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

by

Pulindu Ratnasekera

M.Sc., Simon Fraser University, 2014
B.Sc., University of Sri Jayewardenepura, 2010

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
Department of Statistics and Actuarial Science
Faculty of Science

SIMON FRASER UNIVERSITY
Fall 2023

Declaration of Committee

Name:	**Pulindu Ratnasekera**
Degree:	**Doctor of Philosophy**
Thesis title:	**Inference of gene-environment interaction from heterogeneous case-parent trios**
Committee:	**Chair:** Tim Swartz Professor, Statistics and Actuarial Science
	Brad McNeney Supervisor Associate Professor, Statistics and Actuarial Science
	Jinko Graham Committee Member Professor, Statistics and Actuarial Science
	Lloyd Elliot Examiner Assistant Professor, Statistics and Actuarial Science
	Shelley Bull External Examiner Professor, Dalla Lana School of Public Health University of Toronto

Abstract

Population stratification is a major source of confounding in gene-by-environment ($G \times E$) interaction studies of case-parent trios when the study sample consists of individuals with distinct ancestral backgrounds. This dissertation discusses different ways of controlling for population stratification in order to reduce the chance of false positive signals in $G \times E$ inferences. This work is organized in three parts. First, we investigate the impact of confounding on the results of a genome-wide association analysis by Beaty et al., which identified multiple single nucleotide polymorphisms that appeared to modify the effect of maternal smoking, alcohol consumption, or multivitamin supplementation on risk of cleft palate. The study sample of case-parent trios was primarily of European and East Asian ancestry, and the distribution of all three exposures differed by ancestral group. Such differences raise the possibility that confounders, rather than the exposures, are the risk modifiers and hence that the inference of $G \times E$ interaction may be spurious. Our analyses generally confirmed the result of Beaty et al. and suggest the interaction $G \times E$ is driven by the European trios, whereas the East Asian trios were less informative. Next, we show that current methods to reduce the bias in estimated $G \times E$ interactions from case-parent trio data can only account for simple population structure involving two strata and propose methods to overcome this limitation. Through simulations, we show that our proposed method maintains the nominal type-1 error rate and higher statistical power. The proposed approach was then applied to case-parent trios which consists of cleft-palate-affected children. Consistent with Beaty et al., our results suggest that the gene-environment interaction signal in these data is due to the self-reported European trios. Finally, we discuss methods to infer local ancestry of cleft-palate-affected children and propose methods to control for population stratification via local ancestry. Again, we apply our methods to the case-parent trio data of cleft-palate-affected children to investigate whether local ancestry rather than the genotype at a test locus modifies the association between disease and exposure.

Keywords: gene-environment interaction; case-parent trios; genotype relative risk; population stratification; genome-wide association study; cleft palate; principal components; local ancestry; fastPHASE

Dedication

For my mother, my father and Professor Emeritus R.A Dayananda.

Acknowledgements

First and foremost, I would like to convey my gratitude to my senior supervisor Dr. Brad McNeney for bringing me to the field of Satistical Genetics. Thank you for your encouragement, guidance and patience throughout my PhD studies. I value everything you have done for me throughout my life at Simon Fraser University. Secondly, I would like to thank Dr. Tim Swartz and I am most indebted to him for reposing his trust and confidence in me which enabled me to begin this great educational experience at Simon Fraser University.

My sincere thanks to the examining committee: Dr. Brad McNeney, Dr. Jinko Graham, Dr. Lloyd Elliot and Dr. Shelley Bull for their valuable inputs.

I take this opportunity to convey my gratitude to the faculty in the department of Statistics and Actuarial Science at Simon Fraser University. You provide an incredible learning environment in the department. I am grateful for the financial support provided by the department of Statistics and Actuarial Science. Special thanks to Statistics Workshop Manager Marie Loughin for her support and guidance during my time at Simon Fraser University. I extend my gratitude to Sadika, Charlene, Anna and Caitlin for their kind assistance.

Furthermore, I would like to thank my fellow graduate student colleagues for their friendship and camaraderie and for the fun times we had together. To Lasantha, Harsha, Rajitha and Bhagya my Sri Lankan graduate colleagues at SFU, I want to say thank you for helping me with my studies as well as for enjoyable times we had in Vancouver.

Last but not least, I am indebted to my family in Sri Lanka as well as my wife and children for their patience and blessing throughout my many years at SFU. I appreciate everything you all have done for me.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

Gene and environment factors contribute to etiology of structural birth defects such as oral clefts. Such diseases are thought to result from an interplay between gene and environment factors. It is believed that the study of gene-environment interactions will lead to better understanding of the contribution of genetic and non-genetic factors to the development of complex birth defects.

The case-parent trio design is often used for estimation and testing genetic effects and gene-by-environment interactions for such early onset diseases. The case-parent trio data consist of genotypes (G) of unrelated children affected with a disease and genotypes from their parents. Information may also be collected on non-genetic environmental (E) factors.

The contribution of genetic effects on diseases can be measured by genotype relative risk (GRR) in individuals with an alternate genotype compared to those with some reference genotype. The contribution of statistical interaction between gene-environment ($G \times E$) interaction can be measured by observing variation in GRRs for different values of E.

$G \times E$ inference from case-parent trio data can be subject to bias when trios are pooled from multiple different sites. The pooling of trios from multiple different sites leads to population stratification, which is a major source of confounding bias.

In this thesis, we propose methods to reduce bias in inference of $G \times E$ from case-parent trio data and apply our proposed methods to both simulated and real data from GENEVA Oral Cleft Study. The thesis consists of three projects. The work in Chapters 2 and 3 have been published. As a result, some introductory material are repeated in more than one chapter.

1.2 Organization of this Thesis

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.

Beaty et al. [3] identified multiple single nucleotide polymorphisms that appeared to modify the effect of maternal smoking, alcohol consumption, or multivitamin supplementation on risk of cleft palate. The study sample of case-parent trios was primarily of European and East Asian ancestry, and the distribution of all three exposures differed by ancestral group. Such differences raise the possibility that confounders, rather than the exposures, are the risk modifiers and hence that the inference of gene-environment ($G \times E$) interaction may be spurious. In Chapter 2, we re-analysed the $G \times E$ inferences reported in [3] by introducing the bias-reduced methods proposed by Shin et al. [27] for inference of $G \times E$ interaction from case-parent trio data in the presence of exposure-related population structure. This work has been published in Ratnasekera and McNeney (2020) [19].

The method of [27] can only account for simple population structure involving two strata. To allow for more than two strata, in Chapter 3 we propose to directly accommodate multiple population strata by adjusting for genetic principal components. We evaluate our approach through simulation and illustrate it on data from a study of genetic modifiers of cleft palate. This work has been published in Ratnasekera et al. (2022) [18].

Both Shin et. al. [27] and Ratnasekera et al. [18] look at the global ancestry of affected individuals when making bias-reduced inferences of $G \times E$. However, global ancestry could be less relevant when individuals in the study sample are pooled from multiple different geographic location due to possible admixture. In such circumstances, the global ancestry of a particular individual could be different from the local ancestry at a given haplotype of interest. In chapter 4, we discuss methods to infer local ancestry of individuals and we apply the proposed methods to cleft palate data to infer the local ancestry of individuals at a given haplotype. The inferred local ancestry was used to modify the bias-reduced methods proposed by [18]. The modified bias-reduced methods were then applied to cleft palate data for inference of $G \times E$.

In the last chapter, we make concluding remarks. Some theoretical details and supporting results for Chapters 2, 3 and 4 are provided in Appendices A, B and C, respectively.

Chapter 2

Re-analysis of a genome-wide gene-by-environment interaction study of case parent trios, adjusted for population stratification

2.1 Introduction

In a case-parent trio study we collect genotypes, G, from affected children and their parents. We may also collect environmental exposures or non-genetic attributes, E, from the children. Gene-environment interaction ($G \times E$) is a statistical interaction term in the standard log-linear model of association between covariates and disease status. $G \times E$ interaction can be interpreted as genotypes that modify the effect of an exposure, or as an exposure that modifies the effect of genotypes.

Beaty *et al.* [3] conducted an analysis of the GENEVA Oral Cleft Study data, a genome-wide association study to identify genetic and environmental factors associated with cleft palate (CP). They found multiple single nucleotide polymorphisms (SNPs) appeared to modify the effect of maternal smoking, maternal alcohol consumption or maternal multivitamin supplementation on the risk of CP. The study sample of case-parent trios was primarily of European and East Asian ancestry, and the distribution of all three exposures differ by ancestral group, which raises the possibility of spurious $G \times E$.

When the test locus, G', is not causal, disease risk at G' can appear to be modified by E without $G \times E$ interaction when there is exposure-related population structure [24, 34]. Exposure-related population structure may be thought of as a form of confounding that occurs when both GG' haplotype frequencies and E distributions differ by ancestral group. Differences in GG' haplotypes can lead to differences in G' risk that may be tagged by E, suggesting $G' \times E$ even in the absence of true $G \times E$ interaction.

One of the requirements for exposure-related population structure is different E distributions in the different ancestral groups. Such is the case in the GENEVA Oral Cleft

Study where maternal smoking, maternal alcohol consumption and maternal multivitamin supplementation were all more common in self-reported Europeans than in self-reported East Asian populations (see Table 3.5). If, in addition, haplotype frequencies for markers in the vicinity of a causal SNP also vary by ancestral groups, any inference of $G \times E$ interaction could be spurious. Shin *et al.* [27] proposed bias-reduced methods for inference of $G \times E$ interaction from case-parent trio data in the presence of exposure-related population structure. In this article we use these methods to adjust the analyses of [3] for potential confounding effects of population stratification.

2.2 The GENEVA Oral Cleft Study

The GENEVA Oral Cleft study [1] was comprised of 550 case-parent trios from 13 different sites across the United States, Europe, Southeast and East Asia. For our analyses, 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 included in the GENEVA Oral Cleft study, 462 were available for our analyses. Summaries of the trios by ancestry and sex of the affected child are shown in Table 2.1.

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

Table 2.1: Sex of 462 CP cases in the international consortium study by ancestral group

The objective of the GENEVA Oral Clefts study was 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 2.2. From this table we see the ancestry of the sample is predominantly European (46%) and East Asian (51%), and these three exposures are all more common in Europeans.

	Exposure (%) to Maternal			
Ancestry Group	Alcohol Consumption	Smoking	Vitamin Supplementation	Affected children
European	41%	28%	57%	214
East Asian	4%	3%	21%	234
Other/Afr	14%	7%	71%	14
Total	21%	14%	39%	462

Table 2.2: Exposure rates for maternal alcohol consumption, maternal smoking and maternal vitamin supplementation for the total CP group and for three ancestral groups

Beaty *et al.* found evidence for $G \times E$ interaction for maternal alcohol consumption and SNPs in the genes *MLLT3* and *SMC2*, for maternal smoking and SNPs in the genes *TBK1* and *ZNF236*, and for maternal multivitamin use and SNPs in the *BAALC* gene.

2.3 Models and Methods

$G \times E$ interaction is defined as the statistical interaction term β_{GE} in a log-linear model for the probability of disease:

$$\log[P(D = 1|G = g, E = e)] = \beta_0 + g\beta_G + e\beta_E + ge\beta_{GE}, \tag{2.1}$$

where $D = 1$ indicates the child is affected with cleft palate, G is the child's genotype, coded as 0, 1 or 2 copies of the minor allele, E is an exposure variable and β_E is the environmental main effect, which cannot be directly estimated in the case-parent trio design. To simplify the presentation we assume throughout E is a binary variable with value 1 indicating exposure and 0 indicating no exposure. A non-zero interaction effect, β_{GE}, suggests G modifies the effect of E on disease risk.

Assuming G and E are independent given parental genotypes G_p, and the risk model in equation (2.1), one can derive the conditional distribution of G given $D = 1$, E and G_p, which can be stated in terms of the *genotypic odds*

$$\frac{P(G = g \mid D = 1, E = e, G_p = g_p)}{P(G = g - 1 \mid D = 1, E = e, G_p = g_p)} = \exp\left(k_p + \beta_G + e\beta_{GE}\right)$$

for a constant k_p that depends on g_p (Appendix A.1). The environmental main effect does not appear in this genotypic odds and cannot be estimated from case-parent trio data. In general, any regression effect that does not involve G cannot be estimated from case-parent trio data.

To reduce bias from population stratification, robust methods are required. Robust methods may be classified as design- or data-based. For a binary environmental exposure, the design-based approach of Shi *et al.* [24] augments the basic case-parent trio with exposure information on an unaffected sibling. Shi *et al.* showed including the sibship-averaged exposure over two sibs (affected/unaffected) in the linear model controls this potential bias. Weinberg *et al.* [34] showed all information about interaction in the tetrad design of [24] comes from the siblings, not the parents, which lead them to propose a sibling-augmented case-only design and analysis. Shin *et al.* [27] took a data-based approach and replaced the sibship-averaged exposure of [24] with the predicted exposure given ancestry, as reflected by principal components (PCs) computed from independent genetic markers. Their data-based approach is applicable for arbitrary exposures, including continuous measures, and does not require additional siblings. In this report we consider this data-based approach of [27] to explore whether the $G \times E$ interaction effects reported by [3] could be spurious.
Let X_E be a categorical variable indicating ancestral groups with different E distributions. Shin *et al.* introduce a separate genetic effect for each level of X_E into the risk model Equation 2.1. When X_E is binary, this modified risk model becomes:

$$\log[P(D=1|G=g,E=e,X_E=x)] = \beta_0 + g\beta_G + e\beta_E + x\beta_{X_E} + ge\beta_{GE} + gx\beta_{GX_E} + ex\beta_{EX_E} + gex\beta_{GEX_E}$$

In the above model, β_{GX_E} controls bias by allowing different genetic effects in the two X_E-groups (exposed and unexposed). The term β_{GEX_E} allows for different $G \times E$ effects in the two groups, which can improve power to detect $G \times E$ interaction ([27]).

As X_E is not known, it must be replaced by some surrogate, $\hat{X}_E$. We consider two surrogates, (i) the expectation of E given genetic markers (EEGM) and (ii) self-reported ancestry (SRA). The idea behind the EEGM approach is to distinguish exposure distributions by their mean, which may vary across ancestry groups, S.Though S is not known, it is reflected in the principal components, M, computed from genetic marker data available on all ancestry groups in these data. The expectation of E given M can be estimated by linear regression of E on M when E is continuous, or by logistic regression when E is binary, as in the current study. Thus, for EEGM adjustment, we estimate the expected exposure within ancestral groups with $\hat{X}_E = \widehat{E(E|M)}$ in the risk model shown in Equation 2.2 below. We consider EEGM adjustment to be the gold-standard, because, for the case of two ancestral groups, [27] showed the resulting tests of $G \times E$ interaction achieve the nominal type I error rates. To align this EEGM adjustment results with those based on SRA adjustment, we transform EEGM to the the unit interval by subtracting the minimum value and dividing by the range. The use of SRA as a surrogate for X_E is straightforward and leads to more interpretable models (see below), but is subject to some bias, because self-reported ancestry

may not accurately reflect genetic ancestry [32]. To simplify notation we assume only two self-reported ancestry groups, encoded in $\hat{X}_E$ as as zero or one.

Distributions of EEGM by SRA for maternal exposure to alcohol, smoking and vitamin supplemements, are shown in Figure 2.1. As expected, there is a clear separation of EEGM between self-reported Europeans and East Asians for all three exposures considered here. In our analyses we would therefore expect similar results with adjustment by either EEGM or SRA.

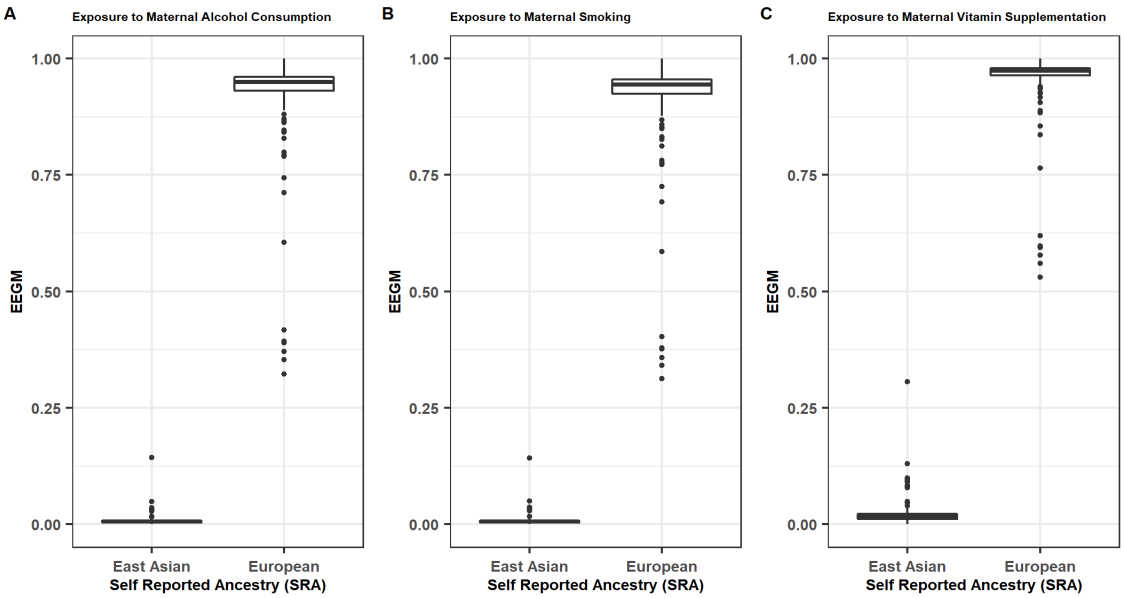

Figure 2.1: EEGM by SRA for each of the three maternal exposures of interest.

For either surrogate $\hat{X}_E$, let $\beta_{G\hat{X}_E}$ denote its coefficient so we can write the risk model as

$$\log[P(D=1|G=g,E=e,\hat{X}_E=x)] = \beta_0 + g\beta_G + e\beta_E + x\beta_{\hat{X_E}} + ge\beta_{GE} + gx\beta_{G\hat{X_E}} + ex\beta_{E\hat{X_E}} + gex\beta_{GE\hat{X_E}}, \tag{2.2}$$

leading to genotypic odds

$$\frac{P(G=g \mid D=1,E=e,\hat{X}_E=x,G_p=g_p)}{P(G=g-1 \mid D=1,E=e,\hat{X}_E=x,G_p=g_p)} = \exp\left(k_p + \beta_G + e\beta_{GE} + x\beta_{G\hat{X}_E} + ex\beta_{GE\hat{X}_E}\right). \tag{2.3}$$

Interpretation of the model in equation 2.2 is simplest when $\hat{X}_E$ is the SRA. If $\hat{X}_E = 0$ for East Asians and 1 for Europeans, then the relative risk of cleft palate due to exposure in self-reported East Asians with $G = g$ is $e^{\beta_E + g\beta_{GE}}$. Thus, $e^{\beta_{GE}}$ is the multiplicative increase in the relative-risk due to exposure for each additional copy of G in self-reported East Asians. Similarly, the relative risk of cleft palate due to exposure in self-reported

Europeans is $e^{\beta_E+\beta_{E\hat{X}_E}+g(\beta_{GE}+\beta_{GE\hat{X}_E})}$, so this latter term $e^{\beta_{GE}+\beta_{GE\hat{X}_E}}$ is the muliplicative increase in the relative-risk due to exposure for each copy of G in self-reported Europeans. To summarize, $e^{\beta_{GE}}$ reflects $G \times E$ in self-reported East Asians and $e^{\beta_{GE}+\beta_{GE\hat{X}_E}}$ reflects $G \times E$ in self-reported Europeans.

Inference can be made from the conditional likelihood based on $P(G|D = 1, E, G_p)$, which can be obtained from the genotypic odds shown in equation 2.3. Parameter estimates are obtained by maximizing the conditional likelihood, and hypothesis tests are obtained from likelihood ratio statistics. In particular, the hypothesis H_O (i.e. not GxE interaction): $\beta_{G\hat{X}_E} = \beta_{GE\hat{X}_E} = 0$, is tested first by fitting models with and without the interaction terms and comparing the resulting likelihood ratio statistic to the χ^2 distribution with 2 degrees of freedom.

2.4 Results

Inference was obtained under the unadjusted risk model in equation (2.1) and the adjusted risk model in equation (2.2) with either SRA or EEGM adjustment. The first analysis was restricted to the six SNPs in the *MLLT3* gene reported by Beaty *et al.* to have significant $G \times E$ with maternal alcohol consumption. P-values for the 1 df (no adjustment) and 2 df tests (with $\hat{X}_E$ adjustment) are shown in Table 2.3. We see the results with adjustment roughly agree, and are only slightly attenuated compared to the results without adjustment for population stratification of exposures.

SNP	Model 1: 1df LRT No adj	Model 2: 2df LRT SRA adj.	Model 2: 2df LRT EEGM adj.
rs4621895	0.0010	0.0043	0.0060
rs4977433	0.0006	0.0035	0.0046
rs6475464	0.0159	0.0128	0.0161
rs668703	0.0014	0.0029	0.0039
rs623828	0.0432	0.1105	0.1358
rs2780841	0.0475	0.1311	0.1683

Table 2.3: P-values of 1 df likelihood ratio test from Model (1), and 2 df likelihood ratio tests from models with SRA and EEGM adjustment for the six SNPs in the *MLLT3* gene on chromosome 9 show significant evidence of interaction with Maternal Alcohol Consumption

Figure 2.2 shows the estimated interaction terms $\hat{\beta}_{G\hat{X}_E}$ and $\hat{\beta}_{G\hat{X}_E} + \hat{\beta}_{GE\hat{X}_E}$ and their corresponding confidence intervals while Table 2.4 shows exponentiated parameter estimates and their confidence intervals, as well as p-values from inference based on the risk model

adjusted by SRA only. The overlap of the confidence intervals for $\hat{\beta}_{GE}$ with zero and the non-overlap of the confidence intervals for $\hat{\beta}_{GE} + \hat{\beta}_{GE\hat{X}_E}$ suggest no evidence of $G \times E$ interaction in self-reported East Asians, but evidence of $G \times E$ interaction in self-reported Europeans. The table of exponentiated parameter estimates quantifies this apparent $G \times E$ in self-reported Europeans. For example, the estimate $e^{\hat{\beta}_{GE}+\hat{\beta}_{GE\hat{X}_E}} = 2.2918$ suggests each copy of the minor allele at SNP rs4621895 more than doubles the relative risk of CP due to maternal alcohol exposure.

SNP	$e^{\hat{\beta}_{GE}}$	95% CI	$e^{\hat{\beta}_{GE}+\hat{\beta}_{GE\hat{X}_E}}$	95% CI	P-value
rs4621895	0.6732	(0.1845, 2.4563)	2.2918	(1.3771, 3.8143)	0.0043
rs4977433	0.8078	(0.2109, 3.0943)	2.3484	(1.4121, 3.9054)	0.0035
rs6475464	0.7507	(0.2228, 2.5296)	2.1621	(1.2792, 3.6545)	0.0128
rs668703	0.5243	(0.1532, 1.7943)	2.2913	(1.3786, 3.8082)	0.0029
rs623828	0.7475	(0.2285, 2.4449)	1.7385	(1.0194, 2.9649)	0.1105
rs2780841	0.5946	(0.1941, 1.8215)	1.6388	(0.9510, 2.8241)	0.1311

Table 2.4: Exponentiated $G \times E$ parameter estimates with corresponding exponentiated 95% Confidence Intervals with SRA adjustment at 6 the SNPs on *MLLT3* (Chr 9) showing significant interaction with Maternal Alcohol Consumption.

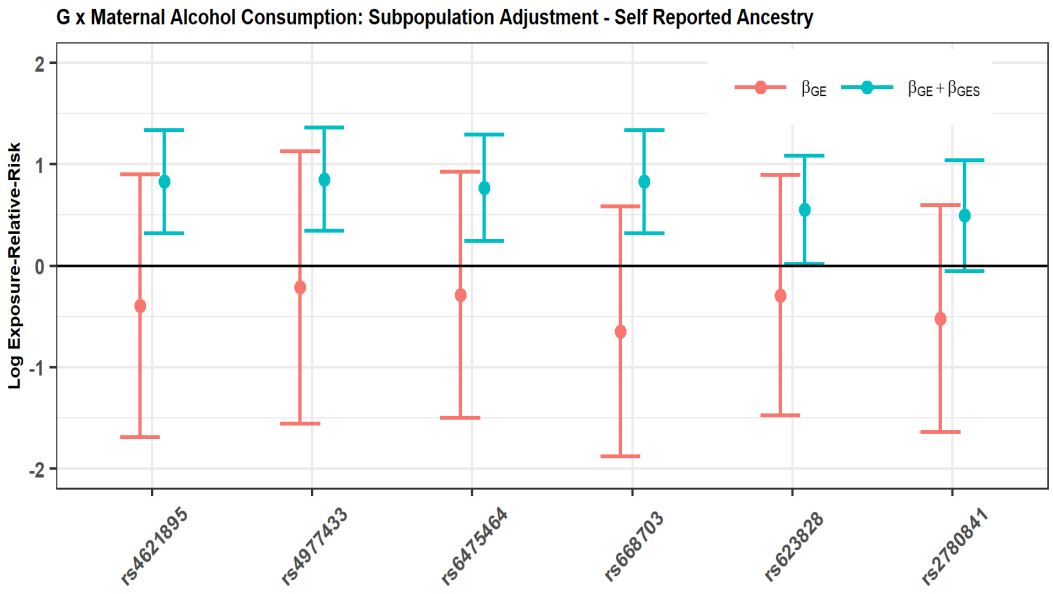

Figure 2.2: SNPs in *MLLT3* gene show significant interaction with Maternal Alcohol Consumption among European CP trios, but not among East Asian CP trios

Beaty *et al.* report that CP-affected European trios were gathered from multiple sites. Given that the $G \times E$ signal for SNPs in *MLLT3* appears to be driven by these trios of European ancestry only, we also investigated whether stratification within Europeans could explain the apparent $G \times E$ interaction. To do this, we computed principal components from genetic markers in all self-reported European trios and performed logistic regression of exposure on PCs computed from 6,231,573 SNPs. However, the PCs were not predictive of exposure, and so there was no evidence of ancestral sub-groups with different E distributions within self-reported Europeans.

We performed similar analyses to the other four genes implicated in the Beaty study with mixed success. In some instances, the SNPs and exposures were so rare in the self-reported East Asian trios the $\beta_{GE\hat{X}_E}$ coefficient in the model (2.2) could not be estimated. In such situations, we must assume a common $G \times E$ effect in both East Asians and Europeans. However, in this extended analysis of the other 4 genes we were still able to find some evidence of $G \times E$ interaction, though, as in Table 2.3, the results were attenuated compared to those when there is no adjustment for confounding. Tables of results are shown in Appendix A.2.

2.5 Discussion and Conclusions

Case-parent trio studies allow robust inference of genetic main effects, but inference of $G \times E$ interaction is susceptible to bias from a particular form of confounding known as exposure-related population structure. In the GENEVA Oral Cleft Study, such confounding would arise from heterogeneity of haplotype and exposure distributions across component sub-populations of the study. This is a concern because of the observed heterogeneity in exposure to maternal tobacco, alcohol and multivitamin use among European and East Asian populations. We emphasize that differences in exposure need not be genetic in nature; indeed, the differences in exposure distributions in the GENEVA Study are presumably cultural. However, if these non-genetic differences coincide with genetic differences in genes such as *MLLT3*, *SMC2*, *TBK1*, *ZNF236* and *BAALC*, they can result in spurious inference of $G \times E$ interaction. We applied the bias-reduced methods of Shin *et al.* to the GENEVA Oral Cleft Study and generally obtained the same conclusions regarding $G \times E$ interaction as an earlier analysis by Beaty *et al.*. However, we found the data only supported genetic modifiers of the exposure effects in self-reported Europeans. By contrast, all three exposures are too rare in self-reported East Asians to draw any conclusions about the presence of such modifiers in that population.

We further investigated whether confounding within the self-reported Europeans could explain these suggestions of $G \times E$ interaction. However, the genetic marker data from the genome-wide marker panel was not predictive of exposure within this population and so the adjustment of Shin *et al.* was not possible. Along the same lines, we investigated whether confounding in self-reported East Asians could explain the $G \times E$ findings of Wu *et al.*, who found evidence of genetic modifiers of the effect of environmental tobacco exposure on CP. The study sample was comprised of trios from Korea, China (Hubei, Sichaun and Shandong provinces), Taiwan and Singapore. Differences in exposure rates and haplotype distributions among these populations could lead to spurious evidence of $G \times E$ interaction. Again, we found genetic markers were not predictive of exposure, so the methods of Shin *et al.* could not be applied directly. Both of these scenarios suggest the need for alternative methods for adjusting for exposure-related stratification among sub-populations.

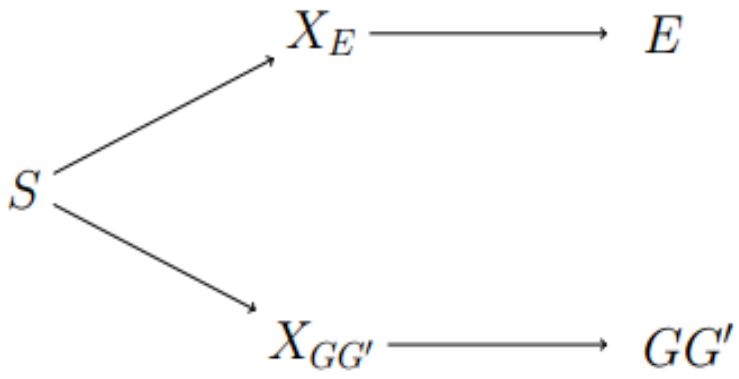

Figure 2.3: Dependence between E and GG' is through latent sub-population S. Latent factors X_E and $X_{GG'}$ indicate different distributions of E and GG', respectively. E and GG' are conditionally independent given any of the three variables on the path between them.

We conclude with a discussion of areas for future work. Exposure-related population structure is illustrated in Figure 2.3. The link between GG' haplotype frequencies and E distributions can be broken by adding any of S, X_E or $X'_{GG}$to the risk model. The approach of Shin et al. is to estimate a surrogate for X_E. As noted above, this approach may not be sensitive enough to adjust for subtle heterogeneity in genetic and exposure distributions. The diagram suggests alternatives, namely conditioning on S or on $X_{GG'}$. The advantage of conditioning on S, or on principal components that reflect S, is its simplicity and familiarity to researchers who study unrelated subjects. The potential advantage of an approach based on inferring $X_{GG'}$ is that it aims to characterize local structure in the genome of study subjects that could be responsible for apparent differences in risk at the test locus caused by differences in Linkage Disequilibrium (LD) with a nearby causal locus. Efforts to develop such methods are ongoing.

Another possible avenue for future research is to develop an approach that is a hybrid of case-parent-trio and case-only designs. The case-only design [16] shares key features with the case-parent trio design. Starting with a log-linear model of disease probabilities, and assuming G and E are independent, $G \times E$ is inferred from an association between G and E in cases. By conditioning on parental genotypes, inference from case-parent trio data is under the weaker assumption that G and E are independent within *families*, rather than in the general population, as in the case-only design. However, the number of cases available for analysis may be increased by dropping the requirement of parental genotype data. Thus, robust methods for inference of $G \times E$ that make use of cases with or without parental genotypes become of interest.

Chapter 3

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

3.1 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 [11]. $G \times E$ is also useful for identifying environmental exposures with greater disease-association in individuals who carry particular alleles at susceptibility loci [30]. 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 [7].

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 [38]. Exposure-related genetic structure occurs when the distribution of E also differs across the population strata [34]. Without some adjustment for the population strata, E will tag the stratum-specific differences in G' risk (Figure 3.1), suggesting that E modifies G' risk, even in the absence of $G \times E$ [24, 34]; we refer to this as spurious $G' \times E$.

Figure 3.1: 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 [33]. 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 [1] 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 [27]. 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 [24, 34]. 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 [24] 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 [34]. Accordingly, they propose a sibling-augmented case-only design and analysis. By contrast, [27] takes a *data*-based approach, replacing the sibship-averaged exposure of [24] 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 [3] 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 Asians [3, 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. [19] focused on the self-reported Europeans and East Asians in the GENEVA Oral Cleft Study data. Applying the approach of [27], they confirmed the gene-environment interaction found by [3], 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 [27]. In Section 3 we present simulations to evaluate the statistical properties of both approaches. In Section 4 we reanalyze the GENEVA data. Section 5 includes a discussion and areas for future work.

3.2 Models and Methods

3.2.1 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, $G \times E$ 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 $G \times E$, 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 $G' \times E$ in the absence of $G \times E$. 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 [27]. We review the rationale for the adjustment used by [27] in this context, and propose an alternative adjustment based on inferred population structure. In particular, we use the method of [8] 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 $G \times E$, but not so much that we compromise power in the presence of $G \times E$. The Models and Methods section concludes with a discussion of the simulation methods used to answer this question.

3.2.2 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 [27] can be obtained from a log-linear model of the GRRs

$$\log GRR_g(e) = \log \frac{P(D = 1|G = g, E = e)}{P(D = 1|G = g - 1, E = e)} = \beta_g + f_g(e) \quad \text{for } g = 1, 2, \tag{3.1}$$

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

$$P(D=1|G=g,E=e) = \eta(e)\prod_{i=0}^{g} GRR_i(e) \quad \text{for } g\text{=0, 1 or 2.} \tag{3.2}$$

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 Appendix B.1, 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.

3.2.3 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

$$P(D=1|G'=g',E=e) = \sum_{g=0}^{2} P(D=1|G=g,E=e)P(G=g|G'=g',E=e). \tag{3.3}$$

Equation (3.3) is a latent-class model [36] with the unobserved causal locus G as the latent class having probabilities $P(G=g|G'=g',E=e)$. Equations (3.3) and (3.2) enable the log-GRRs at G' to be written in terms of the latent-class probabilities and the GRRs at G as follows:

$$\begin{aligned}\log GRR_{g'}(e) &\equiv \log\frac{P(D=1|G'=g',E=e)}{P(D=1|G'=g'-1,E=e)} \\ &= \log\frac{\sum_{g=0}^{2}P(D=1|G=g,E=e)P(G=g|G'=g',E=e)}{\sum_{g=0}^{2}P(D=1|G=g,E=e)P(G=g|G'=g'-1,E=e)} \\ &= \log\frac{\sum_{g=0}^{2}(\prod_{i=0}^{g}GRR_i(e))P(G=g|G'=g',E=e)}{\sum_{g=0}^{2}(\prod_{i=0}^{g}GRR_i(e))P(G=g|G'=g'-1,E=e)}.\end{aligned} \tag{3.4}$$

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 Appendix B.2, 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, equation (3.4) gives insight into how exposure-related genetic structure creates spurious $G' \times E$.

3.2.4 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:

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

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 3.2.3 leads to the following X-adjusted log-GRRs at G':

$$\begin{aligned} \log GRR_{g'}(e,x) &\equiv \log \frac{P(D=1|G'=g',E=e,X=x)}{P(D=1|G'=g'-1,E=e,X=x)} \\ &= \log \frac{\sum_{g=0}^{2}(\prod_{i=0}^{g}\mathrm{GRR}_i(e,x))P(G=g|G'=g',E=e,X=x)}{\sum_{g=0}^{2}(\prod_{i=0}^{g}\mathrm{GRR}_i(e,x))P(G=g|G'=g'-1,E=e,X=x)}. \end{aligned} \tag{3.5}$$

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.

3.2.5 Removing dependence of the latent-class probabilities on E

The diagram in Figure 3.2 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 3.2 is said to be *blocked* by each of the variables X_E, S and $X_{GG'}$ [15, Definition 1]. Therefore, E and GG' are conditionally independent given any of the blocking variables X_E, S or $X_{GG'}$ [14]. As shown in Appendix B.2, a consequence is that conditioning on any of these variables removes the dependence of the

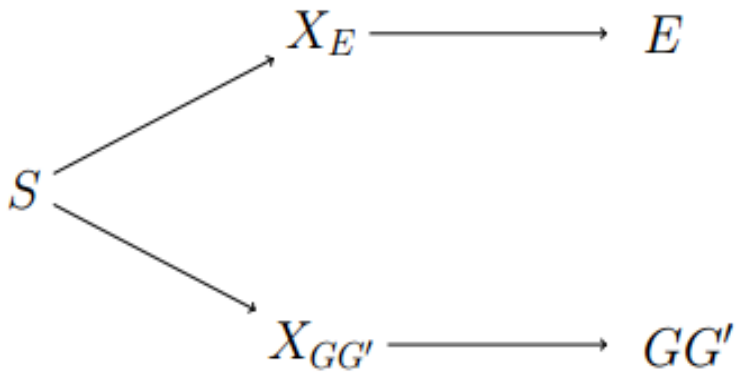

Figure 3.2: Diagram depicting exposure-related genetic structure. The latent population strata S induce dependence beetween E and GG'. Latent factors X_E and $X_{GG'}$ encode different distributions of E and GG', respectively. E and GG' are conditionally independent given any of the three variables that lie on the path between them.

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 equation (3.6),

$$\begin{aligned} \log GRR_{g'}(e,x) &\equiv \log \frac{P(D=1|G'=g',E=e,X=x)}{P(D=1|G'=g'-1,E=e,X=x)} \\ &= \log \frac{\sum_{g=0}^{2}(\prod_{i=0}^{g}\mathrm{GRR}_i(e,x))P(G=g|G'=g',X=x)}{\sum_{g=0}^{2}(\prod_{i=0}^{g}\mathrm{GRR}_i(e,x))P(G=g|G'=g'-1,X=x)}. \end{aligned} \quad (3.6)$$

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

3.2.6 Linear model for the log GRRs

From equation (3.6) 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 [26]. 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:

$$\log GRR_{g'}(e,x) = \beta_{g'} + \beta_{g'X}x + \beta_{g'E} \times e + \beta_{g'EX}x \times e; \quad g' = 1, 2. \quad (3.7)$$

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 [37].

3.2.7 Choice of X

Following [24], [27] 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, [27] 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 $\hat{X}_E = \widehat{E(E|\hat{S})}$. [27] showed that EEGM adjustment works well where there are two population strata, but our simulation results (Section 3.3) 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 $0, \ldots, K$, we let S denote a vector of K dummy variables that distinguish these strata such that that the kth element $S_k = 1$ for trios in stratum $k > 0$ and 0 otherwise, for $k = 1, \ldots, K$.

3.2.8 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 [31] 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 [39] 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 [8] 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 [13] 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.

3.2.9 Simulation methods

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 [27], we generated 5000 data sets of 3000 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 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 [28, 26]. Briefly, GG' haplotypes are first simulated on parents in a random-mating population according to the stratum-specific GG' haplotype distributions in Table 3.1. 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 (3.1). 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 3.1. The GG' distributions for strata $S = 0$ and $S = 1$ are as in [27]. 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 3.1, 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 $S = 3$.

Table 3.1: 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

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 [27].

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.

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 $r^2 \leq 0.1$ [9]. We simulated such panels of markers based on data from the 1000 genomes project [5] 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 [22], and use this fitted model to simulate individual multilocus genotypes using SNPknock [23]. 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 hours 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 5000 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 [8] 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 Sections 3.3.2 and 3.3.3 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.

3.3 Results

3.3.1 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 3000 trios sampled from four unequal-sized strata. For $K+1 = 4$ populations we require $K = 3$ PCs. In 5000 simulation replicates, the method of [8] always selected three, the method of [39] always selected one, and the method of [13] selected three PCs 4942 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 [8] to select PCs.

3.3.2 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 [27], and (iii) PC adjustment. Results for simulated datasets with equal or unequal stratum sizes are shown in Table 3.2. 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.

Table 3.2: 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)

3.3.3 Power

Table 3.3 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.

Table 3.3: 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)

3.4 The GENEVA Oral Cleft study

3.4.1 Data and objectives

The GENEVA Oral Cleft study [1] 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 3.4. From this table we see the ancestry of the sample is predominantly European (46%) and East Asian (51%).

Table 3.4: 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%

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 3.5. 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.

Table 3.5: Exposure rates for maternal alcohol consumption, maternal smoking and maternal vitamin supplementation by self-reported ancestry in affected trios.

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

3.4.2 GENEVA data analysis

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 [8] selects one PC. A plot of the projections of the data onto the first two PCs is shown in Figure 3.3, 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-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.

Inference of $G \times E$

The conditional-likelihood methods outlined in Appendix B.1 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 significant $G \times E$ at the 5% level in the analysis of [3]. Displays of the LD between these SNPs and others nearby [25] are shown in Figure B.1, Appendix B.2.1, for self-reported European subjects and self-reported East Asian subjects. Table 3.6 shows the results of fitting three different log-linear models of $G' \times E$. Following [3], 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 (3.1) were similar (results not shown). The first model, as in [3], 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

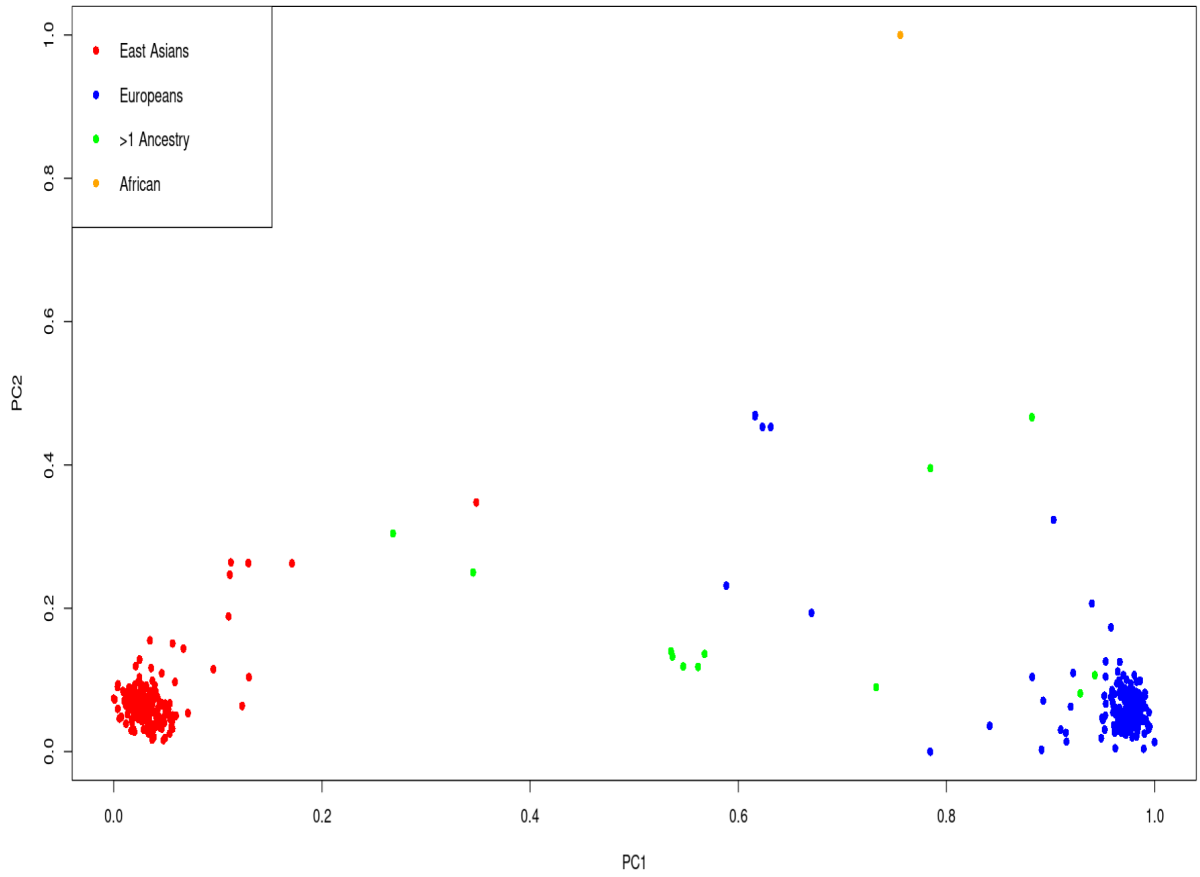

Figure 3.3: Projections of each affected child onto the first two PCs by self-reported ancestry: red=East Asian (234 trios), blue=European (214 trios), orange=African (one trio) and green=multiple ancestry/other (13 trios). Each PC has been shifted and scaled so that a PC1 value near zero corresponds to a hypothetical East Asian and a PC1 value near one corresponds to a hypothetical European.

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 3.6 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.

Table 3.6: 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 [3]. 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

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.

3.5 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 $G' \times E$ 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 [8]. 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 [27, 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 [3]. 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 3.6) and those of [19] 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 [24] and the sibling-augmented case-only approach of [34]. 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 [21]. For continuous exposures, such as those in the simulation study, the analog to the EEGM is a continuous-treatment propensity score [4]. 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 [10] or weights [20] 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.

Chapter 4

Adjustment for population stratification by local ancestry in gene-by-environment interaction studies of case-parent trios

4.1 Introduction

In a case-parent trio study we collect genotypes, G, on affected children and their parents. We may also collect environmental exposures or non-genetic attributes, E, on the children. $G \times E$ which can be interpreted as genes that modify the effect of exposure or exposures that modify the effect of genes.

Population stratification (PS) is a source of confounding in gene-by-environment interaction ($G \times E$) studies of case-parent trios when the study sample is pooled from distinct geographic locations. Specifically, when the test locus, G', is not causal, disease risk at G' can appear to be modified by E without $G \times E$ interaction when there is exposure-related population structure [24, 34]. Exposure-related population structure may be thought of as a form of confounding that occurs when both GG' haplotype frequencies and E distributions differ by ancestral group. Differences in GG' haplotypes can lead to differences in G' risk that may be tagged by E, suggesting $G' \times E$ even in the absence of true $G \times E$ interaction.

The diagram in Figure 4.1 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 variable that depends on S with different levels of X_E corresponding to different E distributions, and $X_{GG'}$ is a variable that depends on S with different levels of $X_{GG'}$ corresponding to different GG' haplotype distributions. In [18] both X_E and $X_{GG'}$ were described as "coarsenings" of S, but this need not be true in general. In this project, following [22], we take $X_{GG'}$ to be haplotype cluster membership. Informally, haplotype clusters are ancestral haplotypes that arose in the past with mutation superposed to allow variation in marker allele and GG'

haplotype frequencies within clusters. The idea is that the frequencies of these haplotype clusters vary by population strata, and hence the dependence of $X_{GG'}$ on S.

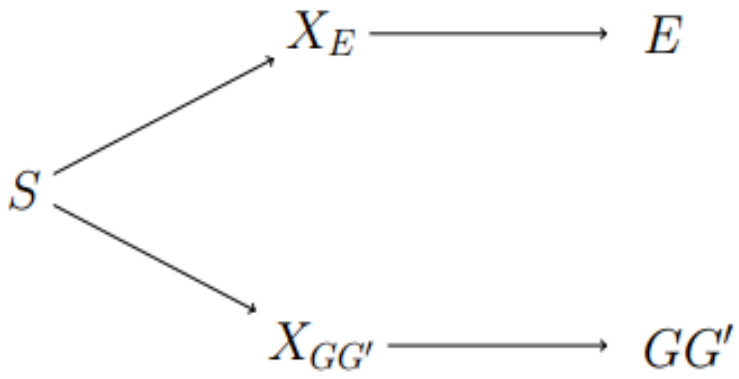

Figure 4.1: Diagram depicting exposure-related genetic structure. The latent population strata S induce dependence between E and GG'. Latent factors X_E and $X_{GG'}$ encode different distributions of E and GG', respectively. E and GG' are conditionally independent given any of the three variables that lie on the path between them.

Shin *et al.* [27] propose methods to reduce bias in $G \times E$ inference due to PS in case-parent trio data when individuals are sampled from two distinct populations. In their work, they make bias-reduced inference of $G \times E$ by conditioning on X_E. They estimate X_E by regressing E on important genetic principal components (PCs). However, their method is limited to two subpopulation strata.

Ratnasekera *et al.* [18] overcomes this limitation and proposes methods to reduce bias in $G \times E$ inference from case-parent trio data when individuals have been recruited from more than two population strata. In their work, they make bias-reduced inference of $G \times E$ by conditioning on S. They infer the global ancestry (S) of each affected individual *via* genetic principal components (PCs).

The use of global ancestry or self reported ancestry for the adjustment on $G \times E$ interaction of case-parent trios has it own limitations. Both global and self-reported ancestry may be less relevant than local ancestry when the individuals in the study sample are pooled from multiple different geographic locations due to possible admixture. In terms of the diagram in Figure 4.1, local ancestry is reflected in the variable $X_{GG'}$ that identifies different GG' haplotype distributions. In this project we make bias-reduced inference of $G \times E$ by conditioning on $X_{GG'}$. We infer local ancestry using a haplotype model proposed by [22]. We apply these ideas to data from the GENEVA Oral Cleft study and compare our results to those of [18].

4.2 GENEVA Oral Cleft Study

The GENEVA Oral Cleft study [1] 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, we studied the 462 trios that were investigated by [18]. Summaries of the trios by ancestry and gender of the affected child are provided in [18, Table 4]. From this table we can see the self-reported ancestry of the sample is predominantly European (46%) and East Asian (51%). In addition to that there are 14 (3%) affected children whose self-reported ancestry is Other or African.

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 [18, Table 5]. From this table we see that the three dichotomous exposures are all more common in Europeans.

Beaty *et al.* [3] found evidence for $G \times E$ interaction for maternal alcohol consumption and SNPs in the genes *MLLT3* and *SMC2*, for maternal smoking and SNPs in the genes *TBK1* and *ZNF236*, and for maternal multivitamin use and SNPs in the *BAALC* gene.

Figure 4.2 [18, Figure 3] presents a plot of the projections of the genome-wide genetic data of of 462 trios in GENEVA study onto the first two PCs, with points colored by self-reported ancestry. The genetic PCs haven been obtained after performing LD pruning of the genome-wide panel of SNPs at an r^2 threshold of 0.1, which yielded $63,694$ markers. Furthermore, 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-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.

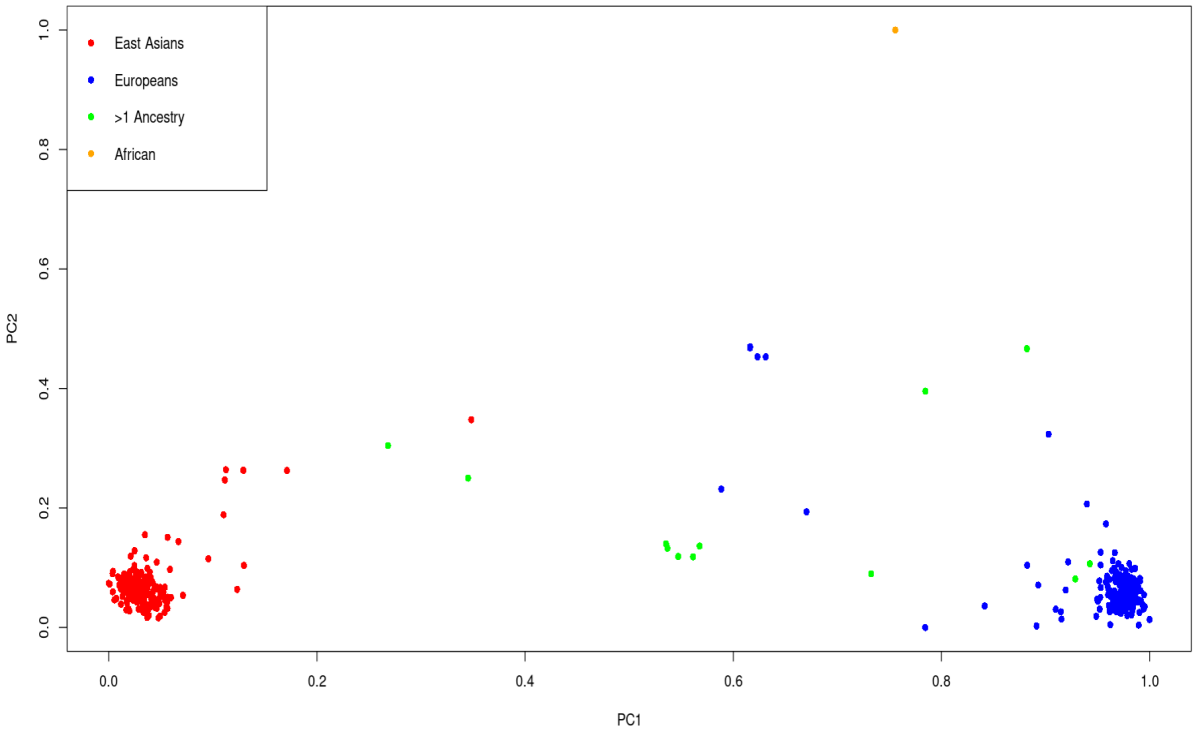

Figure 4.2: Projections of each affected child onto the first two PCs by self-reported ancestry: red=East Asian (234 trios), blue=European (214 trios), orange=African (one trio) and green=multiple ancestry/other (13 trios). Each PC has been shifted and scaled so that a PC1 value near zero corresponds to a hypothetical East Asian and a PC1 value near one corresponds to a hypothetical European.

Figure 4.2 suggests the presence of admixed individuals, whose points on the PC plot lie between the clusters identified as hypothetical East Asian (red) and European (blue). This includes self-reported East Asians, self-reported Europeans, and individuals that identify as being from more than one ancestral group (green). These observations suggest that it is important look at the local ancestry of individuals in addition to self-reported and/or global ancestry to better understand the true population structure of the study sample.

4.3 Methods to Infer Local Ancestry

Inference of local ancestry is challenging. We considered two approaches. The first, due to [6], is a Bayesian model-based clustering algorithm to infer local and global ancestry of individuals based on genome-wide data. Their method identifies the differences in the distribution of genetic variants amongst populations and place samples into groups whose members share similar patterns of variation. However, their methods can only be applied on unlinked or loosely linked panel of genome-wide markers.

The second method we considered is that of [22]. These authors propose a model in which haplotypes are drawn from one of K clusters and cluster membership is allowed to change along the genome. Their method can be applied to tightly-linked markers and is capable of capturing the complex patterns of correlation that exists among such markers. In this study we apply the methods proposed by [22] to infer the local ancestry of a panel of markers from 462 affected children in GENEVA Oral Cleft study.

4.3.1 Inferring local ancestry with fastPHASE

Scheet and Stephens [22] propose a Hidden Markov Model (HMM) to cluster haplotypes of diallelic SNPs. Their starting point is a global-ancestry model that assumes haplotypes are drawn from one of K clusters with cluster-specific allele distributions at each marker. They then incorporate local ancestry by allowing for cluster switching, according to a HMM that captures that idea that alleles at nearby markers are likely to arise from the same cluster [22, Model 3 and 4]. The authors use an expectation-maximization (EM) algorithm to estimate the cluster-specific allele frequencies and the parameters of the HMM, including the transition probabilities which determine whether the adjacent alleles are in the same cluster or in a different cluster.

In our study we are particularly interested in Model 4 of [22], which estimates the cluster memberships of each haplotype. In other words, Model 4 returns estimates of a variable $X_{GG'}$ (Figure 4.1) that tags different GG' haplotype distributions. In this study, our primary goal is to infer $X_{GG'}$ for a given haplotype and make inferences of $G \times E$ for each subpopulation by conditioning on $X_{GG'}$.

The methods described in [22] have been implemented in the software package fastPHASE. In this study we use fastPHASE to infer the $X_{GG'}$ haplotype clusters of 462 cleft palate affected children of GENEVA Oral Cleft study. The fastPHASE software was used with the recommended settings of 20 starts of EM algorithm, with up to 25 iterations per start. We obtained the parameter estimates of HMM based on the parameter values which maximize the likelihood. Furthermore, we fixed the number of clusters K to be 8 as proposed in Sheet *et al.*.

4.3.2 Disease risk model with $X_{GG'}$ adjustment

We apply the disease risk model proposed by Ratnasekera *et al.* [18, Model 7] with a modification in order to make inferences of $G \times E$ adjusted for $X_{GG'}$ haplotype clusters. In particular, we replace the latent variable S of model 7 of [18], estimated by important genetic principal components, with the latent cluster memberships, estimated by fastPHASE. Since the risk model at the test locus is for an individual, and individuals have two chromosomes, we include the combination of inferred clusters at the test locus for each child in the study. In what follows we suppose that there are K such combinations.

The modified log-linear genotype relative risk (GRR) model for G' is,

$$\log GRR_{g'}(e, x^*) = \beta_{g'} + \beta_{g'E} \times e + \sum_{k=1}^{K-1} \beta_{g'k} \times x_k^* + \sum_{k=1}^{K-1} \beta_{g'Ek} \times x_k^* \times e; \quad g' = 1, 2, \quad (4.1)$$

where x_k^*, is an indicator variable of membership in cluster combination k, for $k = 2, ..., K$ and x^* is the vector of $K-1$ indicator variables. The GRR model for G' for a trio in cluster combination $k = 1$ is then

$$\log GRR_{g'}(e, x^*) = \beta_{g'} + \beta_{g'E} \times e$$

and for cluster $k > 1$ is

$$\log GRR_{g'}(e, x^*) = \beta_{g'} + \beta_{g'k} + (\beta_{g'E} + \beta_{g'Ek}) \times e$$

This model may be interpreted as follow. For trios with $G' = g'$ in cluster combination $k = 1$, the log-GRR curve is linear with intercept $\beta_{g'}$ and slope $\beta_{g'E}$. For trios with $G' = g'$ in cluster combination k, $k = 2, \ldots, K$, the log-GRR curve is linear with intercept $\beta_{g'} + \beta_{g'k}$ and slope $\beta_{g'E} + \beta_{g'Ek}$.

To test for $G' \times E$ based on the model (4.1) we use a likelihood-ratio test of the null hypothesis that all curves have zero slope,

$$H_0 : \beta_{g'E} = \beta_{g'E2} = \cdots = \beta_{g'EK} = 0; \quad g' = 1, 2, \quad (4.2)$$

versus the alternative hypothesis that at least one of the slope parameters is non-zero.

4.4 Data Analysis

An analysis was performed with data on 462 children affected with cleft palate (CP) in two phases. Phase one was inference of the local ancestry of affected individuals at given loci with the use of fastPHASE. In phase two, we use the inferred local ancestry to make inference of $G \times E$. Here, we focused on the *MLLT3* gene and inference of $G \times E$ between maternal alcohol consumption and the six SNPs in the *MLLT3* gene that had significant $G \times E$ at the 5% level in the analysis of [3].

4.4.1 Local Ancestry of CP-Affected Children in *MLLT3*

We inferred the local ancestry of tightly linked loci in *MLLT3* with fastPHASE. First we filtered SNPs to those with MAF ≥ 0.05 and with less than 5% missing data. The resulting 364 SNPs were then analysed with fastPHASE using the recommended settings mentioned in section 4.3.1. Figure 4.3 depicts cluster memberships of haplotypes from a selection of nine CP-affected children (18 haplotypes): three of self-reported European ancestry (denoted

EU1 to EU3), one of self-reported African ancestry (denoted AF1), one of self-reported Other ancestry (OT1) and four of self-reported East Asian ancestry (EA1 to EA4). Figure 4.3 is focussed on the 95 markers (out of 364) that span the six SNPs of interest in *MLLT3*. Figure 4.4 shows the nine CP-affected children of Figure 4.3 on the PC plot obtained from a genome-wide panel of markers. On the PC plot, AF1, EA1 and OT1 stand apart from the hypothetical EA and EU individuals based on genome-wide data, represented by {EA2,EA3,EA4} and {EU1,EU2,EU3}, respectively. However, we see some similarities between AF1, EA1, OT1 and the others based on inferred haplotype clusters from *MLLT3*. For example, the inferred clusters of the first (top-most) haplotype from AF1 are like the two haplotypes of EU2, while the inferred cluster of the second (bottom-most) haplotype of AF1 is like haplotypes seen in EU3, EA1, EA3 and EA4. As another example, we see that the inferred clusters of EA1 and OT1 are fairly similar to those of EA3, even though EA1 and OT1 stand apart from {EA2,EA3,EA4} on the PC plot. To summarize, for these data on *MLLT3*, inferred local ancestry is different than global anestry, and we argue that local ancestry is more relevant for inference of $G \times E$.

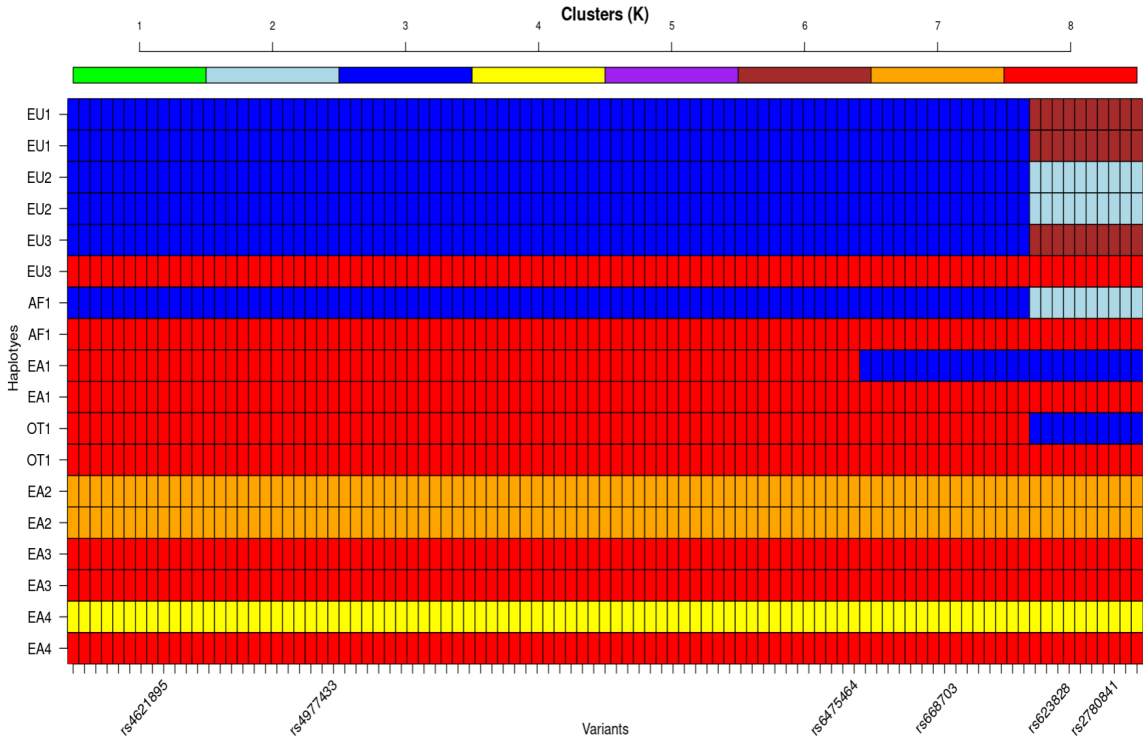

Figure 4.3: Graphical display of cluster memberships of haplotypes from a selection of nine CP-affected children (18 haplotypes). Rows of the plot represents haplotypes and columns represents SNPs. The eight different colors represent estimated cluster membership, which changes as one moves along each haplotype. The locations of the six SNPs of interest in *MLLT3* are shown on the X-axis.

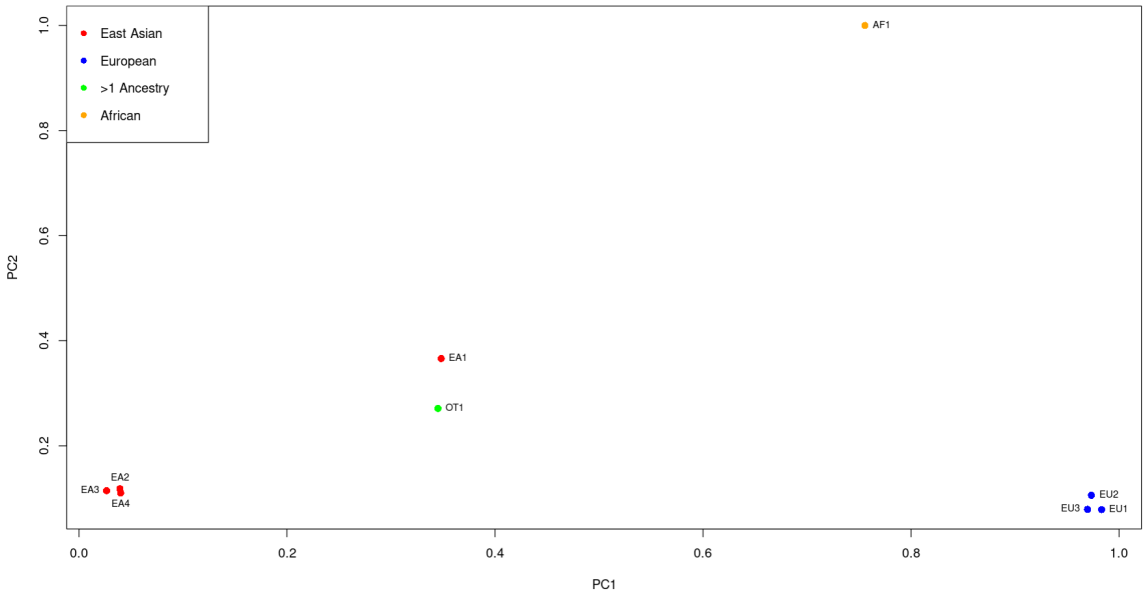

Figure 4.4: Projections of 9 affected children of figure 4.3 onto the first two PCs by self-reported ancestry: red=East Asian (4 trios), blue=European (3 trios), orange=African (one trio) and green=multiple ancestry/other (one trio). Each PC has been shifted and scaled so that a PC1 value near zero corresponds to a hypothetical East Asian and a PC1 value near one corresponds to a hypothetical European.

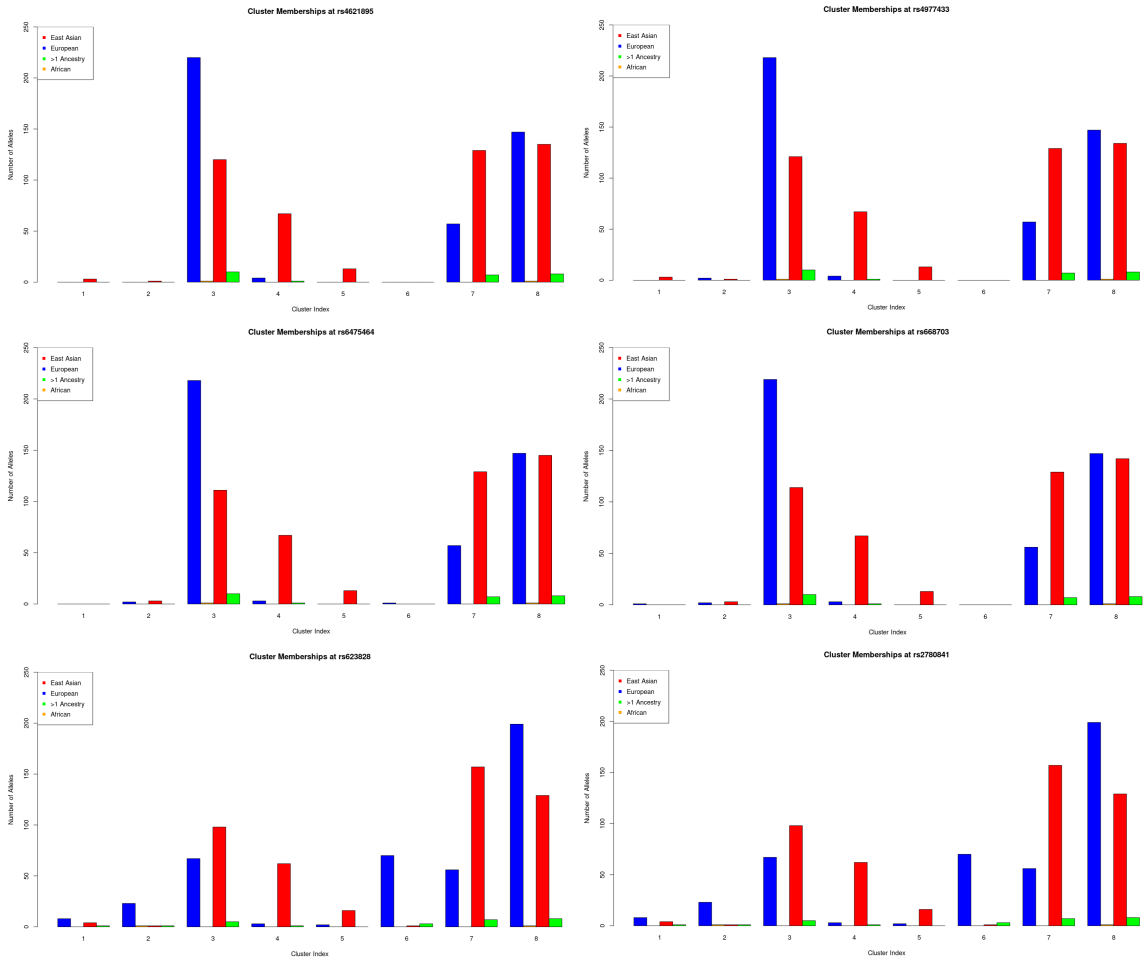

Figure 4.5: The distribution of cluster memberships ($K = 8$) at each of the six SNPs of interest on *MLLT3* gene. The distribution shows the cluster membership of each of the 924 haplotypes at each SNP, with colors indicating self-reported ancestry.

Another way to summarize inferred local ancestry is to look at the distributions of cluster memberships at specific SNPs by self-reported ancestry. Figure 4.5 shows the distributions of cluster membership at each of the six SNPs which were found to significantly modify exposure to maternal alcohol consumption. These distributions of are similar for SNPs $rs4621895$, $rs4977433$, $rs6475464$ and $rs668703$, with a majority of alleles assigned to clusters 3, 4, 7 or 8. Clusters 3 and 8 have more self-reported Europeans, while clusters 4 and 7 have more self-reported East Asians. By contrast, at SNPs $rs623828$ and $rs2780841$ most self-reported Europeans have been assigned clusters 6 and 8 and most self-reported East Asians to clusters 3, 4, 7 and 8. These observations reinforce the idea that local ancestry changes as we move along a particular gene.

4.4.2 $G \times E$ results for CP-affected children at six SNPs from *MLLT3*

We have argued that adjusting for local ancestry is important, but we now show that, in its current form, such adjustment is not useful for inference of $G' \times E$. Table 4.1 shows the p-values from likelihood-ratio test of fitting three different log-linear models of $G' \times E$. Following [3], each is based on an additive genetic model that specifies equal log-GRRs for genotypes $g' = 1$ or 2. The first model, as in [3], makes no adjustment for exposure-related genetic structure in the population, the second uses S adjustment and the third uses $X_{GG'}$ 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. In comparison, the p-values from S adjustments are slightly higher and only four of the six SNPs retain significance at the 5% level. The most striking feature of Table 4.1 is the p-values from $X_{GG'}$ adjustment, which are all approximately one, suggesting that G' does not modify the association between exposure to maternal alcohol consumption. These large p-values are a symptom of failed model fitting, with too little data given the large number of cluster combinations. In particular, the model of equation (4.1) specifies separate log-GRR curves in the exposure E for each pair of haplotype clusters, and we observe pairs of haplotype clusters with few or no exposed individuals. This lack of data leads to a failure of the iterative model-fitting procedure to converge. The lack of convergence is also evident in the estimates and confidence intervals for the interaction terms in model 4.1, shown in Tables C.1 through C.6 in Appendix C.

SNP	No adj.	S adj.	$X_{GG'}$ adj.
rs4621895	0.0006	0.0037	≈ 1
rs4977433	0.0003	0.0028	≈ 1
rs6475464	0.0104	0.0139	≈ 1
rs668703	0.0008	0.0025	≈ 1
rs623828	0.0481	0.1384	≈ 1
rs2780841	0.0417	0.1471	≈ 1

Table 4.1: The P-values of likelihood ratio test from the model with no adjustment, with S adjustment and $X_{GG'}$ adjustment

4.5 Discussion

This study attempts to address the problem of biased inference of $G \times E$ in studies of case-parent trios when trios are recruited from multiple different geographic locations. Exposure-related genetic structure in the study sample can lead to spurious $G' \times E$ at a non-causal test locus G' in linkage-disequilibrium with casual locus G. We obtain inference of $G' \times E$

by introducing two different adjustments to the $G' \times E$ interaction effect along with the no-adjustment model used by Beaty et al. Out of the two adjustments, the PC adjustment works well for multiple ancestral groups as long as there is no admixture in the study sample. Here, we introduce the second adjustment by $X_{GG'}$, which captures local ancestry and should be useful when the study sample includes admixed individuals.

We use fastPHASE to estimate the local ancestry of individuals. FastPHASE infers local ancestry based on a HMM to cluster haplotypes and can be used to capture patterns of variation in a panel of tightly linked markers. We applied fastPHASE to genetic data on 462 Cleft-Palate-affected children from the GENEVA Oral Cleft study. The study sample was collected from multiple different location around the world and the data includes self-reported ancestry in the categories Europeans, East Asians, Africans or Other/more than one ancestry. We applied fastPHASE to 364 diallelic tightly linked markers spanning the *MLLT3* gene and obtained cluster memberships under recommended settings. These cluster memberships were used as the estimates of $X_{GG'}$. We then tried to use estimates of $X_{GG'}$ to fit the model of equation 4.1, but model fitting failed. The model fits separate $G' \times E$ curves for different pairs of haplotype clusters. By assuming eight latent haplotype clusters, as recommended, there are potentially $8 \times (8+1)/2 = 36$ pairs of clusters. Of the 36 we observed 14 pairs at $rs4621895$, 15 at $rs4977433$, 16 at $rs6475464$, 17 at $rs668703$, and 28 at $rs623828$ and $rs2780841$. In addition, the model requires two log-GRR curves, one for $g' = 1$ and one for $g' = 2$, for each pair of haplotype clusters. There are simply too many log-GRR curves to fit with the 462 trios in the GENEVA dataset.

Similar to how the PC-adjustment method required a parsimonious set of PCs, local ancestry adjustment requires a parsimonious set of haplotype clusters. We considered using two or four, rather than eight haplotype clusters, but these more parsimonious latent haplotype cluster models yielded lower likelihoods and are therefore not recommended by [22]. In addition, the classification of self-reported ancestry looks much better when using eight rather than two or four haplotype clusters. A potential problem with fastPHASE (Lloyd Elliot, personal communication, August 16, 2023) is that inference of clusters may not be reversible; that is, inferred clusters could be different depending on whether the algorithm is run forwards or backwards along the chromosome. One area of future work is to re-run fastPHASE backwards on the GENEVA data to see if it changes the cluster memberships. An alternative to fastPHASE for local ancestry inference is RFmix [12], which is used by the Tractor method for GWAS with admixed samples [2]. One direction for future research is to apply RFmix to the GENEVA data to see if it gives a more parsimonious characterization of local ancestry. Another area of future work is to investigate the use of penalized likelihood methods to stabilize parameter estimation.

Another direction for future work is to use inferred pairs of haplotype clusters as multi-allele genotypes. We observed very little variation in SNP genotypes, G', at the test loci once haplotype clusters are considered (results not shown), suggesting that we use haplotype

cluster pairs instead of G' genotypes in the risk model. This requires that we re-work the risk model in terms of a multi-allelic model and then derive the likelihood for trio data, in which the genotype (cluster pair) is the response variable. We could further reduce the number of parameters by specifying an "additive" model in the haplotype clusters, rather than a "co-dominant" model in terms of all haplotype pairs.

Chapter 5

Conclusion

The three projects in this thesis discuss statistical methods to reduce bias in $G \times E$ inference from case-parent trios due to exposure-related genetic population structure. In chapter 2, we re-analysed the $G \times E$ results of case-parent trios in GENEVA Oral Cleft study reported by [3]. Beaty *et al.* reports statistically significant $G \times E$ interaction effects on the risk of cleft palate between exposure to maternal alcohol consumption and SNPs in the *MLLT3* and *SMC2* genes, exposure to maternal smoking and SNPs in the *TBK1* and *ZNF236* genes, and exposure to multivitamin supplementation and SNPs in the *BAALC* gene. The GENEVA study consists primarily of case-parent trios categorised as either self-reported European or self-reported East Asian and the exposure distributions all differ by self-reported ancestry group. If, in addition, haplotype distributions vary with self-reported ancestry we would have exposure-related population structure that could result in biased $G \times E$ inference. We applied the biased-reduced methods of Shin *et al.* to the GENEVA data to investigate potential bias in $G \times E$ results reported by Beaty *et at.* due to exposure-related population structure. Our results generally confirm the results reported in Beaty *et al.* However, the evidence of $G \times E$ was primarily from self-reported European trios, with the exposures being too rare among self-reported East Asians to make any conclusions. A secondary analysis was also performed to investigate whether the evidence for $G \times E$ could be due to population structure within the self-reported Europeans. However, the genetic marker data from the genome-wide marker panel was not predictive of exposure within the European population and so the adjustment of Shin *et al.* was not possible. Furthermore, we also conducted a secondary analysis of the East Asian trios to investigate the potential bias in the $G \times E$ results reported in Wu *et al.* [35]. Wu *et. al.* found evidence of genetic modifiers of the effect of environmental tobacco exposure on CP among East Asian trios. Again, we found genetic markers were not predictive of exposure, so the methods of Shin *et al.* could not be applied.

Exposure-related population structure is illustrated in Figure 5.1. The link between GG' haplotype frequencies and E distributions can be broken by conditioning on any of S, X_E or $X_{GG'}$ to the risk model. The approach of Shin *et al.* is to estimate a surrogate for X_E. Simulation results in chapter 3 show that this method works well when the underlying

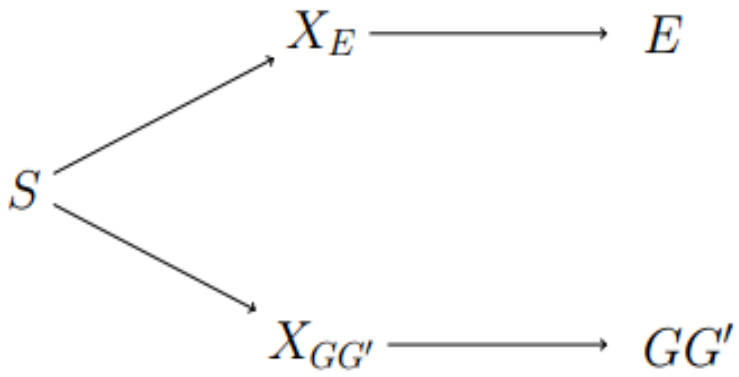

Figure 5.1: Diagram depicting exposure-related genetic structure. The latent population strata S induce dependence between E and GG'. Latent factors X_E and $X_{GG'}$ encode different distributions of E and GG', respectively. E and GG' are conditionally independent given any of the three variables that lie on the path between them.

population structure consists of two strata but not for more than two. In chapter 3, we replace X_E adjustment with S adjustment. The advantage of conditioning on S, or on principal components that reflect S, is its simplicity and familiarity to researchers who study unrelated subjects. We obtained PCs from a genome-wide panel of unlinked markers and select a parsimonious subset of PCs based on the methods proposed by [8]. The subset of PCs were used as the surrogate for S. Our simulations suggest that PC adjustment maintains the nominal type-1 error rate and has nearly identical power to detect $G \times E$ as adjustment with true stratum membership. An analysis of the GENEVA data with S adjustment gave similar results to those with X_E adjustment.

The surrogates for X_E adjustment and S adjustment *via* genetic PCs rely on the global ancestry of individuals. In chapter 4, we introduce the third adjustment using local ancestry in the form of inferred haplotype clusters, $X_{GG'}$. The potential advantage of an approach based on inferring $X_{GG'}$ is that it aims to characterize local structure in the genome of study subjects that could be responsible for apparent differences in risk at the test locus caused by differences in linkage disequilibrium with a nearby causal locus. We used fastPHASE [22] to infer the local ancestry of individuals based on a tightly linked set of 364 genetic markers in the *MLLT3* gene. The inferred local ancestry was then used as the adjustment for $G \times E$ interaction between six SNPs in *MLLT3* and exposure to maternal alcohol consumption. The $X_{GG'}$ adjustment is to fit separate log-GRR curves in the exposure for each observed pair of haplotype clusters. This results in trying to fit a large number of curves with too little data and model fitting failed. Directions for future work are to reduce the number of haplotype clusters used in the adjustment, and re-formulating the risk model in terms of pairs of haplotype clusters instead of genotypes at test loci.

Bibliography

[1] GENEVA Oral Clefts Project Imputation Report - HapMap III reference panel [pdf file], 2010. `https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000094.v1.p1`.

[2] E. G. Atkinson, A. X. Maihofer, M. Kanai, A. R. Martin, K. J. Karczewski, M. L. Santoro, J. C. Ulirsch, Y. Kamatani, Y. Okada, H. K. Finucane, K. C. Koenen, C. M. Nievergelt, M. J. Daly, and B. M. Neale. Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power. *Nat Genet*, 53(2):195–204, Feb 2021.

[3] Terri H. Beaty, Ingo Ruczinski, Jeffrey C. Murray, Mary L. Marazita, Ronald G. Munger, Jacqueline B. Hetmanski, Tanda Murray, Richard J. Redett, M. Daniele Fallin, Kung Yee Liang, Tao Wu, Poorav J. Patel, Sheng-Chih Jin, Tian Xiao Zhang, Holger Schwender, Yah Huei Wu-Chou, Philip K. Chen, Samuel S. Chong, Felicia Cheah, Vincent Yeow, Xiaoqian Ye, Hong Wang, Shangzhi Huang, Ethylin W. Jabs, Bing Shi, Allen J. Wilcox, Rolv T. Lie, Sun Ha Jee, Kaare Christensen, Kimberley F. Doheny, Elizabeth W. Pugh, Hua Ling, and Alan F. Scott. Evidence for gene-environment interaction in a genome wide study of nonsyndromic cleft palate. *Genetic Epidemiology*, 35:469–478, 2011.

[4] D. W. Brown, T. J. Greene, M. D. Swartz, A. V. Wilkinson, and S. M. DeSantis. Propensity score stratification methods for continuous treatments. *Stat Med*, 40(5):1189–1203, Feb 2021.

[5] Laura Clarke, Susan Fairley, Xiangqun Zheng-Bradley, Ian Streeter, Emily Perry, Ernesto Lowy, Anne-Marie Tassé, and Paul Flicek. The international Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data. *Nucleic Acids Research*, 45(D1):D854–D859, Sep 2016.

[6] Daniel Falush, Matthew Stephens, and Jonathan K Pritchard. Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular Ecology Notes*, 7(4):574–578, 2007.

[7] M. Garaulet, C. E. Smith, T. Hernández-González, Y. C. Lee, and J. M. Ordovás. 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(12):1771–1779, Dec 2011.

[8] M. Gavish and D. L. Donoho. The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Transactions on Information Theory*, 60(8):5040–5053, 2014.

[9] Kelsey Grinde. *Statistical Inference in Admixed Populations*. PhD thesis, University of Washington, 2019.

[10] Keisuke Hirano and Guido W. Imbens. *The Propensity Score with Continuous Treatments*, chapter 7, pages 73–84. John Wiley Sons, Ltd, 2004.

[11] D. J. Hunter. Gene-environment interactions in human diseases. *Nat Rev Genet*, 6(4):287–298, Apr 2005.

[12] B. K. Maples, S. Gravel, E. E. Kenny, and C. D. Bustamante. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *The American Journal of Human Genetics*, 93(2):278–288, Aug 2013.

[13] N. Patterson, A. L. Prince, and D. Reich. Population structure and eigenanalysis. *PLoS Genetics*, 12(2):2074–2093, 2006.

[14] Judea Pearl. Graphs, causality, and structural equation models. *Sociological Methods & Research*, 27(2):226–284, 1998.

[15] Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96 – 146, 2009.

[16] Walter W Piegorsch, Clarice R Weinberg, and Jack A Taylor. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Statistics in medicine*, 13(2):153–162, 1994.

[17] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.

[18] Pulindu Ratnasekera, Jinko Graham, and Brad McNeney. Inference of gene-environment interaction from heterogeneous case-parent trios. *Frontiers in Genetics*, 2023.

[19] Pulindu Ratnasekera and Brad McNeney. Re-analysis of a Genome-Wide Gene-By-Environment Interaction Study of Case Parent Trios, Adjusted for Population Stratification. *Frontiers in Genetics*, Jan 2021.

[20] J. M. Robins, M. A. Hernán, and B. Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, Sep 2000.

[21] Paul R. Rosenbaum and Donald B. Rubin. The Central Role of the Propensity Score in Observational Studies for Causal Effects. In *Matched Sampling for Causal Effects*, pages 170–184. Cambridge University Press, 1983.

[22] Paul Scheet and Matthew Stephens. A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *The American Journal of Human Genetics*, 78(4):629–644, 2006.

[23] Matteo Sesia, Chiara Sabatti, and Emmanuel Candès. Gene hunting with hidden markov model knockoffs. *Biometrika*, 106(1):1–18, 2019.

[24] Min Shi, David M. Umbach, and Clarice R. Weinberg. Family-based Gene-by-environment Interaction Studies. *Epidemiology*, 22(3):400–407, May 2011.

[25] J.-H. Shin, S. Blay, B. McNeney, and J. Graham. Ldheatmap: An r function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. *J Stat Soft*, 16:Code Snippet 3, 2006.

[26] J. H. Shin, C. Infante-Rivard, B. McNeney, and J. Graham. A data-smoothing approach to explore and test gene-environment interaction in case-parent trios. *Statistical Applications in Genetics and Molecular Biology*, 13(2):159–171, Apr 2014.

[27] Ji-Hyung Shin, Claire Infante-Rivard, Jinko Graham, and Brad McNeney. Adjusting for spurious gene-by-environment interaction using case-parent triads. *Statistical Applications in Genetics and Molecular Biology*, 11(2), 2012.

[28] Ji-Hyung Shin, Brad McNeney, and Jinko Graham. *trioGxE: A data smoothing approach to explore and test gene-environment interaction in case-parent trio data*, 2013. R package version 0.1-1.

[29] Terry M Therneau. *A Package for Survival Analysis in R*, 2021. R package version 3.2-13.

[30] D. Thomas. Gene–environment-wide association studies: emerging approaches. *Nat Rev Genet*, 11(4):259–272, Apr 2010.

[31] H. Wang, C. A. Haiman, L. N. Kolonel, B. E. Henderson, L. R. Wilkens, L. Le Marchand, and D. O. Stram. Self-reported ethnicity, genetic structure and the impact of population stratification in a multiethnic study. *Hum Genet*, 128(2):165–177, Aug 2010.

[32] Hansong Wang, Christopher A Haiman, Laurence N Kolonel, Brian E Henderson, Lynne R Wilkens, Loïc Le Marchand, and Daniel O Stram. Self-reported ethnicity, genetic structure and the impact of population stratification in a multiethnic study. *Human genetics*, 128:165–177, 2010.

[33] Clarice R. Weinberg. Methods for Detection of Parent-of-Origin Effects in Genetic Studies of Case-Parents Triads. *The American Journal of Human Genetics*, 65(1):229–235, Jul 1999.

[34] Clarice R. Weinberg, Min Shi, and David M. Umbach. A Sibling-augmented case-only Approach for Assessing Multiplicative Gene-Environment Interactions. *American Journal of Epidemiology*, 174(10):1183–1189, Oct 2011.

[35] Tao Wu, Holger Schwender, Ingo Ruczinski, Jeffrey C. Murray, Mary L. Marazita, Ronald G. Munger, Jacqueline B. Hetmanski, Margaret M. Parker, Ping Wang, Tanda Murray, Margaret Taub, Shuai Li, Richard J. Redett, M. Daniele Fallin, Kung Yee Liang, Yah Huei Wu-Chou, Samuel S. Chong, Vincent Yeow, Xiaoqian Ye, Shangzhi Huang, Ethylin W. Jabs, Bing Shi, Allen J. Wilcox, Sun Ha Jee, Alan F. Scott, , and Terri H. Beaty. Evidence of Gene2Environment Interaction for Two Genes on Chromosome 4 and Environmental Tobacco Smoke in Controlling the Risk of Nonsyndromic Cleft Palate. *PLOS ONE*, 9(2), feb 2014.

[36] Gongjun Xu. Identifiability of restricted latent class models with binary responses. *The Annals of Statistics*, 45(2):675–707, Apr 2017.

[37] Z. Yu, M. Demetriou, and D. L. Gillen. Genome-Wide Analysis of Gene-Gene and Gene-Environment Interactions Using Closed-Form Wald Tests. *Genet Epidemiol*, 39(6):446–455, Sep 2015.

[38] D. V. Zaykin and K. Shibata. Genetic flip-flop without an accompanying change in linkage disequilibrium. *Am J Hum Genet*, 82(3):794–796, Mar 2008.

[39] Mu Zhu and Ali Ghodsi. Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics & Data Analysis*, 51(2):918–930, Nov 2006.

Appendix A

Supplementary Material for Chapter 2: *Re-analysis of a genome-wide gene-by-environment interaction study of case parent trios, adjusted for population stratification*

This appendix includes the published supplementary material for the manuscript entitled Re-analysis of a genome-wide gene-by-environment interaction study of case parent trios, adjusted for population stratification.

A.1 Genotypic Odds

Assuming G and E are conditionally independent given parental genotypes G_p, and conditioning on G_p and E we can obtain a likelihood ([27]),

$$P(G = g \mid D = 1, E = e, G_p = g_p) = \frac{P(D = 1 \mid G = g, E = e)P(G = g \mid G_p = g_p)}{\sum_{g^*} P(D = 1 \mid G = g^*, E = e)P(G = g^* \mid G_p = g_p)}$$

We could use this likelihood to define Genotypic Odds,

$$GO_g(e) = \frac{P(G = g \mid D = 1, E = e, G_p = g_p)}{P(G = g - 1 \mid D = 1, E = e, G_p = g_p)}$$

$$GO_g(e) = \frac{P(D = 1 \mid G = g, E = e)}{P(D = 1 \mid G = g - 1, E = e)} \times \frac{P(G = g \mid G_p = g_p)}{P(G = g - 1 \mid G_p = g_p)}$$

$$GO_g(e) = exp\{\beta_G + e\beta_{GE}\} \times exp\{\kappa_p\}$$

where $\kappa_p = ln(2)$ if $g = 1$ and $\kappa_p = -ln(2)$ if $g = 2$ given that parental genotypes, G_p, are both heterozygous.

A.1.1 $G \times E$ parameter estimates and its 95% confidence intervals

We have listed the sub-population specific $G \times E$ interaction parameter estimates and their corresponding 95% confidence intervals with the EEGM adjustment in Tables A.1 to A.5. These five tables correspond to the five genes which Beaty *et al.* found evidence for $G \times E$ for maternal alcohol consumption and SNPs in the genes *MLLT3* and *SMC2*, for maternal smoking and SNPs in the genes *TBK1* and *ZNF236*, and for maternal multivitamin use and SNPs in the *BAALC* gene respectively. At four SNPs shown in Table A.2 and at a one SNP of Table A.3, the exposures were so rare in the self-reported East Asian trios the $\beta_{GE\hat{X}_E}$ coefficient in the model (2.2) could not be estimated. In such situations, we assumed a common $G \times E$ effect in both East Asians and Europeans.

SNP	$e^{\hat{\beta}_{GE}}$	95% CI	$e^{\hat{\beta}_{GE}+\hat{\beta}_{GE\hat{X}_E}}$	95% CI	P-value
rs4621895	0.7430	(0.2038, 2.7084)	2.4089	(1.3935, 4.1640)	0.0060
rs4977433	0.9004	(0.2339, 3.4659)	2.4451	(1.4156, 4.2234)	0.0046
rs6475464	0.8351	(0.2468, 2.8256)	2.2750	(1.2893, 4.0141)	0.0161
rs668703	0.5679	(0.1674, 1.9268)	2.4667	(1.4288, 4.2587)	0.0039
rs623828	0.7785	(0.2367, 2.5608)	1.7828	(1.0062, 3.1589)	0.1358
rs2780841	0.6244	(0.2044, 1.9072)	1.6936	(0.9451, 3.0348)	0.1683

Table A.1: $G \times E$ parameter estimates and its 95% Confidence Intervals obtained with EEGM adjustment of CP children at six SNPs on *MLLT3* gene (Chr 9) which showed evidence significant interaction with Maternal Alcohol Consumption.

SNP	$e^{\hat{\beta}_{GE}}$	95% CI	$e^{\hat{\beta}_{GE}+\hat{\beta}_{GE\hat{X}_E}}$	95% CI	P-value
rs10125685	4.2540	(1.4359, 12.6025)	-	-	0.0056
rs628345	2.1111	(0.9776, 4.5590)	-	-	0.0533
rs630103	1.6991	(0.4002, 7.2138)	0.9415	(0.5505, 1.6100)	0.7645
rs868619	6.0936	(0.2122, 174.9883)	1.1553	(0.6562, 2.0342)	0.3139
rs1536895	5.0690	(1.6175, 15.8859)	-	-	0.0026
rs10217601	2.1076	(0.9758, 4.5523)	-	-	0.0538

Table A.2: $G \times E$ parameter estimates and its 95% Confidence Intervals obtained with EEGM adjustment of CP children at six SNPs on *SMC2* gene (Chr 9) which showed evidence of significant interaction with Maternal Alcohol Consumption.

SNP	$e^{\hat{\beta}_{GE}}$	95% CI	$e^{\hat{\beta}_{GE}+\hat{\beta}_{GE\hat{X}_E}}$	95% CI	P-value
rs1317532	1.3652	(0.2971, 6.2738)	1.5768	(0.8682, 2.8636)	0.2522
rs1317535	1.9270	(0.3337, 11.1292)	1.4510	(0.7517, 2.8011)	0.3361
rs2141765	1.7960	(0.3326, 9.6983)	1.7137	(0.9524, 3.0836)	0.1071
rs7969932	2.1652	(0.2249, 20.8446)	2.0355	(1.0992, 3.7693)	0.0294
rs6581575	2.2341	(0.2336, 21.3651)	2.0304	(1.0964, 3.7598)	0.0289
rs4964110	1.7773	(0.1444, 21.8819)	1.3237	(0.4052, 4.3244)	0.7707
rs10506538	1.8945	(0.6335, 5.6655)	-	-	0.2470
rs4964090	1.8842	(0.4499, 7.8906)	2.0321	(1.0955, 3.7696)	0.0322
rs7963840	1.2831	(0.2771, 5.9417)	1.7063	(0.9226, 3.1556)	0.1867

Table A.3: $G \times E$ parameter estimates and its 95% Confidence Intervals obtained with EEGM adjustment of CP children at 9 SNPs on *TBK1* gene (Chr 12) which showed evidence of significant interaction with Maternal Smoking.

SNP	$e^{\hat{\beta}_{GE}}$	95% CI	$e^{\hat{\beta}_{GE}+\hat{\beta}_{GEX_E}}$	95% CI	P-value
rs8091823	0.6904	(0.1239, 3.8469)	2.2546	(1.1529, 4.4088)	0.0540
rs3752075	2.0202	(0.3139, 13.0032)	2.2164	(1.2009, 4.0906)	0.0144
rs9960774	0.5322	(0.1020, 2.7769)	2.3881	(1.2061, 4.7284)	0.0365
rs486131	0.7675	(0.0480, 12.2659)	1.6425	(0.8236, 3.2756)	0.3583
rs10469070	1.3173	(0.0804, 21.5748)	2.2320	(0.6725, 7.4082)	0.3704
rs470385	0.1785	(0.0217, 1.4702)	1.6756	(0.6281, 4.4697)	0.1191
rs470560	0.5898	(0.1141, 3.0499)	2.7249	(1.4292, 5.1950)	0.0075
rs470563	0.5861	(0.1133, 3.0307)	2.9330	(1.5288, 5.6270)	0.0037
rs470337	1.3210	(0.1235, 14.1317)	0.7956	(0.3298, 1.9195)	0.8692
rs8095808	0.5681	(0.1027, 3.1413)	2.3192	(1.2088, 4.4493)	0.0350

Table A.4: $G \times E$ parameter estimates and its 95% Confidence Intervals obtained with EEGM adjustment of CP children at 10 SNPs on *ZNF236* gene (Chr 18) which showed evidence of significant interaction with Maternal Smoking.

SNP	$e^{\hat{\beta}_{GE}}$	95% CI	$e^{\hat{\beta}_{GE}+\hat{\beta}_{GEX_E}}$	95% CI	P-value
rs963599	1.5394	(0.7968, 2.9741)	1.6378	(0.9228, 2.9068)	0.0891
rs10955309	2.3263	(1.0711, 5.0523)	2.8767	(1.3475, 6.1415)	0.0013
rs1473541	1.3775	(0.7534, 2.5187)	1.2450	(0.6712, 2.3091)	0.4343
rs7814399	1.8379	(0.9504, 3.5543)	1.1470	(0.6365, 2.0669)	0.1590
rs3736042	1.8110	(0.9221, 3.5569)	2.1372	(1.0169, 4.4917)	0.0222
rs2454013	1.2588	(0.6985, 2.2685)	1.7823	(1.0409, 3.0520)	0.0710
rs2935579	1.2062	(0.6617, 2.1986)	1.2443	(0.6398, 2.4201)	0.6568
rs1874091	1.1328	(0.4507, 2.8471)	0.8685	(0.4553, 1.6566)	0.8874
rs1845430	1.3354	(0.7365, 2.4213)	1.7155	(0.9658, 3.0470)	0.1021
rs6468861	1.2080	(0.6531, 2.2345)	1.6268	(0.9390, 2.8185)	0.1699
rs6468862	1.6857	(0.2667, 10.6563)	1.6463	(0.9364, 2.8941)	0.1636

Table A.5: $G \times E$ parameter estimates and its 95% Confidence Intervals obtained with EEGM adjustment of CP children at 11 SNPs on *BAALC* gene (Chr 8) which showed evidence of significant interaction with Maternal Vitamin Supplementation.

Appendix B

Supplementary Material for Chapter 3: *Inference of gene-environment interaction from heterogeneous case-parent trios*

This appendix includes the published supplementary material for the manuscript entitled Inference of gene-environment interaction from heterogeneous case-parent trios.

B.1 Conditional likelihood and analysis

Assuming G and E are independent within families, one can write the conditional probability of the affected child's genotype given E and G_p in terms of the GRRs of equation (3.1), $GRR_g(e) = \exp(\beta_g + f_g(e))$. For example, when both parents are heterozygous, denoted below as $G_p = 3$ [26], one can show that the child's genotype probabilities are:

$$P(G = 0 \mid D = 1, E = e, G_p = 3) = \frac{1}{1 + 2\exp(\beta_1 + f_1(e)) + \exp(\beta_1 + f_1(e) + \beta_2 + f_2(e))},$$

$$P(G = 1 \mid D = 1, E = e, G_p = 3) = \frac{2\exp(\beta_1 + f_1(e))}{1 + 2\exp(\beta_1 + f_1(e)) + \exp(\beta_1 + f_1(e) + \beta_2 + f_2(e))}, \text{ and}$$

$$P(G = 2 \mid D = 1, E = e, G_p = 3) = \frac{\exp(\beta_1 + f_1(e) + \beta_2 + f_2(e))}{1 + 2\exp(\beta_1 + f_1(e)) + \exp(\beta_1 + f_1(e) + \beta_2 + f_2(e))}.$$

A complete list of conditional genotype probabilities for the affected child is given in Table 1 of [26]. For an additive model, in which $\beta_1 = \beta_2 \equiv \beta$ and $f_1(e) = f_2(e) \equiv f(e)$, the model

simplifies considerably; e.g.,

$$P(G = g \mid D = 1, E = e, G_p = 3) = \frac{\exp(O_g + g(\beta + f(e)))}{\sum_{i=0}^{2} \exp(O_i + i(\beta + f(e))},$$

where O_g is an "offset" term that equals $\log 2$ for $g = 1$ and 0 otherwise.

The likelihood is a product of conditional probabilities over all trios in the study, viewed as a function of the parameters β_1, β_2, $f_1(e)$ and $f_2(e)$. Each trio's contribution to the likelihood can be viewed as the contribution of a matched set to a likelihood for a conditional logistic regression, in which the matched set comprises the affected child and other possible offspring of the parents, referred to here as the affected child's pseudo-siblings. After constructing appropriate matched sets, software for conditional logistic regression may be used to maximize the likelihood from a case-parent trio study.

Code in the R environment for statistical computing [17] is available to perform such analyses and may be obtained from the first author upon request. The code sets up a data frame with rows for each affected child and pseudo-sibling, and columns specifying the ID for each trio (`ID`), affection status coded as 1 for the affected child and 0 for pseudo-siblings, an offset variable (`O`) coded as $\log 2$ for a heterozygous offspring of doubly-heterozygous parents and 0 otherwise, and the G, E and PC variables. We then call `clogit()` from the `survival` package [29] to perform the conditional logistic regression. The argument to `clogit()` is a formula that specifies affection status as the response, trio IDs as `strata(ID)`, offsets as `offset(O)` and the other model terms. For an additive model, the other model terms are a main effect for G, two-way interactions between G and E and between G and the PCs, and, finally, a three-way interaction between G, E and the PCs.

B.2 Dependence of latent-class probabilities on E

Write the probabilities in terms of the conditional distribution of GG' given E as

$$P(G = g|G' = g', E = e) = \frac{P(G = g, G' = g'|E = e)}{\sum_{i=0}^{2} P(G = i, G' = g'|E = e)}.$$

Supposing that the numerator and denominator both depend on E, so may their ratio. However, if we condition on the blocking variable X

$$\begin{aligned} P(G = g|G' = g', E = e, X = x) &= \frac{P(G = g, G' = g'|E = e, X = x)}{\sum_{i=0}^{2} P(G = i, G' = g'|E = e, X = x)} \\ &= \frac{P(G = g, G' = g'|X = x)}{\sum_{i=0}^{2} P(G = i, G' = g'|X = x)} \\ &= \frac{P(G = g, G' = g'|X = x)}{P(G' = g'|X = x)} \\ &= P(G = g|G' = g', X = x). \end{aligned}$$

Thus, latent-class probabilities in the model adjusted for X do not depend on E.

B.2.1 LDheatmaps of SNPs in *MLLT3*

LDheatmaps of pairwise R^2 values in and around the six SNPs in the *MLLT3* gene that showed significant $G \times E$ with maternal alcohol consumption in [3] are shown in Figure B.1 for self-reported Europeans and self-reported East Asians. There is generally stronger pairwise LD between SNPs that showed significant $G \times E$ in the self-reported Europeans than in the self-reported East Asians. The $-\log_{10}$ p-values from the PC-adjusted analysis are shown above the self-reported Europeans, who appear to be the drivers of the $G \times E$ signal.

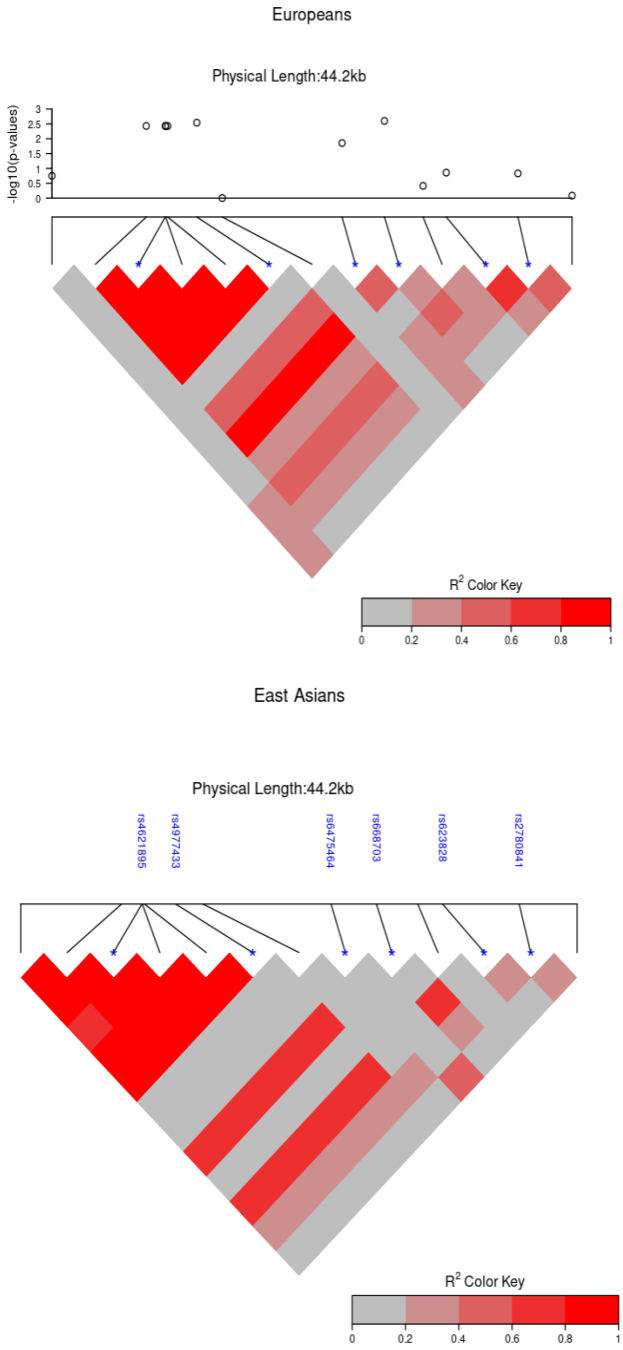

Figure B.1: LDheatmap of pairwise R^2 values in and around the six SNPs in the *MLLT3* gene that showed significant $G \times E$ with maternal alcohol consumption in [3]. Left panel: self-reported Europeans, with p-values from the PC-adjusted analysis shown above. Right panel: self-reported East Asians, with the names of the six SNPs shown above.

Appendix C

Supplementary Material for Chapter 4: *Adjustment for population stratification by local ancestry in gene-by-environment interaction studies of case-parent trios*

Tables C.1 through C.6 show the estimated interaction terms and corresponding 95% confidence intervals based on the risk model adjusted by $X_{GG'}$ for each of the six SNPs of interest from *MLLT3*. The estimate of $\beta_{g'E}$ represent the interaction effect in cluster 1, and $\beta_{g'E} + \beta_{g'E(k-1)}$ represent the interaction effects in cluster combinations $k = 2, ..., K$. In the first column we have listed the cluster memberships at a given locus along with the respective $G' \times E$ estimates. The sparse data and large number of parameters in the model lead to non-convergence of the estimation procedure, which precludes interpretation of the coefficient estimates and confidence intervals. More work on simplifying the model and the use of penalized likelihood methods for estimation is required. However, in what follows we provide example interpretations to illustrate the meaning of the coefficients.

As an example, the first row of table C.1 show the $G' \times E$ interaction of effect of trios whose allele have been assigned to clusters 1 and 3 at $rs4621895$. The corresponding interaction effect is $\beta_{g'E} = 3.3095 \times 10^{-7}$. In other words, the estimate of $e^{\beta_{g'E}} \approx 1$ suggest each copy of the minor allele at SNP $rs4621895$ do not elevate the relative risk of CP due to maternal alcohol exposure. Furthermore, corresponding 95% confidence interval of $[-0.4508, 0.4508]$, confirm that there is no significant interaction between $G' \times E$ among trios whose alleles at $rs4621895$ were assigned to clusters 1 and 3. We can observe a similar interpretation among all the trios who have been assigned to different clusters at SNPs $rs4621895$, $rs4977433$, $rs6475464$ and $rs668703$.

The results of tables C.5 and C.6 show that there is a significant $G' \times E$ interaction effect among trios whose allele have been assigned to clusters 1 and 3. The corresponding estimates of $e^{\beta_{g'E}} = 3.01$ and $e^{\beta_{g'E}} = 2.84$ suggest that each copy of the minor allele at SNPs $rs62382$ and $rs2780841$ more than trebles and doubles the relative risk of CP due to maternal alcohol exposure respectively. The corresponding 95% confidence intervals of $[0.63, 1.57]$ and $[0.54, 1.56]$ too confirm that $G' \times E$ interaction effect is significantly different from zero. However, the 95% confidence intervals of rest of the trios with different cluster assignments do not show significant $G' \times E$ interaction effect at both SNPs. As a result, overall $G' \times E$ interaction effect between $rs62382$ / $rs2780841$ with exposure to maternal alcohol consumption was found to be insignificant as shown in table 4.1.

Clusters	$G' \times E$	Estimate	95% CI
(1,3)	$\beta_{g'E}$	3.3095×10^{-7}	(-0.45, 0.45)
(3,3)	$\beta_{g'E} + \beta_{g'E1}$	-2.5036×10^{-7}	-
(3,4)	$\beta_{g'E} + \beta_{g'E2}$	0.1335	(-2.30, 2.56)
(3,5)	$\beta_{g'E} + \beta_{g'E3}$	3.3095×10^{-7}	(-0.45, 0.45)
(3,7)	$\beta_{g'E} + \beta_{g'E4}$	0.0335	(-0.95, 1.01)
(3,8)	$\beta_{g'E} + \beta_{g'E5}$	0.0985	(-0.59, 0.78)
(4,4)	$\beta_{g'E} + \beta_{g'E6}$	3.3095×10^{-7}	(-0.45, 0.45)
(4,5)	$\beta_{g'E} + \beta_{g'E7}$	3.3095×10^{-7}	(-0.45, 0.45)
(4,7)	$\beta_{g'E} + \beta_{g'E8}$	3.3095×10^{-7}	(-0.45, 0.45)
(4,8)	$\beta_{g'E} + \beta_{g'E9}$	3.3095×10^{-7}	(-0.45, 0.45)
(5,7)	$\beta_{g'E} + \beta_{g'E10}$	3.3095×10^{-7}	(-0.45, 0.45)
(7,7)	$\beta_{g'E} + \beta_{g'E11}$	7.2822×10^{-7}	-
(7,8)	$\beta_{g'E} + \beta_{g'E12}$	7.3221×10^{-7}	-
(8,8)	$\beta_{g'E} + \beta_{g'E13}$	-4.0081×10^{-7}	-

Table C.1: $G \times E$ parameter estimates and corresponding 95% Confidence Intervals of the interaction effect between rs4621895 on MLLT3 (Chr 9) gene and exposure to Maternal Alcohol Consumption at the presence of $X_{GG'}$ adjustment.

Clusters	$G' \times E$	Estimate	95% CI
(1,3)	$\beta_{g'E}$	-2.0107×10^{-6}	(-0.45, 0.45)
(2,3)	$\beta_{g'E} + \beta_{g'E1}$	-2.0107×10^{-6}	(-0.45, 0.45)
(3,3)	$\beta_{g'E} + \beta_{g'E2}$	-6.9191×10^{-8}	-
(3,4)	$\beta_{g'E} + \beta_{g'E3}$	0.1335	(-2.32, 2.58)
(3,5)	$\beta_{g'E} + \beta_{g'E4}$	-2.0107×10^{-6}	(-0.45, 0.45)
(3,7)	$\beta_{g'E} + \beta_{g'E5}$	0.0335	(-0.95, 1.01)
(3,8)	$\beta_{g'E} + \beta_{g'E6}$	0.1131	(-0.59, 0.82)
(4,4)	$\beta_{g'E} + \beta_{g'E7}$	-2.0107×10^{-6}	(-0.45, 0.45)
(4,5)	$\beta_{g'E} + \beta_{g'E8}$	-2.0107×10^{-6}	(-0.45, 0.45)
(4,7)	$\beta_{g'E} + \beta_{g'E9}$	-2.0107×10^{-6}	(-0.45, 0.45)
(4,8)	$\beta_{g'E} + \beta_{g'E10}$	-2.0107×10^{-6}	(-0.45, 0.45)
(5,7)	$\beta_{g'E} + \beta_{g'E11}$	-2.0107×10^{-6}	(-0.45, 0.45)
(7,7)	$\beta_{g'E} + \beta_{g'E12}$	-2.0107×10^{-6}	-
(7,8)	$\beta_{g'E} + \beta_{g'E13}$	-2.0107×10^{-6}	-
(8,8)	$\beta_{g'E} + \beta_{g'E14}$	1.5722×10^{-7}	-

Table C.2: $G \times E$ parameter estimates and corresponding 95% Confidence Intervals of the interaction effect between rs4977433 on MLLT3 (Chr 9) gene and exposure to Maternal Alcohol Consumption at the presence of $X_{GG'}$ adjustment.

Clusters	$G' \times E$	Estimate	95% CI
(2,3)	$\beta_{g'E}$	1.3862×10^{-7}	(-0.49, 0.49)
(2,8)	$\beta_{g'E} + \beta_{g'E1}$	1.3862×10^{-7}	(-0.49, 0.49)
(3,3)	$\beta_{g'E} + \beta_{g'E2}$	-8.6897×10^{-8}	-
(3,4)	$\beta_{g'E} + \beta_{g'E3}$	2.9109×10^{-8}	-
(3,5)	$\beta_{g'E} + \beta_{g'E4}$	1.3862×10^{-7}	(-0.49, 0.49)
(3,7)	$\beta_{g'E} + \beta_{g'E5}$	-2.7573×10^{-8}	-
(3,8)	$\beta_{g'E} + \beta_{g'E6}$	-0.1194	(-0.82, 0.59)
(4,4)	$\beta_{g'E} + \beta_{g'E7}$	1.3862×10^{-7}	(-0.49, 0.49)
(4,5)	$\beta_{g'E} + \beta_{g'E8}$	1.3862×10^{-7}	(-0.49, 0.49)
(4,7)	$\beta_{g'E} + \beta_{g'E9}$	1.3862×10^{-7}	(-0.49, 0.49)
(4,8)	$\beta_{g'E} + \beta_{g'E10}$	1.3862×10^{-7}	(-0.49, 0.49)
(5,7)	$\beta_{g'E} + \beta_{g'E11}$	1.3862×10^{-7}	(-0.49, 0.49)
(5,8)	$\beta_{g'E} + \beta_{g'E12}$	1.3862×10^{-7}	(-0.49, 0.49)
(6,8)	$\beta_{g'E} + \beta_{g'E13}$	1.3862×10^{-7}	(-0.49, 0.49)
(7,7)	$\beta_{g'E} + \beta_{g'E14}$	1.3862×10^{-7}	-
(7,8)	$\beta_{g'E} + \beta_{g'E15}$	-1.4110	(-3.12, 0.29)
(8,8)	$\beta_{g'E} + \beta_{g'E16}$	1.3862×10^{-7}	-

Table C.3: $G \times E$ parameter estimates and corresponding 95% Confidence Intervals of the interaction effect between rs6475464 on MLLT3 (Chr 9) gene and exposure to Maternal Alcohol Consumption at the presence of $X_{GG'}$ adjustment.

Clusters	$G' \times E$	Estimate	95% CI
(1,8)	$\beta_{g'E}$	2.4822×10^{-6}	(-0.43, 0.43)
(2,3)	$\beta_{g'E} + \beta_{g'E1}$	2.4822×10^{-6}	(-0.43, 0.43)
(2,8)	$\beta_{g'E} + \beta_{g'E1}$	2.4822×10^{-6}	(-0.43, 0.43)
(3,3)	$\beta_{g'E} + \beta_{g'E2}$	-6.9468×10^{-8}	-
(3,4)	$\beta_{g'E} + \beta_{g'E3}$	0.0870	(-2.34, 2.52)
(3,5)	$\beta_{g'E} + \beta_{g'E4}$	2.4822×10^{-6}	(-0.43, 0.43)
(3,7)	$\beta_{g'E} + \beta_{g'E5}$	0.1844	(-0.80, 1.16)
(3,8)	$\beta_{g'E} + \beta_{g'E6}$	0.1215	(-0.56, 0.81)
(4,4)	$\beta_{g'E} + \beta_{g'E7}$	2.4822×10^{-6}	(-0.43, 0.43)
(4,5)	$\beta_{g'E} + \beta_{g'E8}$	2.4822×10^{-6}	(-0.43, 0.43)
(4,7)	$\beta_{g'E} + \beta_{g'E9}$	2.4822×10^{-6}	(-0.43, 0.43)
(4,8)	$\beta_{g'E} + \beta_{g'E10}$	2.4822×10^{-6}	(-0.43, 0.43)
(5,7)	$\beta_{g'E} + \beta_{g'E11}$	2.4822×10^{-6}	(-0.43, 0.43)
(5,8)	$\beta_{g'E} + \beta_{g'E12}$	2.4822×10^{-6}	(-0.43, 0.43)
(7,7)	$\beta_{g'E} + \beta_{g'E14}$	2.4822×10^{-6}	-
(7,8)	$\beta_{g'E} + \beta_{g'E15}$	2.4822×10^{-6}	-
(8,8)	$\beta_{g'E} + \beta_{g'E16}$	-6.0946×10^{-8}	-

Table C.4: $G \times E$ parameter estimates and corresponding 95% Confidence Intervals of the interaction effect between rs668703 on MLLT3 (Chr 9) gene and exposure to Maternal Alcohol Consumption at the presence of $X_{GG'}$ adjustment.

Clusters	$G' \times E$	Estimate	95% CI
(1,3)	$\beta_{g'E}$	1.1032	(0.63, 1.57)
(1,6)	$\beta_{g'E} + \beta_{g'E1}$	-42.4057	-
(1,8)	$\beta_{g'E} + \beta_{g'E2}$	-18.7826	-
(2,2)	$\beta_{g'E} + \beta_{g'E3}$	1.1032	-
(2,3)	$\beta_{g'E} + \beta_{g'E4}$	18.2075	-
(2,4)	$\beta_{g'E} + \beta_{g'E5}$	1.1032	(0.63, 1.57)
(2,6)	$\beta_{g'E} + \beta_{g'E6}$	1.1032	-
(2,7)	$\beta_{g'E} + \beta_{g'E7}$	18.9007	-
(2,8)	$\beta_{g'E} + \beta_{g'E8}$	-0.5108	(-1.90, 0.88)
(3,3)	$\beta_{g'E} + \beta_{g'E9}$	-5.3862×10^{-8}	-
(3,4)	$\beta_{g'E} + \beta_{g'E10}$	18.7759	-
(3,5)	$\beta_{g'E} + \beta_{g'E11}$	1.1032	(0.63, 1.57)
(3,6)	$\beta_{g'E} + \beta_{g'E12}$	-19.1234	(-21.26, -16.99)
(3,7)	$\beta_{g'E} + \beta_{g'E13}$	-0.4748	(-1.83, 0.88)
(3,8)	$\beta_{g'E} + \beta_{g'E14}$	0.5968	(-0.52, 1.71)
(4,4)	$\beta_{g'E} + \beta_{g'E15}$	1.1032	(0.63, 1.57)
(4,5)	$\beta_{g'E} + \beta_{g'E16}$	1.1032	(0.63, 1.57)
(4,6)	$\beta_{g'E} + \beta_{g'E17}$	1.1032	(0.63, 1.57)
(4,7)	$\beta_{g'E} + \beta_{g'E18}$	1.1032	(0.63, 1.57)
(4,8)	$\beta_{g'E} + \beta_{g'E19}$	1.1032	(0.63, 1.57)
(5,7)	$\beta_{g'E} + \beta_{g'E20}$	1.1032	(0.63, 1.57)
(5,8)	$\beta_{g'E} + \beta_{g'E21}$	1.1032	(0.63, 1.57)
(6,6)	$\beta_{g'E} + \beta_{g'E22}$	-20.5098	(-22.96, -18.06)
(6,7)	$\beta_{g'E} + \beta_{g'E23}$	0.5108	(-1.19, 2.22)
(6,8)	$\beta_{g'E} + \beta_{g'E24}$	0.1625	(-0.78, 1.10)
(7,7)	$\beta_{g'E} + \beta_{g'E25}$	1.1032	-
(7,8)	$\beta_{g'E} + \beta_{g'E26}$	5.1782×10^{-9}	-
(8,8)	$\beta_{g'E} + \beta_{g'E27}$	-7.4294×10^{-8}	-

Table C.5: $G \times E$ parameter estimates and corresponding 95% Confidence Intervals of the interaction effect between rs623828 on MLLT3 (Chr 9) gene and exposure to Maternal Alcohol Consumption at the presence of $X_{GG'}$ adjustment.

Clusters	$G' \times E$	Estimate	95% CI
(1,3)	$\beta_{g'E}$	1.0454	(0.54, 1.56)
(1,4)	$\beta_{g'E} + \beta_{g'E1}$	1.0454	(0.54, 1.56)
(1,6)	$\beta_{g'E} + \beta_{g'E2}$	-21.2029	(-24.02, -18.38)
(1,8)	$\beta_{g'E} + \beta_{g'E3}$	1.0454	-
(2,3)	$\beta_{g'E} + \beta_{g'E4}$	19.2483	-
(2,4)	$\beta_{g'E} + \beta_{g'E5}$	1.0454	(0.54, 1.56)
(2,6)	$\beta_{g'E} + \beta_{g'E6}$	1.0454	-
(2,7)	$\beta_{g'E} + \beta_{g'E7}$	1.0454	(0.54, 1.56)
(2,8)	$\beta_{g'E} + \beta_{g'E8}$	1.0454	-
(3,3)	$\beta_{g'E} + \beta_{g'E9}$	7.8389×10^{-8}	-
(3,4)	$\beta_{g'E} + \beta_{g'E10}$	3.2214×10^{-7}	-
(3,5)	$\beta_{g'E} + \beta_{g'E11}$	1.0454	(0.54, 1.56)
(3,6)	$\beta_{g'E} + \beta_{g'E12}$	-2.0784×10^{-7}	-
(3,7)	$\beta_{g'E} + \beta_{g'E13}$	0.1082	(-1.26, 1.48)
(3,8)	$\beta_{g'E} + \beta_{g'E14}$	0.2877	(-0.69, 1.27)
(4,4)	$\beta_{g'E} + \beta_{g'E15}$	1.0454	(0.54, 1.56)
(4,5)	$\beta_{g'E} + \beta_{g'E16}$	1.0454	(0.54, 1.56)
(4,6)	$\beta_{g'E} + \beta_{g'E17}$	1.0454	-
(4,7)	$\beta_{g'E} + \beta_{g'E18}$	1.0454	(0.54, 1.56)
(4,8)	$\beta_{g'E} + \beta_{g'E19}$	1.0454	(0.54, 1.56)
(5,7)	$\beta_{g'E} + \beta_{g'E20}$	1.0454	(0.54, 1.56)
(5,8)	$\beta_{g'E} + \beta_{g'E21}$	1.0454	(0.54, 1.56)
(6,6)	$\beta_{g'E} + \beta_{g'E22}$	2.8119×10^{-8}	-
(6,7)	$\beta_{g'E} + \beta_{g'E23}$	1.0498	(-1.16, 3.26)
(6,8)	$\beta_{g'E} + \beta_{g'E24}$	0.1660	(-0.81, 1.15)
(7,7)	$\beta_{g'E} + \beta_{g'E25}$	1.0454	-
(7,8)	$\beta_{g'E} + \beta_{g'E26}$	-5.7947×10^{-9}	-
(8,8)	$\beta_{g'E} + \beta_{g'E27}$	4.1481×10^{-9}	-

Table C.6: $G \times E$ parameter estimates and corresponding 95% Confidence Intervals of the interaction effect between rs2780841 on MLLT3 (Chr 9) gene and exposure to Maternal Alcohol Consumption at the presence of $X_{GG'}$ adjustment.