Graham, J. (2013). trioGxE: A data smoothing approach to explore and test gene-environment interaction in case-parent trio data. R package version 0.1-1

Shin, J. H., Infante-Rivard, C., McNeney, B., and Graham, J. (2014). A data-smoothing approach to explore and test gene-environment interaction in case-parent trios. *Stat. Appl. Genet. Mol. Biol.* 13, 159–171. doi:10.1515/sagmb-2013-0023

Thomas, D. (2010). Gene–environment-wide association studies: Emerging approaches. *Nat. Rev. Genet.* 11, 259–272. doi:10.1038/nrg2764

Wang, H., Haiman, C. A., Kolonel, L. N., Henderson, B. E., Wilkens, L. R., Le Marchand, L., et al. (2010). Self-reported ethnicity, genetic structure and the impact of population stratification in a multiethnic study. *Hum. Genet.* 128, 165–177. doi:10.1007/s00439-010-0841-4

Weinberg, C. R. (1999). Methods for detection of parent-of-origin effects in genetic studies of case-parents triads. *Am. J. Hum. Genet.* 65, 229–235. doi:10.1086/302466

Weinberg, C. R., Shi, M., and Umbach, D. M. (2011). A sibling-augmented case-only approach for assessing multiplicative gene-environment interactions. *Am. J. Epidemiol.* 174, 1183–1189. doi:10.1093/aje/kwr231

Xu, G. (2017). Identifiability of restricted latent class models with binary responses. *Ann. Stat.* 45, 675–707. doi:10.1214/16-aos1464

Yu, Z., Demetriou, M., and Gillen, D. L. (2015). Genome-wide analysis of gene-gene and gene-environment interactions using closed-form wald tests. *Genet. Epidemiol.* 39, 446–455. doi:10.1002/gepi.21907

Zaykin, D. V., and Shibata, K. (2008). Genetic flip-flop without an accompanying change in linkage disequilibrium. *Am. J. Hum. Genet.* 82, 794–796. doi:10.1016/j.ajhg.2008.02.001

Zhu, M., and Ghodsi, A. (2006). Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Comput. Statistics Data Analysis* 51, 918–930. doi:10.1016/j.csda.2005.09.010