Maternal Gene-Environment Effects: An evaluation of statistical approaches to detect effects and an investigation of the effect of violations of model assumptions

Julie Hudson

Thesis submitted to the Faculty of Graduate and Postdoctoral Studies in partial fulfillment
of the requirements for the degree of
Mathematics and Statistics, Specialisation in Biostatistics[1]

Department of Mathematics and Statistics
Faculty of Science
University of Ottawa

[1]The program is a joint program with Carleton University, administered by the Ottawa-Carleton Institute of Mathematics and Statistics

Abstract

Discovering the associations between genetic variables and disease status can help reduce the burden of disease on society. This thesis focuses on the methods required to detect maternal genetic effects (an effect where the genes of the mother affect the disease risk of the child) and interaction effects between these maternal genes and environmental variables in trio data consisting of parents and an affected child. A simulation study was conducted to determine the extent to which testing for these effects is affected by violations to the mating symmetry assumption required for two current methods when control parents are not available.. This study showed that methods for maternal effect estimation are not robust to these violations; however, the interaction test is robust to the violation. Finally, a candidate gene study on orofacial clefts was conducted to evaluate maternal gene-environment interactions in international consortium data. Significant effects were found but the large magnitude of the effect estimates raises concerns about the validity of the results. This thesis tries also discusses the lack of methods and software available to estimate maternal gene environment interactions.

Acknowledgement

Thank you, first and foremost, to my supervisors Dr. Kelly Burkett and Dr. Marie-Hélène Roy-Gagnon without whom this thesis would never have been finished. This has been a difficult journey but having supportive, intelligent and helpful female role models has made it possible.

I would also like to thank the friends I made in Ottawa who helped make it my home for the short time I was there. I sometimes miss the city because of them.

Finally, my family, friends and partner Felipe have all supported and loved me in ways no one expects to be and for that I will be forever blessed.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Detecting associations between genetic variables and disease status is an important endeavor in health research. In the last 30 years, conducting genetic association studies has become more feasible thanks to advancements in genotyping technology [43]. However, with this huge increase in genotype data comes a need for novel statistical methods to be developed and for detailed evaluation of the plausibility of assumptions underlying these methods.

This work will focus on a powerful study design used to detect genetic associations using familial data: the parent-offspring trio design. This is a commonly used design wherein the two biologic parents and an offspring affected by the disease in question are sampled. This trio design is most commonly used to find genetic variants that are preferentially passed down to a child using tests such as the Transmission Disequilibrium Test (TDT) [53]. The TDT is an application of the McNemar test, where a chi-square statistic is calculated to evaluate whether the risk variant is passed to the affected offspring more often than the non-risk variant.This test is more robust than other, more traditional methods such as case-control studies because it is robust to population stratification, which is a known confounder in genetic association studies. This concept will be discussed further in Chapter 2.

Another advantage of the trio design is that it can also be used, as it will be in this work, to detect maternal genetic effects. These effects occur when the increased risk of disease is due to the mother's genetic variant. These variants are important because the mother provides the gestational environment in which the child develops. For disorders that begin in early life, the genetic variables of the mother may be the key in understanding the source. Some diseases with known maternal genetic contributions to disease risk are orofacial clefts [50], spina bifida [11] and schizophrenia [25].

Prenatal non-genetic exposures for a child include the state of the environment provided by the mother, both from her genetic makeup and her behaviours/exposures. For the rest of this work, the maternal exposures and behaviours during pregnancy will be referred to as the environmental variables. Environmental exposures are also known to contribute to disease risk. Just as in other statistical analyses, it is important to check whether there exists any interactions between the exposures. In this context, that means testing for gene-environment interactions. The specifics of maternal gene-environment interactions will be described in Chapter 2. To illustrate what is meant by this sort of interaction, imagine a disease associated with a genetic marker G where the variant causes blood pressure to increase. If a person with this G variant also smokes, the two factors might have a synergistic effect in increasing blood pressure (smoking is known to cause increased blood pressure [42]). Alone, each factor might increase risk for heart failure by 20%, but combined that risk is increased to 70%. This same concept applies to the genes of a mother and the environmental variables she is exposed to during pregnancy and shortly after. Should a child or fetus develop in an environment with exposures that interact with the genetic factors from the mother, a maternal gene-environment interaction is said to exist. The environmental exposures can be either protective or destructive. This work will look at smoking and prenatal vitamin supplementation.

Statistical methods to detect maternal effects beyond any child effect are limited. In 1998 and beyond, Weinberg et al. [64] [65] [66] published a series of papers using nested log-linear models designed to test whether maternal variants were overrepresented in the sampled genetic information of mothers, conditional on the child's genetic information. Multinomial modelling was also used by Ainsworth et al. [2], and Cordell et al. [22] to develop a direct maximization method to tease apart these effects. Both of these methods have been implemented in freely-available software packages. The two software used in this work are EMIM, a multinomial modelling software designed to detect indirect genetic effects at the parental level;and LEM, a generic log-linear regression software. However, neither approach is sufficient for analysing the dataset that motivated this work. EMIM does not take as input environmental information, so post-hoc tests of heterogeneity are conducted instead; LEM is not user friendly and the parameterization by Weinberg et al. used here needs to be implemented in more current software. Therefore, in this work, we test and expand on these methods and in Chapter 6, this lack of adequate methods will be discussed.

When examining these methods, it becomes clear that one of the modelling assumptions that is relied upon heavily is that the probability of a particular genotype assignment within a mate-pair is equal to the opposite assignment (e.g. the probability of the mother having one risk variant and the father having none is equal to the probability of the mother having no risk variants and the father having one). This concept of mating symmetry has been shown to not always hold due to various biologic factors such as assortative mating and other differential mating patterns [5] [47]. Distinguishing between a maternal effect and mating asymmetry is not possible when the study contains only case family data since it is precisely this overrepresentation of maternal risk variants that we expect to see in a maternal effect. This will be explored in Chapter 3, where additional details will be provided.

The main goals of this thesis are to investigate how to detect maternal effects in the

presence of mating asymmetry (MA) and to extend the existing methods to detect maternal gene-environment interactions. The motivation for this work came from data from an international study on orofacial clefts. The researchers collected genome-wide data on affected offspring and their parents (some incomplete trios were included as well). The primary analysis conducted on these data was to investigate child genetic effects [3] and they found two potential risk variants. Since that publication, others have investigated maternal effects, so we are looking at the potential for the maternal gene-environment interaction. As mentioned in the previous paragraph, the lack of control trio data means that mating symmetry will be assumed in any analysis we do.

Highlighting the scarcity of resources to detect these indirect genetic effects (both maternal and maternal gene-environment) is a secondary goal for this work. The methods described are far from perfect: the first method does not allow for the interaction effect to be estimated, and the second method, in the real data analysis section, found unreliable estimates and suggests perhaps a software or model fitting flaw. This work can be used as motivation for further methods development to take place.

This thesis is organized as follows. In Chapter 2, the background genetic and epidemiological information is explained. This includes genetic transmission of disease, sources of confounding, and an explanation of the study design. In Chapter 3 the statistical methods are described, along with their software implementations.

The focus of Chapter 4 is on evaluating the effect of mating asymmetry on the performance of statistical models to detect maternal effects and maternal gene-environment interaction effects. Type I error, power and bias are estimated through simulation. Trio data were simulated under scenarios of mating symmetry and mating asymmetry with maternal and child genetic effects. The methods of Cordell et al. and Weinberg et al. were

compared and used to evaluate the presence a maternal gene-environment variable as well. We show that, under the weaker assumption of independence between environment and mating asymmetry, the estimation of maternal gene-environment interaction will be unaffected by MA.

In Chapter 5, the methods described in Chapter 3 are applied to the orofacial cleft study mentioned previously. Genetic variables within five genes previously associated with orofacial clefts were genotyped in a sample of European families. As the data were collected only on case families, our analysis must assume mating symmetry. We find an unusual number of highly significant interaction effects, which could be explained by mating asymmetry (lack of independence between the environment variable and mating symmetry) or a poor fitting model. In Chapter 6, I conclude the thesis with a discussion of the limitations of the modelling and with future directions for research.

Chapter 2

Genetic and Epidemiological Background

In this chapter concepts in genetics and epidemiology will be introduced. The methods evaluated in this thesis are used for identifying maternal genetic effects and they are susceptible to confounding in many ways.

2.1 Terminology

As I will be investigating methods for genetic data analysis, I now introduce the genetic terminology that will be used throughout the thesis. For a more thorough background, Laird and Lange (2011), Chapters 1-3, provides a good summary of these concepts for a statistical audience [28].

Human genetic information is encoded within the human genome. The human genome consists of 23 pairs of chromosomes which are composed of protein and deoxyribonucleic acid (DNA). The DNA molecule is made up of four different types of subunits or bases - adenine (A), guanine (G), cytosine (C) or thymine (T). These nucleotides pair with their

complementary base (A pairs with T, C pairs with G) to form base-pairs, of which there are around three billion in the genome.

Genes are subunits of DNA that are composed of a varying length of bases. Genes encode the information for creating the proteins that play important functional and structural roles within an organism. Current estimates for the number of genes in the human genome place the total around 20,000. Studies of genetic information tend to focus on the approximately 0.1% of the genome that varies between individuals. Genetic loci (singular: locus) are locations along the genome that have been observed to be vary between individuals. These are called polymorphisms, mutations or markers. The most common type of polymorphism evaluated in human genetic studies is the single nucleotide polymorphism (SNP), which is a substitution of one base for another at a single nucleotide. For example, in a large sample, we might observe 40% of the sample has an A base and 60% has the C base.

The different variants observed in the population at a locus are called alleles. For a diallelic marker, two variants are found and are generally noted as A for the more common allele and a for the less common or minor allele. This notation must not be confused with the abbreviation for adenine "A".

SNPs are diallelic markers; that is, in the human population only two alleles are typically observed. At each SN,P the pair of alleles inherited from the father and mother is called the genotype. If a person carries two identical allelles (i.e., AA or aa), the genotype is homozygous. If the alleles are different (i.e., Aa), the genotype is heterozygous. We often will look at the number of minor alleles present in the genotype; the genotype can then be coded as 0, 1, or 2 for the count of number of minor alleles.

As this work deals specifically with the pattern of genotypes observed in family units,

it is important to understand the process of transmission of alleles from parent to child. The term "Mendelian transmission" is used to indicate a locus that is inherited according to Mendel's first law of inheritance: One allele from each parent is randomly and independently selected from the parent's two alleles. That is, the probability a parent transmits a particular allele to their child is $\frac{1}{2}$, and the allele transmitted from one parent is independent of the allele transmitted from the other. Therefore, based on the genotypes of the parents, the probability model of the genotype of the child is known. Conditional upon parental genotypes, Table 2.1 shows the discrete distribution of offpring genotypes. For example,

$$\Pr(C = aa|M = aa, F = Aa...) = \Pr(a\ from\ M)\Pr(a\ from\ F) = 1 \times 1/2$$

.

Table 2.1: Probability distribution of offspring's genotype conditional on parental genotypes

Maternal Genotype	Paternal Genotype	Offspring Genotype		
		aa	Aa	AA
aa	aa	1	0	0
aa	Aa	$\frac{1}{2}$	$\frac{1}{2}$	0
aa	AA	0	1	0
Aa	aa	$\frac{1}{2}$	$\frac{1}{2}$	0
Aa	Aa	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
Aa	AA	0	$\frac{1}{2}$	$\frac{1}{2}$
AA	aa	0	1	0
AA	Aa	0	$\frac{1}{2}$	$\frac{1}{2}$
AA	AA	0	0	1

This allele transmission probability model is the foundation for many statistical genetic analyses; in section 3.1, I will show how the transmission probabilities shown in Table 2.1 are used in deriving the multinomial probabilities which are used to estimate relative risk parameters with case trio data.

A population genetics principle that is commonly assumed is called Hardy Weinberg Equilibrium (HWE). HWE is stated as follows: in a population free from selection, mutation, migration, and random mating, with an infinite population size, discrete generations, and equal allele frequencies in males and females, it can be shown that allele frequencies from one generation to another will not change and the genotypes of the children's generation have the following probabilities:

$$P(AA) = p^2$$

$$P(Aa) = 2pq$$

$$P(aa) = q^2$$

where p is the allele frequency for A and q is the allele frequency for a. Of course, the conditions described above do not hold fully in any population. Nevertheless, the principle provides a good approximation for population genotype frequencies and a model for estimating allele frequencies.

To estimate the A allele frequency in a sample population, let n be the number of unrelated individuals in the sample, n_{AA} be the number of homozygous A individuals, and n_{Aa} be the number of heterozygotes. Then,

$$p_A = (2n_{AA} + n_{Aa})/2n.$$

At a SNP where two alleles exist, the proportion of a alleles will be $p_a = 1 - p_A$. This is a

maximum likelihood estimate from the multinomial distribution.

An assumption of Hardy Weinberg Equilibrium (HWE) is often made in genetic data analysis. Note that one of the assumptions required for HWE to apply is that of random mating. Random mating means that all individuals of the opposite sex are equally likely to be a mating partner. Forms of non-random mating include inbreeding.

In studies that determine if there is a genetic basis for various response variables, geneticists call these outcome variables the phenotype. A phenotype can be any measurable trait in an individual: disease status, eye colour, weight, etc. Phenotypes can be continuous, categorical, or dichotomous. The goal of genetic studies is to detect association between genotype and phenotype. This task is complicated by various sources of confounding, interaction, limited statistical methodologies, and other factors which will be addressed in subsequent sections.

2.2 Types of Genetic Effects

When considering the genetic contributions to disease, all types of genetic effects must be considered. Primary consideration within genetic association studies is to identify variants within an individual's genome that will alter their risk for disease. This is a direct genetic effect. Although this type is the most studied and, as a result, a plethora of methods to detect individual genetic risk have been developed, these do not describe all genetic associations to disease that are detectable. It is here where indirect genetic effects fit in. Indirect genetic effects arise from a gene expressed in an interacting individual and not in the individual whose phenotype is measured [67]. These indirect genetic effects are most often studied in mothers of offspring as mothers provide the gestational environment and post-natal care.

When discussing maternal genetic effects as compared to individual effects, it is important to recognize how a maternal effect may be mistaken for a child effect and vice versa. Assume only a true maternal effect exists so the true relative risk of the child allele is 1. In this scenario, the risk allele will be enriched in the mothers of disease affected children as compared to the mothers of unaffected children. Let D be the disease status so $D = 1$ represents having the disease, and M, F and C be the risk allele count for the mother, father and child respectively. If, say, $P(D = 1|M = 1) = 1$, this means that the allele being present in the mother guarantees her offspring will have the disease. Recall that the true child $RR = 1$, so any presence of the allele in the child genotype has no bearing on their disease susceptibility. But, by Mendelian transmission, we know that $P(C = 1|M = 1) = 0.5$ so these offspring will have a higher risk allele frequency than the control offspring since $P(D = 0|M = 0) = 1$, therefore the control mothers are not passing on the risk allele at all.

In an association study which examines only the offspring, the higher frequency of the risk allele among affected individuals might be high enough to be detected with statistical methods. The effect size would be smaller than the true maternal effect size because the cases would have fewer risk alleles than their mothers. A strong child effect may also be mistaken for a weaker maternal effect using the same logic. Taking this into account, all maternal effects will be detected conditionally on a child effect in the analyses presented in this thesis. That is to say, a maternal effect will only be considered present if it exists above and beyond a child effect. This caution is taken because a maternal effect is less likely than an effect from one's own alleles.

Another class of gene effects that are important to account for are gene-environment interactions. Studying environmental effects, both main and interaction, is important because some exposures are modifiable. For example, if a child is homozygous for the phenylketonuria

gene variant, he or she will have severe developmental delays if their diet includes the amino acid phenylalanine. However, development is normal if the diet is restricted to not include phenylalanine. Gene-environment interactions can obscure both environmental and genetic effects when the disease risk is altered only when both the exposure and the risk allele are present. This is often a simple linear interaction, shown in Figure 2.1a. Three non-linear interaction effect scenarios (exclusive or, conditional dominant and small marginal effect) will only be captured when two separate risk parameters are estimated for carrying one or two alleles. This will be discussed in Chapter 3 when deciding whether restricting the number of fitted parameters in the allele model is appropriate.

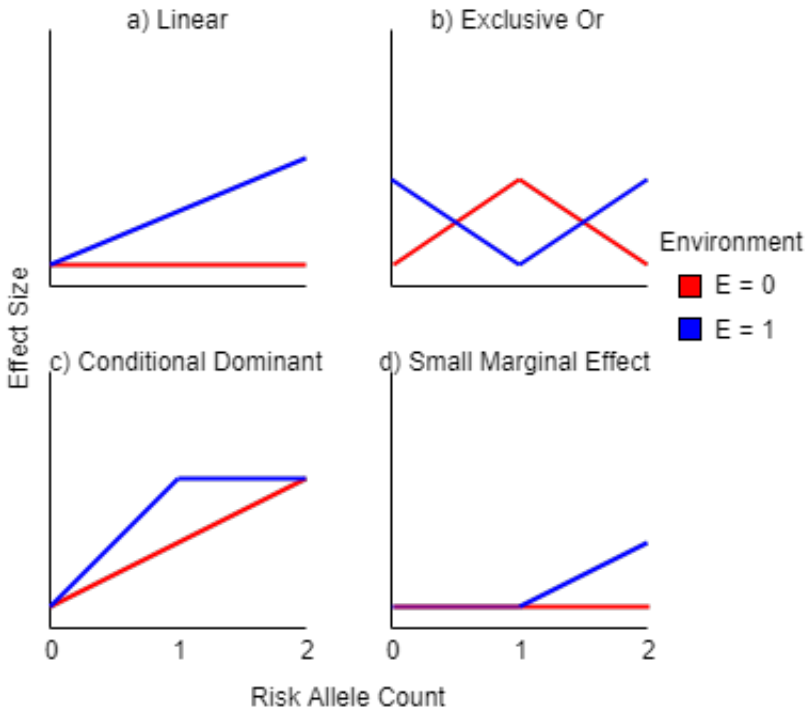


Figure 2.1: Scenarios for gene-environment interactions

2.3 Confounding

When a statistical association is detected between a covariate and a response variable, one must consider whether the association is actually due to confounding. A variable is considered a confounder if it meets three properties: (1) it is associated with the disease, independently of the risk factor (in this case, the locus), (2) it is associated with the risk factor, and (3) it is not an intermediary step in the causal pathway between the risk factor and the disease. Since confounders are necessarily not on the causal pathway of interest, they are a nuisance effect that can lead to faulty conclusions, either of a significant genetic association when one is not present (positive confounding) or the obfuscation of a true effect (negative confounding). In Figure 2.2, the association of interest is represented by the horizontal arrow (left to right to indicate that the risk factor is causing the disease). The confounder, at bottom, is associated with both the risk factor and the disease but the direction of these associations are unknown (represented by question marks). Is the risk factor causing the person to exhibit the confounder or is the confounder causing the risk factor? This unknown direction of effect creates the problem of an unknown true causal relationship. To account for confounding, data is collected on any known confounding variables and they are typically included as variables in the statistical modeling.

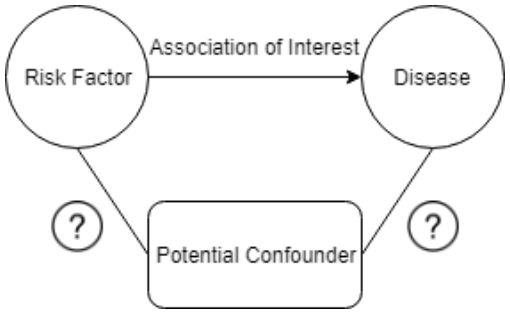


Figure 2.2: Diagram of Confounding

Confounding in the context of genetic association studies will come about through any process that distorts allelic frequency or transmission that is not accounted for. In population-based association studies, the main source of confounding is population stratification. Confounding due to population stratification can occur if the study is composed of genetic subpopulations and each subpopulations has a different disease prevalence. If not accounted for, any genetic variant that has different allele frequencies in the different subpopulations may be significantly associated with disease. Maternal effect tests are susceptible to confounders through mating asymmetry. These two concepts will be described in the next two subsections.

2.4 Population Stratification

As previously described, population stratification occurs when there is a difference in disease prevalence between cases and controls, and a difference in frequencies of genetic variation between groups. Two instances where population stratification may be present are in international studies, where cases and controls are recruited from populations around the world, as well as in studies with genetic admixture, which occurs when isolated populations begin interbreeding. This genetic admixture can be easily seen in studies of North Americans, where there has been on average only six generations since African and European populations came into contact in North America [51]. A classic example of confounding due to population stratification and genetic admixture comes from a study of type 2 diabetes mellitus in the Pima and Papago tribes of Native Americans. Initial analyses of the collected genetic information found a strong negative association (prevalence ratio = 0.27) between type 2 diabetes mellitus and the Gm haplotype $Gm^{3;5,13,14}$ (a haplotype is a collection of alleles inherited from one parent). However, when the analysis was stratified based on the fraction of ancestry being Native American, the association was no longer present [24].

The severity of confounding due to population stratification in association studies is debated in the field of genetic epidemiology ([6] argues it is a minimal issue, and [56] explains how crucial it is to consider this source of confounding). However, it is generally agreed that with proper study design and methodology, this problem can be addressed. In case-control study designs, the most important step taken to prevent this bias is by collecting ancestry-related data and accounting for ancestry in the analysis. This can be done using self-reported ethnicity data (though this can sometimes be an unreliable method [63]) or ethnicity defined by the top principal components of a principal component analysis on genomic SNP data [27]. Study design might also restrict study participants to those of one ethnic group. This helps reduce the likelihood of population stratification but causes some ethnic groups to be under-studied. Another method to control for population stratification is by using family-based study designs. Briefly, family-based studies are robust to population stratification because instead of using unrelated controls, the untransmitted alleles from parents are used as pseudo-controls. This creates a perfectly matched control.

2.5 Genetic Association Studies

Studies of genetic association compare the allele and genotype frequencies between cases and controls (or within families) to see whether the number of variants is higher or lower than unexpected under a hypothesis of no association. This would suggest that there exists some relationship between the disease present in the cases and the genetic component. The relative risk is a measure used to describe the increased or decreased risk of disease based on a particular genotype. By risk we mean the likelihood of developing a disorder. The relative risk (RR) is defined to be the ratio between the risk of disease among exposed ($E = 1$) and unexposed ($E = 0$) groups where in this case, exposure will be the number of risk alleles:

$$RR = \frac{P(D = 1|E = 1)}{P(D = 1|E = 0)}$$

where E is the exposure and D is the disease. If RR = 1 then the risk is equal in exposed and unexposed.

If RR > 1, the risk is higher among the exposed - i.e., the probability of disease is higher in those exposed relative to those that aren't exposed - and if RR < 1, the risk is lower among the exposed.

2.6 Trio Study Design

There exist many study designs to conduct analysis on human genetic data with the goal of finding the underlying cause of various traits and diseases. Which design is most appropriate depends on the context of the study question. As discussed in Section 2.4, case-control studies are susceptible to confounding by population stratification, whereas family-based designs are more robust to population stratification.

This work will focus on the trio study design. The members of the trio are a child and the child's biological mother and father. A systematic analysis of 93 case-control studies and family-based studies found no heterogeneity of estimates for child genetic effects [12], with no observable tendency towards over- or underestimation using either design. When considering the power added per genotyped individual, trio designs will have lower power [20]. This is because some families will contribute non-informative genetic information (if, say, the parent of origin cannot be ascertained in a maternal effect study) and in traditional case-control studies where you might have one non-case for every case, a trio study has at least two non-cases for every case. When control-trios are collected, this increases to 5 non-cases for every case. The primary reason for the use of a trio design here is the ability to detect maternal genetic effects with this design; this analysis cannot be conducted with a case-control study.

Case-only trio genetic studies will have 15 cell counts to estimate risk parameters as well as parameters corresponding to parental genotypes (the mating-type parameters). As will be discussed in the statistical background (Chapter 3) and investigated through simulation in Chapter 4, an assumption of mating symmetry is required for case-only trio studies. In the next section we will define the additional parameters, which will be called mating-type parameters, that need to be included in order to relax the assumption of mating symmetry. These mating-type parameters need to be estimated in a non-diseased population so that they describe mating and aren't instead capturing a true maternal effect. That is, this confounding can only be addressed in the statistical analysis when control families are available. Note that for logistic and financial reasons, it can be more difficult to justify the recruitment of control trios as compared to case trios. There are therefore some hybrid study designs that are used to include both case trios and unrelated control groups. For instance, Vermeulen et al (2009) augments case trios with control-mother dyads [60], and Weinberg and Umbach (2005) study case trios and the parents of controls to reduce the genotyping costs [64].

2.7 Mating Asymmetry

Mating symmetry (MS) is the assumption that the probability of a genotype for a certain mate pair is equal to the opposite assignment within the pair. That is, letting M denote the genotype of the mother and F the genotype of the father, MS is defined as $Pr(M = a, F = b) = Pr(M = b, F = a)$, where a and b are two possible genotypic states.

MA is a problem of an unknown magnitude. This is due in large part to the fact that only a small subset of genetic study designs, such as parental effect studies, will be affected by it. To quantify mating asymmetry, Bourgey et al. introduced Mating asymmetry Statistics

(MaS) [5]. MaS is a measure of the difference between the observed proportion of informative mate-pairs and the expected proportion under a hypothesis of MS. When Bourgey et al. computed MaS on publicly-available trio data from the HapMap project, a strong background level of asymmetry with peaks of high MA scattered throughout the genome was found.

No underlying biological mechanisms of MA are conclusively known, but some have been hypothesized. For example, assortative mating is a form of non-random mating that occurs when individuals with similar phenotypes or genotypes have an increased tendency to mate. Evidence of assortative mating was found in the Framingham Heart Study population, a longitudinal study looking at cardiovascular outcomes in individuals sampled from Framingham, Massachusetts [47]. In particular, individuals from the same ethnic group tended to form couples. Since ethnicity is at least partly related to genetics, this resulted in an excess of mate-pairs with the same genotypes, which is an example of positive assortative mating [47]. Additionally, there are known disassortative mating preferences within humans, such as major histocompatibility complex (MHC) genes: Women are more likely to pick MHC-dissimilar males as partners. Other potential sources of genotype frequency differences between sexes might be sex-specific migration, sexual selection, bottlenecks or sex-directed admixture [5]. These are cryptic population forces that are difficult to ascertain.

In defining the mating-type parameters, The order in which the genotypes are written is crucial. Let the first genotype be the female genotype and the second be the male. When random mating occurs the nine parental genotypic probabilities are given by six mating-type parameters (μ_i):

$$\mu_0 = Pr(M = AA \times F = AA)$$

$$\mu_1 = Pr(M = AA \times F = Aa) = Pr(M = Aa \times F = AA)$$

$$\mu_2 = Pr(M = AA \times F = aa) = Pr(M = aa \times F = AA)$$

$$\mu_3 = Pr(M = Aa \times F = Aa)$$

$$\mu_4 = Pr(M = Aa \times F = aa) = Pr(M = aa \times F = Aa)$$

$$\mu_5 = Pr(M = aa \times F = aa)$$

Under HWE, the genotype frequencies are as follows:

$$p_{AA} = p_A^2$$

$$p_{Aa} = 2p_A p_a$$

$$p_{aa} = p_a^2$$

where p_A is the major allele frequency and $p_a = 1 - p_A$. So under random mating and HWE (which implies mating symmetry):

$$p_A^4 = \mu_0$$

$$2p_A^3 p_a = \mu_1$$

$$p_A^2 p_a^2 = \mu_2$$

$$4p_A^2 p_a^2 = \mu_3$$

$$2p_A p_a^3 = \mu_4$$

$$p_a^4 = \mu_5$$

To relax the assumption of mating symmetry, three asymmetry coefficients ($C_1,\ C_2,\ C_4$) were introduced by Bourgey et al. [5] to describe the imbalance of mate-pair frequencies

within these three groups. Now,

$$Pr(M = AA \times F = Aa) = \mu_1 \cdot C_1$$

$$Pr(M = Aa \times F = AA) = \mu_1 \cdot (2 - C_1)$$

$$Pr(M = AA \times F = aa) = \mu_2 \cdot C_2$$

$$Pr(M = aa \times F = AA) = \mu_2 \cdot (2 - C_2)$$

$$Pr(M = Aa \times F = aa) = \mu_4 \cdot C_4$$

$$Pr(M = aa \times F = Aa) = \mu_4 \cdot (2 - C_4)$$

The asymmetry coefficients are bounded between 0 and 2, with $C_i = 1$ being MS for that pair. Though HWE is not required for this definition, it is preserved by this parameterization as the overall frequency for $p_{AA} \times p_{Aa}$, $p_{AA} \times p_{aa}$ and $p_{Aa} \times p_{aa}$ and their inverses will be as expected. Note however that this does not keep sex-specific genotype frequencies equal. Consider μ_1: if $C_1 > 1$, then the maternal AA, paternal Aa genotype pair will occur more often than the reverse, which will increase the A allele frequency among women as compared to men. If only case parent trios are available, this increased frequency of the A allele among case-mothers could be detected as a maternal effect of the A allele. Table 2.2 gives mate-pair frequencies under MS and MA assumptions.

The measure of asymmetry is based on the scaled differences between the observed and expected mating-type frequencies in Table 2.2:

$$M = \sum_{i \ in \ 1,2,4} \frac{[(C_i \cdot \mu_i) - \mu_i]^2 + [((2 - C_i) \cdot \mu_i) - \mu_i]^2}{\mu_i}$$

$$= 2 \times \sum_{i \ in \ 1,2,4} [\mu_i \cdot (C_i - 1)^2]$$

where the values of μ_i are given by the HW proportions (Table 2.2). The MaS estimator is computed by dividing the M estimator by the maximum theoretical value of M for a given allele frequency.

$$MaS = \frac{M}{M_{max}}$$

MaS is bounded by 0 (complete MS) and 1 (complete MA). The probability models given in Table 2.2 can be used to conduct a likelihood ratio test for mating symmetry with three degrees of freedom (one for each asymmetry coefficient). This likelihood ratio statistic is as follows:

$$LRT = 2[L(C_1 = \hat{C}_1, C_2 = \hat{C}_2; C_4 = \hat{C}_4) - L(C_1 = C_2 = C_4 = 1)]$$

where

$$L(counts) = n_{AA\times AA} log(\hat{\mu}_{AA\times AA}) + n_{AA\times Aa} log(C_1 * \hat{\mu}_{AA\times Aa}) + ... + n_{aa\times aa} log(\hat{\mu}_{aa\times aa})$$

Table 2.2: Mating Type Probabilities under MS and MA model, assuming HWE

M	F	Probability Under		Probability assuming 0.3 MAF		
		Symmetry (MS)	Asymmetry Model (MA)	MS	MA C1 = C2 = 1.05	MA C1 = C2 = 1.5
AA	AA	$p^4 = \mu_0$	μ_0	0.2401	0.2401	0.2401
Aa	AA	$2p^3q = \mu_1$	$\mu_1 C_1$	0.2058	0.2161	0.3087
AA	Aa	$2p^3q = \mu_1$	$\mu_1(2 - C_1)$	0.2058	0.1955	0.1029
aa	AA	$p^2q^2 = \mu_2$	$\mu_2 C_2$	0.0441	0.0463	0.0662
AA	aa	$p^2q^2 = \mu_2$	$\mu_2(2 - C_2)$	0.0441	0.0419	0.0221
Aa	Aa	$4p^2q^2 = \mu_3$	μ_3	0.1764	0.1764	0.1764
aa	Aa	$2pq^3 = \mu_4$	$\mu_4 C_4$	0.0378	0.0397	0.0567
Aa	aa	$2pq^3 = \mu_4$	$\mu_4(2 - C_4)$	0.0378	0.0359	0.0189
aa	aa	$q^4 = \mu_5$	μ_5	0.0081	0.0081	0.0081

In the final three columns of Table 2.2 we see the effect mating asymmetry can have on our multinomial distribution at 2 levels of asymmetry. Column 6 shows $C_1 = C_2 = 1.05$ which is mild asymmetry. The cell probabilities are only changed slightly from the mating symmetry levels (column 5). When MA is more severe as in column 7, the mating asymmetry distorts the cell probabilities from equality between rows 2 and 3 (0.0441) to a 3-fold difference (0.3087 vs. 0.1029). This would suggest that a maternal Aa is more common than a paternal Aa, which would affect any testing of maternal effects.

Chapter 3

Statistical Methods for Trio Data

Two statistical techniques were used for testing whether there are maternal gene-environment effects: multinomial modelling (described in [2] and implemented in EMIM) and log-linear regression (described in [64] [65] [66] and implemented in LEM). These two methods produce equivalent estimates but make different assumptions and use different maximization techniques. This equivalence is explored in Section 3.4 after the two methods are introduced.

3.1 Count Data as Multinomial Probabilities

First, we consider only the genotypic information on case-parent trios. For each diallelic SNP, there are 15 possible patterns of mother, father and child's genotypes (as shown in Table 2.1). Letting the minor allele be coded as a lower case a, the genotype of each individual is either 0 (AA), 1 (Aa) or 2 (aa) for the number of minor alleles they carry.

Since the data were collected from a fixed number of families with the recruitment criterion being that the child is a case, Bayes rule is used to model the probability of a (M,F,C) genotypic category conditional on disease status as follows:

$$\begin{aligned} P(M,F,C|D) &= \frac{P(M,F,C,D)}{P(D)} \\ &= \frac{P(D|M,F,C) \times P(M,F,C)}{P(D)} \\ &= \frac{P(D|M,F,C) \times P(C|M,F) \times P(M,F)}{P(D)} \end{aligned} \tag{3.1.1}$$

Notice the denominator is $P(D)$, which is the disease prevalence. This is computed by summing over all the possible numerator values, or

$$\begin{aligned} P(D) &= \sum_{M,F,C} P(D|M,F,C) \times P(M,F,C) \\ &= \sum_{M,F,C} P(D|M,F,C) \times P(C|M,F) \times P(M,F) \\ &= \sum_{M,F} P(M,F) \times \sum_{C} P(D|M,F,C) \times P(C|M,F) \end{aligned}$$

$P(D|M,F,C)$ is our disease model conditional on genotype. This term will be discussed on page 26. The probability of a child's genotype given the parental genotypes ($P(C|M,F)$) was described in the Genetic Background on Mendelian Transmission (see Table 2.1). The parental genotype probabilities, $P(M,F)$, are called the mate-pair probabilities. In Table 3.1, three different models corresponding to different mate-pair frequency assumptions are presented. The second column shows the mating-type probabilities under the strongest assumption of HWE and random mating; here, only one mating-type parameter (MTP), the allele frequency p, is estimated. This is the strongest assumption. Under mating symmetry (column 3), there are 6 mating-type parameters. This is a weaker assumption as it does not require HWE. Finally, under a model making no assumptions about HW or mating symmetry, there are 9 MTPs (column 4). This final model makes no assumptions on mating forces; note, however, that these parameters cannot be estimated from the data if control trios are not available due to overparameterization.

Table 3.1: Mate-pair probability models ($P(M, F)$)

Mating Type (M, F)	HWE	Mating Symmetry	Mating Asymmetry
AA × AA	$(1-p)^4$	μ_0	μ_0
AA × Aa	$2p(1-p)^3$	μ_1	μ_1
Aa × AA	$2p(1-p)^3$	μ_1	μ_2
aa × AA	$p^2(1-p)^2$	μ_2	μ_3
AA × aa	$p^2(1-p)^2$	μ_2	μ_4
Aa × Aa	$4p^2(1-p)^2$	μ_3	μ_5
Aa × aa	$2p^3(1-p)$	μ_4	μ_6
aa × AA	$2p^3(1-p)$	μ_4	μ_7
aa × aa	p^4	μ_5	μ_8

In order to model $P(D|C, M, F)$, a suitable choice of parameters must be made. There are many types of genetic effects that can be fit but we will focus solely on maternal, child and environmental effects. Table 3.2 introduces the model parameters used in our modelling procedures. Here, let α be the baseline probability of disease for a child who is part of a trio composed of homozygotes for the non-risk allele (i.e. $AA \times AA \times AA$), R_i be the relative risk for the child carrying 1 or 2 risk alleles (where i equals 1 or 2), and S_i be the relative risk for the mother carrying 1 or 2 risk alleles. Table 3.3 gives the model $P(D|C, M, F)$ as a function of the parameters α, R_1, R_2, S_2, and S_2; this parameterization was initially described in [65]. If we assume that only maternal effects are present, then this can be simplified to $P(D|M, F, C) = P(D|M)$. Similarly, if we assume only child effects then $P(D|M, F, C) = P(D|C)$. These simplifications have the implicit assumption of independence between the effect of child genotype and maternal genotype. That is, $R_i \perp\!\!\!\perp S_i$ once inheritance has been accounted for in $P(C|M, F)$. Recall that a maternal genetic effect would be one where risk is increased or decreased when carried by the mother but no additional risk exists when the allele is transmitted to the offspring. This means that one would expect a higher frequency of risk alleles among the mothers of cases as compared to

the fathers of cases [66].

Table 3.2: Model parameters for multinomial and loglinear models

Parameter	Description
α	Baseline Probability of Disease
R_1	Child has one minor allele (child genotype effect)
R_2	Child has two minor alleles (child genotype effect)
S_1	Mother has one minor allele (mother genotype effect)
S_2	Mother has two minor alleles (mother genotype effect)
E	Environmental exposure present (environment effect)

Table 3.3: Probability that child has disease given case-parent trio genotypes and assuming maternal and child genetic effects only (no interaction effects)

Maternal Genotype	Paternal Genotype	Child Genotype	$P(D\|M,F,C)$
aa	aa	aa	$\alpha S_2 R_2$
aa	aa	Aa	$\alpha S_2 R_2$
aa	Aa	Aa	$\alpha S_2 R_1$
Aa	aa	aa	$\alpha S_1 R_2$
Aa	aa	Aa	$\alpha S_1 R_1$
aa	AA	Aa	$\alpha S_2 R_1$
AA	aa	Aa	αR_1
Aa	Aa	aa	$\alpha S_1 R_2$
Aa	Aa	Aa	$\alpha S_1 R_1$
Aa	Aa	AA	αS_1
Aa	AA	Aa	$\alpha S_1 R_1$
Aa	AA	AA	αS_1
AA	Aa	Aa	αR_1
AA	Aa	AA	α
AA	AA	AA	α

A model which includes the binary environmental variable E will have the same $P(C|M,F)$ probabilities, the relative risk parameters would include an E term for those with the exposure, and there would be twice as many mating-type parameters (MTP) in our mate-pair probability models, so for the mating symmetry model there are 12 MTPs and for the mating asymmetry model there are 18. Unfortunately, as will be discussed in subsequent sections, one of the implementations of the method (EMIM) being investigated cannot be extended to include an environmental variable so instead a method based on stratifying by the level of the environmental factor and testing for heterogeneity is used.

The two methods being compared use slightly different approaches for estimating the model parameters and testing the various genetic effects given observed counts of the 15 (M,F,C) genotype categories. In Table 3.2, the possible genetic effects of interest and their corresponding parameter in the multinomial model are listed. With Tables 2.1, 3.1 and 3.3, the multinomial probabilities are fully defined for case trios. The full table corresponding to the multinomial model for case and control trios can be found in the next chapter in Table 4.1. In the next two sections, we describe the two different approaches to estimate the model parameters used by the Ainsworth and the Weinberg groups, respectively.

3.2 Estimating Model Parameters by Direct Maximization of the Likelihood

The method described here was first described in Ainsworth et al. (2011) [2] and implemented in the EMIM program [22].

With the 15-category multinomially distributed random variable defined above, the Ainsworth approach directly maximizes the corresponding multinomial likelihood defined

as:

$$\prod_{i=1}^{15}\{P(M_i, F_i, C_i|child\ diseased)\}^{n_i}$$

where n_i are the observed genotype counts for the 15 possible trio configurations.

The authors note that this method is computationally intensive. This is due to the heterozygous trio category (ie, $(M, F, C) = (Aa, Aa, Aa)$). In this category, it is impossible to distinguish from which parent the child inherited the risk allele. This piece of missing information is dealt with by considering that cell probability as the sum of two cells: one where the risk allele was inherited from the mother and the other from the father. With this consideration, the cell probabilities are no longer purely products of our risk parameters but include cells which contain sums of the risk parameters. These are not straightforward to fit in standard statistical software since the MLEs cannot be computed analytically and so numerical methods are needed [2]. This problem is exacerbated when incomplete trios are included in analysis since that is additional missing data. A common approach is to use the EM algorithm to maximize the likelihood; this is the approach used for the log-linear model of Weinberg et al. Instead, the EMIM implementation directly maximizes the multinomial likelihood.

Direct numerical maximization methods do not require calculation of derivatives. Instead, a direct search algorithm will examine the points surrounding the current point, seeking the "best" value. Here, since the goal is maximizing a likelihood, the algorithm searches for the largest likelihood in the region surrounding the current parameter values. EMIM uses the MAXFUN subroutine to solve this problem [46]. Unsurprisingly, this type of algorithm may take some time to converge and will only be an approximation to the true parameter value. Comparisons between direct search and other maximization techniques, including multinomial methods, show similar results [32].

It is important to note that this method is not robust to population stratification when an assumption of HWE and random mating is made. This is evident as population stratification is a violation of the assumptions required for HWE. Ainsworth et al also found that using external control samples to estimate mating-type parameters failed, likely due to these frequencies being different between populations. Therefore methods to account for population stratification need to be used. In the candidate gene study performed in Chapter 5, only trios of European ancestry are included to reduce confounding by population stratification.

3.2.1 Software Implementation: EMIM

Overview

EMIM (Estimation of Maternal, Imprinting and interaction effects using Multinomial modelling) is software created by Howey and Cordell [22] that can be used to test for and estimate allele relative risk parameters through maximum likelihood estimation of up to seven of the meaningful estimable parameters in Table 3.2 as well as mating-type parameters. This program runs on Windows and Linux through the command line.

EMIM uses standard format pedigree files with parent-offspring trios (both cases and controls) and any subsets of this type. As expected, incomplete trios contribute far less information to the study than complete. EMIM has a companion program PREMIM that takes these standard format pedigrees and converts them to EMIM input files (one for each provided type of trio subset) with each line representing a SNP and counts of the maximum 15 possible genotype combinations. PREMIM assumes the risk allele to be the minor allele, however this can be changed by the user. PREMIM will also estimate the allele frequency at each marker, a necessary input for EMIM as well. These allele frequencies can be provided separately from PREMIM through estimates from HapMap or prior genetic studies.

With these files, EMIM reads in a parameter file that outlines which input file types need to be included, how many SNPs are in each file, as well as which parameters should be estimated. These include all the parameters in Table 3.2 as well as any restrictions on the parameters, such as $R_2 = R_1$ (forcing a recessive effect) or $R_2 = R_1^2$ (forcing a multiplicative effect). These restrictions will reduce the number of estimated parameters, allowing for more parameters to be included in the model or to increase the degrees of freedom. It is in this parameter file where model assumptions regarding mating are imposed. In decreasing order of restriction, there are options for:

1. HWE and random mating (one allele frequency parameter estimated)
2. Conditional on exchangeable parental genotypes with parental allelic exchangeability (five mating-type parameters estimated with the restriction of $\mu_4 = \mu_3$)
3. Conditional on exchangeable parental genotypes (six mating-type parameters estimated)
4. Conditional on parental genotypes (nine mating-type parameters estimated)

It is known that increasing the number of estimated parameters will decrease the power for a test; however, the least restricted model is the only method that properly accounts for mating asymmetry. The first, third and fourth methods are compared later in this work under varying levels of mating asymmetry and number of control trios.

EMIM will fit the specified model and return effect estimates on the specified parameters, along with standard errors and log likelihoods. The log likelihoods can be used to test the significance of maternal effects and to detect a maternal gene-environment interaction, as described in the following two subsections.

Maternal Effect Allowing for Child Effect

To ensure a true maternal genetic effect is detected, as opposed to a strong child effect being detected as a weak maternal effect, a likelihood ratio test is used to compare the fit of two nested models: one with only child effects and the second with both child and maternal effects. Specifically, this test has the following hypotheses: $H_0 : S_1 = S_2 = 1$ vs $H_1 : at\ least\ one\ of\ S_1\ S_2 \neq 1$. This test may have one or two degrees of freedom, depending on whether the model parameterization assumes that $S_2 = S_1^2$. This test is important to support any claim of true maternal genetic effect. See Figure 3.1 for a description of the nested modeling procedure. On the Exposed side of the diagram, this test is represented by the one degree of freedom LRT oval.

Maternal Gene-Environment Interaction Test

The EMIM implementation in its current state cannot take exposure information as input and therefore does not conduct the desired maternal gene-environment test. Note that the model could be extended to include such terms (as is possible with LEM), which would be a worthwhile addition to this software. Instead, a method described in Tai et al [55] that is borrowed from meta analysis can be used to conduct an interaction test. In this approach, the model is fit separately in the exposed and unexposed categories. Cochran's Q test is a test of heterogeneity of effect estimates between the two strata of environmental exposure. It can be conducted using the R package rmeta [31]. Cochran's Q test summarizes variation around a weighted mean $\bar{x}_w$. In the meta analysis setting, the test is used to determine whether there is a difference in effect estimates from different studies. Each study provides an effect estimate, x_i, and standard error s_i. The weight for study i is the reciprocal of the estimated variance $w_i = 1/s_i^2$. The weighted mean is calculated as:

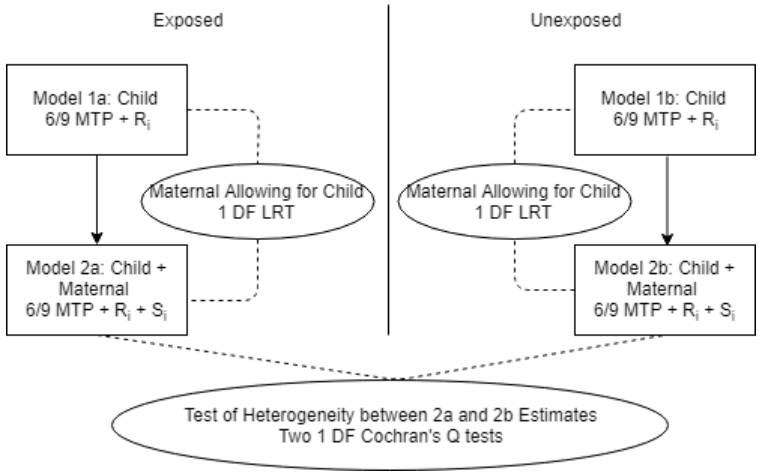
$$\bar{x}_w = \frac{\sum_{i=1}^{k} w_i x_i}{\sum_{i=1}^{k} w_i}$$

Cochran's Q statistic is:

$$Q = \sum_{i=1}^{k} w_i (x_i - \bar{x}_w)^2$$

Under the null hypothesis, it is chi-square distributed with k-1 degrees of freedom. Low p-values would suggest heterogeneity of estimates. In Figure 3.1, the interaction test is shown at the bottom: Maternal effect estimates from models 2a and 2b are compared using the test of heterogeneity. In the case of the current study where maternal genetic effect estimates and standard errors are provided by EMIM, a maternal gene-environment interaction is suggested if the null hypothesis is rejected. Unfortunately, this is the only information provided by this test; no estimate for this interaction effect is given. This is an obvious disadvantage of this method.

Figure 3.1: Nested Models for Hypothesis Testing in EMIM



MTP: Mating-type parameter; DF: Degree of freedom; LRT: Likelihood ratio test

3.3 Log-Linear Regression

The second method we examine is based on the log-linear modelling described by Weinberg et al. in a series of papers [64] [65] [66].

Log-linear models are categorized as generalized linear models (GLM). A GLM is defined by three components: a random component which identifies the response variable Y and its probability distribution; a systematic component $\eta_i = \sum \beta_j x_{ij}$, that defines the relationship between the linear predictor η_i and the explanatory variables x_{ij}; and a link function, $g(E[Y]) = \eta_i$, to specify the relationship between $E(Y)$ and the systematic component.

In log-linear regression, the random component Y is assumed to be Poisson distributed. The density of the Poisson distribution is:

$$f(y;\mu) = \frac{e^{-\mu}\mu^y}{y!}$$

To go from the multinomial model defined above to Poisson distributed counts, we need to show the relationship between the multinomial and Poisson distributions. When sampling from c independent Poisson random variables, the total n_i per group ($i = 1, ..., c$) is random. For c independent Poisson variables with $E(Y_i) = \mu_i$, the conditional distribution of the counts $Y_1, ..., Y_c$ given $\sum Y_i = n$ has distribution:

$$\begin{aligned} P(Y_1 = n_1, Y_2 = n_2, ..., Y_c = n_c | \sum Y_i = n) &= \frac{P(Y_1 = n_1, Y_2 = n_2, ..., Y_c = n_c)}{P(\sum Y_i = n)} \\ &= \frac{\prod_i [exp(-\mu_i)\mu_i^{n_i}/n_i!]}{exp(-\sum \mu_j)(\sum \mu_j)^n/n!} \\ &= \frac{n_i!}{\prod_i n_i!} \prod_i \pi_i^{n_i} \end{aligned}$$

where $\pi_i = \mu_i/(\sum_j \mu_j)$. The last line is the definition of the multinomial distribution with sample size n and cell probabilities π_i [1]. In our case, $E(Y) = E[n_{M,F,C}|D]$ is the expected

count for a particular cell.

Recall that the Poisson distribution is a member of the exponential family so a GLM can be fit. The natural link function for Poisson count data is the natural log function. A feature of this link is that it guarantees that the estimated parameter obeys the positivity constraint of the Poisson rate (a rate parameter cannot be negative). This leaves as the final model being fitted:

$$\begin{aligned} ln(E(n_{M,F,C}|D)) = \mu_i + ln(R_1)I_{[C=1]} + ln(R_2)I_{[C=2]} + \\ ln(S_1)I_{[M=1]} + ln(S_2)I_{[M=2]} + \\ ln(E)I_{[e=1]} + ln(E_{int})I_{[e=1]}I_{[M>0]} + ln(2)I_{[M=1,F=1,C=1]} \end{aligned}$$

where I is the indicator function. This comes directly from taking the log of the values in the final column of Table 3.3 for each cell. Notice the final term is $ln(2)I_{[M=1,F=1,C=1]}$. This is an offset term to account for the heterozyous trio ($Aa \times Aa \times Aa$) having twice the expected frequency as the other cells. Additionally, in the implementation of the model, we need to provide information about the structural zeros caused by the constraints of inheritance.

3.3.1 Software Implementation: LEM

LEM, which stands for 'log-linear and event history analysis with missing data using the EM algorithm', is a very general program by Vermunt [61] that is used for the analysis of categorical data. LEM was written in 1997 for the Windows OS, and it currently only runs on Windows 1998. It can be used to obtain parameter estimates from many types of models and analyses in addition to log-linear models.

LEM requires an input file defining the model to be fit. In this file, one defines the types

of variables to be expected in the dataset, how to handle missing data, the parameters to be fit using a dummy design matrix, and a weight vector (See Appendix A for an example input file); an example of a mother and child effect model assuming MS is shown in Figure A.1. The input begins with a description of the type of variables in the dataset. Included in this section: the type of variable (manifest, latent, continuous, etc), the dimension of each variable, and the name of each variable.

For this work the model by Weinberg, Wilcox and Lie [66] was implemented and extended to include the maternal gene-environment interaction. As was done in Weinberg and Umbach [64], we choose to switch the optimization from EM to Newton-Raphson after 10 iterations.

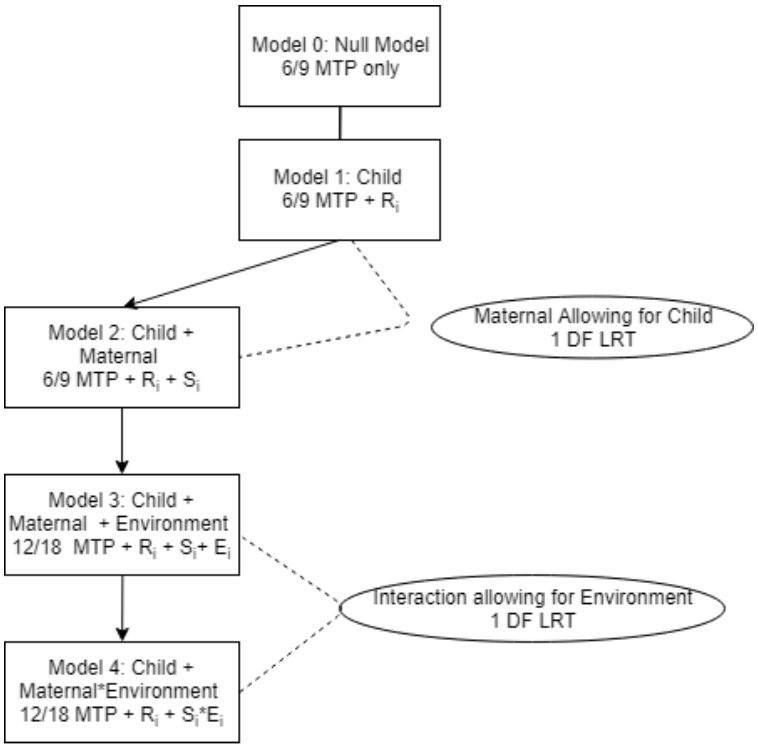
Model Fitting

To test for significant maternal effects above and beyond any overrepresentation of minor alleles that would be present should there be a child effect, we conduct a similar one degree of freedom LRT as with EMIM that compares the fit of the child effects only model to the one that also has maternal effects (Models 1 and 2 in Figure 3.2). If significant maternal effects are found after that test, models 3 and 4 (Figure 3.2) are fit and compared using a similar LRT to determine whether a maternal gene-environment interaction effect is present.

When including a gene environment interaction in the model, it is necessary to include two sets of mating-type parameters (MTP) (thus the difference in numbers of MTP between models 2 and 3). This is similar to what was done with the models being fit in EMIM: Models are fit twice based on exposure status and therefore two sets of MTP are considered. This prevents a dependence between mating-type frequencies and exposure from obfuscating true associations.

Models 3 and 4 are compared similarly to models 1 and 2 using a 1 degree of freedom LRT (2 if S_2 is not restricted). This test will determine whether the interaction is present above and beyond any environmental or maternal main effect.

Figure 3.2: Nested Models for Hypothesis Testing in LEM



MTP: Mating-type parameter; DF: Degree of freedom; LRT: Likelihood ratio test

Chapter 4

Mating Asymmetry and Maternal Effects Testing: A Simulation Study

Recall from Section 2.7 that mating symmetry (MS) is the probability that a certain mother-father genotype assignment is equal to that of the reverse father-mother assignment. When this is violated, it is called mating asymmetry (MA) and may lead to spurious results. For example, an overrepresentation of maternal minor alleles could be due to asymmetry rather than increased risk of disease. This confounding can be accounted for with control trio data where the mating parameters can be estimated using these data and then compared to the case trios (similar to how case-control studies work). In this simulation study we aim to find how power, type I error and estimation bias are affected by varying numbers of control trios. Since genetic data collection is costly, it is useful to know how much data is enough to control this problem.

Methods to control for MA can be included in existing maternal effect methods through the inclusion of three additional mating-type parameters. However, in order to estimate these additional mating-type parameters, control trios must have also been collected. Otherwise these nuisance parameters are confounded with the risk parameters the model is designed to

detect. The existing methods implemented in EMIM and LEM both can accomodate control trios and subsets thereof. The purpose of this chapter is two-fold: first, what effect does the MA violation have on the two methods we are comparing when we erroneously assume MS, and secondly, if we have control trios and analyse the data without the assumption of MS, how well do the two methods perform in terms of their power and bias.

4.1 Data Simulation Strategy

Consider a diallelic marker in trios of mother, father and child (M, F, C) where the child is either a case or a control. Baseline disease penetrance α remains constant throughout the simulations at $\alpha = 0.2$. Relative risk parameters R and S denoting child and mother risks respectively, were set to $R_1 = S_1 = 1.3$ for the risk for a single risk allele, and $R_2 = R_1^2$ and $S_2 = S_1^2$. The maternal gene-environment risk parameter E was set to 1.5. The minor allele frequency $q = 1 - p$ was set at 0.3. To simulate MA, the model from Bourgey et al. 2011 was used [5]. This model incorporates three asymmetry parameters, C_1, C_2 and C_4, to depart from symmetry while maintaining Hardy-Weinberg proportions. Each of the C_i takes a value between 0 and 2, where the extremes are complete asymmetry and 1 is symmetry. These apply to the three informative mating types. Mating-type frequencies are shown in Table 2.2 with the assumption of Mendelian transmission and HWE. We simulated 1000 datasets consisting of 650 case trios for each of the varying number of control trios and 6 levels of asymmetry, covering a range of small to medium asymmetry levels according to the range of the Bourgey et al. mating asymmetry statistic: $C1 = C2 = (0.5, 0.9, 0.95, 1.05, 1.1, 1.5)$.

Table 4.1 gives the probabilities used to simulate the case and control trios. These probabilities were derived as in Ainsworth et al. (2011) with the mating asymmetry parameterization described in Chapter 3. An additional risk parameter E is included corresponding to the environment interaction. Column 8 gives the multinomial distribution from which

the case trios are sampled and column 10 is the distribution for the control trios. In both columns the denominator is $P(D)$, or the probability of being a case or a control. This probability is computed by summing over $P(D|M, F, C)$.

The R code used to sample from the multinomial is included in Appendix A.

4.2 Methods

We are interested in evaluating, for each method, power, type I error and relative bias of estimation of the effects. For both LEM and EMIM methods, we consider a Conditional on Parental Genotype (CPG) model and a Conditional on Exchangeable Parental Genotype (CEPG) model. CEPG fits 6 MTP and CPG fits 9 MTP. For EMIM, the HWE + random mating model is also considered.

Power and type I error are estimated by the fraction of replicates where the test accurately found a simulated genetic effect (Power), and the test inadvertently found an effect where none is present (Type I error). For EMIM, the main maternal effect tests are conducted using a 1-df LRT and the interaction tests are conducted using the test of heterogeneity described in Section 3.2. LEM uses a LRT for all tests.

Relative bias for the maternal effect estimate is calculated as:

$$B = \frac{S_{obs} - S_{sim}}{S_{sim}} \times 100$$

Since the method of maternal gene-environment interaction testing for EMIM does not produce effect estimates, we cannot estimate the bias.

Table 4.1: Multinomial Probabilities for Case and Control Trios

M	F	C	$P(C\|M,F)$	$P(M,F)$	$P(D=1\|M,F,C)$	$P(M,F,C,D=1)$	$P(M,F,C\|D=1)$	$P(D=0,M,F,C)$	$P(M,F,C\|D=0)$
aa	aa	aa	1	p_a^4	$\alpha S_2 R_2$	$\alpha S_2 R_2 p_a^4$	$\frac{\alpha S_2 R_2 p_a^4}{P(D=1)}$	$(1-\alpha S_2 R_2)p_a^4$	$\frac{(1-\alpha S_2 R_2)p_a^4}{P(D=0)}$
aa	Aa	aa	0.5	$C_4 2 p_A p_a^3$	$\alpha S_2 R_2$	$\alpha S_2 R_2 C_4 p_A p_a^3$	$\frac{\alpha S_2 R_2 C_4 p_A p_a^3}{P(D=1)}$	$(1-\alpha S_2 R_2)C_4 p_A p_a^3$	$\frac{(1-\alpha S_2 R_2)C_4 p_A p_a^3}{P(D=0)}$
aa	Aa	Aa	0.5	$C_4 2 p_A p_a^3$	$\alpha S_2 R_1$	$\alpha S_2 R_1 C_4 p_A p_a^3$	$\frac{\alpha S_2 R_1 C_4 p_A p_a^3}{P(D=1)}$	$(1-\alpha S_2 R_1)C_4 p_A p_a^3$	$\frac{(1-\alpha S_2 R_1)C_4 p_A p_a^3}{P(D=0)}$
Aa	aa	aa	0.5	$(2-C_4)2 p_A p_a^3$	$\alpha S_1 R_2$	$\alpha S_1 R_2 (2-C_4) p_A p_a^3$	$\frac{\alpha S_1 R_2 (2-C_4) p_A p_a^3}{P(D=1)}$	$(1-\alpha S_1 R_2)(2-C_4) p_A p_a^3$	$\frac{(1-\alpha S_1 R_2)(2-C_4) p_A p_a^3}{P(D=0)}$
Aa	aa	Aa	0.5	$(2-C_4)2 p_A p_a^3$	$\alpha S_1 R_1$	$\alpha S_1 R_1 (2-C_4) p_A p_a^3$	$\frac{\alpha S_1 R_1 (2-C_4) p_A p_a^3}{P(D=1)}$	$(1-\alpha S_1 R_1)(2-C_4) p_A p_a^3$	$\frac{(1-\alpha S_1 R_1)(2-C_4) p_A p_a^3}{P(D=0)}$
aa	AA	Aa	1	$C_2 p_A^2 p_a^2$	$\alpha S_2 R_1$	$\alpha S_2 R_1 C_2 p_A^2 p_a^2$	$\frac{\alpha S_2 R_1 C_2 p_A^2 p_a^2}{P(D=1)}$	$(1-\alpha S_2 R_1)C_2 p_A^2 p_a^2$	$\frac{(1-\alpha S_2 R_1)C_2 p_A^2 p_a^2}{P(D=0)}$
AA	aa	Aa	1	$(2-C_2)p_A^2 p_a^2$	αR_1	$\alpha R_1 (2-C_2) p_A^2 p_a^2$	$\frac{\alpha R_1 (2-C_2) p_A^2 p_a^2}{P(D=1)}$	$(1-\alpha R_1)(2-C_2) p_A^2 p_a^2$	$\frac{(1-\alpha R_1)(2-C_2) p_A^2 p_a^2}{P(D=0)}$
Aa	Aa	aa	0.25	$4 p_A^2 p_a^2$	$\alpha S_1 R_2$	$\alpha S_1 R_2 p_A^2 p_a^2$	$\frac{\alpha S_1 R_2 p_A^2 p_a^2}{P(D=1)}$	$(1-\alpha S_1 R_2) p_A^2 p_a^2$	$\frac{(1-\alpha S_1 R_2) p_A^2 p_a^2}{P(D=0)}$
Aa	Aa	Aa	0.5	$4 p_A^2 p_a^2$	$\alpha S_1 R_1$	$\alpha S_1 R_1 2 p_A^2 p_a^2$	$\frac{\alpha S_1 R_1 2 p_A^2 p_a^2}{P(D=1)}$	$(1-\alpha S_1 R_1) 2 p_A^2 p_a^2$	$\frac{(1-\alpha S_1 R_1) 2 p_A^2 p_a^2}{P(D=0)}$
Aa	Aa	AA	0.25	$4 p_A^2 p_a^2$	αS_1	$\alpha S_1 p_A^2 p_a^2$	$\frac{\alpha S_1 p_A^2 p_a^2}{P(D=1)}$	$(1-\alpha S_1) p_A^2 p_a^2$	$\frac{(1-\alpha S_1) p_A^2 p_a^2}{P(D=0)}$
Aa	AA	Aa	0.5	$C_1 2 p_A^3 p_a$	$\alpha S_1 R_1$	$\alpha S_1 R_1 C_1 p_A^3 p_a$	$\frac{\alpha S_1 R_1 C_1 p_A^3 p_a}{P(D=1)}$	$(1-\alpha S_1 R_1) C_1 p_A^3 p_a$	$\frac{(1-\alpha S_1 R_1) C_1 p_A^3 p_a}{P(D=0)}$
Aa	AA	AA	0.5	$C_1 2 p_A^3 p_a$	αS_1	$\alpha S_1 C_1 p_A^3 p_a$	$\frac{\alpha S_1 C_1 p_A^3 p_a}{P(D=1)}$	$(1-\alpha S_1) C_1 p_A^3 p_a$	$\frac{(1-\alpha S_1) C_1 p_A^3 p_a}{P(D=0)}$
AA	Aa	Aa	0.5	$(2-C_1)2 p_A^3 p_a$	αR_1	$\alpha R_1 (2-C_1) p_A^3 p_a$	$\frac{\alpha R_1 (2-C_1) p_A^3 p_a}{P(D=1)}$	$(1-\alpha R_1)(2-C_1) p_A^3 p_a$	$\frac{(1-\alpha R_1)(2-C_1) p_A^3 p_a}{P(D=0)}$
AA	Aa	AA	0.5	$(2-C_1)2 p_A^3 p_a$	α	$\alpha (2-C_1) p_A^3 p_a$	$\frac{\alpha (2-C_1) p_A^3 p_a}{P(D=1)}$	$(1-\alpha)(2-C_1) p_A^3 p_a$	$\frac{(1-\alpha)(2-C_1) p_A^3 p_a}{P(D=0)}$
AA	AA	AA	1	p_A^4	α	αp_A^4	$\frac{\alpha p_A^4}{P(D=1)}$	$(1-\alpha) p_A^4$	$\frac{(1-\alpha) p_A^4}{P(D=0)}$

4.3 Results

In this section, we summarize the results of the simulation studies in terms of power, type I error and bias. When we refer to the model, we are referring to the analysis model and not the model used to simulate the data.

Before showing results for the simulation under MA, the methods were verified under mating symmetry to ensure they performed as expected. Using EMIM, we can see in Tables 4.2 and 4.3 that the model behaves as expected. Power is low for detecting the maternal gene-environment interaction but type I error is contained both with and without controls.

Table 4.2: EMIM Main Maternal Effect Performance with No Asymmetry

# Control Trios	Maternal Estimate	Type I Error	Power
0	1.32	0.046	0.580
650	1.38	0.043	0.893

Table 4.3: EMIM Interaction Effect Performance with No Asymmetry

# Control Trios	Type I Error	Power
0	0.048	0.371
650	0.039	0.610

LEM also produces results as expected (see Tables 4.4 and 4.5), with type I error being slightly higher than the desired 0.05 but still low.

Table 4.4: LEM Main Maternal Effect Performance with No Asymmetry

# Control Trios	Maternal Estimate	Type I Error	Power
0	1.481	0.046	0.997
650	1.736	0.036	1.000

Table 4.5: LEM Interaction Effect Performance with No Asymmetry

# Control Trios	Interaction Estimate	Type I Error	Power
0	1.635	0.079	0.385
650	1.803	0.076	0.932

4.3.1 EMIM Results

Detection of Maternal Genetic Main Effect

Type I Error In both the HWE+RM (fig. 4.1) and CEPG models (fig. 4.2), type I error is extremely inflated. When minimal MA is present, type I error is still around 0.15-0.2. With extreme MA, type I error increases to almost 1. Therefore both of these models are not reliable; when the data is erroneously analysed assuming MA, the probability of falsely concluding a main maternal effect is much higher than the desired value of 0.05.

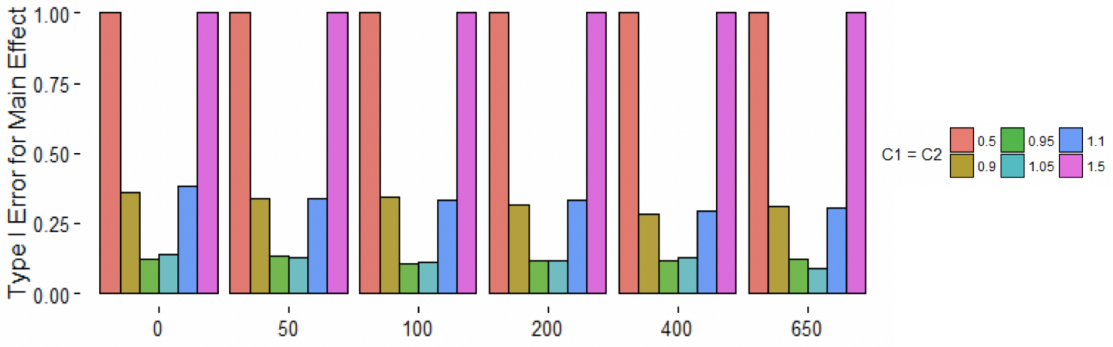


Figure 4.1: Type I error results using EMIM when data is analysed assuming HWE and Random Mating

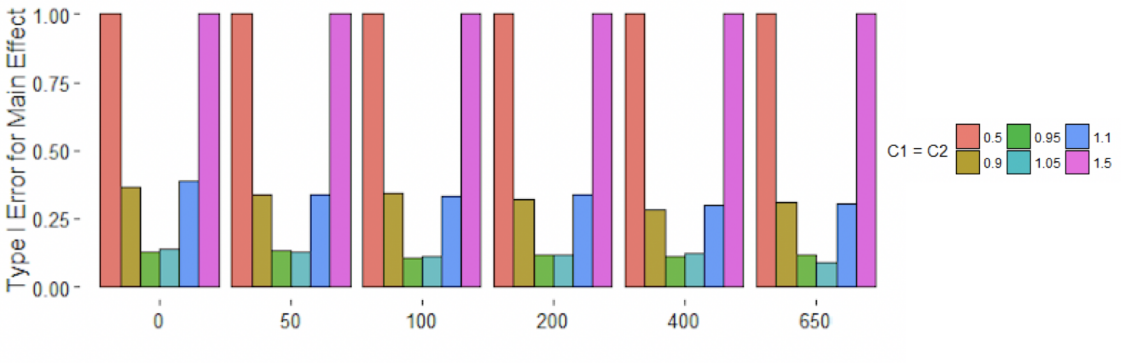


Figure 4.2: Type I error results using EMIM when data is analysed assuming CEPG

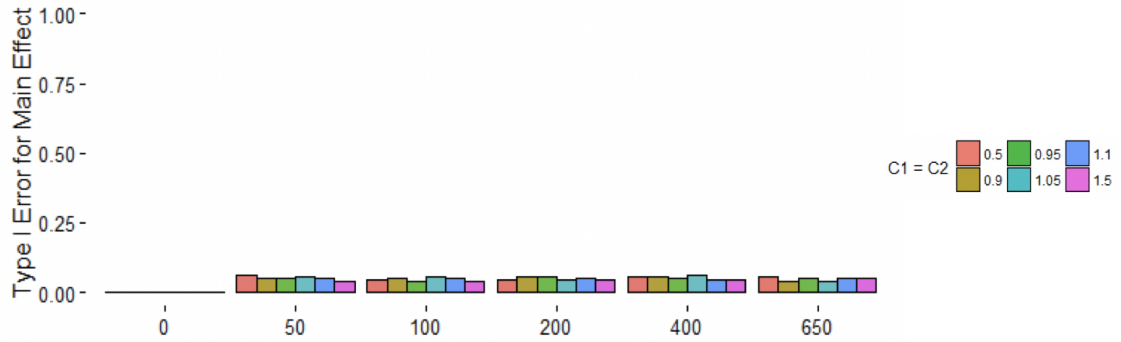


Figure 4.3: Type I error results using EMIM when data is analysed assuming CPG

On the other hand, with CPG (fig. 4.3), the estimated type I error rate is close to 0.05.

Power Figure 4.4 shows the power to detect a main maternal effect allowing for a child effect under the model with assumed random mating and HWE (HWE+RM). As expected, with increased numbers of control trios the power increases. The two extreme values of MA ($C1 = C2 = 0.5$ and $C1 = C2 = 1.5$) have recorded 100% power, which is unrealistic for any likelihood ratio test. The middle values of MA ($C1 = C2 = 0.9$, $C1 = C2 = 0.95$, $C1 = C2 = 1.05$ and $C1 = C2 = 1.1$) have power increasing in the direction of simulated asymmetry as a higher C1 and C2 will increase the number of minor alleles. In other words, this is confirming that increasing asymmetry is confounded with the maternal effect.

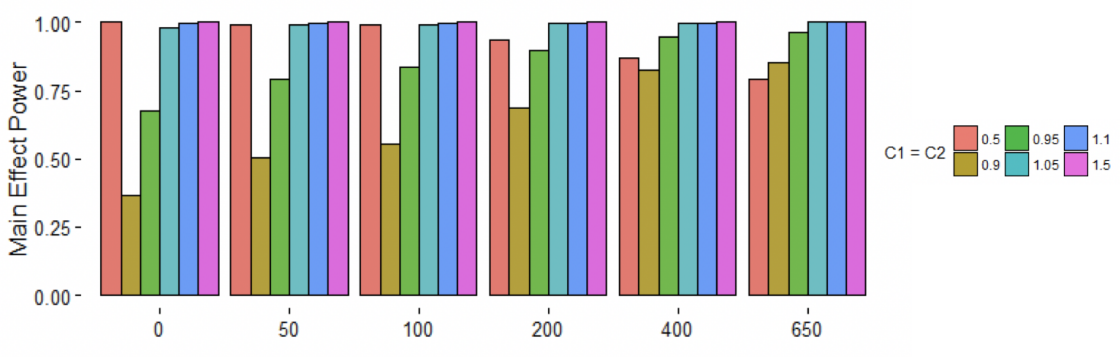


Figure 4.4: Estimated power using EMIM with analysis model that assumes HWE and Random Mating

A similar pattern is seen with the CEPG model (see Figure 4.5). This shows that estimating the extra mating-type parameters does not significantly reduce power.

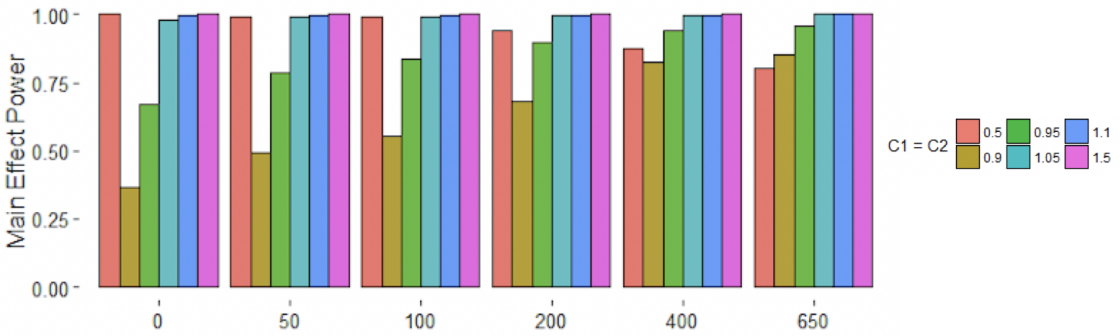


Figure 4.5: Estimated power using EMIM with analysis model that assumes CEPG

Figure 4.6 shows power results for the CPG model, the least restricted model which controls for mating asymmetry by estimating additional mating-type parameters. Note that this method requires control parents to estimate these extra parameters so there is no power with 0 control trios. As desired, the level of mating asymmetry is no longer affecting the power; instead power increases only with additional controls. This is shown by the constant bar height within each level of control trio numbers. When the number of control trios matches the number of case trios power is estimated to be nearly 1.

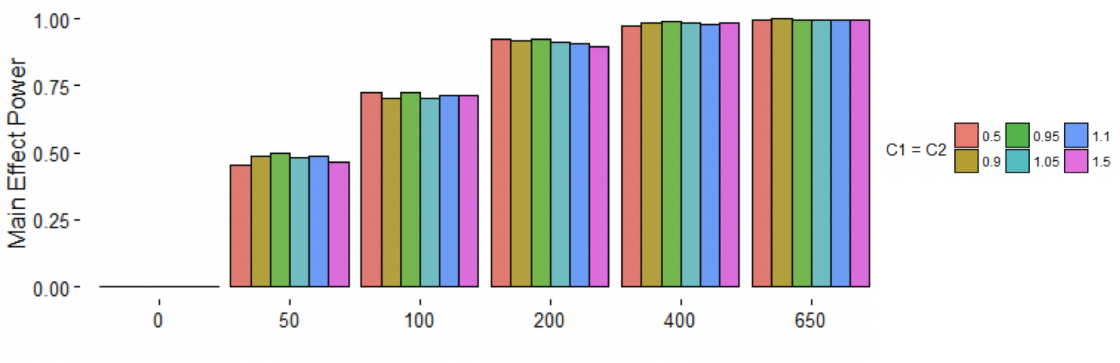


Figure 4.6: Estimated power using EMIM with analysis model that assumes CPG

Relative Bias In the first two figures for HWE+RM (4.7) and CEPG (4.8), relative bias follows with the direction of simulated asymmetry, where overestimation occurs when the asymmetry model creates an over-representation of minor alleles.

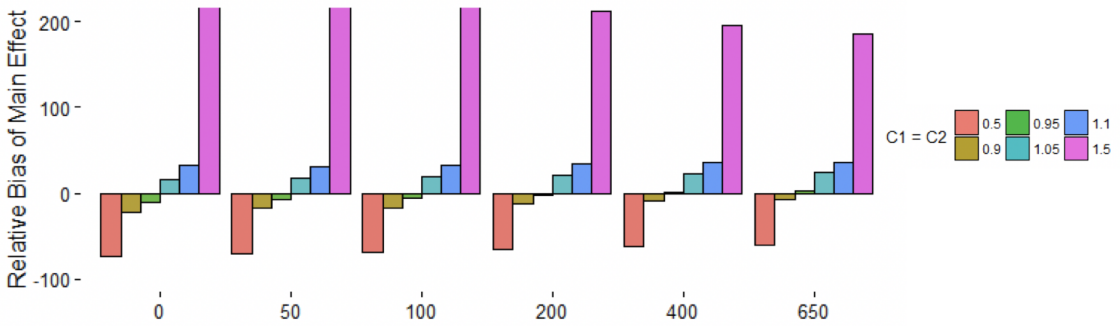


Figure 4.7: Relative bias results using EMIM when data is analysed assuming HWE and Random Mating

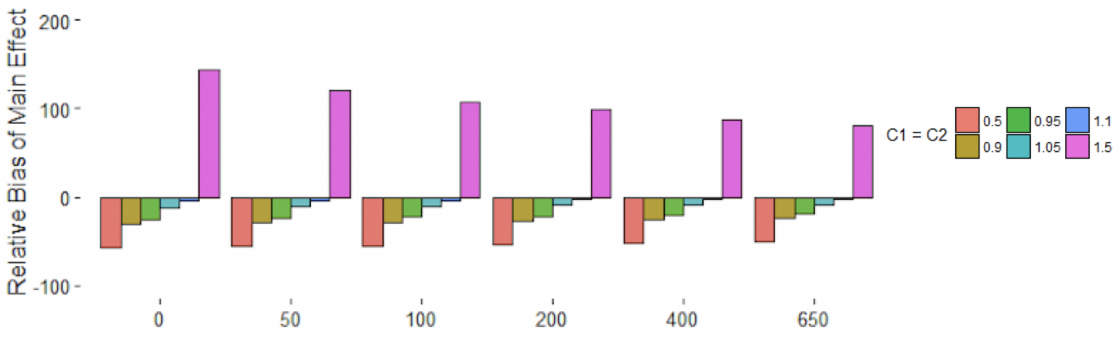


Figure 4.8: Relative bias results using EMIM when data is analysed assuming CEPG

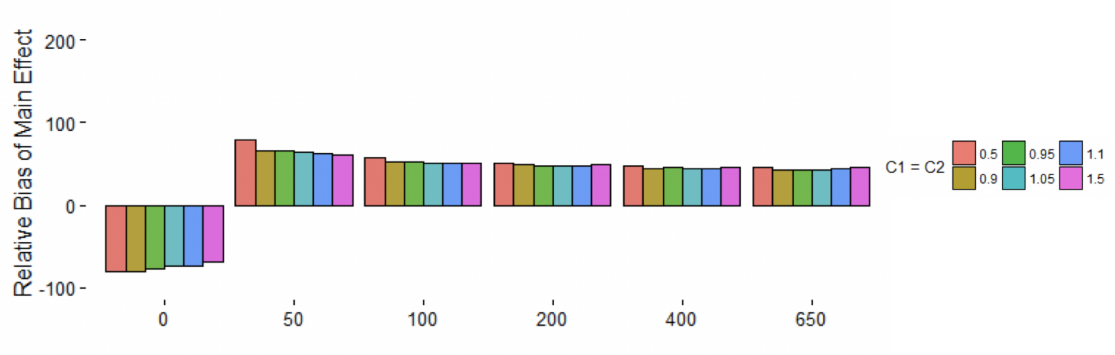


Figure 4.9: Relative bias results using EMIM when data is analysed assuming CPG

For the CPG model, shown in Figure 4.9, all of the effect estimates are biased. Relative bias is unaffected by the amount of MA. The amount of relative bias decreases slightly when the number of control trios is increased. The reason for this bias is not clear; this will be discussed further in Chapter 6.

Maternal Gene-Environment Interaction

Type I Error Type I error is estimated to be close to the nominal value of 0.05 for all models (Figures 4.10, 4.11, 4.12). This is, again, due to the independence of environment and MA; the exposed and unexposed trios will display the same patterns of asymmetry and therefore the asymmetry will not affect the test.

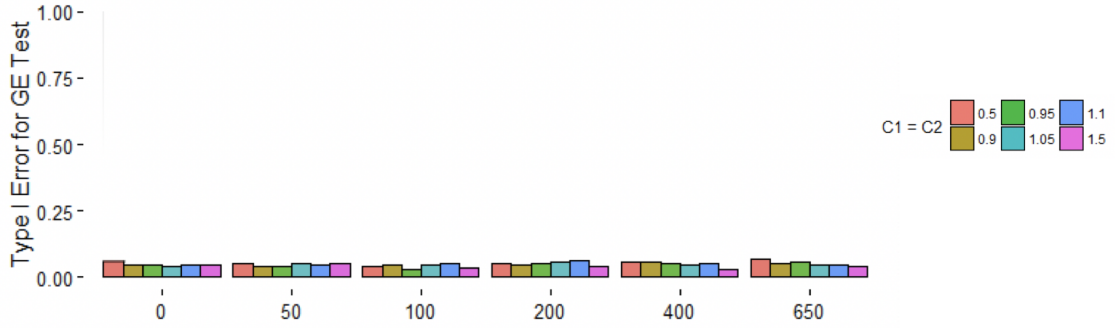


Figure 4.10: GE type I error results using EMIM when data is analysed assuming HWE and Random Mating

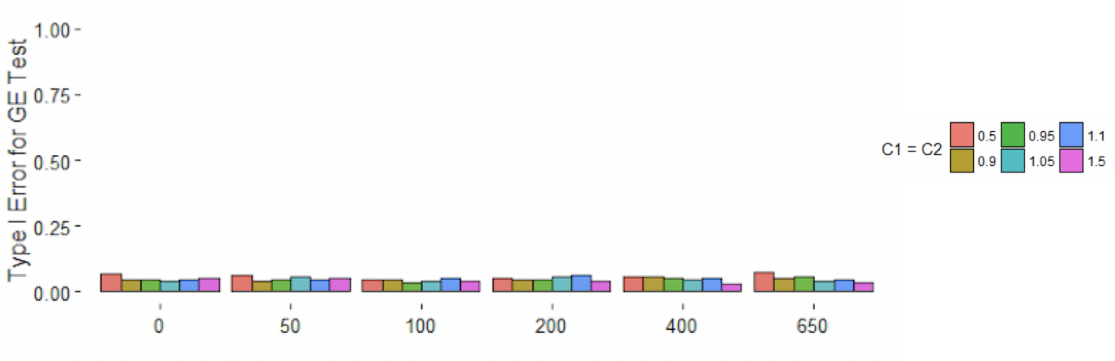


Figure 4.11: GE type I error results using EMIM when data is analysed assuming CEPG

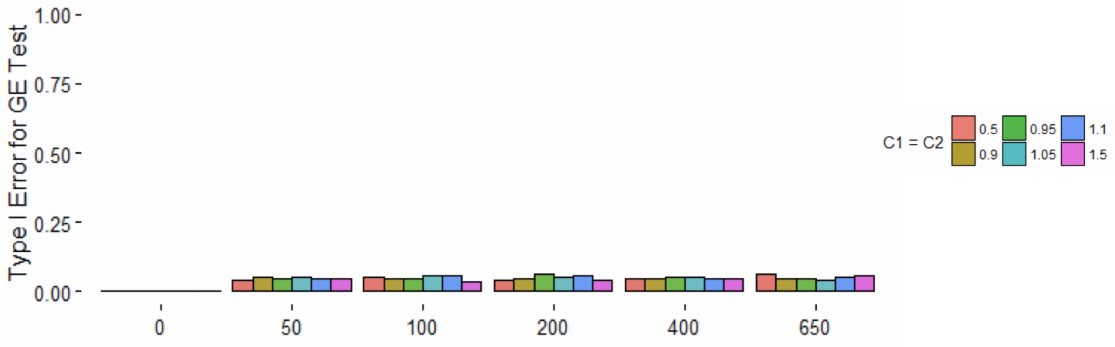


Figure 4.12: GE type I error results using EMIM when data is analysed assuming CPG

Power Under HWE+RM (Figure 4.13) and CEPG (Figure 4.14) assumptions, power increases with the number of controls. This is expected as the environment effect was simulated independently of MA so it should not affect the test of heterogeneity greatly. The observed power in the first two models is higher than in the CPG model (Figure 4.15) which also is to be expected as it requires estimation of additional parameters.

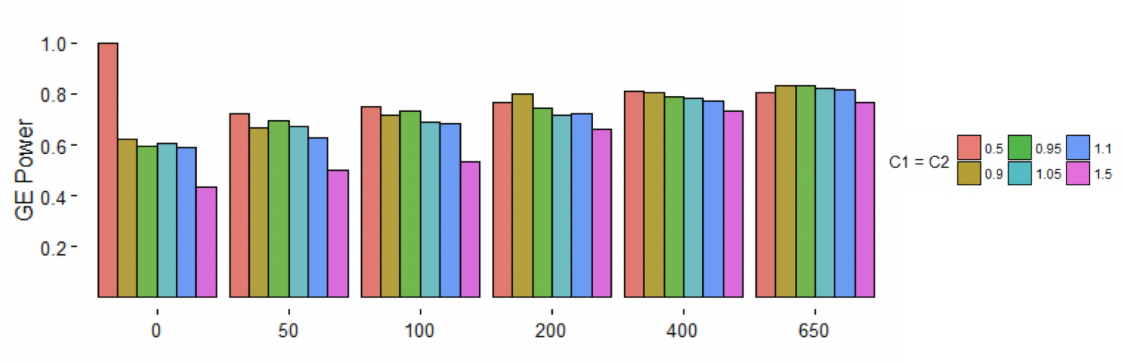


Figure 4.13: Estimated GE power using EMIM with analysis model that assumes HWE and Random Mating

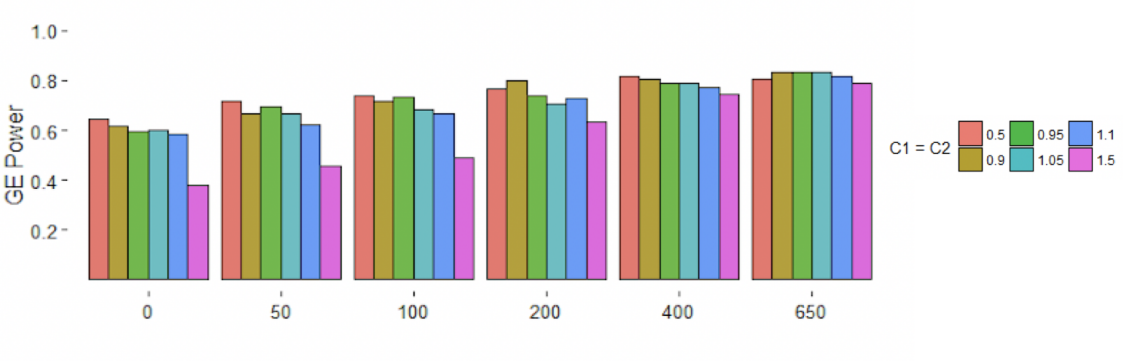


Figure 4.14: Estimated GE power using EMIM with analysis model that assumes CEPG

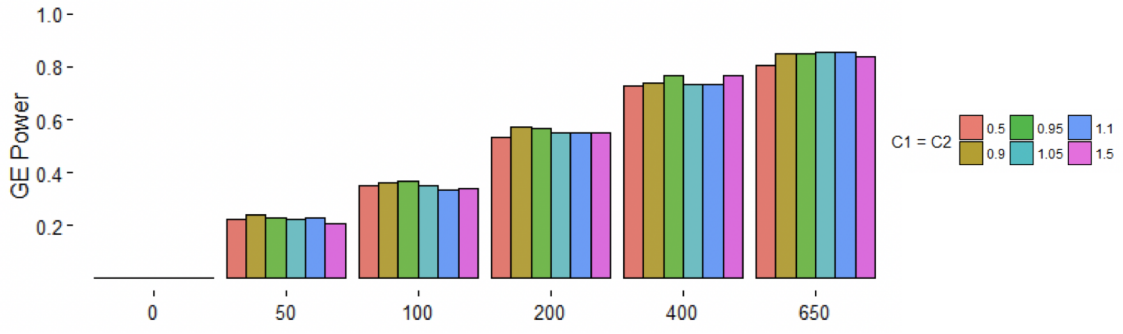


Figure 4.15: Estimated GE power using EMIM with analysis model that assumes CPG

Relative Bias As previously mentioned, a drawback of using this method for conducting a maternal gene-environment test is that no effect estimate is computed and therefore

bias is cannot be measured.

4.3.2 LEM Results

Main Effect Detection

Type I Error Type I errors in CEPG are very inflated for moderate to high asymmetry ($C1 = C2 \geq |0.1|$) and are still inflated, around 0.1, for low asymmetry (fig. 4.16). The CPG parameterization adequately controls the Type I error rate (fig. 4.17).

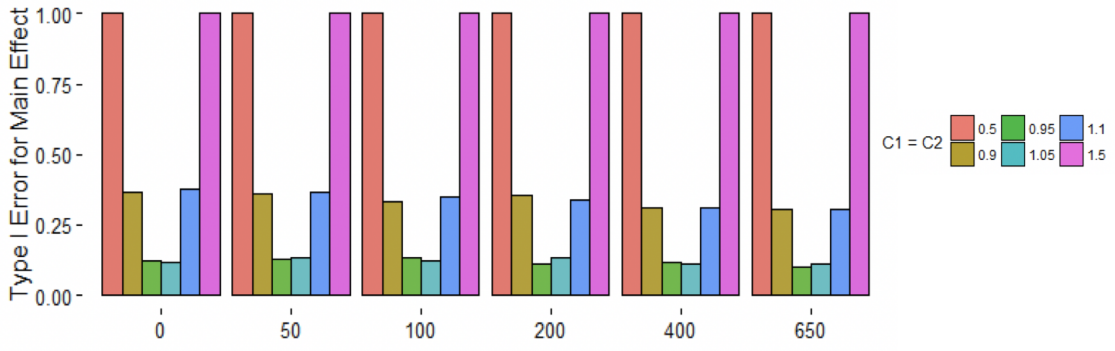


Figure 4.16: Type I error results using LEM when data is analysed assuming CEPG

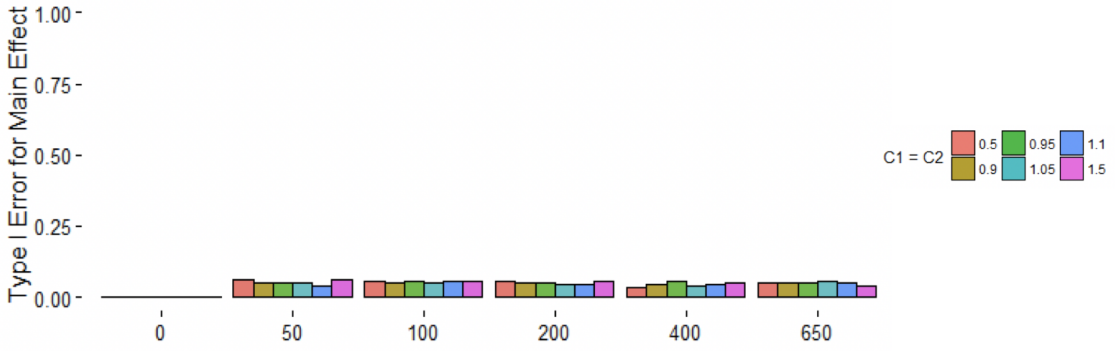


Figure 4.17: Type I error results using LEM when data is analysed assuming CPG

Power When the Type I error rate is inflated, we are not concerned about maintaining power for the maternal effect test, as they are in the CEPG model shown in Figure 4.18. In

the CPG model, power to detect a main maternal effect is high for all levels of asymmetry (fig. 4.19).

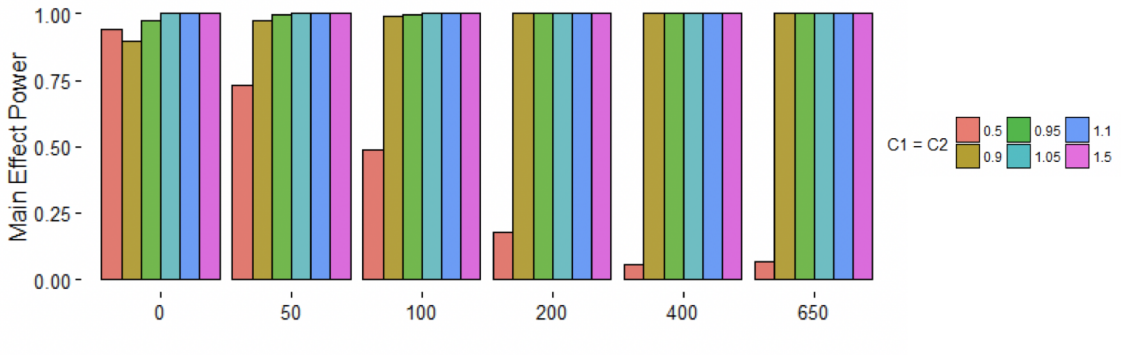


Figure 4.18: Estimated power using LEM with analysis model that assumes CEPG

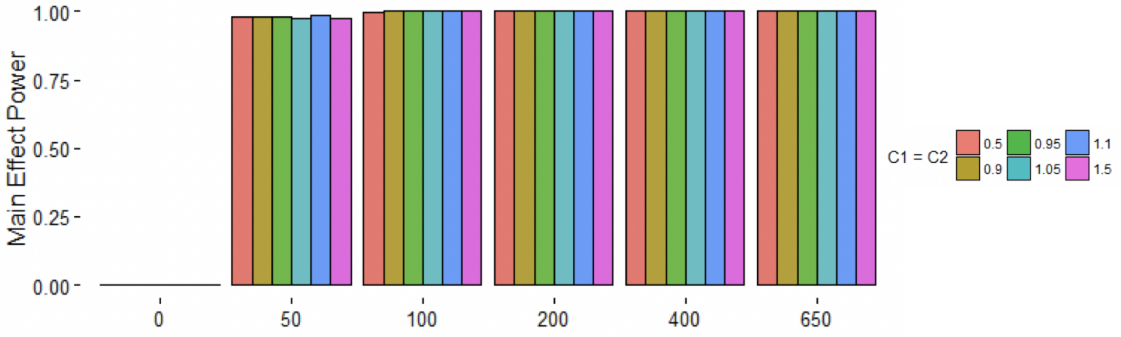


Figure 4.19: Estimated power using LEM with analysis model that assumes CPG

Relative Bias Relative bias is low for both CEPG and CPG models, with the amount of bias increasing as the number of controls increases in the CEPG in Figure 4.20.

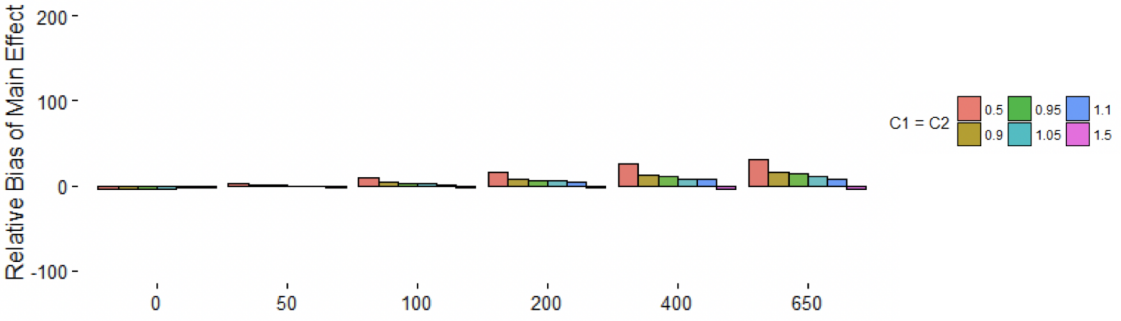


Figure 4.20: Relative bias results using LEM when data is analysed assuming CEPG

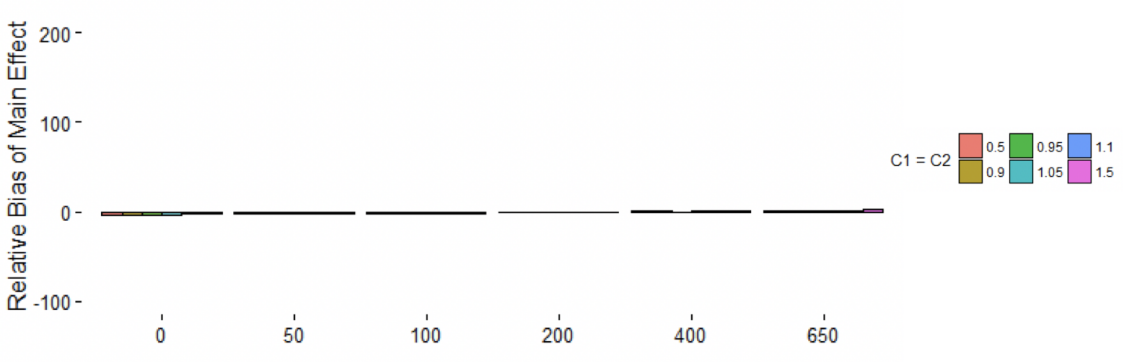


Figure 4.21: Relative bias results using LEM when data is analysed assuming CPG

Maternal Gene-Environment Interaction

Type I Error Type I error is estimated to be inflated with the CEPG model (Figure 4.22) whereas it is not inflated in CPG parameterization (Figure 4.22).

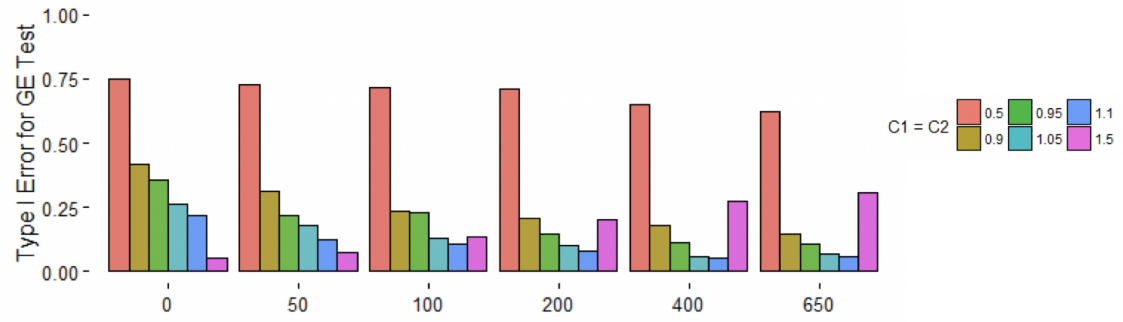


Figure 4.22: GE type I error results using LEM when data is analysed assuming CEPG

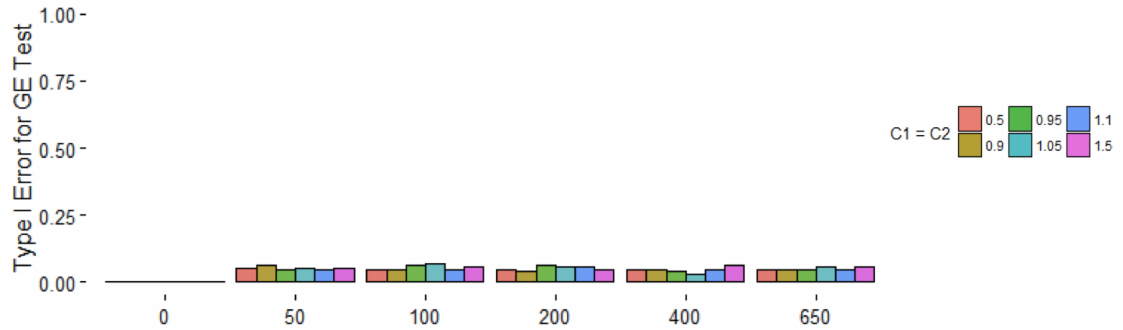


Figure 4.23: GE type I error results using LEM when data is analysed assuming CPG

Power Figure 4.24 shows that the LRT for a maternal gene-environment interaction has lower power as the mating asymmetry coefficients increase, with the difference being more pronounced with fewer controls.

In the CPG model (Figure 4.25), the test requires around 400 control trios to attain 80% power. When no controls are present the over-parameterization causes no test to show significant differences.

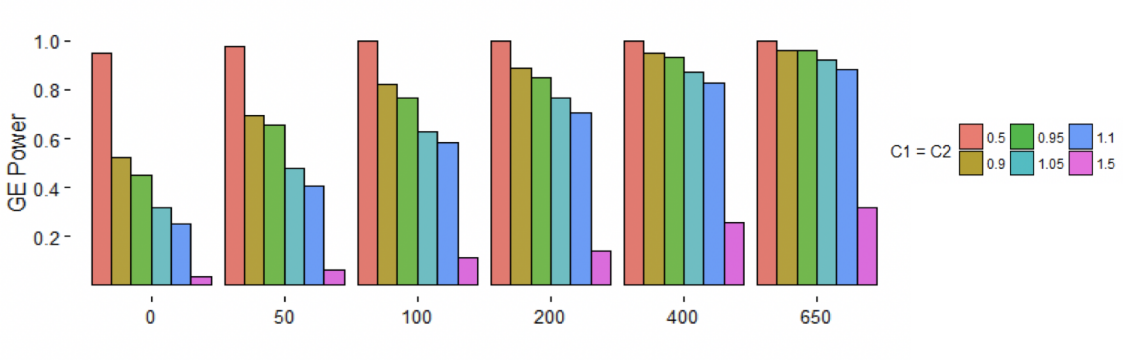


Figure 4.24: Estimated GE power using LEM with analysis model that assumes CEPG

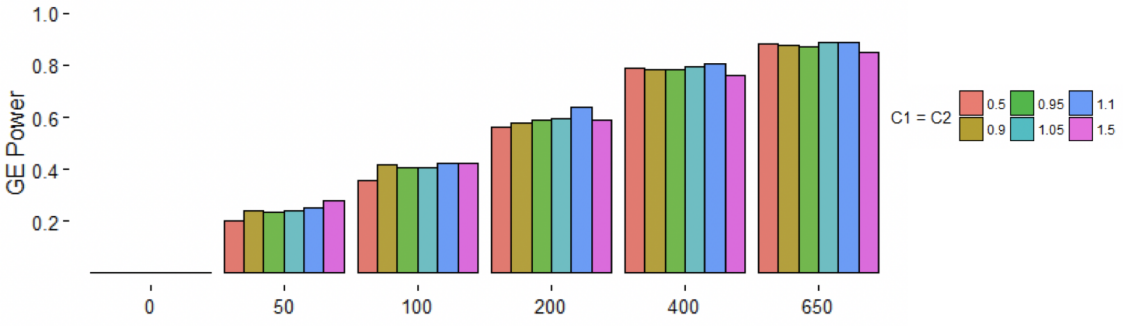


Figure 4.25: Estimated GE power using LEM with analysis model that assumes CPG

Relative Bias Relative bias of the interaction effect follows in the direction of simulated asymmetry for the CEPG parameterization (fig. 4.26). Estimates are highly biased for the CPG model, regardless of the degree of asymmetry and only slightly lower when more

controls are present (fig. 4.27).

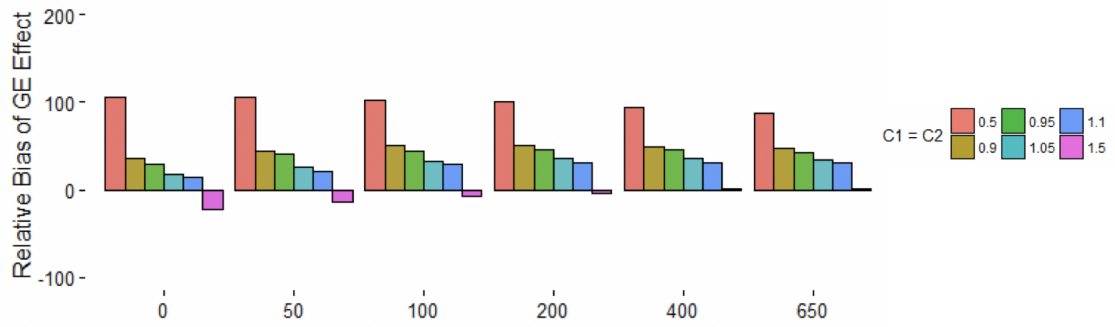


Figure 4.26: GE relative bias results using LEM when data is analysed assuming CEPG

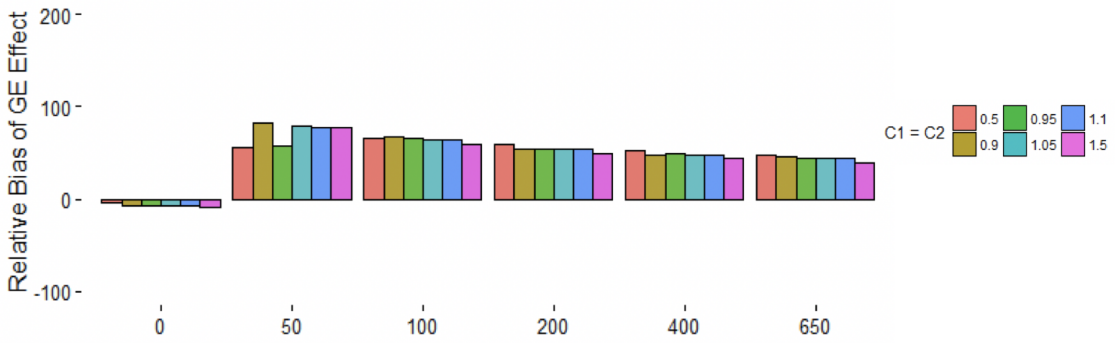


Figure 4.27: GE relative bias results using LEM when data is analysed assuming CPG

4.4 Discussion

In this chapter, I have described a simulation study to investigate the performance of LEM and EMIM in estimating both maternal effects and maternal GE effects under scenarios with and without mating asymmetry. The results highlight that care is needed when estimating maternal effects or maternal gene-environment effects using trio data.

This study confirms a risk of estimating maternal effects using case trio data only: there is no way to detect or control for mating asymmetry. To circumvent this problem, exter-

nal control trios could be taken from other sources, such as HapMap [9] or 1000 Genomes [8]. However, MA may be population-specific so precise population matching will be crucial to avoid any spurious results due to population stratification [19]. If control trios are not properly population matched, the estimation of the maternal genetic effect is subject to confounding by population stratification. This is because mating-type frequency estimates from the external control trios will be different due to allele frequency differences across population, which could introduce a spurious maternal effect. Therefore, collecting control trios from the same population is important if a goal of the study is to detect maternal effects.

A drawback to EMIM is that it does not allow exposure data variables and so the gene-environment test is done through a post-hoc test of heterogeneity. It is for this reason that the second program, LEM, was evaluated. However, LEM is not very intuitive, its computational speed is slower than more modern programs, and it requires a long obsolete operating system. There is a shortage of available resources to answer questions of this nature. Our investigation highlights the need for the development of an alternative or an update in log-linear modeling resources to detect maternal effects.

Even more concerning is the bias of interaction effects estimated by LEM and of main maternal effect estimation with EMIM. Even under low to moderate asymmetry, the estimates are not reliable. The problem of bias does not disappear when additional controls are used to estimate mate-pair frequencies. This may be due to part of the Maternal GE effect being captured by the maternal genetic effect estimated by EMIM. This will be discussed further in Chapter 6.

Chapter 5

Candidate Gene Study

5.1 Introduction

Orofacial clefts (OFC), the group of conditions including cleft lip and cleft palate, are one of the most common birth defects in the world at a rate of about 1.7 per 1000 live births [37]. There is wide variability in rates of OFCs based on geographic regions, racial groups, socioeconomic status and environmental exposures. There are two forms of OFCs called syndromic and non-syndromic; we focus on the non-syndromic form of cleft lip (around 70% of all cleft lip [10]) in this analysis, since the different forms may have different genetic mechanisms. OFCs occur when the lips or mouth do not form properly during pregnancy. These malformations have an effect on the child's speech, hearing, appearance, and psychological well-being, and requires life-long intervention and care. Although treatments are improving the quality of life for children born with OFCs, prevention is still the most successful intervention.

There is strong evidence that cleft lip with or without cleft palate (CL/P) has a genetic component as well as environmental risk factors. In this chapter, I describe statistical analysis of a candidate gene study designed to find interaction effects between maternal genes

and maternal environmental variables during pregnancy. Two known environmental factors that modify risk for CL/P are maternal smoking, which increases the risk, and folic acid supplementation, which decreases risk; genetic risk factors might modify the effect of either smoking or folic acid supplementation.

We analyse data from genetic variants selected from five candidate genes that were previously shown to be associated with cleft lip or cleft palate when the mother carried the variant. We aim to replicate the previous findings and determine whether the variants within these genes have any synergistic effects with the binary exposure variables.

After considering the results from the simulation study in Chapter 4, we will be careful in interpreting the magnitude of any effects found. As this dataset did not include controls, the mode of analysis will require the assumption of mating symmetry. We will discuss consequences of this assumed independence between MS and environmental variables in the conclusion.

5.2 Methods

5.2.1 Study Population

The OFC data that we analyse is from the "International Consortium to Identify Genes and Interactions Controlling Oral Clefts", a part of the Gene Environment Association Studies (GENEVA) consortium [3]. These data were collected over the period of multiple years and across 13 international sites.

Children with OFCs and their parents were recruited into the study. We had access to data on 7089 individuals. These individuals comprise 88 singletons, 401 parent-offspring

pairs, 2029 parent-offspring trios and 25 assorted nuclear and extended families of size 3 to 6. The self-reported race of individuals shows that most participants are either Asian (49%) or White (46%). This investigation into maternal effects is not robust to population stratification so we consider only individuals who satisfy a principal component-based European ancestry, as determined by Beaty et al. [3]. After filtering out family units where the ancestry is not European and incomplete trios, the final sample consists of 664 case-parent trios.

5.2.2 Candidate Genes

Genetic variants from five candidate genes were selected for analysis based on previous significant results. The five genes are named FLNB, HIC1, ZNF189, MTHFR, and RFC1. In total, 69 SNPs from these genes were captured by the GWAS from which these data came. Biological details on these genes are available in Appendix B; details on previous associations between these genes and orofacial cleft phenotypes are described below.

FLNB Jugessur et al. (2010) found an association between FLNB and isolated cleft palate only (CP) in two European populations [23]. Although the results were not replicated in the CL/P group, we wish to investigate whether there is an interaction between maternal FLNB variants and the exposures under study.

HIC1 In Jugessur et al. (2010), a strong association was found between maternal HIC1 and CP in the Norwegian population and a weaker association in the Danish population [23].

ZNF189 This gene was found to have an association with CP with one analytical method [23]. Although this was the first study to link this gene to cleft palate, the gene is located in a region of the genome that has been linked to oral clefts using another type of genetic data analysis called linkage analysis [33] [34] . This gene was included to see if this dataset

will find any association with CL/P where the Jugessur paper did not, potentially due to smaller sample sizes than in the GENEVA study.

MTHFR Mutations in the MTHFR gene are known genetic risk factors for many diseases including: vascular disease [14], male infertility [4], and other birth defects such as neural tube defects [58] [39]. Mutations in this gene have been associated with CL/P, both as maternal [40] [15] [35] [41] and child effects [36]. It is hypothesized that, like nutritional supplementation of folic acid, a variant that impedes the activity of a folate metabolism enzyme might increase the susceptibility for oral clefts [15].

RFC1 The reduced folate carrier 1 gene (RFC1, also known as SLC19A1) is located on chromosome 21 and has been found to be associated with non-syndromic CL/P in multiple studies ([29] [49] [52] [62]). These studies were conducted in populations of different origin and using different study designs and analytical methods. This association has been extensively replicated.

5.2.3 Environmental Variables

In human genetic studies, any variable that is not a genetic variable is called an environmental variable. Two maternal environmental variables were included in the analysis. 461

Maternal Smoking The first variable being considered is maternal smoking. Smoking during pregnancy is known to be associated with both CL/P and CP. A systematic review identified positive odds ratio for maternal smoking 1.28 (95% CI 1.20 – 1.36) [18], and a relative risk of 1.34, (95% CI 1.25 - 1.44) [30]. In addition to the nicotine and carbon monoxide in cigarettes, there are over 4000 other chemicals, including known carcinogens, toxic heavy metals, and many chemicals untested for developmental toxicity [45].These may cross the

placental barrier and be harmful to the fetus.

Participants were asked: Did the mother smoke in the perinatal period (3 months prior through 3rd month of pregnancy)? In the European CL/P subgroup 461 families had no self-reported perinatal maternal smoking, 200 did, and 3 did not respond.

Prenatal Vitamin Usage The second variable evaluated for a maternal gene-environment interaction is perinatal vitamin usage. Prenatal vitamin supplementation is used to prevent neural tube defects and has also been associated with a reduced risk of OFCs. In one study, daily folic acid supplementation, starting pre-conception, found a risk reduction of 47% as compared to those who did not use any supplements [59]. Other studies found similar risk reduction, however they suggest that this may be instead due to other correlated behaviours that women who use multivitamins may exhibit [48].

Specifically, the respondent was asked: Did the mother take multivitamins or prenatal vitamins in the perinatal period (3 months prior through 3rd month of pregnancy)? In the European CL/P subgroup 240 responded no to supplementation with multivitamins, 422 yes, and 2 responses were missing.

5.2.4 Statistical Methods

As introduced in Chapter 3, these data were analysed using EMIM and LEM. In EMIM, the data were stratified by exposure type and a model including a child and a maternal effect was fit. The effect estimates in the two environmental variable groups were then compared using a test of heterogeneity. The same effects were included in the log linear analysis fit using LEM; however, the GE interaction was also included in the model. A 1 degree of freedom likelihood ratio test was used to test for the interaction term. Multiplicative relative risk

terms are assumed.

For both analysis strategies, we must assume mating symmetry (CEPG model) as control trios are not available. This means that, in total, EMIM and LEM will estimate 6 mating-type parameters by 2 environment groups (12 nuisance parameters total) and 1 parameter for maternal effects since we assume $S_2 = S_1^2$. Two methods to adjust for multiple comparisons are used: first, a conservative Bonferroni threshold and the less strict False Discovery Rate (represented by Q values) [54]. The Bonferroni threshold will be $\frac{0.05}{69} = 0.0007$, and the Q value is significant at 0.05.

5.3 Results

In this section, results are presented for the 5 most significant SNPs for each model and each method. Results for all 69 SNPs can be found in Appendix B.

5.3.1 Testing for Maternal Gene - Maternal Smoking Interactions

EMIM After adjustment for multiple testing (both using the Bonferroni correction and FDR methods), none of the SNPs remain significant at a 0.05 level. All of the top five SNPs are located in the RFC1 gene. The main genetic effects stratified by smoking status do not conform with the expectation. It was hypothesized that smoking would confer a higher risk for CL/P as compared to the non-smoking group; the interaction suggests that smoking acts as a protective factor for mothers with the minor alleles. The effect estimates are similar across all the SNPs. This is likely due to one variant with a true association, where the other SNPs show a significant effect simply because they are within the same region as the causal variant.

Table 5.1: Top 5 most significant SNPs for Maternal Gene - Maternal Smoking Interaction using EMIM

Gene	rs Number	Effect Estimate for No Smoking	Effect Estimate for Smoking	P-value for Heterogeneity	Q-value
RFC1	rs4819130	1.209	0.794	0.0157	0.305
RFC1	rs2330183	1.152	0.788	0.0249	0.305
RFC1	rs1051266	1.162	0.809	0.0316	0.305
RFC1	rs914232	1.151	0.809	0.0353	0.305
RFC1	rs2838956	1.182	0.828	0.0368	0.305

Columns 1 & 2 describe the gene and SNP. Column 3 is the relative risk estimate for each SNP among the non-smoking subgroup and column 4 is for the smoking subgroup. Column 5 shows the p-value for the test for heterogeneity using the EMIM results. Column 6 shows the q-value, which adjusts for multiple comparison

LEM The fourteen most significant SNPs are shown in Table 5.2. The top 7 are significant after a Bonferroni correction and FDR, whereas the final 7 do not meet the stricter Bonferonni threshold. The LEM analysis produced different results than the EMIM analysis; none of the SNPs identified in RFC1 using EMIM were significant using LEM and the most significant SNPs were from the FLNB gene. The effect estimates for the interaction are all quite large; however, some caution is needed when interpreting these results as the effect estimates are very high, something we expected based on our previous simulation study.

Table 5.2: Top 14 most significant SNPs for Maternal Gene - Maternal Smoking Interaction using LEM

Gene	rs Number	Effect Estimate for Maternal Effect	Effect Estimate for Interaction	P-Value for Interaction	Q-Value for Interaction
MTHFR	rs17367504	1.139	6.343	0.00003 *	0.001
FLNB	rs9880603	1.002	13900	0.00007 *	0.005
FLNB	rs1623879	0.972	13.748	0.00017 *	0.005
FLNB	rs839225	1.128	7.910	0.00027 *	0.005
FLNB	rs9884098	1.285	4.128	0.00037 *	0.005
RFC1	rs7278425	1.046	5.134	0.00043 *	0.005
FLNB	rs2177153	1.076	3.572	0.00062 *	0.006
FLNB	rs1658347	1.120	6.972	0.001	0.007
FLNB	rs13092880	0.893	7.126	0.001	0.008
FLNB	rs6787425	1.259	3.280	0.002	0.012
ZNF189	rs2253258	1.025	5.146	0.002	0.012
RFC1	rs12483553	1.224	6.249	0.002	0.012
RFC1	rs3788205	1.085	2.382	0.008	0.042
FLNB	rs2362907	1.165	2.698	0.009	0.044

* : significance after Bonferroni correction

The two methods have no agreement on the results. Considering also that the EMIM test for heterogeneity produced no significant hits after correction for multiple testing, we do not have conclusive evidence for any interactions between maternal genes and smoking during pregnancy.

5.3.2 Perinatal Vitamin Usage

EMIM The top five SNPs are shown in Table 5.3. None of these 5 SNPs are significant at a 0.05 level after a Bonferroni correction or using the FDR procedure (shown with high Q-values).

Table 5.3: Top 9 most significant SNPs for Maternal Gene - Maternal Vitamin Supplementation Interaction using EMIM

Gene	rs Number	Effect Estimate for No Supplementation	Effect Estimate for Supplementation	P-value for Heterogeneity	Q-value
FLNB	rs9822918	1.241	0.879	0.0369	0.922
FLNB	rs13073391	1.282	0.919	0.0559	0.922
FLNB	rs1718483	0.794	1.089	0.0586	0.922
FLNB	rs2033739	1.283	0.929	0.0647	0.922
FLNB	rs839241	1.102	0.832	0.1037	0.922

LEM The SNPs with the nine lowest p-values are shown in Table 5.4. The top 4 SNPs are significant after a Bonferroni correction, and all nine are significant using the FDR method. However, these parameter estimates are very large and have not been reported elsewhere. Therefore, this could be a novel interaction discovery or a false positive. The model is fitting unbelievable relative risk estimates.

Table 5.4: Top 9 most significant SNPs for Maternal Gene - Maternal Vitamin Supplementation Interaction using LEM

Gene	rs Number	Effect Estimate for Maternal Effect	Effect Estimate for Interaction	P-Value for Interaction	Q-Value for Interaction
FLNB	rs13092880	1.167	7.227	0.0002	0.009
ZNF189	rs2253258	1.327	6.752	0.0004	0.009
FLNB	rs1623879	1.591	13.748	0.0004	0.009
FLNB	rs7627013	1.870	3.472	0.0006	0.012
RFC1	rs7278425	1.599	4.169	0.0009	0.012
MTHFR	rs9651118	1.891	2.749	0.002	0.017
FLNB	rs839225	1.591	4.876	0.001	0.012
FLNB	rs1658347	1.777	4.806	0.004	0.031
FLNB	rs9880603	1.041	5.842	0.002	0.017

Although both methods identified SNPs within the FLNB gene, both methods fail to identify the same SNP. This combined with the lack of significance from EMIM, suggests a possibility for an interaction between perinatal vitamin usage and maternal genes but it is far from definitive evidence.

5.4 Discussion

Our analysis does not show conclusive evidence for the presence of maternal gene environment interaction effects. Results were inconsistent between the two analysis methods selected. In addition, the effect estimates given by LEM for the interaction effect were all larger than 1. We therefore have some concern that LEM produces biased results for the models we analysed. Additionally, no significant main maternal effects were detected. This was expected as these data have been analysed previously as part of the original study [3]

and none were reported.

To understand what may be happening with the modelling using LEM, the R package Haplin [16] was used to fit the Weinberg et al model to the dataset for SNP rs1623879 in the FLNB gene. Recall that the interaction effect estimate for this SNP was 13.748. The loglinear model was fit separately within each exposure group and the test of heterogeneity used with EMIM was performed. The cell counts are shown in Table 5.5. The large number of zeroes sugggests that estimation of an interaction parameter might not be possible. This SNP, using the EMIM method, found nothing significant. When Haplin was used, the results obtained with EMIM were corroborated. This leads to a conclusion that the LEM results estimating the maternal gene-environment interaction may not be trustworthy.

Table 5.5: Example Cell Counts for Extreme Estimate

M	F	C	Unxposed Count	Exposed Count
aa	aa	aa	0	0
aa	Aa	aa	0	0
aa	Aa	Aa	0	0
Aa	aa	aa	0	0
Aa	aa	Aa	0	0
aa	AA	Aa	2	0
AA	aa	Aa	2	1
Aa	Aa	aa	5	2
Aa	Aa	Aa	11	3
Aa	Aa	AA	4	1
Aa	AA	Aa	27	13
Aa	AA	AA	35	18
AA	Aa	Aa	33	14
AA	Aa	AA	48	22
AA	AA	AA	294	126

It may be worthwhile to further investigate whether the FLNB does have any synergistic interaction effects with multivitamin supplementation as there was some evidence of interaction with SNPs in FLNB for both the EMIM and LEM approaches. This gene has

previously only been associated with cleft palate [23]. The association with the gene could be explained by the function that the corresponding protein plays in skeletal development. Multivitamin supplementation was studied primarily for the known protective effect of folic acid, but other compounds within the supplements such as calcium or vitamin D may in fact be interacting with this gene.

Chapter 6

Conclusion and Discussion

This thesis served to highlight how methods for maternal effect testing and maternal effect interaction testing require further development. In Chapter 2 we introduced concepts in genetics and epidemiology to build up to the concepts of maternal genetic effects and mating asymmetry. The parameterization for mating types was carried throughout the rest of the chapters.

In Chapter 3 we reviewed the statistical methods used for parent-offspring trio studies, focusing on the estimation of main maternal and interaction effects. The fifteen categories of mother-father-offspring genotype assignments were presented and we described the multinomial model for the counts in the fifteen categories. Multinomial modelling under models that do not include interactions with environmental variables, as is done by Ainsworth et al. using EMIM, and log-linear regression, as used by Weinberg et al. using LEM, are equivalent strategies for analysis of main maternal effects, but the implementations differ slightly so the estimations will not be identical. From where Ainsworth and Weinberg left off, we endeavoured to extend their maternal genetic effect methods to include an interaction with environmental variables. As the EMIM software does not allow for this interaction term to be included directly, a method from meta-analysis was used, as recommended in [55]. LEM

is a generic log-linear modelling software so we could take the Weinberg parameterization and include the interaction term directly.

Chapter 4 is a simulation study where we assessed the effect of mating asymmetry on statistical methods to detect maternal effect and maternal gene-environment interactions. We found that mating asymmetry increases type 1 error for the detection of maternal effects, but not maternal gene-environment effects when the environmental factor is independent of the mating asymmetry. We believe this is a realistic setting because in studies where a rare disease assumption is made (this assumption was not necessary for this work but could be made as CL/P and many disorders of early life are rare), researchers commonly assume independence between genotype and exposure [57], which would imply also independence from mating type. The addition of control families allows for all nine mating types to be estimated and therefore effects of MA are minimized.

In Chapter 5, a candidate gene study was performed to determine whether any maternal gene-environment interaction effects are present in the orofacial cleft data. As this study was conducted without the collection of control trios, an assumption of mating symmetry was required. No conclusive evidence was found to suggest any interactions are present within the selected SNPs.

The Weinberg et al. [65] model for detecting maternal effects was implemented using LEM, a computationally slow software that requires some effort to run since it has not been updated for modern operating systems. Knowledge of batch scripting is required to use LEM on a large scale as it requires that, for each SNP, a separate data file and LEM run is required. This program is a point-and-click type that can, with difficulty, be controlled through a batch script in Windows. To improve the usability of the log-linear model for maternal effects, the Weinberg approach should be implemented in modern statistical software,

such as R. If trios are not complete, the EM algorithm described in [64] can be implemented in R.

A challenge in interpreting the results described here is that it is not possible to fit the same model in EMIM. The EMIM approach is limited in that it handles only genetic covariates so any environment interactions can only be evaluated by using the meta-analysis technique suggested in [55]. As mentioned earlier, this does not provide any point estimate for this effect and so we can't compare point estimates to LEM. Since the LEM estimates are suspect, having an alternate method of estimation would have been helpful. Although EMIM is easy to use, this software does not conduct the desired GE test so an extension to include environmental factors into EMIM should be a priority. When the Weinber log-linear model was used to conduct the same test of heterogeneity as with EMIM, the results supported each other. Since the LEM method is giving unbelievable results, the conclusion must then be that the estimation of this GE parameter is still not reliably doable.

A method by Yang and Lin [68] was not included in our methods comparison. This method, called LIME (partial Likelihood approach for detecting Imprinting and Maternal Effects), uses a partial likelihood approach to circumvent the need to assume mating symmetry and is therefore a potential alternate method to the two discussed in this work. LIME still requires control family units: they match the case families to control families with the same structure and stratify according to familial genotype; this allows the mating-type parameters to be factored out of the likelihood and so they don't need to be fit. This both makes moot the assumption of mating symmetry and reduces the problem of overparameterization that may be the cause of such biased effect estimates for the interaction terms. However, the model used in LIME did not include gene-environment effects. Future work could be done to extend the LIME approach to estimation of GE effects.

The candidate gene study clearly highlighted the fact that the Weinberg approach to estimating GE interactions may be unreliable. This could either be due to the LEM software or the model fitting itself. Two reasons for such an ill-fitting model could be the sampling zeros or overparameterization.

Sampling zeros, discussed in the background on log-linear regression, occur when cells in the multinomial have zero counts. With these data, both simulated and from the candidate gene study, sampling zeros occur frequently, especially in the homozygous for the minor allele cells. The combination of low sample size for environmental exposures, low allele frequencies, and a large 30 cell multinomial table nearly guarantees sampling zeroes. If, for example, neither exposed nor unexposed cell for a homozygous mother are observed, how can this parameter be estimated? By restricting risk parameter estimation to be multiplicative ($S_2 = S_1^2$), this problem is minimized but does not disappear completely. Should this assumed multiplicative risk model not hold, estimation is biased as well. This assumption was made to decrease the number of parameters in an already highly parameterized model. This may still not have been sufficient.

A model with this many parameters (recall, that there are 12 mating-type parameters, 1 child parameter, 1 maternal parameter and 1 interaction parameter) has low power for finding small effects, which is all we expected to uncover within this study. Should collinearity exist between the variables, the estimation provided by the model could be wildly off. Each cell may be informative to only a certain number of parameters, and it is possible this model is attempting to draw more information than is available. In fact, David Freedman [13] showed that when the number of variables is similar to the number of data points, predictor variables not associated with the dependent variable are more likely to be found significant, which we found in the models which did not account for MA. Whatever the reason for the model being unreliable in fitting, this method can not be trusted for peer-reviewed publi-

cation at this point. A deeper investigation into what is occurring will be done before any further real data analyses are conducted. This will include performing a simulation study where protective effects (relative risks below 1) are simulated to see whether these effects can be teased out under perfect conditions. Additional inquiry into the collinearity of the estimated interaction effect with the 12 mating-type parameters should be done as well. When the loglinear model was fit using Haplin and compared with the test of heterogeneity, the results from EMIM were supported. This points to the fact that the loglinear model with this maternal gene-environment interaction is not suitably implemented in LEM currently.

The body of this thesis, in Chapters 4 and 5, endeavoured to show the gaps in appropriate methodology present in the current technology. The presence of mating asymmetry needs to be accepted in the field of maternal genetic studies, and accounted for using methods that do not require an assumption of mating symmetry. All future methods that are to be implemented should encourage the use of this loosened assumption. Additionally, these new methods need to include the important environmental exposure data to detect additional measurable indirect effects that may affect disease risk.

To determine genetic risk factors for human traits, especially those present at birth, it is necessary to consider whether the genes of the mother are conveying any additional risk to the child and whether environmental factors modify the genetic risk. We verified that mating asymmetry biases maternal genetic effect estimates and tests; however, we showed that it does not bias maternal gene environment effect tests. Our work also highlights potential deficiencies with methods of estimating maternal gene environment effects with existing software. This area of study is still under-developed and work should be done to create reliable and user-friendly statistical tools to estimate maternal gene environment effects.

Appendix A

R Code and Software Inputs

A.1 R Code

```
## Data Simulation (formatted for reading by EMIM) ##
ma_sim_cells_ENV = function(ncase, E, ncon, nrep, maf, C, file_path)
    {
alpha = 0.2
# Parameter setting
R1 = 1.3
R2 = 1.69
S1 = 1.3*E
S2 = 1.69*E
# Mating-type probabilities assuming HWE
maprobs = c()
maprobs[1] = (1 - maf) ^ 4
maprobs[2] = (2 - C[1]) * (1 - maf) ^ 2 * 2 * maf * (1 - maf)
maprobs[3] = C[1] * (1 - maf) ^ 2 * 2 * maf * (1 - maf)
maprobs[4] = 2 * maf * (1 - maf) * 2 * maf * (1 - maf)
maprobs[5] = (2 - C[2]) * (1 - maf) ^ 2 * maf ^ 2
maprobs[6] = C[2] * (1 - maf) ^ 2 * maf ^ 2
maprobs[7] = C[3] * 2 * maf * (1 - maf) * maf ^ 2
maprobs[8] = (2 - C[3]) * 2 * maf * (1 - maf) * maf ^ 2
maprobs[9] = maf ^ 4

# P(D = 1| M,F, C)
p_dmfc = c()
p_dmfc[1] = 1 * maprobs[9] * S2 * R2 * alpha
p_dmfc[2] = 0.5 * maprobs[7] * S2 * R2 * alpha
p_dmfc[3] = 0.5 * maprobs[7] * S2 * R1 * alpha
```

```
p_dmfc[4] = 0.5 * maprobs[8] * S1 * R2 * alpha
p_dmfc[5] = 0.5 * maprobs[8] * S1 * R1 * alpha
p_dmfc[6] = 1 * maprobs[6] * S2 * R1 * alpha
p_dmfc[7] = 1 * maprobs[5] * R1 * alpha
p_dmfc[8] = 0.25 * maprobs[4] * R2 * S1 * alpha
p_dmfc[9] = 0.5 * maprobs[4] * S1 * R1 * alpha
p_dmfc[10] = 0.25 * maprobs[4] * S1 * alpha
p_dmfc[11] = 0.5 * maprobs[3] * S1 * R1 * alpha
p_dmfc[12] = 0.5 * maprobs[3] * S1 * alpha
p_dmfc[13] = 0.5 * maprobs[2] * R1 * alpha
p_dmfc[14] = 0.5 * maprobs[2] * alpha
p_dmfc[15] = 1 * maprobs[1] * alpha

p_d = sum(p_dmfc)
# P(M,F,C | D = 1)
cell_probs_cases = p_dmfc / p_d

# P(D = 0| M,F, C)
p_dmfc_Con = c()
p_dmfc_Con[1] = 1 * maprobs[9] * (1 - S2 * R2 * alpha)
p_dmfc_Con[2] = 0.5 * maprobs[7] * (1 - S2 * R2 * alpha)
p_dmfc_Con[3] = 0.5 * maprobs[7] * (1 - S2 * R1 * alpha)
p_dmfc_Con[4] = 0.5 * maprobs[8] * (1 - S1 * R2 * alpha)
p_dmfc_Con[5] = 0.5 * maprobs[8] * (1 - S1 * R1 * alpha)
p_dmfc_Con[6] = 1 * maprobs[6] * (1 - S2 * R1 * alpha)
p_dmfc_Con[7] = 1 * maprobs[5] * (1 - R1 * alpha)
p_dmfc_Con[8] = 0.25 * maprobs[4] * (1 - R2 * S1 * alpha)
p_dmfc_Con[9] = 0.5 * maprobs[4] * (1 - S1 * R1 * alpha)
p_dmfc_Con[10] = 0.25 * maprobs[4] * (1 - S1 * alpha)
p_dmfc_Con[11] = 0.5 * maprobs[3] * (1 - S1 * R1 * alpha)
p_dmfc_Con[12] = 0.5 * maprobs[3] * (1 - S1 * alpha)
p_dmfc_Con[13] = 0.5 * maprobs[2] * (1 - R1 * alpha)
p_dmfc_Con[14] = 0.5 * maprobs[2] * (1 - alpha)
p_dmfc_Con[15] = 1 * maprobs[1] * (1 - alpha)

p_nod = sum(p_dmfc_Con)

# P(M,F,C | D = 0)
cell_probs_cons = p_dmfc_Con / p_nod
cell_probs_con2 = c(
cell_probs_cons[1],
cell_probs_cons[2] + cell_probs_cons[3],
cell_probs_cons[6],
cell_probs_cons[4] + cell_probs_cons[5],
cell_probs_cons[8] + cell_probs_cons[9] + cell_probs_cons[10],
cell_probs_cons[11] + cell_probs_cons[12],
cell_probs_cons[7],
```

```
cell_probs_cons[13] + cell_probs_cons[14],
cell_probs_cons[15]
)

p_dmfc_ne = c()
p_dmfc_ne[1] = 1 * maprobs[9]
p_dmfc_ne[2] = 0.5 * maprobs[7]
p_dmfc_ne[3] = 0.5 * maprobs[7]
p_dmfc_ne[4] = 0.5 * maprobs[8]
p_dmfc_ne[5] = 0.5 * maprobs[8]
p_dmfc_ne[6] = 1 * maprobs[6]
p_dmfc_ne[7] = 1 * maprobs[5]
p_dmfc_ne[8] = 0.25 * maprobs[4]
p_dmfc_ne[9] = 0.5 * maprobs[4]
p_dmfc_ne[10] = 0.25 * maprobs[4]
p_dmfc_ne[11] = 0.5 * maprobs[3]
p_dmfc_ne[12] = 0.5 * maprobs[3]
p_dmfc_ne[13] = 0.5 * maprobs[2]
p_dmfc_ne[14] = 0.5 * maprobs[2]
p_dmfc_ne[15] = 1 * maprobs[1]

cell_probs_ne2 = c(
p_dmfc_ne[1],
p_dmfc_ne[2] + p_dmfc_ne[3],
p_dmfc_ne[6],
p_dmfc_ne[4] + p_dmfc_ne[5],
p_dmfc_ne[8] + p_dmfc_ne[9] + p_dmfc_ne[10],
p_dmfc_ne[11] + p_dmfc_ne[12],
p_dmfc_ne[7],
p_dmfc_ne[13] + p_dmfc_ne[14],
p_dmfc_ne[15]
)

# Sampling from the multinomials defined above (exposed cases,
   exposed controls, unexposed cases, unexposed controls)
cases = rmultinom(n = nrep, size = ncase, prob = cell_probs_cases)
cons = rmultinom(n = nrep, size = ncon, prob = cell_probs_con2)
ne_case = rmultinom(n = nrep, size = ncase, prob = p_dmfc_ne)
ne_con = rmultinom(n = nrep, size = ncon, prob = cell_probs_ne2)

for (i in 1:nrep) {
cptrio = data.frame(snp = numeric(0), c1 = numeric(0), c2 = numeric(
   0), c3 = numeric(0), c4 = numeric(0), c5 = numeric(0),
c6 = numeric(0), c7 = numeric(0), c8 = numeric(0), c9 = numeric(0),
   c10 = numeric(0), c11 = numeric(0),
c12 = numeric(0), c13 = numeric(0), c14 = numeric(0), c15 = numeric(
   0))
```

```
cptrio[1,] = c(1, cases[, i])
cptrio[2,] = c(2, ne_case[, i])

conparents = data.frame(snp = numeric(0), c1 = numeric(0), c2 =
   numeric(0), c3 = numeric(0), c4 = numeric(0), c5 = numeric(0),
c6 = numeric(0), c7 = numeric(0), c8 = numeric(0), c9 = numeric(0))

conparents[1,] = c(1, cons[, i])
conparents[2,] = c(2, ne_con[, i])

# Write to file
cat("snp␣␣␣␣␣␣␣␣cellcount␣1-15\n", file = paste(file_path, "Env_",
   i, "_", "cptrio.dat", sep = ""))
write.table(cptrio, file = paste(file_path, "Env_", i, "_", "cptrio.
   dat", sep = ""),
sep = "␣", quote = FALSE, col.names = FALSE, row.names = FALSE,
   append = TRUE)

if (ncon > 0 ) {
cat("snp␣␣␣␣␣␣␣␣cellcount␣1-9\n", file = paste(file_path, "Env_", i,
    "_", "conparents.dat", sep = ""))
write.table(conparents, file = paste(file_path, "Env_", i, "_", "
   conparents.dat", sep = ""),
sep = "␣", quote = FALSE, col.names = FALSE, row.names = FALSE,
   append = TRUE)
}
}
}

## Conversion from EMIM input to LEM input ##

setwd(paste(file_path, "LEMEMIM_Equivalence/", sep = ""))
l = list.files()
for (file in l) {

q = paste("Z:/LEMEMIM_Equivalence/", file, sep = "")
df <- read.table(q, quote="\", comment.char="")

distr = c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
distr3 = c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)

for (i in 1:4000) {
if (i < 2001) {
if (df$V7[3 * i - 2] == "A" && df$V8[3 * i - 2] == "A") {
if (df$V7[3 * i - 1] == "A" && df$V8[3 * i - 1] == "A") {
if (df$V7[3 * i] == "A" && df$V8[3 * i] == "A") {
```

```
distr[15] = distr[15] + 1
}
}

if (df$V7[3 * i - 1] == "A" && df$V8[3 * i - 1] == "C") {
if (df$V7[3 * i] == "A" && df$V8[3 * i] == "A") {
distr[14] = distr[14] + 1
}
if (df$V7[3 * i] == "A" && df$V8[3 * i] == "C") {
distr[13] = distr[13] + 1
}
}
if (df$V7[3 * i - 1] == "C" && df$V8[3 * i - 1] == "C") {
distr[7] = distr[7] + 1
}
}
if (df$V7[3 * i - 2] == "A" && df$V8[3 * i - 2] == "C") {
if (df$V7[3 * i - 1] == "A" && df$V8[3 * i - 1] == "A") {
if (df$V7[3 * i] == "A" && df$V8[3 * i] == "A") {
distr[12] = distr[12] + 1
}
if (df$V7[3 * i] == "A" && df$V8[3 * i] == "C") {
distr[11] = distr[11] + 1
}
}
if (df$V7[3 * i - 1] == "A" && df$V8[3 * i - 1] == "C") {
if(df$V7[3 * i] == "A" && df$V8[3 * i] == "A") {
distr[10] = distr[10] + 1
}
if(df$V7[3 * i] == "A" && df$V8[3 * i] == "C") {
distr[9] = distr[9] + 1
}
if(df$V7[3 * i] == "C" && df$V8[3 * i] == "C") {
distr[8] = distr[8] + 1
}
}
if (df$V7[3 * i - 1] == "C" && df$V8[3 * i - 1] == "C") {
if (df$V7[3 * i] == "A" && df$V8[3 * i] == "C") {
distr[5] = distr[5] + 1
}
if (df$V7[3 * i] == "C" && df$V8[3 * i] == "C") {
distr[4] = distr[4] + 1
}
}

}
if (df$V7[3 * i - 2] == "C" && df$V8[3 * i - 2] == "C") {
```

```
if (df$V7[3 * i - 1] == "A" && df$V8[3 * i - 1] == "A") {
distr[6] = distr[6] + 1
}
if(df$V7[3 * i - 1] == "A" && df$V8[3 * i - 1] == "C") {
if (df$V7[3 * i] == "A" && df$V8[3 * i] == "C") {
distr[3] = distr[3] + 1
}
if (df$V7[3 * i] == "C" && df$V8[3 * i] == "C") {
distr[2] = distr[2] + 1
}
}
if (df$V7[3 * i - 1] == "C" && df$V8[3 * i - 1] == "C") {
distr[1] = distr[1] + 1
}
}
}

if (i > 2000) {
if (df$V7[3 * i - 2] == "A" && df$V8[3 * i - 2] == "A") {
if (df$V7[3 * i - 1] == "A" && df$V8[3 * i - 1] == "A") {
if (df$V7[3 * i] == "A" && df$V8[3 * i] == "A") {
distr3[15] = distr3[15] + 1
}
}
if (df$V7[3 * i - 1] == "A" && df$V8[3 * i - 1] == "C") {
if (df$V7[3 * i] == "A" && df$V8[3 * i] == "A") {
distr3[14] = distr3[14] + 1
}
if (df$V7[3 * i] == "A" && df$V8[3 * i] == "C") {
distr3[13] = distr3[13] + 1
}

}
if (df$V7[3 * i - 1] == "C" && df$V8[3 * i - 1] == "C") {
distr3[7] = distr3[7] + 1
}
}
if (df$V7[3 * i - 2] == "A" && df$V8[3 * i - 2] == "C") {
if (df$V7[3 * i - 1] == "A" && df$V8[3 * i - 1] == "A") {
if (df$V7[3 * i] == "A" && df$V8[3 * i] == "A") {
distr3[12] = distr3[12] + 1
}
if (df$V7[3 * i] == "A" && df$V8[3 * i] == "C") {
distr3[11] = distr3[11] + 1
}
}
if (df$V7[3 * i - 1] == "A" && df$V8[3 * i - 1] == "C") {
```

```
if(df$V7[3 * i] == "A" && df$V8[3 * i] == "A") {
distr3[10] = distr3[10] + 1
}
if(df$V7[3 * i] == "A" && df$V8[3 * i] == "C") {
distr3[9] = distr3[9] + 1
}
if(df$V7[3 * i] == "C" && df$V8[3 * i] == "C") {
distr3[8] = distr3[8] + 1
}
}
if (df$V7[3 * i - 1] == "C" && df$V8[3 * i - 1] == "C") {
if (df$V7[3 * i] == "A" && df$V8[3 * i] == "C") {
distr3[5] = distr3[5] + 1
}
if (df$V7[3 * i] == "C" && df$V8[3 * i] == "C") {
distr3[4] = distr3[4] + 1
}
}

}
if (df$V7[3 * i - 2] == "C" && df$V8[3 * i - 2] == "C") {
if (df$V7[3 * i - 1] == "A" && df$V8[3 * i - 1] == "A") {
distr3[6] = distr3[6] + 1
}
if(df$V7[3 * i - 1] == "A" && df$V8[3 * i - 1] == "C") {
if (df$V7[3 * i] == "A" && df$V8[3 * i] == "C") {
distr3[3] = distr3[3] + 1
}
if (df$V7[3 * i] == "C" && df$V8[3 * i] == "C") {
distr3[2] = distr3[2] + 1
}
}
if (df$V7[3 * i - 1] == "C" && df$V8[3 * i - 1] == "C") {
distr3[1] = distr3[1] + 1
}
}
}
}

snp = data.frame(case = integer(), M = integer(), P = integer(), C =
    integer(), freq = integer())

snp[1, ] = c(2, 3, 3, 3,distr[1])
snp[2, ] = c(2, 3, 2, 3, distr[2])
snp[3, ] = c(2, 3, 2, 2, distr[3])
snp[4, ] = c(2, 2, 3, 3, distr[4])
snp[5, ] = c(2, 2, 3, 2, distr[5])
```

```
snp[6, ] = c(2, 3, 1, 2, distr[6])
snp[7, ] = c(2, 1, 3, 2, distr[7])
snp[8, ] = c(2, 2, 2, 3, distr[8])
snp[9, ] = c(2, 2, 2, 2, distr[9])
snp[10, ] = c(2, 2, 2, 1, distr[10])
snp[11, ] = c(2, 2, 1, 2, distr[11])
snp[12, ] = c(2, 2, 1, 1,distr[12])
snp[13, ] = c(2, 1, 2, 2,distr[13])
snp[14, ] = c(2, 1, 2, 1, distr[14])
snp[15, ] = c(2, 1, 1, 1, distr[15])

snp[16, ] = c(1, 3, 3, 3, distr3[1])
snp[17, ] = c(1, 3, 2, 3, distr3[2])
snp[18, ] = c(1, 3, 2, 2, distr3[3])
snp[19, ] = c(1, 2, 3, 3, distr3[4])
snp[20, ] = c(1, 2, 3, 2, distr3[5])
snp[21, ] = c(1, 3, 1, 2, distr3[6])
snp[22, ] = c(1, 1, 3, 2, distr3[7])
snp[23, ] = c(1, 2, 2, 3, distr3[8])
snp[24, ] = c(1, 2, 2, 2, distr3[9])
snp[25, ] = c(1, 2, 2, 1, distr3[10])
snp[26, ] = c(1, 2, 1, 2, distr3[11])
snp[27, ] = c(1, 2, 1, 1, distr3[12])
snp[28, ] = c(1, 1, 2, 2, distr3[13])
snp[29, ] = c(1, 1, 2, 1, distr3[14])
snp[30, ] = c(1, 1, 1, 1, distr3[15])

i = which(l == file)
write.table(snp, file = paste(file_path, i, "_", "cptrio2.dat", sep
   = ""),
sep = "␣", quote = FALSE, col.names = FALSE, row.names = FALSE)
}

## Maternal Gene-Environment Interaction Test of Homogeneity ##

library(rmeta)
ge_test_9 = function(file_path) {

setwd(paste(file_path, "EnvNeg/9mu/RS/", sep = ""))
l = list.files()

for (file in l){
temp_dataset <-read.table(file, header=TRUE)
if(!exists("dataset")) {
dataset = temp_dataset
}
else{
```

```
dataset<-rbind(dataset, temp_dataset)
rm(temp_dataset)
}
}
DataNeg = dataset
rm(dataset)

setwd(paste(file_path, "EnvPlus/9mu/RS/", sep = ""))
l = list.files()

for (file in l){
temp_dataset <-read.table(file, header=TRUE)
if(!exists("dataset")) {
dataset = temp_dataset
}
else{
dataset<-rbind(dataset, temp_dataset)
rm(temp_dataset)
}
}

DataPlus = dataset
rm(dataset)

Data2 = rbind(DataPlus, DataNeg)
nad = c(rep("P", 1000), rep("N", 1000))
nums = c(1, 4, 5, 8, 9, 12, 13, 16, 17)
Data3 = subset(Data2, select = nums)

DataNeg2 = subset(DataNeg, select = nums)

snp1_neg = DataNeg2[DataNeg2$snp == 1,]
snp2_neg = DataNeg2[DataNeg2$snp == 2,]

DataPlus2 = subset(DataPlus, select = nums)

snp1_plus = DataPlus2[DataPlus2$snp == 1,]
snp2_plus = DataPlus2[DataPlus2$snp == 2,]

GE_summary1 =data.frame(ID = numeric(2), R1 = numeric(2), R2 =
   numeric(2), S1 = numeric(2), S2 = numeric(2))
GE_summary1$ID = c(1, 2)
for (j in 1:4) {
h1 = c()
h2 = c()

for (i in 1:1000) {
```

```
g = c(snp1_neg[i, 2*j], snp1_plus[i, 2*j], snp1_neg[i, 2*j+1], snp1_
   plus[i, 2*j+1])
if (!is.na(g[2]) && !is.na(g[4])) {
f = meta.summaries(d = g[1:2], se = g[3:4], logscale = TRUE, weights
    = c(1, 1))
h1 = c(h1, f[[6]][3])
}

g = c(snp2_neg[i, 2*j], snp2_plus[i, 2*j], snp2_neg[i, 2*j+1], snp2_
   plus[i, 2*j+1])
if (!is.na(g[2]) && !is.na(g[4])) {
f = meta.summaries(d = g[1:2], se = g[3:4], logscale = TRUE, weights
    = c(1, 1))
h2 = c(h2, f[[6]][3])
}

}
GE_summary1[ ,j+1] = c(sum(h1<0.05)/length(h1), sum(h2<0.05)/length(
   h2))
}
return(GE_summary1)
}
```

A.2 LEM Input

```
man 5              * allow 5 categorical predictor variables
                   *  ('manifest' variables to LEM)
dim 2 3 3 3 2     * dimension table for these 5 variables:
                   *  'pattern variable' followed by 4 predictors
lab D M P C E     * label the variables in order dimensions given:
                   *  D=disease M=mother P=father C=child E=Environment

* model

mod  DMPCE {cov(MP,5) * 5 mating type parameters (6 via default intercept)
            cov(D, 1) * 1 case/control parameter
            cov(DC,1)   * child log-additive genetic effect apply only to cases
            cov(DM,1)   * maternal log-additive genetic effect apply only to cases
           wei(DMPC)} * weight vector for structural zeros and offsets

* data format

rec 60              * expect 60 records in data file
rco                 * records contain a count
dat cptrio.dat           * specify file containing data

* design matrix and parameter specification
des [ 0 0 0  0 0 0  0 0 1   * mt=1; m,p=2,2
      0 0 0  0 0 1  0 1 0   * mt=2: m,p=2,1 or 1,2
      0 0 1  0 0 0  1 0 0   * mt=3: m,p=2,0 or 0,2
      0 0 0  0 1 0  0 0 0   * mt=4: m,p=1,1
      0 1 0  1 0 0  0 0 0   * mt=5: m,p=0,1 or 1,0
      0 1                          * Case/control parameter
      0 0 0  0 1 2                 * child effect: log-additive
      0 0 0  0 1 2 ]               * maternal effect: log-additive

*  values for weight vector

sta wei(DMPC)
   [1 0 0  1 1 0  0 1 0  1 1 0  1 2 1  0 1 1  0 1 0  0 1 1  0 0 1
    1 0 0  1 1 0  0 1 0  1 1 0  1 2 1  0 1 1  0 1 0  0 1 1  0 0 1]
```

Figure A.1: Example (truncated) input file for LEM (CEPG model with child and maternal effects considered)

Appendix B

Candidate Genes Study: Genes and SNP Results

B.1 Description of Genes

FLNB The filamin B gene is located on chromosome 3 and is a member of the filamin family. It encodes for the filamin B protein, an actin-binding protein that stabilizes the cytoskeleton of a cell. These act as scaffolding upon which intracellular signaling and protein trafficking pathways are organized. This cell adhesion gene is involved in the development of the skeleton before birth. FLNB mutations disrupt vertebral segmentation, joint formation, and skeletogenesis [26]. FLNB-related disorders (such as spondylocarpotarsal synostosis syndrome and Larsen syndrome), have cleft palate as a related phenotype [44], which gives a reason to investigate whether an association between maternal FLNB and CL/P may be present.

HIC1 Hypermethylated in cancer 1 (HIC1) is a gene located on chromosome 17 and acts as a growth regulatory and tumor suppressor gene. It encodes a zinc-finger transcription factor and is associated with Miller-Dieker lissencephaly syndrome (MDS) in both mice and humans [7] [21] . This syndrome is characterized, among other traits, by facial abnormalities. In studies of mice, researchers have found that HIC1 is expressed in the connective tissues of cranio-facial regions and may correlate with defective development of the nose, jaw, extremities, gastrointestinal tract and kidneys in MDS patients [17].

ZNF189 The third gene selected for inclusion in this candidate gene study that came from Jugessur et al. (2010) is the zinc finger protein 189 (ZNF189). ZNF189 encodes for a protein with 16 zinc fingers and it belongs to the Krüppel-associated box-containing group of zinc finger proteins [38].

MTHFR Methylenetetrahydrofolate reductase (MTHFR) is a gene on chromosome 1 which triggers the production of the MTHFR enzyme in the body. This enzyme plays a

large role in folate metabolism [40]. A common mutation in this gene (C677T) decreases the activity of the enzyme by 35 % among heterozygotes and 70 % among homozygotes for the risk mutation[14].

RFC1 The RFC family is the major transporter of folates in cells and tissue, as well as specialized tissue functions including transplacental transport of folates. This points to the fact that low levels of RFC1 can contribute to pathophysiological states associates with folate deficiency. In this study of oral clefts, folate deficiency is an important factor to study as folate is a well documented risk factor. This will be explained further in the subsequent section.

B.2 SNP-Level Results for Maternal Gene-Environment Interaction Tests

B.2.1 Maternal Gene-Smoking Interaction Results

Table B.1: EMIM Results

Gene	rs Number	Effect Estimate for No Smoking	Effect Estimate for Smoking	P-value for Heterogeneity	Q-Value
MTHFR	rs6696752	0.918	0.786	0.376	0.910
MTHFR	rs4846048	0.948	0.790	0.296	0.910
MTHFR	rs2274976	0.893	1.005	0.752	0.910
MTHFR	rs1476413	0.869	0.823	0.766	0.910
MTHFR	rs1801131	0.838	0.910	0.634	0.910
MTHFR	rs6541003	0.847	0.890	0.757	0.910
MTHFR	rs1801133	1.011	0.960	0.758	0.910
MTHFR	rs9651118	1.251	1.182	0.760	0.910
MTHFR	rs17367504	0.784	1.150	0.075	0.647
MTHFR	rs3737964	1.044	0.817	0.165	0.910
MTHFR	rs12404124	0.927	0.912	0.921	0.948
FLNB	rs11130605	1.104	1.116	0.952	0.966
FLNB	rs9828717	0.961	0.963	0.990	0.990
FLNB	rs839239	1.013	0.889	0.439	0.910
FLNB	rs839241	0.911	1.014	0.547	0.910
FLNB	rs1718459	1.021	0.931	0.578	0.910
FLNB	rs1718483	0.994	0.879	0.458	0.910
FLNB	rs1718481	1.029	1.123	0.604	0.910
FLNB	rs1623879	0.784	0.826	0.850	0.910
FLNB	rs839230	1.053	1.179	0.525	0.910
FLNB	rs12634644	0.922	0.884	0.817	0.910
FLNB	rs839225	0.874	0.912	0.857	0.910
FLNB	rs1658347	0.826	1.006	0.429	0.910

FLNB	rs9809315	1.089	1.193	0.622	0.910
FLNB	rs1718454	0.984	1.142	0.360	0.910
FLNB	rs1718480	1.028	1.151	0.487	0.910
FLNB	rs9822918	1.007	1.100	0.584	0.910
FLNB	rs13073391	1.012	1.168	0.417	0.910
FLNB	rs2033739	1.012	1.194	0.348	0.910
FLNB	rs939882	0.905	0.841	0.662	0.910
FLNB	rs6445945	1.090	0.985	0.521	0.910
FLNB	rs7638552	0.959	0.792	0.325	0.910
FLNB	rs2177153	1.015	1.244	0.267	0.910
FLNB	rs2259091	0.951	0.696	0.090	0.669
FLNB	rs7615893	0.942	0.780	0.332	0.910
FLNB	rs9880603	0.754	0.712	0.843	0.910
FLNB	rs12488642	0.854	0.775	0.574	0.910
FLNB	rs2362904	0.919	0.807	0.495	0.910
FLNB	rs2362907	0.927	0.815	0.505	0.910
FLNB	rs2001972	0.938	1.046	0.518	0.910
FLNB	rs4284952	1.105	1.071	0.846	0.910
FLNB	rs13095822	0.943	1.058	0.499	0.910
FLNB	rs4681795	1.125	1.065	0.732	0.910
FLNB	rs9884098	1.073	1.276	0.361	0.910
FLNB	rs6787425	0.960	0.926	0.853	0.910
FLNB	rs13092880	0.765	0.846	0.686	0.910
FLNB	rs7627013	0.928	0.763	0.329	0.910
FLNB	rs6445954	1.150	1.072	0.672	0.910
FLNB	rs10470697	0.844	0.859	0.915	0.948
ZNF189	rs2253258	0.825	1.175	0.190	0.910
ZNF189	rs10819926	1.019	0.917	0.538	0.910
ZNF189	rs3739719	1.045	0.973	0.674	0.910
ZNF189	rs546577	1.059	0.969	0.559	0.910
HIC1	rs11870150	0.847	1.033	0.242	0.910
RFC1	rs11702425	1.049	0.799	0.097	0.669
RFC1	rs1556329	1.196	1.126	0.830	0.910
RFC1	rs2236475	1.033	0.921	0.516	0.910
RFC1	rs3753019	1.138	0.840	0.072	0.647
RFC1	rs2236483	1.115	0.790	0.050	0.575
RFC1	rs12483553	0.942	1.033	0.726	0.910
RFC1	rs2838950	1.071	0.968	0.584	0.910
RFC1	rs7278425	0.949	1.190	0.321	0.910
RFC1	rs3788190	1.104	0.868	0.153	0.910
RFC1	rs2838956	1.182	0.828	0.037	0.511
RFC1	rs914232	1.151	0.809	0.035	0.511
RFC1	rs2330183	1.152	0.788	0.025	0.511
RFC1	rs1051266	1.163	0.801	0.032	0.511
RFC1	rs4819130	1.209	0.794	0.016	0.511

RFC1	rs3788205	0.854	1.009	0.338	0.910

Table B.2: LEM Results

Gene	rs Number	Effect Estimate for Maternal Effect	Effect Estimate for Interaction	P-Value for Interaction	Q-Value for Interaction
MTHFR	rs6696752	1.210	1.279	0.438	0.593
MTHFR	rs4846048	1.276	1.176	0.605	0.700
MTHFR	rs2274976	1.055	5.681	0.041	0.105
MTHFR	rs1476413	1.141	2.150	0.025	0.091
MTHFR	rs1801131	1.152	1.452	0.214	0.369
MTHFR	rs6541003	1.014	0.903	0.712	0.779
MTHFR	rs1801133	1.167	1.407	0.249	0.383
MTHFR	rs9651118	1.640	2.465	0.012	0.055
MTHFR	rs17367504	1.139	6.343	0.000	0.001
MTHFR	rs3737964	1.436	1.033	0.923	0.937
MTHFR	rs12404124	1.150	0.765	0.326	0.472
FLNB	rs11130605	1.282	1.211	0.522	0.639
FLNB	rs9828717	1.231	2.456	0.015	0.065
FLNB	rs839239	1.175	1.018	0.953	0.953
FLNB	rs839241	1.117	1.976	0.033	0.099
FLNB	rs1718459	1.177	1.130	0.665	0.746
FLNB	rs1718483	1.116	1.211	0.498	0.632
FLNB	rs1718481	1.023	1.403	0.220	0.370
FLNB	rs1623879	0.972	NA	0.000	0.004
FLNB	rs839230	1.180	1.761	0.069	0.149
FLNB	rs12634644	1.223	1.853	0.063	0.140
FLNB	rs839225	1.128	7.911	0.000	0.005
FLNB	rs1658347	1.120	6.972	0.001	0.007
FLNB	rs9809315	1.331	2.074	0.036	0.099
FLNB	rs1718454	1.108	1.306	0.367	0.517
FLNB	rs1718480	1.166	1.447	0.213	0.369
FLNB	rs9822918	1.014	1.453	0.195	0.354
FLNB	rs13073391	1.133	1.915	0.054	0.128
FLNB	rs2033739	1.153	2.086	0.031	0.099
FLNB	rs939882	1.192	1.648	0.120	0.244
FLNB	rs6445945	1.257	0.919	0.754	0.813
FLNB	rs7638552	1.251	2.249	0.036	0.099
FLNB	rs2177153	1.046	3.572	0.001	0.006
FLNB	rs2259091	1.122	1.264	0.456	0.605
FLNB	rs7615893	1.205	2.386	0.024	0.091
FLNB	rs9880603	1.002	NA	0.000	0.005
FLNB	rs12488642	1.077	1.492	0.181	0.338
FLNB	rs2362904	1.165	2.368	0.020	0.081
FLNB	rs2362907	1.165	2.698	0.009	0.044

FLNB	rs2001972	1.067	1.998	0.032	0.099
FLNB	rs4284952	1.284	0.958	0.892	0.919
FLNB	rs13095822	1.077	1.971	0.036	0.099
FLNB	rs4681795	1.323	0.834	0.504	0.632
FLNB	rs9884098	1.285	4.128	0.000	0.000
FLNB	rs6787425	1.259	3.280	0.002	0.012
FLNB	rs13092880	0.893	7.126	0.001	0.008
FLNB	rs7627013	1.169	2.203	0.057	0.131
FLNB	rs6445954	1.439	1.272	0.388	0.535
FLNB	rs10470697	0.978	1.189	0.528	0.639
ZNF189	rs2253258	1.025	5.146	0.002	0.012
ZNF189	rs10819926	1.274	1.329	0.328	0.472
ZNF189	rs3739719	1.487	1.130	0.670	0.746
ZNF189	rs546577	1.381	0.738	0.243	0.383
HIC1	rs11870150	0.929	1.649	0.084	0.176
RFC1	rs11702425	1.132	0.802	0.466	0.607
RFC1	rs1556329	1.535	3.741	0.041	0.105
RFC1	rs2236475	1.284	1.055	0.871	0.911
RFC1	rs3753019	1.538	0.625	0.124	0.244
RFC1	rs2236483	1.308	0.578	0.053	0.128
RFC1	rs12483553	1.224	6.249	0.002	0.012
RFC1	rs2838950	1.548	1.094	0.795	0.844
RFC1	rs7278425	1.076	5.134	0.000	0.005
RFC1	rs3788190	1.275	0.676	0.150	0.288
RFC1	rs2838956	1.386	0.732	0.250	0.383
RFC1	rs914232	1.328	0.747	0.283	0.425
RFC1	rs2330183	1.344	0.724	0.235	0.383
RFC1	rs1051266	1.334	0.869	0.609	0.700
RFC1	rs4819130	1.411	0.856	0.579	0.689
RFC1	rs3788205	1.085	2.382	0.008	0.042

B.2.2 Maternal Gene-Vitamin Interaction Results

Table B.3: EMIM Results

Gene	rs Number	Effect Estimate for No Smoking	Effect Estimate for Smoking	P-value for Heterogeneity	Q-Value
MTHFR	rs6696752	0.824	0.862	0.793	0.995
MTHFR	rs4846048	0.856	0.881	0.870	0.995
MTHFR	rs2274976	0.964	1.035	0.855	0.995
MTHFR	rs1476413	0.888	0.760	0.401	0.922
MTHFR	rs1801131	0.850	0.792	0.678	0.995
MTHFR	rs6541003	0.840	0.785	0.682	0.995
MTHFR	rs1801133	1.105	1.058	0.795	0.995
MTHFR	rs9651118	1.057	1.319	0.240	0.922

MTHFR	rs17367504	0.941	0.775	0.366	0.922
MTHFR	rs3737964	0.973	0.934	0.820	0.995
MTHFR	rs12404124	0.926	0.839	0.533	0.995
FLNB	rs11130605	1.090	1.121	0.868	0.995
FLNB	rs9828717	1.124	0.844	0.131	0.922
FLNB	rs839239	0.842	1.095	0.116	0.922
FLNB	rs839241	1.102	0.832	0.104	0.922
FLNB	rs1718459	0.890	1.082	0.234	0.922
FLNB	rs1718483	0.794	1.089	0.059	0.922
FLNB	rs1718481	1.177	1.001	0.343	0.995
FLNB	rs1623879	1.000	0.721	0.267	0.922
FLNB	rs839230	1.112	1.072	0.829	0.995
FLNB	rs12634644	0.984	0.883	0.555	0.995
FLNB	rs839225	1.019	0.790	0.295	0.922
FLNB	rs1658347	0.977	0.819	0.481	0.995
FLNB	rs9809315	1.255	1.022	0.252	0.922
FLNB	rs1718454	1.190	0.926	0.123	0.922
FLNB	rs1718480	1.188	0.969	0.217	0.922
FLNB	rs9822918	1.241	0.879	0.037	0.922
FLNB	rs13073391	1.282	0.919	0.056	0.922
FLNB	rs2033739	1.283	0.929	0.065	0.922
FLNB	rs939882	0.867	0.923	0.702	0.995
FLNB	rs6445945	0.891	1.095	0.193	0.922
FLNB	rs7638552	0.899	0.966	0.694	0.995
FLNB	rs2177153	1.328	1.016	0.142	0.922
FLNB	rs2259091	0.832	0.936	0.506	0.995
FLNB	rs7615893	0.917	0.930	0.936	0.995
FLNB	rs9880603	0.732	0.791	0.786	0.995
FLNB	rs12488642	0.817	0.880	0.659	0.995
FLNB	rs2362904	0.926	0.900	0.875	0.995
FLNB	rs2362907	0.924	0.917	0.966	0.995
FLNB	rs2001972	1.081	0.954	0.452	0.995
FLNB	rs4284952	0.967	1.122	0.360	0.922
FLNB	rs13095822	1.107	0.958	0.385	0.922
FLNB	rs4681795	0.996	1.097	0.548	0.995
FLNB	rs9884098	1.319	1.095	0.322	0.922
FLNB	rs6787425	0.946	0.974	0.875	0.995
FLNB	rs13092880	0.844	0.811	0.872	0.995
FLNB	rs7627013	0.839	0.931	0.584	0.995
FLNB	rs6445954	0.957	1.123	0.333	0.922
FLNB	rs10470697	0.850	0.876	0.846	0.995
ZNF189	rs2253258	0.804	0.965	0.484	0.995
ZNF189	rs10819926	0.885	1.042	0.348	0.922
ZNF189	rs3739719	0.894	1.095	0.235	0.922
ZNF189	rs546577	0.928	1.058	0.385	0.922

HIC1	rs11870150	0.979	0.842	0.388	0.922
RFC1	rs11702425	0.958	0.950	0.957	0.995
RFC1	rs1556329	1.414	1.028	0.253	0.922
RFC1	rs2236475	1.127	0.956	0.366	0.922
RFC1	rs3753019	1.015	0.983	0.848	0.995
RFC1	rs2236483	0.910	1.049	0.398	0.922
RFC1	rs12483553	0.994	0.980	0.957	0.995
RFC1	rs2838950	0.977	1.027	0.790	0.995
RFC1	rs7278425	1.086	1.029	0.810	0.995
RFC1	rs3788190	1.021	1.020	0.996	0.996
RFC1	rs2838956	1.050	1.059	0.961	0.996
RFC1	rs914232	1.018	1.029	0.951	0.995
RFC1	rs2330183	1.022	1.020	0.992	0.996
RFC1	rs1051266	1.010	1.024	0.934	0.995
RFC1	rs4819130	1.036	1.077	0.820	0.995
RFC1	rs3788205	0.947	0.846	0.526	0.995

Table B.4: LEM Results

Gene	rs Number	Effect Estimate for Maternal Effect	Effect Estimate for Interaction	P-Value for Interaction	Q-Value for Interaction
MTHFR	rs6696752	1.122	1.475	0.195	0.376
MTHFR	rs4846048	1.141	1.212	0.517	0.595
MTHFR	rs2274976	1.271	4.529	0.095	0.243
MTHFR	rs1476413	1.224	1.839	0.045	0.173
MTHFR	rs1801131	1.217	1.363	0.288	0.452
MTHFR	rs6541003	0.940	0.791	0.399	0.540
MTHFR	rs1801133	1.356	1.408	0.233	0.423
MTHFR	rs9651118	1.891	2.749	0.002	0.017
MTHFR	rs17367504	1.437	2.167	0.054	0.181
MTHFR	rs3737964	1.307	1.406	0.258	0.424
MTHFR	rs12404124	1.006	0.594	0.058	0.182
FLNB	rs11130605	1.435	1.352	0.283	0.452
FLNB	rs9828717	1.574	1.384	0.295	0.452
FLNB	rs839239	1.309	1.734	0.061	0.183
FLNB	rs839241	1.596	1.212	0.512	0.595
FLNB	rs1718459	1.393	1.783	0.048	0.174
FLNB	rs1718483	1.251	2.096	0.011	0.069
FLNB	rs1718481	1.078	1.018	0.946	0.946
FLNB	rs1623879	1.591	NA	0.000	0.009
FLNB	rs839230	1.374	1.380	0.257	0.424
FLNB	rs12634644	1.368	1.467	0.185	0.375
FLNB	rs839225	1.591	4.876	0.001	0.012
FLNB	rs1658347	1.777	4.806	0.004	0.031
FLNB	rs9809315	1.419	1.255	0.432	0.552

FLNB	rs1718454	1.192	0.805	0.414	0.549
FLNB	rs1718480	1.286	0.916	0.740	0.790
FLNB	rs9822918	1.185	0.917	0.735	0.790
FLNB	rs13073391	1.362	1.180	0.556	0.623
FLNB	rs2033739	1.407	1.218	0.487	0.582
FLNB	rs939882	1.231	1.241	0.432	0.552
FLNB	rs6445945	1.295	1.277	0.380	0.535
FLNB	rs7638552	1.590	2.098	0.019	0.098
FLNB	rs2177153	1.368	1.776	0.055	0.181
FLNB	rs2259091	1.512	1.934	0.026	0.112
FLNB	rs7615893	1.549	1.966	0.033	0.134
FLNB	rs9880603	1.041	5.842	0.002	0.017
FLNB	rs12488642	1.247	1.314	0.331	0.484
FLNB	rs2362904	1.527	1.787	0.066	0.189
FLNB	rs2362907	1.537	2.047	0.026	0.112
FLNB	rs2001972	1.217	1.347	0.302	0.453
FLNB	rs4284952	1.251	1.502	0.139	0.319
FLNB	rs13095822	1.239	1.320	0.337	0.484
FLNB	rs4681795	1.184	1.079	0.770	0.805
FLNB	rs9884098	1.549	1.642	0.121	0.297
FLNB	rs6787425	1.568	2.100	0.020	0.099
FLNB	rs13092880	1.167	7.227	0.000	0.009
FLNB	rs7627013	1.870	3.472	0.001	0.012
FLNB	rs6445954	1.576	1.482	0.161	0.337
FLNB	rs10470697	1.277	1.532	0.111	0.274
ZNF189	rs2253258	1.327	6.752	0.000	0.009
ZNF189	rs10819926	1.555	1.974	0.019	0.099
ZNF189	rs3739719	1.516	1.635	0.085	0.226
ZNF189	rs546577	1.231	1.025	0.923	0.936
HIC1	rs11870150	0.969	1.394	0.227	0.423
RFC1	rs11702425	1.374	1.394	0.248	0.424
RFC1	rs1556329	2.283	2.881	0.079	0.218
RFC1	rs2236475	1.492	1.199	0.560	0.623
RFC1	rs3753019	1.314	1.080	0.795	0.819
RFC1	rs2236483	1.255	1.208	0.489	0.582
RFC1	rs12483553	1.678	3.859	0.010	0.069
RFC1	rs2838950	1.456	1.581	0.157	0.337
RFC1	rs7278425	1.599	4.169	0.001	0.012
RFC1	rs3788190	1.119	1.091	0.744	0.789
RFC1	rs2838956	1.248	1.262	0.391	0.540
RFC1	rs914232	1.180	1.208	0.487	0.582
RFC1	rs2330183	1.197	1.229	0.447	0.561
RFC1	rs1051266	1.187	1.427	0.196	0.376
RFC1	rs4819130	1.227	1.377	0.243	0.424
RFC1	rs3788205	1.302	1.501	0.148	0.329

Bibliography

[1] Alan Agresti. Introduction: Distributions and Inference for Categorical Data. In *Categorical Data Analysis*, pages 1–35. Wiley-Blackwell, 2003.

[2] Holly F Ainsworth, Jennifer Unwin, Deborah L Jamison, and Heather J Cordell. Investigation of Maternal Effects, Maternal-Fetal Interactions and Parent-of-Origin Effects (Imprinting), Using Mothers and Their Offspring. *Genetic Epidemiology*, 35(1):19–45, January 2011.

[3] Terri H Beaty, Jeffrey C Murray, Mary L Marazita, Ronald G Munger, Ingo Ruczinski, Jacqueline B Hetmanski, Kung Yee Liang, Tao Wu, Tanda Murray, M Daniele Fallin, Richard A Redett, Gerald Raymond, Holger Schwender, Shin C Jin, Margaret E Cooper, Martine Dunnwald, Maria A Mansilla, Elizabeth Leslie, Stephen Bullard, Andrew C Lidral, Lina M Moreno, Renato Menezes, Alexandre R Vieira, Aline Petrin, Allen J Wilcox, Rolv T Lie, Ethylin W Jabs, Yah Huei Wu-Chou, Philip K Chen, Hong Wang, Xiaoqian Ye, Shangzhi Huang, Vincent Yeow, Samuel S Chong, Sun Ha Jee, Bing Shi, Kaare Christensen, Doheny Kimberly, W Pugh Elizabeth, Ling Hua, E Castilla Eduardo, Andrew E Czeizel, Lian Ma, L Leigh Field, Lawrence Brody, Faith Pangilinan, James L Mills, Anne M Molloy, Peadar N Kirke, John M Scott, Mauricio Arcos-Burgos, and Alan F Scott. A genome-wide association study of cleft lip with and without cleft palate identifies risk variants near MAFB and ABCA4. *Nature genetics*, 42(6):525–529, June 2010.

[4] Guntram Bezold, Monika Lange, and Ralf Uwe Peter. Homozygous Methylenetetrahydrofolate Reductase C677t Mutation and Male Infertility. *New England Journal of Medicine*, 344(15):1172–1173, April 2001.

[5] Mathieu Bourgey, Jasmine Healy, Pascal Saint-Onge, Hugues Massé, Daniel Sinnett, and Marie-Hélène Roy-Gagnon. Genome-wide detection and characterization of mating asymmetry in human populations. *Genetic Epidemiology*, 35(6):526–535, September 2011.

[6] Lon R Cardon and Lyle J Palmer. Population stratification and spurious allelic association. 361(9357):598–604.

[7] Mark G. Carter, Margaret A. Johns, Xiaobei Zeng, Li Zhou, M. Christine Zink, Joseph L. Mankowski, David M. Donovan, and Stephen B. Baylin. Mice deficient in the candidate tumor suppressor gene Hic1 exhibit developmental defects of structures affected in the Miller–Dieker syndrome. *Human Molecular Genetics*, 9(3):413–419, February 2000.

[8] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

[9] International HapMap Consortium et al. The international hapmap project. *Nature*, 426(6968):789, 2003.

[10] Michael J. Dixon, Mary L. Marazita, Terri H. Beaty, and Jeffrey C. Murray. Cleft lip and palate: understanding genetic and environmental influences. *Nature Reviews Genetics*, 12(3):167–178, March 2011.

[11] Marie-Therese Doolin, Sandrine Barbaux, Maeve McDonnell, Katy Hoess, Alexander S. Whitehead, and Laura E. Mitchell. Maternal genetic effects, exerted by genes involved in homocysteine remethylation, influence the risk of spina bifida. *The American Journal of Human Genetics*, 71(5):1222 – 1226, 2002.

[12] Evangelos Evangelou, Thomas A. Trikalinos, Georgia Salanti, and John P. A. Ioannidis. Family-Based versus Unrelated Case-Control Designs for Genetic Associations. *PLOS Genetics*, 2(8):e123, August 2006.

[13] David A. Freedman. A note on screening regression equations. *The American Statistician*, 37(2):152–155, 1983.

[14] P. Frosst, H. J. Blom, R. Milos, P. Goyette, C. A. Sheppard, R. G. Matthews, G. J. H. Boers, M. den Heijer, L. a. J. Kluijtmans, L. P. van den Heuve, and R. Rozen. A candidate genetic risk factor for vascular disease: a common mutation in methylenetetrahydrofolate reductase. *Nature Genetics*, 10(1):111–113, May 1995.

[15] Dinamar A. Gaspar, Sergio R. Matioli, Rita de Cássia Pavanello, Belmino C. Araújo, Nivaldo Alonso, Diego Wyszynski, and Maria Rita Passos-Bueno. Maternal MTHFR interacts with the offspring's BCL3 genotypes, but not with TGFA, in increasing risk to nonsyndromic cleft lip with or without cleft palate. *European Journal of Human Genetics*, 12(7):521–526, July 2004.

[16] Håkon K Gjessing and Rolv Terje Lie. Case-parent triads: Estimating single- and double-dose effects of fetal and maternal disease gene haplotypes. *Annals of Human Genetics*, 70(3):382–396, 2006.

[17] C. Grimm, R. Spörle, T. E. Schmid, I. D. Adler, J. Adamski, K. Schughart, and J. Graw. Isolation and embryonic expression of the novel mouse gene Hic1, the homologue of HIC1, a candidate gene for the Miller-Dieker syndrome. *Human Molecular Genetics*, 8(4):697–710, April 1999.

[18] Allan Hackshaw, Charles Rodeck, and Sadie Boniface. Maternal smoking in pregnancy and birth defects: a systematic review based on 173 687 malformed cases and 11.7 million controls. *Human Reproduction Update*, 17(5):589–604, September 2011.

[19] Jasmine Healy, Mathieu Bourgey, Chantal Richer, Daniel Sinnett, and Marie-Helene Roy-Gagnon. Detection of Fetomaternal Genotype Associations in Early-Onset Disorders: Evaluation of Different Methods and Their Application to Childhood Leukemia. *BioMed Research International*, 2010. DOI: 10.1155/2010/369534.

[20] P. Hintsanen, P. Sevon, P. Onkamo, L. Eronen, and H. Toivonen. An empirical comparison of case-control and trio based study designs in high throughput association mapping. *Journal of Medical Genetics*, 43(7):617–624, July 2006.

[21] Shinji Hirotsune, Svetlana D. Pack, Samuel S. Chong, Christiane M. Robbins, William J. Pavan, David H. Ledbetter, and Anthony Wynshaw-Boris. Genomic Organization of the Murine Miller–Dieker/Lissencephaly Region: Conservation of Linkage with the Human Region. *Genome Research*, 7(6):625–634, June 1997.

[22] Richard Howey and Heather J. Cordell. PREMIM and EMIM: tools for estimation of maternal, imprinting and interaction effects using multinomial modelling. *BMC Bioinformatics*, 13:149, June 2012.

[23] Astanand Jugessur, Min Shi, Håkon Kristian Gjessing, Rolv Terje Lie, Allen James Wilcox, Clarice Ring Weinberg, Kaare Christensen, Abee Lowman Boyles, Sandra Daack-Hirsch, Truc Trung Nguyen, Lene Christiansen, Andrew Carl Lidral, and Jeffrey Clark Murray. Maternal genes and facial clefts in offspring: A comprehensive search for genetic associations in two population-based cleft studies from scandinavia. *PLoS One*, 5(7), 2010.

[24] W C Knowler, R C Williams, D J Pettitt, and A G Steinberg. Gm3;5,13,14 and type 2 diabetes mellitus: an association in american indians with genetic admixture. 43(4):520–526.

[25] Peter Kraft, Christina G. S. Palmer, Arthur J. Woodward, Joni A. Turunen, Sonia Minassian, Tiina Paunio, Jouko Lönnqvist, Leena Peltonen, and Janet S. Sinsheimer.

RHD maternal–fetal genotype incompatibility and schizophrenia: extending the MFG test to include multiple siblings and birth order. 12(3):192–198.

[26] Deborah Krakow, Stephen P. Robertson, Lily M. King, Timothy Morgan, Eiman T. Sebald, Cristina Bertolotto, Sebastian Wachsmann-Hogiu, Dora Acuna, Sandor S. Shapiro, Toshiro Takafuta, Salim Aftimos, Chong Ae Kim, Helen Firth, Carlos E. Steiner, Valerie Cormier-Daire, Andrea Superti-Furga, Luisa Bonafe, John M. Graham Jr, Arthur Grix, Carlos A. Bacino, Judith Allanson, Martin G. Bialer, Ralph S. Lachman, David L. Rimoin, and Daniel H. Cohn. Mutations in the gene encoding filamin B disrupt vertebral segmentation, joint formation and skeletogenesis. *Nature Genetics*, 36(4):405–410, April 2004.

[27] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38:904–9, 09 2006.

[28] Nan M. Laird and Christoph Lange. *The Fundamentals of Modern Statistical Genetics*. Statistics for Biology and Health. Springer-Verlag, New York, 2011.

[29] Bhaskar Lakkakula, Jyotsna Murthy, and Venkatesh Babu Gurramkonda. Relationship between reduced folate carrier gene polymorphism and non-syndromic cleft lip and palate in Indian population. *The Journal of Maternal-Fetal & Neonatal Medicine: The Official Journal of the European Association of Perinatal Medicine, the Federation of Asia and Oceania Perinatal Societies, the International Society of Perinatal Obstetricians*, 28(3):329–332, February 2015.

[30] Julian Little, Amanda Cardy, and Ronald G. Munger. Tobacco smoking and oral clefts: a meta-analysis. *Bulletin of the World Health Organization*, 82:213–218, March 2004.

[31] Thomas Lumley. *rmeta: Meta-Analysis*, 2018. R package version 3.0.

[32] Iain L. MacDonald. Numerical Maximisation of Likelihood: A Neglected Alternative to EM?: Numerical Maximisation of Likelihood and EM. *International Statistical Review*, 82(2):296–308, August 2014.

[33] Mary L. Marazita, Andrew C. Lidral, Jeffrey C. Murray, L.Leigh Field, Brion S. Maher, Toby Goldstein McHenry, Margaret E. Cooper, Manika Govil, Sandra Daack-Hirsch, Bridget Riley, Astanand Jugessur, Temis Felix, Lina Morene, M.Adela Mansilla, Alexandre R. Vieira, Kim Doheny, Elizabeth Pugh, Consuelo Valencia-Ramirez, and Mauricio Arcos-Burgos. Genome Scan, Fine-Mapping, and Candidate Gene Analysis of Non-Syndromic Cleft Lip with or without Cleft Palate Reveals Phenotype-Specific Differences in Linkage and Association Results. *Human Heredity*, 68(3):151–170, July 2009.

[34] Mary L. Marazita, Jeffrey C. Murray, Andrew C. Lidral, Mauricio Arcos-Burgos, Margaret E. Cooper, Toby Goldstein, Brion S. Maher, Sandra Daack-Hirsch, Rebecca Schultz, M. Adela Mansilla, L. Leigh Field, You-e Liu, Natalie Prescott, Sue Malcolm, Robin Winter, Ajit Ray, Lina Moreno, Consuelo Valencia, Katherine Neiswanger, Diego F. Wyszynski, Joan E. Bailey-Wilson, Hasan Albacha-Hejazi, Terri H. Beaty, Iain McIntosh, Jacqueline B. Hetmanski, Gökhan Tunçbilek, Matthew Edwards, Louise Harkin, Rodney Scott, and Laurence G. Roddick. Meta-analysis of 13 genome scans reveals multiple cleft lip/palate genes with novel loci on 9q21 and 2q32-35. *American Journal of Human Genetics*, 75(2):161–173, August 2004.

[35] M. Martinelli, L. Scapoli, F. Pezzetti, F. Carinci, P. Carinci, G. Stabellini, L. Bisceglia, F. Gombos, and M. Tognon. C677t variant form at the MTHFR gene and CL/P: A risk factor for mothers? *American Journal of Medical Genetics*, 98(4):357–360, February 2001.

[36] James L. Mills, Peadar N. Kirke, Anne M. Molloy, Helen Burke, Mary R. Conley, Young Jack Lee, Philip D. Mayne, Donald G. Weir, and John M. Scott. Methylenete-

trahydrofolate reductase thermolabile variant and oral clefts. *American Journal of Medical Genetics*, 86(1):71–74, September 1999.

[37] Peter A Mossey, Julian Little, Ron G Munger, Mike J Dixon, and William C Shaw. Cleft lip and palate. *The Lancet*, 374(9703):1773–1785, November 2009.

[38] Jacob Odeberg, Øystein Røsok, Gudmundur H. Gudmundsson, Afshin Ahmadian, Leyla Roshani, Cecilia Williams, Catharina Larsson, Fredrik Pontén, Mathias Uhlén, Hans-Christian Åsheim, and Joakim Lundeberg. Cloning and Characterization of ZNF189, a Novel HumanKrüppel-like Zinc Finger Gene Localized to Chromosome 9q22–q31. *Genomics*, 50(2):213–221, June 1998.

[39] C. Y. Ou, R. E. Stevenson, V. K. Brown, C. E. Schwartz, W. P. Allen, M. J. Khoury, R. Rozen, G. P. Oakley, and M. J. Adams. 5,10 Methylenetetrahydrofolate reductase genetic polymorphism as a risk factor for neural tube defects. *American Journal of Medical Genetics*, 63(4):610–614, June 1996.

[40] F. Pezzetti, M. Martinelli, L. Scapoli, F. Carinci, A. Palmieri, J. Marchesini, P. Carinci, E. Caramelli, R. Rullo, F. Gombos, and M. Tognon. Maternal mthfr variant forms increase the risk in offspring of isolated nonsyndromic cleft lip with or without cleft palate. *Human Mutation*, 24(1):104–105, 2004.

[41] N. J. Prescott, R. M. Winter, and S. Malcolm. Maternal MTHFR genotype contributes to the risk of non-syndromic cleft lip and palate. *Journal of Medical Genetics*, 39(5):368–369, May 2002.

[42] Paola Primatesta, Emanuela Falaschetti, Sunjai Gupta, Michael G. Marmot, and Neil R. Poulter. Association between smoking and blood pressure. *Hypertension*, 37(2):187–193, 2001.

[43] Jiannis Ragoussis. Genotyping technologies for genetic research. *Annual Review of Genomics and Human Genetics*, 10(1):117–133, 2009. PMID: 19453250.

[44] Stephen Robertson. FLNB-Related Disorders. In Margaret P. Adam, Holly H. Ardinger, Roberta A. Pagon, Stephanie E. Wallace, Lora JH Bean, Karen Stephens, and Anne Amemiya, editors, *GeneReviews®*. University of Washington, Seattle, Seattle (WA), 1993.

[45] John M. Rogers. Tobacco and pregnancy. *Reproductive Toxicology (Elmsford, N.Y.)*, 28(2):152–160, September 2009.

[46] S.A.G.E. Statistical analysis for genetic epidemiology. *http://darwin.cwru.edu*, 2016.

[47] Ronnie Sebro, Thomas J. Hoffman, Christoph Lange, John J. Rogus, and Neil J. Risch. Testing for non-random mating: Evidence for ancestry-related assortative mating in the framingham heart study. 34(7):674–679.

[48] G. M Shaw, C. R Wasserman, C. D O'Malley, M. M Tolarova, and E. J Lammer. Risks of orofacial clefts in children born to women using multivitamins containing folic acid periconceptionally. *The Lancet*, 346(8972):393–396, August 1995.

[49] Gary M. Shaw, Huiping Zhu, Edward J. Lammer, Wei Yang, and Richard H. Finnell. Genetic Variation of Infant Reduced Folate Carrier (A80g) and Risk of Orofacial and Conotruncal Heart Defects. *American Journal of Epidemiology*, 158(8):747–752, October 2003.

[50] Janet S. Sinsheimer, Christina G.S. Palmer, and J. Arthur Woodward. Detecting genotype combinations that increase risk for disease: Maternal-fetal genotype incompatibility test. 24(1):1–13.

[51] Michael W. Smith, Nick Patterson, James A. Lautenberger, Ann L. Truelove, Gavin J. McDonald, Alicja Waliszewska, Bailey D. Kessing, Michael J. Malasky, Charles Scafe, Ernest Le, Philip L. De Jager, Andre A. Mignault, Zeng Yi, Guy de Thé, Myron Essex, Jean-Louis Sankalé, Jason H. Moore, Kwabena Poku, John P. Phair, James J. Goedert,

David Vlahov, Scott M. Williams, Sarah A. Tishkoff, Cheryl A. Winkler, Francisco M. De La Vega, Trevor Woodage, John J. Sninsky, David A. Hafler, David Altshuler, Dennis A. Gilbert, Stephen J. O'Brien, and David Reich. A high-density admixture map for disease gene discovery in african americans. 74(5):1001–1013.

[52] Behnoosh Soghani, Asghar Ebadifar, Hamid Reza Khorram Khorshid, Koorosh Kamali, Roya Hamedi, and Fatemeh Aghakhani Moghadam. The study of association between reduced folate carrier 1 (RFC1) polymorphism and non-syndromic cleft lip/palate in Iranian population. *BioImpacts : BI*, 7(4):263–268, 2017.

[53] Richard S Spielman, Ralph E McGinnis, and Warren J Ewens. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (iddm). *American journal of human genetics*, 52(3):506, 1993.

[54] John D. Storey. The positive false discovery rate: a bayesian interpretation and the q -value. *Ann. Statist.*, 31(6):2013–2035, 12 2003.

[55] Caroline G. Tai, Rebecca E. Graff, Jinghua Liu, Michael N. Passarelli, Joel A. Mefford, Gary M. Shaw, Thomas J. Hoffmann, and John S. Witte. Detecting Gene-Environment Interactions in Human Birth Defects: Study Designs and Statistical Methods. *Birth defects research. Part A, Clinical and molecular teratology*, 103(8):692–702, August 2015.

[56] Duncan C. Thomas and John S. Witte. Point: population stratification: a problem for case-control studies of candidate-gene associations? 11(6):505–512.

[57] David M. Umbach and Clarice R. Weinberg. Designing and analysing case-control studies to exploit independence of genotype and exposure. *Statistics in Medicine*, 16(15):1731–1743, August 1997.

[58] Nathalie M. J. van der Put, Fons Gabreëls, Erik M. B. Stevens, Jan A. M. Smeitink, Frans J. M. Trijbels, Tom K. A. B. Eskes, Lambert P. van den Heuvel, and Henk J.

Blom. A Second Common Mutation in the Methylenetetrahydrofolate Reductase Gene: An Additional Risk Factor for Neural-Tube Defects? *The American Journal of Human Genetics*, 62(5):1044–1051, May 1998.

[59] Iris A. L. M. van Rooij, Marga C. Ocké, Huub Straatman, Gerhard A. Zielhuis, Hans M. W. M. Merkus, and Régine P. M. Steegers-Theunissen. Periconceptional folate intake by supplement and food reduces the risk of nonsyndromic cleft lip with or without cleft palate. *Preventive Medicine*, 39(4):689–694, October 2004.

[60] S.H. Vermeulen, M. Shi, C.R. Weinberg, and D.M. Umbach. A hybrid design: case-parent triads supplemented by control-mother dyads. *Genetic epidemiology*, 33(2):136, February 2009.

[61] J.K. Vermunt. Lem: A general program for the analysis of categorical data. *Department of Methodology and Statistics, Tilburg University*, 1997.

[62] A.R. Vieira, M.E. Cooper, M.L. Marazita, E.E. Castilla, and I.M. Orioli. Reduced folate carrier 1 (RFC1) is associated with cleft of the lip only. *Brazilian Journal of Medical and Biological Research*, 41:689 – 693, 08 2008.

[63] Hansong Wang, Christopher A. Haiman, Laurence N. Kolonel, Brian E. Henderson, Lynne R. Wilkens, Loic Le Marchand, and Daniel O. Stram. Self-reported ethnicity, genetic structure and the impact of population stratification in a multiethnic study. *Human Genetics*, 128(2):165–177, Aug 2010.

[64] C. R. Weinberg and D. M. Umbach. A hybrid design for studying genetic influences on risk of diseases with onset early in life. 77(4):627–636.

[65] C R Weinberg, A J Wilcox, and R T Lie. A log-linear approach to case-parent-triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. 62(4):969–978.

[66] Allen J. Wilcox, Clarice R. Weinberg, and Rolv Terje Lie. Distinguishing the effects of maternal and offspring genes through studies of "case-parent triads". 148(9):893–901.

[67] Jason B. Wolf, Edmund D. Brodie III, James M. Cheverud, Allen J. Moore, and Michael J. Wade. Evolutionary consequences of indirect genetic effects. 13(2):64–69.

[68] Jingyuan Yang and Shili Lin. Robust partial likelihood approach for detecting imprinting and maternal effects using case-control families. *Ann. Appl. Stat.*, 7(1):249–268, 03 2013.