TriLLIEM: Triad Log-Linear modelling of Imprinting, Environmental interactions, and Maternal effects

Kevin Zhao

Thesis submitted to the University of Ottawa in partial fulfillment of the requirements for the degree of

Master of Science Mathematics and Statistics*

Department of Mathematics and Statistics

Faculty of Science

University of Ottawa

*The M.Sc. program is a joint program with Carleton University, administered by the Ottawa-Carleton Institute of Mathematics and Statistics

Abstract

The case-parents design, where an affected child and their parents are genotyped, allows for estimation of disease risk due to either the child's or mother's genes through log-linear modelling. It is robust to the confounding effects of population subdivision, but cannot account for non-exchangeability of parental genotypes. Factors such as gene-environment interactions and imprinting, where risk depends on which parent contributed a genetic factor, can also be estimated. Existing analytical software are either deprecated (LEM) or have limited modelling capabilities (**Haplin**, **EMIM**). We introduce our R package, **TriLLIEM**, which implements the log-linear approach to address these limitations. The software includes options for gene-environment interactions, imprinting by environment interactions, and non-exchangeable parental genotypes if parents of controls are also available. We use these features in a simulation study to assess the accuracy of **TriLLIEM** and to show how population stratification confounds the model for gene-environment interactions when control-triads are included.

Acknowledgements

I would like to thank my supervisor, Dr. Kelly Burkett, who has provided me with so much guidance and support throughout my time at the University of Ottawa. You helped guide me through so many fascinating problems and challenges in the field of statistical genetics, and have always put my development as a student and researcher first. Thank you for always supporting and encouraging me as I work towards my dreams and aspirations.

I would also like to thank the members of our lab, who provided me a comfortable and welcoming environment as I first entered this new field. This work would not have been possible without everyone's expertise and feedback as I progressed through my research.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The goal of gene-disease association studies is to identify genetic risk factors for disease. These genetic risk factors can act through the child's genetic makeup. There are also examples of the mother's genetic makeup affecting her child in the womb [10, 15], and examples where the child's genetic risk of disease depended on which parent contributed the risk factor to the child [5]. To capture the relevant genetic effects while avoiding confounding due to a phenomenon called population stratification, we use a study design that samples and genotypes triads composed of case children and their parents. Weinberg et al. [27] proposed a log-linear model for testing various genetic effects and estimating their relative risk using triad data. This model has been expanded to allow for gene-environment interactions [23] and control-triads for modelling various assumptions about population mating patterns [25]. However, an official software package for their log-linear model was never released, and the existing software (**Haplin** [7], **EMIM** [11]) have limitations in their features when accounting for gene-environment interactions.

In this thesis, we present **TriLLIEM**: Triad Log-Linear modelling of Imprinting, Environmental interactions, and Maternal effects, a freely available R package for fitting the log-linear model for trio data. **TriLLIEM** uses the generalized linear model (glm) object in R, to estimate the risk parameters and standard errors, which can then be used for the

construction of confidence intervals and for significance testing. The package is designed to be comprehensive and flexible; it allows users to specify precisely which effects to include in their model, directly includes gene-environment effects in the log-linear model, and analyzes data both with and without controls.

In Chapter 2, we provide a brief overview of the genetic concepts required to understand the statistical methods used throughout this thesis. We then describe the case-triad design, where genotype data is collected on a sample of cases from the population, and their parents. The inclusion of parents is advantageous as we can model genetic effects acting through the parents. This includes maternal effects, where the fetal environment affects the child's disease risk (e.g. spina bifida [10]), maternal gene-environment interactions, where environmental stimuli during pregnancy impact maternal effects (e.g. consumption of phenylalanine in mothers with phenylketonuria [15]), and imprinting effects, where diseases could be caused by genes activated based on parent-of-origin (e.g. orofacial clefts [5]). Finally, we describe the Weinberg group's log-linear model [23, 25–27], a statistical model for estimation of genetic effects using triad data. The parameters estimated by the model can be interpreted as relative risk for disease based on a given genetic factor. Fitting a model with imprinting requires the use of an Expectation-Maximization algorithm, which is also summarized in this section. We also describe the hybrid model for triad data where control trios composed of a non-affected child and their parents are recruited. With the hybrid design, the assumption of a population with mating symmetry can be relaxed to acommodate non-exchangeable parental genotypes [25]; however, the approach loses its robustness to population stratification.

With a thorough understanding of the log-linear model, in Chapter 3 we describe how this model has been implemented in various software package. The oldest of these packages is LEM [24], released in 1997, a general purpose log-linear modelling software which implemented the Expectation-Maximization (EM) algorithm to handle missing category counts while accounting for the uncertainty due to the missing counts [4, 14, 16]. LEM would be sufficient to capture all the features of the log-linear model, but unfortunately it is now

deprecated and does not run on modern operating systems. **Haplin** [7], released in 2006, is an R package that fits log-linear models to trio data. It was designed to estimate multi-SNP (haplotype effects) but it can also be used for simpler single-SNP models. **Haplin** is convenient and accessible to most researchers as an R package, but makes very strong assumptions on how the genotypes of mothers and fathers are distributed in a population, which can be violated in real-world populations [2]. **Haplin** offers many of the features of the log-linear model, albeit with a modified parameterization for imprinting, but is limited to a stratified approach for gene-environment interaction testing. **EMIM** [11], released in 2012, is another software package that can be used to estimate genetic and imprinting effects with triad data. The **EMIM** approach directly estimates the parameters of the multinomial model using a direct search algorithm. While **EMIM** incorporates many of the features of the log-linear model and is very fast, it is a standalone software package written in Fortran and some users may need to compile before being able to use it; this can be a barrier to those who are less technologically oriented. In addition, results must be imported into R or other statistical software for subsequent processing. **EMIM** is also limited in that it does not include an option for estimation of interaction effects; instead, users must write their own scripts to estimate and perform statistical tests on environmental interactions, via the same stratified approach used in **Haplin**. We aimed to address the limitations of all these software packages while preserving the advantages of the log-linear model: **TriLLIEM** is an R package that fits the log-linear model to genotype data from triads. It offers a robust set of options for users, such as the assumptions they wish to make about the distribution of parental genotypes in the population, and the ability to easily acquire point estimates and perform significance tests on effects in the model using an interface that will be familiar to R `glm()` function users. **TriLLIEM** can directly include gene by environment interactions in the model, unlike **EMIM** and **Haplin** which require fitting the model on the data stratified by the environment variable; this best exploits the features of the log-linear model without over-parameterizing the model. However, to allow for equivalent estimation approaches

as **Haplin** and **EMIM**, both the stratified and non-stratified methods are implemented in **TriLLIEM**.

In Chapter 4, we conducted two simulation studies, where we used **TriLLIEM** to create many simulated datasets. First, we assessed the accuracy of **TriLLIEM** by comparing its results to those of **EMIM** and **Haplin** under a wide assortment of conditions. We found that **TriLLIEM**, **EMIM** and **Haplin** were able to produce essentially identical results (ignoring minor floating point differences) in all applicable conditions. We also compared the performance of the stratified and non-stratified methods for environmental interactions in **TriLLIEM**, and found that the non-stratified method had higher power than the stratified method when testing for environmental interactions. In our second simulation study, we examined model performance when sampling from a population composed of two hidden sub-populations, with distinct demographic parameters such as disease prevalence and environmental exposure rates. Data was simulated and analyzed in **TriLLIEM**, and we found that in the hybrid design, the model's Type 1 error rate at a siginificance level of $\alpha < 0.05$ rose to more than 10% for main effects. Type 1 error for gene-environment interactions remained within the 5% – 6.5% range.

Last of all, in Chapter 5, we summarize and discuss the results from this thesis, and provide an outline of future work that could be done to improve the log-linear methods in **TriLLIEM**.

Chapter 2

Background to Gene-Disease Associations

In this chapter, we discuss the background necessary to understand the theory behind the log-linear model **TriLLIEM** uses and the analyses in this thesis. In Section 2.1, we introduce the basic principles of genetics and inheritance, which are the foundations of statistical genetics. We then explain methods for gathering data to study associations between genes and diseases in Section 2.2, and lastly, introduce the theory behind the log-linear model in Section 2.3.

2.1 Foundations of Genetics

The human genome comprises billions of nucleotides, which are transmitted from parent to child [17]. This genetic material is typically stored as 23 pairs of homologous chromosomes in humans, with each individual inheriting one set of 23 chromosomes from each parent. Each chromosome holds two strands of DNA, whose nucleotides pair up with one another. Variations in the sequence of nucleotides at specific regions (loci) of the genome can lead to differences in biology. An example of a locus is shown in Figure 2.1. Researchers focus on loci believed to code for particular traits, such as blood type or the synthesis of a specific

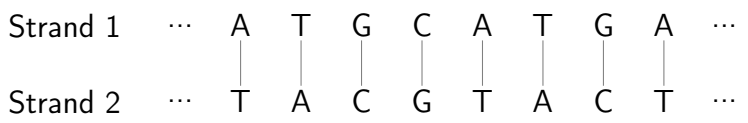


Figure 2.1: Example of nucleotides (A, T, G, C) at a locus in DNA. Individuals may be categorized by their nucleotide sequence at such loci to find associations with diseases.

protein in a cell. We refer to different variants of a locus as alleles, and in the context of this thesis, we assume that a locus of interest has been identified. We only consider loci having two alleles observed in the general population. The allele that is less frequent in the population is called the minor allele. Since each individual receives half of their genetic material from either parent, each locus is paired with one from an individual's mother and the other from their father. We measure the *genotype* (genetic information for the locus of interest) by the count of minor alleles the individual has (0, 1, or 2).

2.1.1 Inheritance

As mentioned, each parent contributes one allele at each locus to a child, which forms the basis of inheritance. We can quantify this with the Mendelian inheritance model, based on the research of Gregor Mendel in 1865. In Mendelian inheritance, since each parent has two copies of their genes (received from the respective grandparents of the child), we assume both copies are equally likely to be transmitted to the offspring, and that the genes transmitted maternally and paternally are independent [22]. With this information, we may construct a probability distribution for a gene of an offspring based solely on parental alleles. Letting M, F, and C represent the number of copies of an allele of interest a mother, father, and child have, respectively, the marginal distribution of C given (M, F) is shown in Table 2.1.

Table 2.1: Marginal distribution of a child's genotype (C) based on the genotype of their mother (M) and father (F).

(M, F)	$\Pr(C = 0 \mid M, F)$	$\Pr(C = 1 \mid M, F)$	$\Pr(C = 2 \mid M, F)$
$(0, 0)$	1	0	0
$(0, 1)$	$1/2$	$1/2$	0
$(1, 0)$	$1/2$	$1/2$	0
$(0, 2)$	0	1	0
$(2, 0)$	0	1	0
$(1, 1)$	$1/4$	$1/2$	$1/4$
$(1, 2)$	0	$1/2$	$1/2$
$(2, 1)$	0	$1/2$	$1/2$
$(2, 2)$	0	0	1

2.2 Studies on Gene-Disease Associations

An individual's genetic makeup has been shown to influence disease risk, especially in congenital diseases such as spina bifida [10]. A common approach for discovering gene-disease associations is the case-control study. In a case-control study, individuals with and without the disease are recruited, and we compare the frequencies of genotypes between cases and controls. However, case-control studies are susceptible to confounding from hidden population substructures (e.g. population stratification) and are unable to detect maternal effects.

Population stratification occurs when the population is composed of genetically distinct sub-populations, such as when groups are descending from various ancestries [9]. This leads to spurious associations [3] when one sub-population has a higher disease prevalence than another while simultaneously having a larger frequency for the minor allele. In the case-control study, the case group would oversample individuals from the former sub-population, while the control group would oversample individuals from the latter sub-population, leading to an incorrect conclusion that the larger prevalence of the minor allele in the case group is what increases disease risk.

Maternal effects, where the mother's genotype impacts the child's disease risk, have

been observed empirically in congenital diseases associated with the fetal environment. For example, the genetic disorder phenylketonuria (PKU) in a pregnant mother can lead to heart disease in the child, hence there is a maternal effect where the variant of the allele coding for PKU in the mother increases the child's risk of disease [15]. In the case of spina bifida, the disease is associated with folate deficiency during pregnancy, and it was found that alongside the genes of the fetus, the mother's genes involved in the metabolic pathway for synthesizing folate affect the risk of spina bifida in the child [10].

These considerations motivate the use of family-based genetic association studies, where genotype data is collected on multiple family members. Family-based designs are more robust to confounding by population stratification. In addition, maternal effects and parent-of-origin effects can be estimated if parents are collected. The case-triad design is a type of family-based sampling design where an affected child and both parents are sampled and genotype. With this study design, the Weinberg group developed a log-linear approach [25, 27, 28] to analyze the data. This approach has been extended to test for child and maternal effects, alongside various other genetically plausible factors, and to include control-trios, where the child is not known to be affected with the disease.

Although trio-based designs are more robust to confounding by population stratification, they are not immune to the problem. In particular, testing for gene-environment interactions or including control-trios can lead to scenarios where spurious associations are identified. When an environmental variable's status is considered (e.g. when the child's sex could influence genetic effects) with sub-populations that have different environmental variable frequencies (e.g. one sub-population has proportionally more men than another), we would oversample different sub-populations in each environmental strata, leading to spurious associations when disease prevalence also varies between groups [23].

2.3 The Log-Linear Model

Assume we genotype a sample of case-triads, leading to 15 (M, F, C) categories (first column in Table 2.3). The probabilities for being in one of these 15 categories can be modelled using a multinomial distribution. We use the relationship between the multinomial distribution and the Poisson distribution when we condition on a fixed sample size n to motivate the use of the log-linear model. Given k independent Poisson variables with means $\lambda_1, \ldots, \lambda_k$, the conditional distribution of the counts $Y_1, \ldots, Y_k$ given $\sum_{i=1}^{k} Y_i = Y = n$ is

$$\begin{aligned}
\Pr(Y_1 = n_1, \ldots, Y_k = n_k \mid Y = n) &= \frac{\Pr(Y_1 = n_1, \ldots, Y_k = n_k)}{\Pr(Y = n)}, \\
&= \frac{\prod_{i=1}^{n} e^{-\lambda_i} \lambda_i^{n_i} / n_i!}{e^{-\sum_{i=1}^{k} \lambda_i} \left(\sum_{i=1}^{k} \lambda_i\right)^n / n!}, \\
&= \frac{n!}{\prod_{i=1}^{k} n_i!} \prod_{i=1}^{k} \left(\frac{\lambda_i}{\sum_{j=1}^{k} \lambda_j}\right)^{n_i},
\end{aligned}$$

which is a multinomial distribution of sample size n and k categories, with respective probabilities $\pi_i = \lambda_i / \sum_{j=1}^{k} \lambda_j$ for $i \in \{1, \ldots, k\}$ [12]. Assuming only case-triads, we wish to model the counts in each category with probabilities $\pi_{\mathrm{M,F,C}} = \Pr(\mathrm{M, F, C} \mid \mathrm{D} = 1)$, where D is a binary variable representing the child's disease status. We can write

$$\Pr(\mathrm{M, F, C} \mid \mathrm{D} = 1) = \frac{\Pr(\mathrm{D} = 1 \mid \mathrm{M, F, C}) \Pr(\mathrm{C} \mid \mathrm{M, F}) \Pr(\mathrm{M, F})}{\Pr(\mathrm{D} = 1)}. \tag{2.3.1}$$

Taking the natural logarithm of both sides, we can express this as a log-linear model,

$$\ln(\pi_{\mathrm{M,F,C}}) = \ln\left(\Pr(\mathrm{D} = 1 \mid \mathrm{M, F, C})\right) + \ln\left(\Pr(\mathrm{C} \mid \mathrm{M, F})\right) + \ln\left(\Pr(\mathrm{M, F})\right) - \ln\left(\Pr(\mathrm{D} = 1)\right). \tag{2.3.2}$$

The terms in this model are readily interpretable. $\Pr(\mathrm{D} = 1 \mid \mathrm{M, F, C})$ is the probability of disease based on the trio's genotypes; we describe how to model this term in Section 2.3.2. $\Pr(\mathrm{C} \mid \mathrm{M, F})$ is the Mendelian inheritance probabilities (given in Table 2.1); these probability

values are known constants and so this term is represented by an "offset" when fitting the log-linear model. $\Pr(\mathrm{M}, \mathrm{F})$ is the probability of parental genotypes, which we call mating type probability, whose distribution is given in Section 2.3.1. The final term, $\Pr(\mathrm{D} = 1) = \alpha$, is the prevalence of disease.

2.3.1 Mating Type Models

To express the mating type probabilities, we must provide a mating type model which captures our assumptions about how genotypes from paired individuals are distributed in the population, independent of disease status. We outline three population mating type models, from weakest to strongest in terms of assumptions: Hardy-Weinberg Equilibrium (HWE), Mating Symmetry (MS), and Mating Asymmetry (MaS).

The HWE model assumes independence between the parental genotypes (random mating) and independent inheritance of the two alleles making up each parent's genotype. Letting p represent the frequency of the minor allele, $1 - p$ is the frequency of the other allele (assuming a biallelic SNP locus). Then a population under HWE has genotype probabilities $(1-p)^2$, $2p(1-p)$, and p^2 for individuals with 0, 1, and 2 copies of the minor allele. Assuming the distributions of M and F are identical and independent, $\Pr(\mathrm{M}, \mathrm{F}) = \Pr(\mathrm{M}) \Pr(\mathrm{F})$. For example, a population under the HWE model would have $\Pr(\mathrm{M} = 1, \mathrm{F} = 1) = (2p(1-p))^2 = 4p^2(1-p)^2$ for the proportion of parents who each have 1 copy of the variant allele.

In the MS model, we relax our assumptions from the HWE model to only require that $\Pr(\mathrm{M} = a, \mathrm{F} = b) = \Pr(\mathrm{M} = b, \mathrm{F} = a)$ for all $a, b \in \{0, 1, 2\}$. Hence, knowing the proportion of $(\mathrm{M}, \mathrm{F}) = (a, b)$ only tells us the proportion of $(\mathrm{M}, \mathrm{F}) = (b, a)$, and not the probabilities of any other mating combinations. There are 6 unique combinations of (M, F), hence we introduce μ_i for $i \in \{0, \ldots, 5\}$ as mating type parameters to represent $\Pr(\mathrm{M}, \mathrm{F})$.

Lastly, in the MaS model, we allow $\Pr(\mathrm{M} = a, \mathrm{F} = b) \neq \Pr(\mathrm{M} = b, \mathrm{F} = a)$, hence we introduce ν_j for $j \in \{0, \ldots, 8\}$ as 9 mating asymmetry parameters for the 9 distinct

(M, F) tuples. Unaccounted mating asymmetry can lead to spurious maternal effects. For instance, when $\Pr(\mathrm{M} > \mathrm{F}) > \Pr(\mathrm{M} < \mathrm{F})$, case-triad samples will oversample triads with $\mathrm{M} > \mathrm{F}$, leading to overestimated maternal effects. Evidence of mating asymmetry in human populations has been previously documented [2, 8].

We summarize the proportions of every (M, F) category under each mating type model in Table 2.2.

Table 2.2: $\Pr(\mathrm{M}, \mathrm{F})$ in each given mating type model.

Mating type: (M, F)		$\Pr(\mathrm{M}, \mathrm{F})$		
		HWE	MS	MaS
0:				
	$(2, 2)$	p^4	μ_0	ν_0
1:				
	$(2, 1)$	$2p^3(1-p)$	μ_1	ν_1
	$(1, 2)$	$2p^3(1-p)$	μ_1	ν_2
2:				
	$(2, 0)$	$p^2(1-p)^2$	μ_2	ν_3
	$(0, 2)$	$p^2(1-p)^2$	μ_2	ν_4
3:				
	$(1, 1)$	$4p^2(1-p)^2$	μ_3	ν_5
4:				
	$(1, 0)$	$2p(1-p)^3$	μ_4	ν_6
	$(0, 1)$	$2p(1-p)^3$	μ_4	ν_7
5:				
	$(0, 0)$	$(1-p)^4$	μ_5	ν_8

2.3.2 Expressing Disease Risk Based on Genotype

For our log-linear model in Equation (2.3.2), $\Pr(\mathrm{D} = 1 \mid \mathrm{M}, \mathrm{F}, \mathrm{C})$ encompasses how the variant allele could influence the child's risk for disease. Based on our discussion in Section 2.2, we can introduce child and maternal effects by using the multiplicative model

$$\Pr(\mathrm{D} = 1 \mid \mathrm{M}, \mathrm{F}, \mathrm{C}) = \kappa R^{\mathrm{C}} S^{\mathrm{M}},$$

where κ is the baseline probability of disease when $\mathrm{M} = \mathrm{C} = 0$ (and is uninformative), and R and S can be interpreted as the relative risk parameters representing the child and maternal effects, respectively. This multiplicative model allows us to conveniently estimate the R and S relative risk parameters with the log-linear model, since substituting this into Equation (2.3.2) gives

$$\ln\left(\pi_{\mathrm{M,F,C}}\right) = \ln(\kappa/\alpha) + \ln\left(\Pr(\mathrm{C} \mid \mathrm{M}, \mathrm{F})\right) + \ln\left(\Pr(\mathrm{M}, \mathrm{F})\right) + r\mathrm{C} + s\mathrm{M}, \tag{2.3.3}$$

where $r = \ln(R)$ and $s = \ln(S)$ are parameters to be estimated by the model. While other models for these effects can be accomodated, such as a model for a recessive variant allele, where risk of disease is only modified when the appropriate individual has 2 copies of the variant allele, in this thesis we focus on the multiplicative model only. Pulling all components together, Table 2.3 summarizes the cell probabilities for each $(\mathrm{M}, \mathrm{F}, \mathrm{C})$ category. Taking the natural log of the cell probabilities yields Equation (2.3.3).

2.3.3 Extensions to the Model

The log-linear model allows us to make several modifications to the study design and/or the model relatively intuitively. For example, we could extend the model to account for maternal and paternal imprinting (also called parent-of-origin effects), where risk of disease is conditional on which parent passed the minor allele to their child [26]. Imprinting is known to impact the expression of various alleles, which is suspected to play a role in the development of orofacial clefts for instance, another congenital disease [5, 19]. We can write the log-linear model with imprinting as follows

$$\ln\left(\pi_{\mathrm{M,F,C,O}}\right) = \ln(\kappa/\alpha) + \ln\left(\Pr(\mathrm{C} \mid \mathrm{M}, \mathrm{F})\right) + \ln\left(\Pr(\mathrm{M}, \mathrm{F})\right) + r\mathrm{C} + s\mathrm{M} + i_{\mathrm{M}} I_{\{\text{M-derived}\}} + i_{\mathrm{F}} I_{\{\text{F-derived}\}},$$

Table 2.3: Multinomial distribution for the MS model with maternal and child effects.

Triad genotype category: (M, F, C)	Pr(M, F, C | D)
0:	
(2, 2, 2)	$\kappa\mu_0 R^2 S^2/\alpha$
1:	
(2, 1, 2)	$\kappa\mu_1 R^2 S^2/(2\alpha)$
(2, 1, 1)	$\kappa\mu_1 R S^2/(2\alpha)$
(1, 2, 2)	$\kappa\mu_1 R^2 S/(2\alpha)$
(1, 2, 1)	$\kappa\mu_1 R S/(2\alpha)$
2:	
(2, 0, 1)	$\kappa\mu_2 R S^2/\alpha$
(0, 2, 1)	$\kappa\mu_2 R/\alpha$
3:	
(1, 1, 2)	$\kappa\mu_3 R^2 S/(4\alpha)$
(1, 1, 1)	$\kappa\mu_3 R S/(2\alpha)$
(1, 1, 0)	$\kappa\mu_3 S/(4\alpha)$
4:	
(1, 0, 1)	$\kappa\mu_4 R S/(2\alpha)$
(1, 0, 0)	$\kappa\mu_4 S/(2\alpha)$
(0, 1, 1)	$\kappa\mu_4 R/(2\alpha)$
(0, 1, 0)	$\kappa\mu_4/(2\alpha)$
5:	
(0, 0, 0)	$\kappa\mu_5/\alpha$

where O is a categorical variable representing the originator(s) of the child's minor allele(s), and $I_{\text{M-derived}}$ and $I_{\text{F-derived}}$ are dummy variables which are 1 if the child inherited a variant allele from the mother or father, respectively, and 0 otherwise. The model is not statistically identifiable when child effects and both imprinting effects are included, as $\text{C} = I_{\{\text{M-derived}\}} + I_{\{\text{F-derived}\}}$, hence r, i_{M}, and i_{F} are collinear. To address this, we impose that in any of these models, at most two of R, I_{M}, and I_{F} may be estimated simultaneously. We can use $I_{\text{M}} = e^{i_{\text{M}}}$ and $I_{\text{F}} = e^{i_{\text{F}}}$ to represent the relative risk of disease from maternal and paternal imprinting, respectively. Knowing (M, F, C) is sufficient to determine O in all circumstances except when $\text{M} = \text{F} = \text{C} = 1$, in which case the counts for this category contain both maternally and paternally inherited variant alleles. Since an allele's parent-of-origin is often not reported in case-triad studies, determining the number of these counts

which are maternally and paternally inherited becomes a missing data problem. Weinberg et al. [27] use the Expectation-Maximization (EM) algorithm to handle these missing counts, which we summarize as follows:

1. Augment the observed data matrix by dividing the $(\mathrm{M}, \mathrm{F}, \mathrm{C}) = (1, 1, 1)$ category into a category for the counts of maternal parent of origin and a category for paternal parent of origin. Set an initial value for $m_p = I_{\mathrm{M}} / (I_{\mathrm{M}} + I_{\mathrm{F}})$, the proportion of $(1, 1, 1)$ counts which are maternally inherited. For example, if there was no effect due to parental origin, then $I_{\mathrm{M}} = I_{\mathrm{F}} = 1$ and $m_p = 1/2$.

2. Letting $n_{1,1,1}$ be the observed number of $(\mathrm{M}, \mathrm{F}, \mathrm{C}) = (1, 1, 1)$ triads in our study, $n_{1,1,1} \cdot m_p$ and $n_{1,1,1} \cdot (1 - m_p)$ are the expected counts of maternally and paternally inherited $(\mathrm{M}, \mathrm{F}, \mathrm{C}) = (1, 1, 1)$ counts (Initial expectation step).

3. Assuming these expected counts are the true counts, fit the log-linear model to obtain the maximum likelihood estimates of the model parameters, including the imprinting parameters, $\hat{I}_{\mathrm{M}}$ and $\hat{I}_{\mathrm{F}}$ (Maximization).

4. Use the parameters estimated from the previous step to update $m_p = \hat{I}_{\mathrm{M}} / \left(\hat{I}_{\mathrm{M}} + \hat{I}_{\mathrm{F}} \right)$ and the expected counts (Expectation).

5. Repeat steps 3–4 until m_p converges.

In simulations, this process was not sensitive to initial values of m_p and usually converged within 12 iterations [27]. If we are only modelling one of the imprinting parameters, we only need to modify this procedure by setting the other parameter to 1, as we are assuming that it contributes no risk for disease.

Another extension to the log-linear model is gene-environment interactions [12]. Recalling the example of PKU, it was hypothesized that maternal PKU led to child disease outcomes since the mother's diet contained phenylalanine before and during pregnancy, which

cannot be properly broken down by those with PKU [15]. Statistical methods are therefore needed to assess whether maternal exposure to phenylalanine during pregnancy AND maternal PKU status are the true cause of this increased risk. With the addition of information on environmental exposure status, E, we modify our multinomial probability by taking

$$\begin{aligned}\Pr(\mathrm{M},\mathrm{F},\mathrm{C},\mathrm{E} \mid \mathrm{D}=1) &= \frac{\Pr(\mathrm{D}=1 \mid \mathrm{M},\mathrm{F},\mathrm{C},\mathrm{E})\Pr(\mathrm{C} \mid \mathrm{M},\mathrm{F},\mathrm{E})\Pr(\mathrm{M},\mathrm{F},\mathrm{E})}{\Pr(\mathrm{D}=1)} \\ &= \frac{\Pr(\mathrm{D}=1 \mid \mathrm{M},\mathrm{F},\mathrm{C},\mathrm{E})\Pr(\mathrm{C} \mid \mathrm{M},\mathrm{F})\Pr(\mathrm{M},\mathrm{F},\mathrm{E})}{\Pr(\mathrm{D}=1)}.\end{aligned}$$

We must make the assumption that C and E are conditionally independent given (M, F) in order to have $\Pr(\mathrm{C} \mid \mathrm{M},\mathrm{F},\mathrm{E}) = \Pr(\mathrm{C} \mid \mathrm{M},\mathrm{F})$ in the above equation [23]. In practice, this is often the case as we examine environmental interactions with parents, with exposure having occurred while the child was in the womb (such as with maternal PKU). We use $\Pr(\mathrm{M},\mathrm{F},\mathrm{E})$ as mating type parameters could vary between the exposed and unexposed sub-populations. This is modelled as an interaction between E and the mating type variable. Inclusion of this interaction ensures that the log-linear model is not susceptible to the confounding effects of population stratification [21]. We model an interaction term between the maternal genotype and the environmental exposure (alongside child and maternal effects) in the log-linear model by including an E by M interaction. The corresponding log-linear model when there are maternal effect by environment interactions is

$$\ln\left(\pi_{\mathrm{M,F,C,E}}\right) = \ln(\kappa/\alpha) + \ln\left(\Pr(\mathrm{C} \mid \mathrm{M},\mathrm{F})\right) + \ln\left(\Pr(\mathrm{M},\mathrm{F} \mid \mathrm{E})\right) + r\mathrm{C} + s\mathrm{M} + v\mathrm{E}\cdot\mathrm{M},$$

We can interpret $V = e^{v}$ as the relative risk of disease for each additional copy of the minor allele in the mother when the environmental exposure is present. Note that the environmental interaction need not be with maternal effects, for example, we could examine an imprinting by environment interaction as well or a child genetic effect by environment interaction. The EM algorithm for imprinting would remain unchanged, only we would require extending the

algorithm for the missing $(\mathrm{M}, \mathrm{F}, \mathrm{C}) = (1, 1, 1)$ counts when $E = 1$.

Lastly, the Weinberg group [25] extended the log-linear model to a hybrid model that includes control families. They only genotype the parents of controls, as the child's genotype is uninformative [1, 11, 25], though we equivalently include the child's genotype. To model control-triads, we make a rare disease assumption, and hence

$$\pi_{\mathrm{M,F}} = \Pr(\mathrm{M}, \mathrm{F} \mid \mathrm{D} = 0) \approx \Pr(\mathrm{M}, \mathrm{F}).$$

An example of such a hybrid model would be

$$\ln(\pi_{\mathrm{M,F,C,D}}) = \ln(\kappa/\alpha) \cdot \mathrm{D} + \ln(\Pr(\mathrm{C} \mid \mathrm{M}, \mathrm{F})) + \ln(\Pr(\mathrm{M}, \mathrm{F})) + r\mathrm{C} \cdot \mathrm{D} + s\mathrm{M} \cdot \mathrm{D},$$

where $\pi_{\mathrm{M,F,C,D}} = \Pr(\mathrm{M}, \mathrm{F}, \mathrm{C} \mid \mathrm{D})$. The hybrid model is useful to model mating asymmetry [25] because the 9 parameter mating type model is identifiable only when control parents are available. Using a mating type model that fails to account for mating asymmetry has been shown to lead to spurious conclusions about maternal effects [8]. By including control-triads, however, the model again becomes vulnerable to population stratification, for the same reasons as with case-control studies.

All of these extensions can be simultaneously implemented in the log-linear model. For example, we can write a hybrid model for child, maternal imprinting, and maternal imprinting by environment effects as

$$\begin{aligned}\ln(\pi_{\mathrm{M,F,C,O,E,D}}) = \ln(\kappa/\alpha) \cdot \mathrm{D} + \ln(\Pr(\mathrm{C} \mid \mathrm{M}, \mathrm{F})) + \ln(\Pr(\mathrm{M}, \mathrm{F} \mid \mathrm{E})) + r\mathrm{C} \cdot \mathrm{D} \\ + i_{\mathrm{M}} I_{\{\text{M-derived}\}} \cdot \mathrm{D} + v\mathrm{E} \cdot I_{\{\text{M-derived}\}} \cdot \mathrm{D}. \quad (2.3.4)\end{aligned}$$

We give an abbreviated example of the multinomial distribution for the model in Equation (2.3.4) in Table 2.4.

Table 2.4: Examples of probabilities in the multinomial distributions for $\Pr(\mathrm{M},\mathrm{F},\mathrm{C},\mathrm{O},\mathrm{E}\mid\mathrm{D})$ based on the model in Equation (2.3.4).

					$\Pr(\mathrm{M},\mathrm{F},\mathrm{C},\mathrm{O},\mathrm{E}\mid\mathrm{D})$		
Mating type	(M, F, C)	O	E	D	HWE	MS	MaS
0	(2, 2, 2)	M, F	1	1	$\kappa p_{\mathrm{E}=1}^4 R^2V^2I_{\mathrm{M}}/\alpha$	$\kappa\mu_{0,\mathrm{E}=1}R^2V^2I_{\mathrm{M}}/\alpha$	$\kappa\nu_{0,\mathrm{E}=1}R^2V^2I_{\mathrm{M}}/\alpha$
1	(2, 1, 2)	M, F	1	1	$\kappa p_{\mathrm{E}=1}^3(1-p_{\mathrm{E}=1})R^2V^2I_{\mathrm{M}}/\alpha$	$\kappa\mu_{1,\mathrm{E}=1}R^2V^2I_{\mathrm{M}}/(2\alpha)$	$\kappa\nu_{1,\mathrm{E}=1}R^2V^2I_{\mathrm{M}}/(2\alpha)$
1	(2, 1, 1)	M	1	1	$\kappa p_{\mathrm{E}=1}^3(1-p_{\mathrm{E}=1})RV^2I_{\mathrm{M}}/\alpha$	$\kappa\mu_{1,\mathrm{E}=1}RV^2I_{\mathrm{M}}/(2\alpha)$	$\kappa\nu_{1,\mathrm{E}=1}RV^2I_{\mathrm{M}}/(2\alpha)$
1	(1, 2, 2)	M, F	1	1	$\kappa p_{\mathrm{E}=1}^3(1-p_{\mathrm{E}=1})R^2VI_{\mathrm{M}}/\alpha$	$\kappa\mu_{1,\mathrm{E}=1}R^2VI_{\mathrm{M}}/(2\alpha)$	$\kappa\nu_{2,\mathrm{E}=1}R^2VI_{\mathrm{M}}/(2\alpha)$
1	(1, 2, 1)	F	1	1	$\kappa p_{\mathrm{E}=1}^3(1-p_{\mathrm{E}=1})RV/\alpha$	$\kappa\mu_{1,\mathrm{E}=1}RV/(2\alpha)$	$\kappa\nu_{2,\mathrm{E}=1}RV/(2\alpha)$
2	(2, 0, 1)	M	1	1	$\kappa p_{\mathrm{E}=1}^2(1-p_{\mathrm{E}=1})^2RV^2I_{\mathrm{M}}/\alpha$	$\kappa\mu_{1,\mathrm{E}=1}RV^2I_{\mathrm{M}}/\alpha$	$\kappa\nu_{3,\mathrm{E}=1}RV^2I_{\mathrm{M}}/\alpha$
2	(0, 2, 1)	F	1	1	$\kappa p_{\mathrm{E}=1}^2(1-p_{\mathrm{E}=1})^2R/\alpha$	$\kappa\mu_{1,\mathrm{E}=1}R/\alpha$	$\kappa\nu_{4,\mathrm{E}=1}R/\alpha$
3	(1, 1, 2)	M, F	1	1	$\kappa p_{\mathrm{E}=1}^2(1-p_{\mathrm{E}=1})^2R^2VI_{\mathrm{M}}/\alpha$	$\kappa\mu_{3,\mathrm{E}=1}R^2VI_{\mathrm{M}}/(4\alpha)$	$\kappa\nu_{5,\mathrm{E}=1}R^2VI_{\mathrm{M}}/(4\alpha)$
3	(1, 1, 1)	M	1	1	$\kappa p_{\mathrm{E}=1}^2(1-p_{\mathrm{E}=1})^2RVI_{\mathrm{M}}/\alpha$	$\kappa\mu_{3,\mathrm{E}=1}RVI_{\mathrm{M}}/(4\alpha)$	$\kappa\nu_{5,\mathrm{E}=1}RVI_{\mathrm{M}}/(4\alpha)$
3	(1, 1, 1)	F	1	1	$\kappa p_{\mathrm{E}=1}^2(1-p_{\mathrm{E}=1})^2RV/\alpha$	$\kappa\mu_{3,\mathrm{E}=1}RV/(4\alpha)$	$\kappa\nu_{5,\mathrm{E}=1}RV/(4\alpha)$
3	(1, 1, 0)		1	1	$\kappa p_{\mathrm{E}=1}^2(1-p_{\mathrm{E}=1})^2V/\alpha$	$\kappa\mu_{3,\mathrm{E}=1}V/(4\alpha)$	$\kappa\nu_{5,\mathrm{E}=1}V/(4\alpha)$
4	(1, 0, 1)	M	1	1	$\kappa p_{\mathrm{E}=1}(1-p_{\mathrm{E}=1})^3RVI_{\mathrm{M}}/\alpha$	$\kappa\mu_{4,\mathrm{E}=1}RVI_{\mathrm{M}}/(2\alpha)$	$\kappa\nu_{6,\mathrm{E}=1}RVI_{\mathrm{M}}/(2\alpha)$
4	(1, 0, 0)		1	1	$\kappa p_{\mathrm{E}=1}(1-p_{\mathrm{E}=1})^3V/\alpha$	$\kappa\mu_{4,\mathrm{E}=1}V/(2\alpha)$	$\kappa\nu_{6,\mathrm{E}=1}V/(2\alpha)$
4	(0, 1, 1)	F	1	1	$\kappa p_{\mathrm{E}=1}(1-p_{\mathrm{E}=1})^3R/\alpha$	$\kappa\mu_{4,\mathrm{E}=1}R/(2\alpha)$	$\kappa\nu_{7,\mathrm{E}=1}R/(2\alpha)$
4	(0, 1, 0)		1	1	$\kappa p_{\mathrm{E}=1}(1-p_{\mathrm{E}=1})^3/\alpha$	$\kappa\mu_{4,\mathrm{E}=1}/(2\alpha)$	$\kappa\nu_{7,\mathrm{E}=1}/(2\alpha)$
5	(0, 0, 0)		1	1	$\kappa(1-p_{\mathrm{E}=1})^4/\alpha$	$\kappa\mu_{5,\mathrm{E}=1}/\alpha$	$\kappa\nu_{8,\mathrm{E}=1}/\alpha$
0	(2, 2, 2)	M, F	0	1	$\kappa p_{\mathrm{E}=0}^4R^2I_{\mathrm{M}}/\alpha$	$\kappa\mu_{0,\mathrm{E}=0}R^2I_{\mathrm{M}}/\alpha$	$\kappa\nu_{0,\mathrm{E}=0}R^2I_{\mathrm{M}}/\alpha$
1	(2, 1, 2)	M, F	0	1	$\kappa p_{\mathrm{E}=0}^3(1-p_{\mathrm{E}=0})R^2I_{\mathrm{M}}/\alpha$	$\kappa\mu_{1,\mathrm{E}=0}R^2I_{\mathrm{M}}/(2\alpha)$	$\kappa\nu_{1,\mathrm{E}=0}R^2I_{\mathrm{M}}/(2\alpha)$
1	(2, 1, 1)	M	0	1	$\kappa p_{\mathrm{E}=0}^3(1-p_{\mathrm{E}=0})RI_{\mathrm{M}}/\alpha$	$\kappa\mu_{1,\mathrm{E}=0}RI_{\mathrm{M}}/(2\alpha)$	$\kappa\nu_{1,\mathrm{E}=0}RI_{\mathrm{M}}/(2\alpha)$
1	(1, 2, 2)	M, F	0	1	$\kappa p_{\mathrm{E}=0}^3(1-p_{\mathrm{E}=0})R^2I_{\mathrm{M}}/\alpha$	$\kappa\mu_{1,\mathrm{E}=0}R^2I_{\mathrm{M}}/(2\alpha)$	$\kappa\nu_{2,\mathrm{E}=0}R^2I_{\mathrm{M}}/(2\alpha)$
1	(1, 2, 1)	F	0	1	$\kappa p_{\mathrm{E}=0}^3(1-p_{\mathrm{E}=0})R/\alpha$	$\kappa\mu_{1,\mathrm{E}=0}R/(2\alpha)$	$\kappa\nu_{2,\mathrm{E}=0}R/(2\alpha)$
2	(2, 0, 1)	M	0	1	$\kappa p_{\mathrm{E}=0}^2(1-p_{\mathrm{E}=0})^2RI_{\mathrm{M}}/\alpha$	$\kappa\mu_{1,\mathrm{E}=0}RI_{\mathrm{M}}/\alpha$	$\kappa\nu_{3,\mathrm{E}=0}RI_{\mathrm{M}}/\alpha$
2	(0, 2, 1)	F	0	1	$\kappa p_{\mathrm{E}=0}^2(1-p_{\mathrm{E}=0})^2R/\alpha$	$\kappa\mu_{1,\mathrm{E}=0}R/\alpha$	$\kappa\nu_{4,\mathrm{E}=0}R/\alpha$
3	(1, 1, 2)	M, F	0	1	$\kappa p_{\mathrm{E}=0}^2(1-p_{\mathrm{E}=0})^2R^2I_{\mathrm{M}}/\alpha$	$\kappa\mu_{3,\mathrm{E}=0}R^2I_{\mathrm{M}}/(4\alpha)$	$\kappa\nu_{5,\mathrm{E}=0}R^2I_{\mathrm{M}}/(4\alpha)$
3	(1, 1, 1)	M	0	1	$\kappa p_{\mathrm{E}=0}^2(1-p_{\mathrm{E}=0})^2RI_{\mathrm{M}}/\alpha$	$\kappa\mu_{3,\mathrm{E}=0}RI_{\mathrm{M}}/(4\alpha)$	$\kappa\nu_{5,\mathrm{E}=0}RI_{\mathrm{M}}/(4\alpha)$
3	(1, 1, 1)	F	0	1	$\kappa p_{\mathrm{E}=0}^2(1-p_{\mathrm{E}=0})^2R/\alpha$	$\kappa\mu_{3,\mathrm{E}=0}R/(4\alpha)$	$\kappa\nu_{5,\mathrm{E}=0}R/(4\alpha)$
3	(1, 1, 0)		0	1	$\kappa p_{\mathrm{E}=0}^2(1-p_{\mathrm{E}=0})^2/\alpha$	$\kappa\mu_{3,\mathrm{E}=0}/(4\alpha)$	$\kappa\nu_{5,\mathrm{E}=0}/(4\alpha)$
4	(1, 0, 1)	M	0	1	$\kappa p_{\mathrm{E}=0}(1-p_{\mathrm{E}=0})^3RI_{\mathrm{M}}/\alpha$	$\kappa\mu_{4,\mathrm{E}=0}RI_{\mathrm{M}}/(2\alpha)$	$\kappa\nu_{6,\mathrm{E}=0}RI_{\mathrm{M}}/(2\alpha)$
4	(1, 0, 0)		0	1	$\kappa p_{\mathrm{E}=0}(1-p_{\mathrm{E}=0})^3/\alpha$	$\kappa\mu_{4,\mathrm{E}=0}/(2\alpha)$	$\kappa\nu_{6,\mathrm{E}=0}/(2\alpha)$
4	(0, 1, 1)	F	0	1	$\kappa p_{\mathrm{E}=0}(1-p_{\mathrm{E}=0})^3R/\alpha$	$\kappa\mu_{4,\mathrm{E}=0}R/(2\alpha)$	$\kappa\nu_{7,\mathrm{E}=0}R/(2\alpha)$
4	(0, 1, 0)		0	1	$\kappa p_{\mathrm{E}=0}(1-p_{\mathrm{E}=0})^3/\alpha$	$\kappa\mu_{4,\mathrm{E}=0}/(2\alpha)$	$\kappa\nu_{7,\mathrm{E}=0}/(2\alpha)$
5	(0, 0, 0)		0	1	$\kappa(1-p_{\mathrm{E}=0})^4/\alpha$	$\kappa\mu_{5,\mathrm{E}=0}/\alpha$	$\kappa\nu_{8,\mathrm{E}=0}/\alpha$
0	(2, 2, 2)	M, F	1	0	$p_{\mathrm{E}=1}^4$	$\mu_{0,\mathrm{E}=1}$	$\nu_{0,\mathrm{E}=1}$
					$\vdots$		
5	(0, 0, 0)		1	0	$(1-p_{\mathrm{E}=1})^4$	$\mu_{5,\mathrm{E}=1}$	$\nu_{8,\mathrm{E}=1}$
0	(2, 2, 2)	M, F	0	0	$p_{\mathrm{E}=0}^4$	$\mu_{0,\mathrm{E}=0}$	$\nu_{0,\mathrm{E}=0}$
					$\vdots$		
5	(0, 0, 0)		0	0	$(1-p_{\mathrm{E}=0})^4$	$\mu_{5,\mathrm{E}=0}$	$\nu_{8,\mathrm{E}=0}$

Chapter 3

Computational Methods

In this chapter, we discuss the features and limitations of the currently available software packages for modelling case-triad data, **Haplin** and **EMIM**. We then introduce our software package, **TriLLIEM**, which we designed to overcome these limitations. We show that **TriLLIEM**, **EMIM**, and **Haplin** produce essentially identical statistical inferences when equivalent models are fit, to justify the validity of **TriLLIEM**'s results. We also highlight models that can be fit in **TriLLIEM** but not **Haplin** and/or **EMIM**.

3.1 Modelling Software

Methods of fitting the parameters in the multinomial and log-linear models that we describe in Section 2.3 have been described in several papers [1, 23, 25–28]. Unfortunately, no official software for the log-linear model proposed by the Weinberg group was ever made. Instead, the Weinberg papers [23, 25–27] suggested LEM [24] for their analyses. LEM was a DOS program released in 1997 that allowed users to fit their own log-linear models and run the EM algorithm; however, it is now deprecated. Two other programs, **Haplin** [7] and **EMIM** [11], have been developed to fit some of the log-linear and/or multinomial models for trio-based analyses.

3.1.1 Haplin

Gjessing et al.'s **Haplin** [7], released in 2006, is an R package that fits the log-linear model. It extends the model to allow for multiple alleles and haplotypes, though for our work, we only examine its performance in the biallelic case. A major limitation of **Haplin** is that it only fits the model with HWE mating type frequencies (see Table 2.2). **Haplin** can model case-triad data with maternal effects, gene-environment interactions, and parent-of-origin effects simultaneously [6, 7, 13]. It also allows for optional independent control-triads [13]. However, since **Haplin** is limited to a HWE mating type model, these controls only allow us to estimate main effects corresponding to the environmental variable [13]. **Haplin** makes use of the `glm()` function in R [18] for point and interval estimation, as well as Wald-based p-values [7].

Haplin's implementation of environmental interactions uses a stratified approach through its `haplinStrat()` and `gxe()` functions [7, 13]. We briefly describe how **Haplin** performs this approach for the case of two exposure states. `haplinStrat()` partitions the data by E and independently fits the log-linear model to each partitioned data set. This gives us results for the respective fitted log-linear models, as well as the model for the entire data set without environmental effects. `gxe()` takes the output from `haplinStrat()` and performs a Wald chi-squared significance test for whether child, maternal, and/or imprinting effects differ between the two E strata (when the respective effect is in the model). To perform this Wald test, suppose we wish to compare the estimates for some effect in the log-linear model (e.g., $r = \ln(R)$) across the E = 0 and 1 strata, which we label $\beta_{\mathrm{E}=0}$ and $\beta_{\mathrm{E}=1}$ respectively. As maximum likelihood estimates, $\hat{\beta}_{\mathrm{E}=0}$ and $\hat{\beta}_{\mathrm{E}=1}$ asymptotically follow normal distributions with respective means $\beta_{\mathrm{E}=0}, \beta_{\mathrm{E}=1}$ and variances σ_0^2, σ_1^2. These distributions are not correlated as the environmental strata are independent. We are testing

$$H_0 : \beta_{\mathrm{E}=1} - \beta_{\mathrm{E}=0} = 0, H_1 : \beta_{\mathrm{E}=1} - \beta_{\mathrm{E}=0} \neq 0,$$

with the Wald test statistic $w = \frac{\left(\hat{\beta}_{\mathrm{E}=1} - \hat{\beta}_{\mathrm{E}=0}\right)^2}{\hat{\sigma}_0^2 + \hat{\sigma}_1^2}$ asymptotically following a $\chi^2(1)$ distribution under the null hypothesis. Point estimates and standard deviations are extracted from the fitted log-linear models, allowing us to use w for significance tests [6].

For parent-of-origin effects, **Haplin** uses a similar parameterization to the one described in Weinberg et al. [27] that we describe in Section 2.3.3. However, instead of allowing up to two of C, I_{M}, and I_{F} to be estimated, **Haplin** does not fit child effects and always estimates both I_{M} and I_{F} when parent-of-origin effects are enabled [6]. Fitting both parent imprinting effects gives us information about the child's risk of disease, given they have the minor allele, based on the parent-of-origin; hence, not having child effects is not a major drawback, other than leading to a different model interpretation. However, not being able to force one of these imprinting effects to be null could impact the model's power. **Haplin** uses the EM algorithm (described in Section 2.3.3) to estimate parental transmission in the $(\mathrm{M}, \mathrm{F}, \mathrm{C}) = (1, 1, 1)$ category, and also uses it to handle other missing genotype data (e.g., when we only have mother-child dyads without data for the father) [6]. **Haplin** can also perform parent-of-origin by environment interaction tests by comparing the relative risk ratio $I_{\mathrm{M}}/I_{\mathrm{F}}$ between the environmental strata. This is done by testing

$$H_0 : \ln\left(\frac{I_{\mathrm{M,E}=0}}{I_{\mathrm{F,E}=0}}\right) = \ln\left(\frac{I_{\mathrm{M,E}=1}}{I_{\mathrm{F,E}=1}}\right) \quad \text{vs.} \quad H_1 : \ln\left(\frac{I_{\mathrm{M,E}=0}}{I_{\mathrm{F,E}=0}}\right) \neq \ln\left(\frac{I_{\mathrm{M,E}=1}}{I_{\mathrm{F,E}=1}}\right).$$

This hypothesis has similar Wald test statistic

$$w = \frac{\left(\ln\left(I_{\mathrm{M,E}=0}\right) - \ln\left(I_{\mathrm{M,E}=1}\right) + \ln\left(I_{\mathrm{F,E}=1}\right) - \ln\left(I_{\mathrm{F,E}=0}\right)\right)^2}{\hat{\sigma}_0^2 + \hat{\sigma}_1^2},$$

but with $\hat{\sigma}_i^2 = \hat{\sigma}_{\mathrm{M},i}^2 + \hat{\sigma}_{\mathrm{F},i}^2 - 2\hat{\rho}_{\mathrm{M,F},i}\hat{\sigma}_{\mathrm{M},i}\hat{\sigma}_{\mathrm{F},i}$, where $\hat{\sigma}_{\mathrm{M},i}$ and $\hat{\sigma}_{\mathrm{F},i}$ are the respective variances for I_{M} and I_{F} in stratum i and $\hat{\rho}_{\mathrm{M,F},i}$ is the correlation between the imprinting parameters in stratum i [6]. Under the null hypothesis, w again follows a $\chi^2(1)$ distribution.

3.1.2 EMIM

Howey and Cordell's **EMIM** [11], released in 2012, is a standalone FORTRAN program that fits triad data using the multinomial model. The software is available as precompiled binaries for Windows and Linux, or can be manually compiled from the source code. **EMIM** is run through the command line, and has no official interface with common statistical software like R [18], which limits its accessibility to end users who have less programming experience, or who are unable to compile the program on a different operating system (e.g. macOS). However, **EMIM**'s features are highly comprehensive, and being compiled from FORTRAN code makes it orders of magnitude faster than software written only in R [11].

EMIM has customizable flags that allow users to model child, maternal, and/or either parent-of-origin effect. It also automatically incorporates control triads using the hybrid multinomial model [25]. A major advantage it has over **Haplin** is that users can choose to use the HWE, MS, or MaS mating type models, allowing users to model populations under mating asymmetry. The software produces a text file summarizing the multinomial model, including point estimates and standard errors for the model parameters, which users must parse manually. Instead of using the EM algorithm to impute missing parental transmission counts, **EMIM** uses a direct search algorithm (using the MAXFUN subroutine from S.A.G.E. [20]) to maximize the likelihood of the data [11]. **EMIM** does not have any features for including environmental interactions. However, a user with statistical expertise can use **EMIM** output to implement their own stratified analysis, using the same stratified approach as **Haplin**.

3.2 TriLLIEM

As we show later in this chapter, **Haplin** and **EMIM** produce identical estimates and p-values when applied to the same data sets (up to rounding error). However, they both

have severe limitations. **Haplin** forcing the HWE assumption invalidates its inferences in populations where this assumption is not appropriate, such as in admixed populations. **EMIM**'s lack of accessibility, not just with the initial setup but also the lack of formal environmental interaction tests, limits its utility. Furthermore, the stratified approach to environmental interactions does not allow us to simultaneously estimate main effects (e.g., estimating R and V in a maternal gene-environment interaction model), and could have limitations on model power as each model parameter is fit twice on smaller subsets of the data. To address these limitations and make full use of the log-linear model, we developed the R package **TriLLIEM**. Like **Haplin**, **TriLLIEM** gives users access to modelling child and maternal effects in case-triad and control-triad data using the log-linear model in R directly. However, **TriLLIEM** also allows users to choose the mating type model to use (HWE, MS, or MaS), and allows users to customize how imprinting and child effects are fit (unlike **Haplin**'s approach to parent-of-origin effects). Results from **TriLLIEM** use the machinery of the `glm()` function in R, which many base R users would be comfortable with interpreting. Currently, **TriLLIEM** cannot handle sporadically missing genotype data like **Haplin**, though this feature is planned for the future.

3.2.1 Gene by Environment Interactions in TriLLIEM

Haplin and **EMIM**'s models' stratified approach to analyzing environmental interactions is equivalent to fitting the log-linear model in **TriLLIEM** with environmental interaction terms for all the main effects in the model. For example, in a MS mating type model with child and maternal effects and environmental exposures, there are 8 main effects (6 mating type parameters and 2 genetic effects), and so the stratified model would also include 8 environmental interaction terms. The E : M interaction term in this model is equivalent to the ratio between the maternal effects in the E = 1 and E = 0 strata that we obtain from **Haplin** and **EMIM**'s stratified approach.

A weakness of the stratified approach is that we are forced to fit additional interaction terms, which may not be of statistical or biological interest. In the example above, if we were only interested in maternal gene-environment interactions, the model would also estimate E : C, potentially reducing power. While this is impossible to do in **Haplin** and **EMIM** since the model is fit twice, in **TriLLIEM**, this is as simple as removing the E : C interaction from the model effects. Note that mating type by environment interaction terms should always be included since they account for population stratification where exposure prevalence varies by sub-population [21, 23].

3.2.2 Parent-of-Origin Effects in TriLLIEM

TriLLIEM, like **Haplin**, uses an EM algorithm to model imprinting effects, where each iteration fits a generalized linear model to expected data (where the $(\mathrm{M}, \mathrm{F}, \mathrm{C}) = (1, 1, 1)$ counts are split into two $(\mathrm{M}, \mathrm{F}, \mathrm{C}, \mathrm{O})$ categories). This is implemented by augmenting the 15-row data matrix with one extra row to account for the two $(\mathrm{M}, \mathrm{F}, \mathrm{C}, \mathrm{O})$ category expected counts. However, this change means that the standard errors computed by `glm()` are not correct, as they treat these expected counts as having been truly observed.

Lang [14] gives a method for finding the observed information and covariance matrices based on Louis' method [16], which we implemented in **TriLLIEM**. To summarize their method in the context of the log-linear model with imprinting, in each iteration the EM algorithm estimates $X = (X_1, X_2, \ldots, X_{16})$, the complete counts for each $(\mathrm{M}, \mathrm{F}, \mathrm{C}, \mathrm{O})$ category, using just the observed data $Y = (Y_1, Y_2, \ldots, Y_{15}) = y$ and current parameter estimates can relate X to Y using $Y = LX$, where L is a 15×16 matrix of 0's and 1's written in block form as

$$L = \begin{bmatrix} I_8 & 0_{8\times1} & 0_{8\times1} & 0_{8\times6} \\ 0_{1\times8} & 1 & 1 & 0_{1\times6} \\ 0_{6\times8} & 0_{6\times1} & 0_{6\times1} & I_6 \end{bmatrix}.$$

This L gives $Y_1 = X_1, Y_2 = X_2, \ldots, Y_8 = X_8, Y_9 = X_9 + X_{10}, Y_{10} = X_{11}, \ldots, Y_{15} = X_{16}$. Note that the 9th row of Table 2.3 corresponds to the $(\mathrm{M}, \mathrm{F}, \mathrm{C}) = (1, 1, 1)$ category where parent-of-origin is not known. Lang [14] showed that the observed information matrix, I_Y, is given by

$$I_Y = I_X - I_{X|LX=y},$$

where I_X is the information matrix of the 16-row expected data, which is obtained from the `glm()` function output, and $I_{X|LX=y}$ is an adjustment matrix for the fact that we only observed y. This adjustment matrix is expressed as $I_{X|LX=y} = Z^T \mathrm{Cov}(X|LX = y)Z$, where Z is the known $16 \times p$ model matrix for the p parameters in the log-linear model. The covariance matrix $\mathrm{Cov}(X|LX = y)$ is a 16×16 matrix whose elements represent the covariances between the elements of X conditional on $LX = y$. $LX = y$ tells us the exact values of all elements of X aside from X_9, X_{10}, which must satisfy $X_9 + X_{10} = y_9$. All conditional covariances are 0 except for $\mathrm{Cov}(X_9, X_9), \mathrm{Cov}(X_9, X_{10})$, and $\mathrm{Cov}(X_{10}, X_{10})$. Since $X_9 = y_9 - X_{10}$, it can be shown that $\mathrm{Cov}(X_9, X_9) = \mathrm{Cov}(X_{10}, X_{10}) = -\mathrm{Cov}(X_9, X_{10}) = \theta$. To estimate θ, we take

$$\hat{\theta} = \frac{y_9 \cdot \hat{x}_9 \cdot \hat{x}_{10}}{(\hat{x}_9 + \hat{x}_{10})^2},$$

where $\hat{x}_9$ and $\hat{x}_{10}$ are the model's fitted counts for X_9 and X_{10} from the final iteration of the EM algorithm [14, eq. 3.2]. Thus, the covariance matrix is a 16×16 matrix that is zero everywhere except for the 2×2 submatrix in the 9th and 10th rows and columns, whose diagonal elements are $\hat{\theta}$ and whose off-diagonal elements are $-\hat{\theta}$. From this, we calculate the adjustment matrix, and subsequently compute the observed information matrix I_Y, which we can use to compute the model's $p \times p$ variance-covariance matrix by taking I_Y^{-1}.

This approach is easily extended to the model with environmental exposures and/or control-triads. **Haplin** also implements this correction in their log-linear model [13, p. 253].

3.2.3 Simulating Triad Data in TriLLIEM

TriLLIEM also has a function for simulating triad data, which is highly customizable. Users may set the following parameters:

- Study design parameters (total number of triads, number of control-triads),
- Demographic parameters (minor allele frequency, proportion of cases/controls who are exposed, MaS coefficients),
- Genetic main effects (R, S, I_{M}, I_{F}),
- Gene-Environment interactions (V, effects undergoing environmental interaction, e.g., M or I_{M}).

Currently, to simulate data under MaS, the simulation function assumes HWE in the general population with MaS coefficients (C_1, C_2, C_4) modifying the parental genotype distribution based on the Bourgey et al. [2] parameterization for MaS. These coefficients are an alternative parameterization to MaS compared to the one shown in Table 2.2, where we modify the MS mating type model as shown in Table 3.1.

In this MaS parameterization, each coefficient must be between 0 and 2, with larger coefficients indicating that the given mating type has a higher probability of M > F. Using the specified parameters, the simulation function randomly determines the number of exposed and unexposed cases/controls to be generated based on the exposure proportions, and then samples the counts for each $(\mathrm{M}, \mathrm{F}, \mathrm{C})$ category from the appropriate multinomial distribution. Other parameters for simulating population stratification are also available, which we outline in Section 4.2.

Table 3.1: Pr(M, F) in the Bourgey et al. MaS [2] mating asymmetry parameterization, with respect to the MS mating type model.

Mating type: (M, F)	Pr(M, F)	
	MS	Bourgey et al. MaS [2]
0:		
$(2,2)$	μ_0	μ_0
1:		
$(2,1)$	μ_1	$\mu_1 C_1$
$(1,2)$	μ_1	$\mu_1 (1 - C_1)$
2:		
$(2,0)$	μ_2	$\mu_2 C_2$
$(0,2)$	μ_2	$\mu_2 (1 - C_2)$
3:		
$(1,1)$	μ_3	μ_3
4:		
$(1,0)$	μ_4	$\mu_4 C_4$
$(0,1)$	μ_4	$\mu_4 (1 - C_4)$
5:		
$(0,0,0)$	μ_5	μ_5

3.3 TriLLIEM Setup and Usage

To get started with using **TriLLIEM**, we first install and load the package from the package repository on GitHub via the commands:

```
devtools::install_github("KevinHZhao/TriLLIEM")
library(TriLLIEM)
```

Simulating data in **TriLLIEM** is performed through the `simulateData()` function. For example, we can simulate $1,000$ case-triads from a population under HWE with a minor allele frequency of 30%, $S = 1.2$ and no other genetic factors by using the commands:

```
set.seed(123) # setting random seed for reproducibility
matdat <-
  simulateData(
```

```
    ntrios = 1000,
    maf = 0.3,
    mtCoef = c(1,1,1),
    S = c(1, 1.2, 1.2^2)
  )$dat4R # extract data formatted for TriLLIEM
## simulateData also returns data formatted for EMIM and Haplin
```

More complex simulations can also be run. For example, consider the following conditions:

- $1,000$ case-triads
- $1,000$ control-triads
- 30% minor allele frequency
- $(C_1, C_2, C_4) = c(0.5, 0.5, 0.5)$
- $S = 1.2$
- $I_{\mathrm{M}} = 1.4$
- 30% environmental exposure frequency in the population
- $V = 1.5$ with maternal imprinting by environment interactions

We can obtain the corresponding simulated data by running:

```
impdat <-
  simulateData(
    ntrios = 2000,
    maf = 0.3,
    S = c(1, 1.2, 1.2^2),
    V = c(1, 1.5, 1.5^2),
```

```
    mtCoef = c(0.5, 0.5, 0.5),
    Im = 1.4,
    includeE = TRUE,
    Einteraction = "Im",
    propE = 0.3,
    includeControl = TRUE,
    nControl = 1000
  )$dat4R
```

Analyzing triad data sets is performed through the `TriLLIEM()` function. This function takes a vector of strings `effects` to include in the model. It also allows users to specify the mating type model to use, and whether to use the stratified or non-stratified approach for environmental interactions (non-stratified by default). With the first simulated data set above, an appropriate model would use an HWE mating type model with maternal effects only, hence we would run the command:

```
matres <-
  TriLLIEM(
    mtmodel = "HWE",
    effects = c("M"),
    dat = matdat
  )
```

The `TriLLIEM()` function returns an object of class `TriLLIEM`, which inherits from the `glm` object in R. This means users can apply many of the usual functions for interpreting results from the `glm()` function. To interpret `matres`, we would run:

```
matres |> summary() |> coef()
```

This outputs a matrix equivalent to Table 3.2.

Table 3.2: Output of analysis on `matres`.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	5.3642380	0.0537314	99.834238	0.0000000
HWgeno	-0.8640207	0.0489599	-17.647514	0.0000000
M	0.2140671	0.0679399	3.150829	0.0016281

The "Estimate" column for the M row represents our point estimate for s, which we can exponentiate to obtain the equivalent risk parameter, S. Based on these results, the model estimates $\hat{S} = 1.24$ with a p-value of 0.00163. Since the data were simulated for a population with $S = 1.2$, this level of significance is expected.

Likewise, for the `impdat` data, we would use an MaS mating type model, with S, I_{M}, V as genetic effects in the hybrid model, where V is the maternal imprinting by environment interaction. The code to do this is shown below, with output omitted for brevity.

```
impres <-
  TriLLIEM(
    mtmodel = "MaS",
    effects = c("M", "Im"),
    dat = impdat,
    includeE = TRUE,
    Einteraction = "Im",
    includeD = TRUE
  )
impres |> summary() |> coef()
```

From the results for genetic effects, the model estimates $\hat{S} = 1.19, \hat{I}_{\mathrm{M}} = 1.33, \hat{V} = 1.50$ with respective p-values $0.0594, 0.0199, 0.0269$. Note that these results use the non-stratified approach. For results equivalent to the stratified approach that are equivalent to **EMIM** and **Haplin**, we use:

```
impres_strat <-
  TriLLIEM(
    mtmodel = "MaS",
    effects = c("M", "Im"),
    dat = impdat,
    includeE = TRUE,
    Einteraction = "Im",
    Estrat = TRUE,
    includeD = TRUE
  )
impres_strat |> summary() |> coef()
```

This model (which adds a E : M effect due to stratification), has $\hat{S} = 1.18$, $\hat{I}_{\mathrm{M}} = 1.34$, and $\hat{V} = 1.46$, with respective p-values $0.0135, 0.0237, 0.0121$. Note that for comparison with **EMIM** and **Haplin**, we can only compare the $\hat{V}$ estimates, since the stratified approach does not produce equivalent estimates for the main effects.

For convenience, from this point onward we use C, M rather than S, V when discussing child and maternal effects in **TriLLIEM** and other software packages.

3.4 Software Comparison

To demonstrate that **EMIM**, **Haplin**, and **TriLLIEM** produce consistent results we simulated 6 data sets with $1,000$ case-triads and $1,000$ control-triads, minor allele frequency of 0.3, environmental exposure prevalence of 0.3, and with no MaS. The first data set was simulated with no genetic effects (i.e., $R, S, V, I_{\mathrm{M}}, I_{\mathrm{F}}$ all equal to 1), and the remaining 5 were simulated with one of $R, S, I_{\mathrm{M}}, I_{\mathrm{F}}, V$ equal to 1.5 respectively, and no other genetic effects (under a E : I_{M} gene-environment interaction when applicable). This was to verify that

models between the 3 software were equivalent regardless of how the data was simulated. As an example of what a data set looks like, the simulated data set with no genetic effects is shown in Table 3.3. For each data set, we fit 5 separate genetic effect models, with included effects:

1. C, M,
2. $\mathrm{M}, I_{\mathrm{M}}, I_{\mathrm{F}}$ (Haplin parent-of-origin parameterization),
3. $\mathrm{C}, \mathrm{M}, I_{\mathrm{M}}$,
4. $\mathrm{C}, \mathrm{M}, \mathrm{E} : \mathrm{C}, \mathrm{E} : \mathrm{M}$ (stratified approach focussing on the maternal gene-environment interaction),
5. $\mathrm{C}, \mathrm{M}, I_{\mathrm{M}}, \mathrm{E} : \mathrm{C}, \mathrm{E} : \mathrm{M}, \mathrm{E} : I_{\mathrm{M}}$ (stratified approach focussing on the maternal imprinting by environment interaction).

Each genetic effect model was fit 3 times, once using each mating type model. We can compare the results from **TriLLIEM**, **EMIM**, and **Haplin** to verify that they are fitting the same model. Note Haplin can only fit models $1, 2$, and 4 under the HWE mating type model.

We plot all the point estimates and p-values for the relevant genetic effects and interactions in Figures 3.1 and 3.2. Note that the point estimates and p-values for each simulated data set are aggregated together, since the purpose of this study was to show that each method produced the same results. We assess model validity later in Chapter 4.

From Figure 3.1, we can see that in all models fit and in every simulated data set, **TriLLIEM**, **EMIM**, and **Haplin** produce point estimates that are essentially identical up to rounding error. The point estimates all fall on the $y = x$ line. A similar conclusion can be made based on the p-values for these estimates in Figure 3.2. Overall, Figures 3.1 and 3.2

show that **TriLLIEM**, **EMIM**, and **Haplin** indeed provide equivalent results when they are expected to.

Table 3.3: Simulated null data set for comparison between **EMIM**, **Haplin**, and **TriLLIEM**.

				D = 0		D = 1	
Mating type	M	F	C	E = 0	E = 1	E = 0	E = 1
1	0	0	0	169	74	154	78
2	0	1	0	81	34	70	28
2	0	1	1	64	38	78	33
2	1	0	0	82	32	86	27
2	1	0	1	75	25	71	23
3	0	2	1	24	6	26	5
3	2	0	1	32	11	32	11
4	1	1	0	23	17	29	14
4	1	1	1	55	31	63	17
4	1	1	2	37	16	33	14
5	1	2	1	7	4	19	12
5	1	2	2	11	6	18	8
5	2	1	1	17	9	16	12
5	2	1	2	11	4	11	4
6	2	2	2	4	1	7	1

Figure 3.1: Point estimates of various effects from equivalent models fit in **TriLLIEM**, **EMIM**, and **Haplin** between all 6 simulated data sets. The x-axis corresponds to the **TriLLIEM** point estimate while the y-axis corresponds to the **Haplin** and **EMIM** point estimates. The line $y = x$ is plotted in red and indicates equal point estimates between methods. From top to bottom, the rows correspond to fitting a HWE, MS, and MaS mating type model. Since **Haplin** only allows HWE, **Haplin** results are only in the top row. From left to right, the first 3 columns represent models that use: child and maternal effects; maternal effects and both imprinting effects (equivalent to Haplin's parent-of-origin parameterization); and child, maternal, and maternal imprinting effects. The fourth column represents results for *only* the interaction term in two different stratified gene-environment interaction models: one where child, maternal effects, and maternal gene-environment interactions are present (green), and another where child, maternal, maternal imprinting effects, and maternal imprinting by environment interactions are present (yellow).

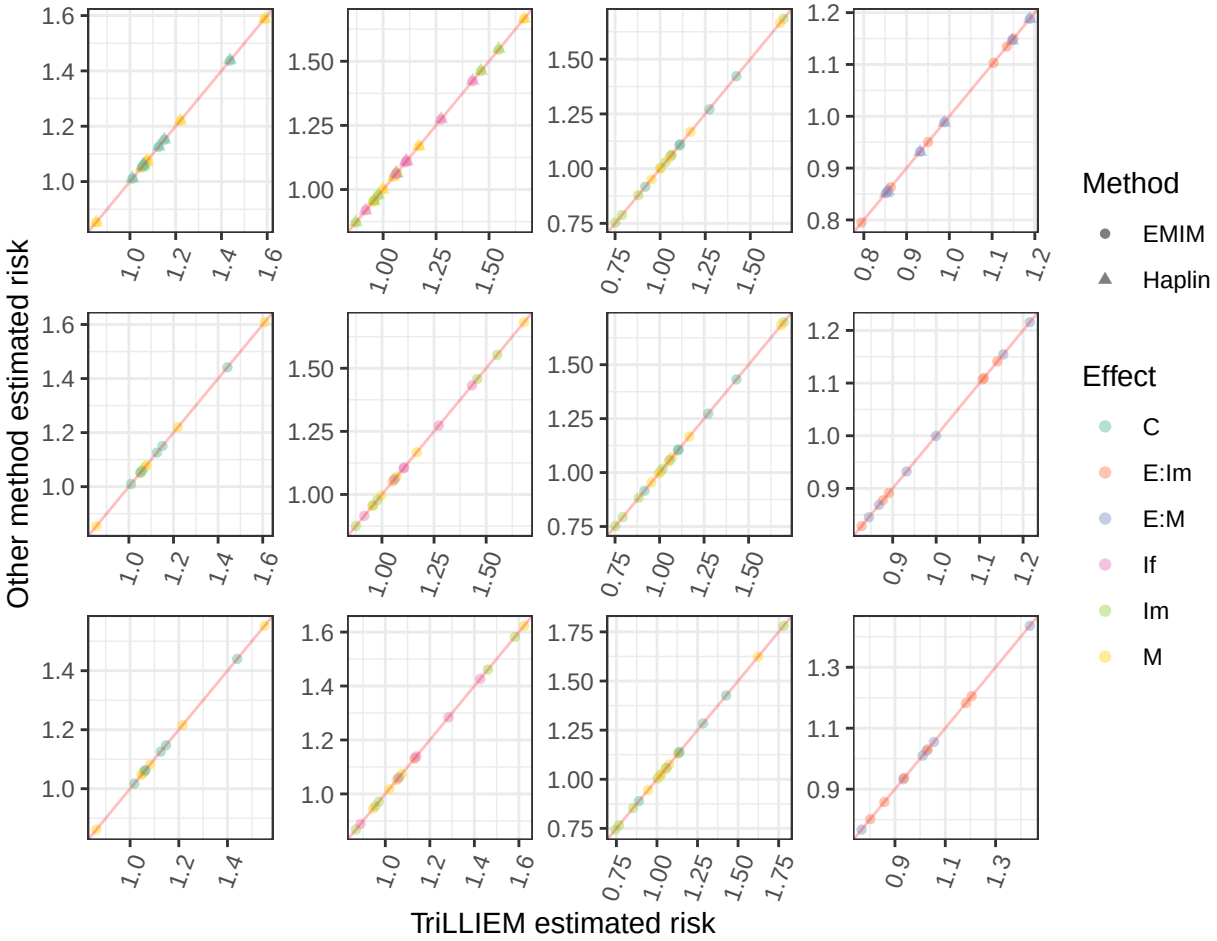
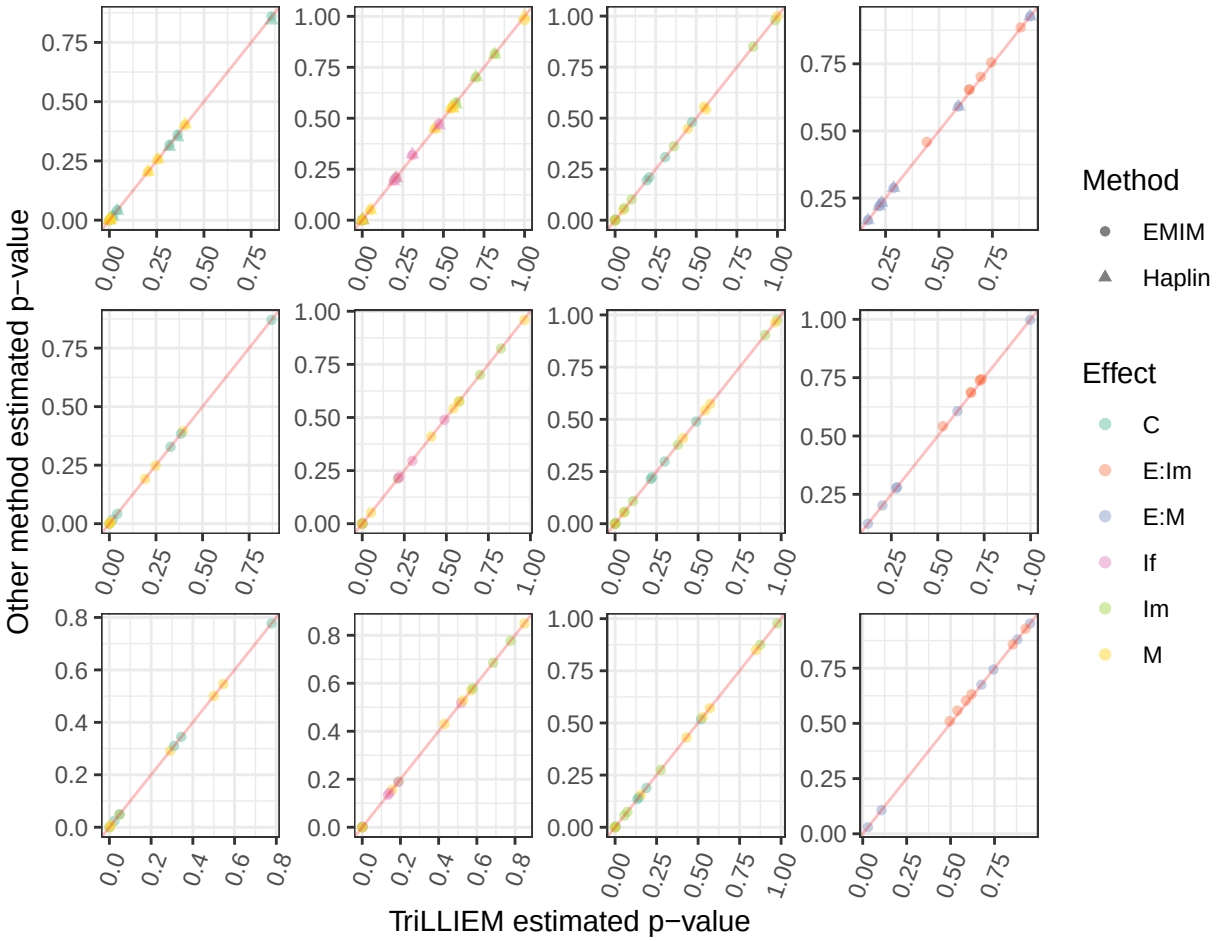


Figure 3.2: Wald based p-values of various effects from equivalent models in **TriLLIEM**, **EMIM**, and **Haplin** between all 6 simulated data sets. The line $y = x$ is plotted in red and indicates equal p-values between methods. The plot layout follows the same row-column layout described in the caption of Figure 3.1.



Chapter 4

Simulation Study

In this chapter, we use **TriLLIEM** to conduct several simulation studies. First, we validate the log-linear model's performance by simulating and analyzing data under a large variety of conditions using **TriLLIEM**, evaluating model power and Type 1 error rate. We then perform a similar study on data simulated with population stratification to verify whether the log-linear model produces spurious gene-environment interactions, with a focus on imprinting by environment interactions.

4.1 Model Validation Simulations

4.1.1 Data Simulation Conditions

Having shown that **TriLLIEM** is consistent with **EMIM** and **Haplin**, we utilize its ability to easily simulate and analyze data to validate the log-linear model's performance. We focus on the performance of **TriLLIEM**'s imprinting model, since it is more flexible than Haplin's. Another major point of interest is the performance of the non-stratified gene-environment interaction model, which is not available in **EMIM** or **Haplin**. With these considerations in mind, we simulated data under 72 distinct conditions, summarized in Table 4.1. For each

Table 4.1: Parameter values in our model validation simulation study. Only 1 of the main effects ($R, S, V, I_{\mathrm{M}}, I_{\mathrm{F}}$) were non-null in any given simulation model. All mating asymmetry parameters were set to be equal. In total, this comes out to 72 different scenarios with parameterizations. For an exact table of all simulated scenarios, see Table A.1 in the Appendix.

Parameter	Assigned values
R	$1, 1.15$
S	1
V	$1, 1.5, 1.6$
I_{M}	$1, 1.2, 1.4$
I_{F}	$1, 1.2, 1.4$
Environmental variable frequency	$0.3, 0.5$
Mating asymmetry parameters (C_1, C_2, C_4)	$0.85, 1, 1.15$
Environmental interaction parameter	$I_{\mathrm{F}}, I_{\mathrm{M}}$

condition, we simulated $2,000$ data sets with $2,000$ case-triads to test the case-triad model, and another $2,000$ data sets with $1,000$ case and $1,000$ control-triads to test the hybrid model. In every simulation, we set the minor allele frequency to be 0.3. We considered several different cases for the mating asymmetry coefficients, C_1, C_2, and C_4 (see Table 3.1), to compare model performance under varying mating asymmetry levels. We did not simulate true maternal effects ($S \neq 1$) or maternal gene-environment interactions, as the performance of the log-linear model under these conditions has been previously examined by Hudson [12].

4.1.2 Statistical Analysis

We analyzed each data set using **TriLLIEM** under the HWE, MS, and MaS (when controls were available) mating type models. For the effects included in the model, we always had C and M, as well as the appropriate environmental interaction term (unless the model did not consider environmental exposures). I_{M} or I_{F} was included when the model conditions set $I_{\mathrm{M}} \neq 1$ or $I_{\mathrm{F}} \neq 1$, respectively, *or* if the underlying scenario uses it as the environmental

interaction term. Note that this means we do not consider cases where the model was improperly defined (e.g., using maternal imprinting by environment interactions when the data was simulated with paternal imprinting by environment interactions), as the purpose of this study is to validate the model's accuracy when correctly defined. We also used a model that did not consider the environmental variable, a model that used the single log-linear model based on the Weinberg approach (which will be referred to as the non-stratified model) and the stratified model based on the Haplin approach. However, to fit the stratified model in **TriLLIEM**, we do not need to run the model twice, as with **EMIM** and **Haplin**. Instead, we use the equivalent approach of including an interaction of the environmental variable with all main effects, regardless of whether the interaction is truly of interest. For example, one of our simulation scenarios had $R = S = I_{\mathrm{M}} = I_{\mathrm{F}} = 1$, $C_1 = C_2 = C_4 = 1.15$, an environmental factor proportion of 0.5, and a maternal imprinting by environment interaction of $V = 1.6$. To analyze the case-triad and hybrid data for this scenario, for the stratified approach we must include C, M, I_{M}, E : C, E : M, and E : I_{M} as terms in the model whereas in **TriLLIEM** we can fit the unstratified true model with terms $\mathrm{C}, \mathrm{M}, I_{\mathrm{M}}$, and $\mathrm{E} : I_{\mathrm{M}}$. For this example, the model that does not consider the environmental variable only uses the terms C, M, and I_{M}, though the results in this case would not be very useful since the data were simulated with a true maternal imprinting by environment interaction which cannot be captured by this model.

From the analysis of each simulated data set, we return the bias of each effect estimate and the result of the significance test on the appropriate terms. We calculate the relative bias for each term in the model using

$$\text{Relative Bias} = \frac{\hat{\beta} - \beta}{\beta}, \tag{4.1.1}$$

where $\hat{\beta}$ is the estimated risk parameter, and β is the true value of the risk parameter used to simulate the data. Averaging the relative bias of each term in the model across all $2,000$

simulated data sets, we then determine the proportion of the $2,000$ data sets where the hypothesis test was rejected for each parameter at the 0.05 significance level.

4.1.3 Model Validation

In this section, we summarize major observations from the results of our simulation study. Since we ran 72 different scenarios, we will not include detailed results for all scenarios in this thesis. Supplementary results for models with maternal imprinting, environmental variable frequency 30%, and MaS coefficients either 0.85 or 1 are given in Tables A.2 to A.5 in the Appendix. Complete results are available at our GitHub page.

4.1.3.1 Mating Type Model Selection

The presence of mating asymmetry in the simulated population has a major impact on the model's performance when we do not use the mating asymmetry model. In our simulations with $C_1 = C_2 = C_4 = 0.85$, relative bias for the maternal effect, M, when the mating type model assumed HWE or MS was observed to always fall between -0.22 and -0.2 for the case-triad model, and between -0.13 and -0.11 when control-triads were present (see, for example, the rows with ID 1–4 in Tables A.2 and A.4). Similar results, but with bias in the opposite direction, were observed for the 1.15 MaS coefficient scenarios; these results are not displayed in the Appendix and are available on GitHub. This bias is due to the unaccounted mating asymmetry; when analyzing the data under the correct MaS mating type model, the largest observed absolute relative bias was approximately 0.6%, which is in the same range as the estimated bias when the data were simulated without mating asymmetry and either HWE or MS was assumed. The deviation from the true value is also evident from the higher false rejection rates for M. When the mating asymmetry parameters are not 1, but data are analyzed assuming a HWE or MS mating type model, the false rejection rates were always $> 15\%$, far exceeding the nominal Type 1 error rate of 5% (see, for example, the rows with

ID 1–4 in Tables A.3 and A.5).

Not accounting for mating asymmetry when it is present in the population appears to bias the estimates of M, but it also appears to further impact models for imprinting by environment interactions when controls are present. For example, with data simulated with no true effects and an environmental variable frequency of 30%, the results of models with $\mathrm{C}, \mathrm{M}, I_{\mathrm{M}}, \mathrm{E} : I_{\mathrm{M}}$ terms in the model also have inflated rejection rates for C and I_{M} when there are control-triads and the HWE or MS mating type models are assumed (alongside the inflated rates for M that we previously discussed). The rejection rates for the non-stratified model are shown in Figure 4.1. Estimated Type 1 error rates assuming the HWE and MS mating type models are higher than 18% for data with simulated mating asymmetry. The inflated rates for C and I_{M} are less pronounced when analyzing the respective data with mating asymmetry parameters of 1, as the false rejection rates are less than 8% for both. This Type 1 error rate is still higher than the nominal rate of 5%, which could be due to having too many parameters in the model, or due to the correlation between C and I_{M} effects, which univariate Wald-based tests do not account for. With only case-triad data, we no longer see the large inflation when mating asymmetry is present, despite the model's genetic effects being the same.

Interestingly, when interchanging maternal imprinting with paternal imprinting, the inflated rejection rates for I_{F} remain, but C's rejection rate now remains less than 8% (Figure 4.2). This exception only occurs when controls are present and a HWE or MS model is assumed and child effects are only affected when maternal imprinting is included in the model rather than paternal. A hypothesis for why this occurs could be because I_{M} is highly dependent on both C and M (as the mother and child both having the minor allele is necessary for a maternal imprinting effect to be present), whereas I_{F} is only correlated with C. Since our Wald based p-values do not take these correlations into account, it could result in incorrect behaviour when examining the p-values of correlated effects. This is a pattern across other similar conditions of our data when comparing imprinting models.

Figure 4.1: Average rejection rates for genetic effects when modelling $\mathrm{C}, \mathrm{M}, I_{\mathrm{M}}, \mathrm{E} : I_{\mathrm{M}}$ in null scenarios with various degrees of mating asymmetry. The red line represents the desired Type 1 error rate of 5%. Note that the y-axis scaling is different between rows.

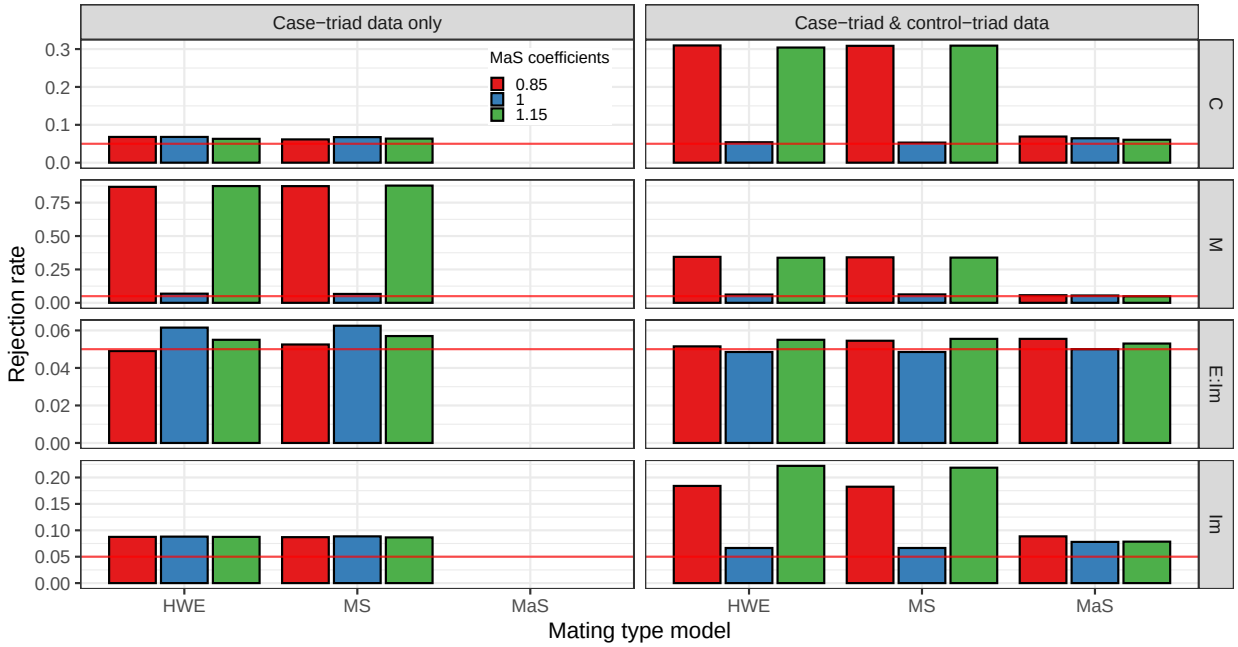
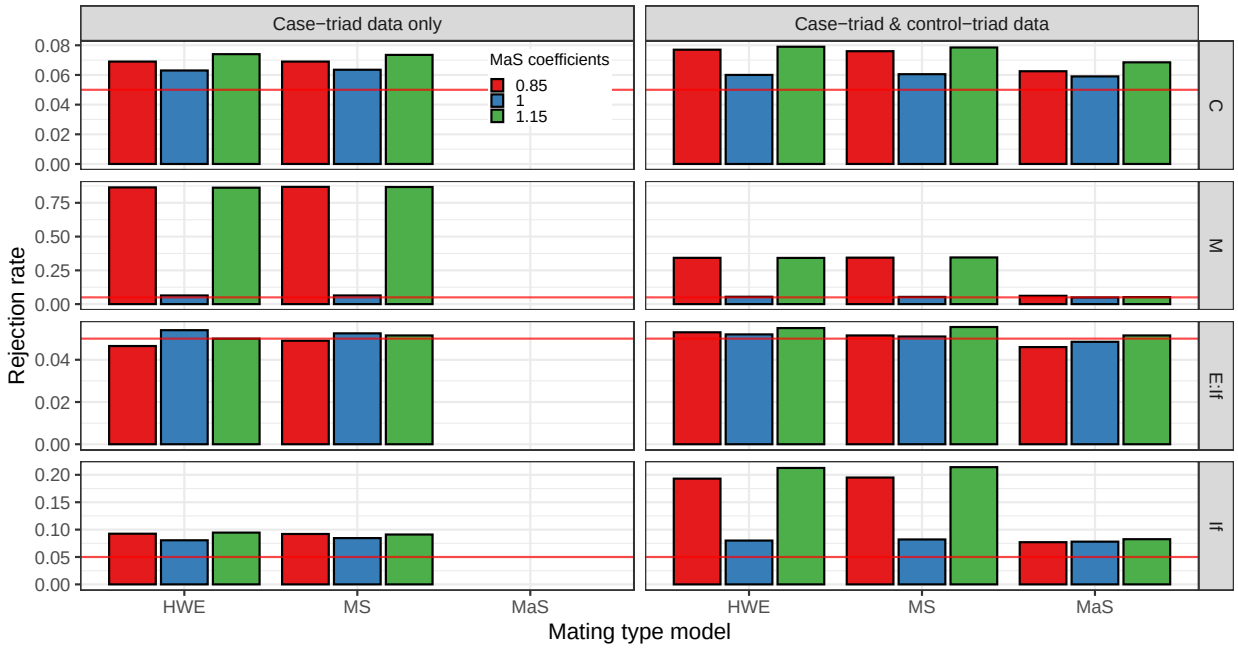


Figure 4.2: Average rejection rates for genetic effects when modelling $\mathrm{C}, \mathrm{M}, I_{\mathrm{F}}, \mathrm{E} : I_{\mathrm{F}}$ in null scenarios with various degrees of mating asymmetry. The red line represents the desired Type 1 error rate of 5%. Note that the y-axis scaling is different between rows.



We can conclude from these results that mating asymmetry does not just impact the model's performance in detecting maternal effects, which is well-known [2], but can also influence the estimation of other genetic effects depending on whether control-trios are sampled and environmental interactions are in the model.

4.1.3.2 Stratified vs. Non-Stratified Environmental Interactions

Our simulations with environmental interactions were run using both the stratified and non-stratified approaches. We compare the two approaches' Type 1 error and power when detecting parent-of-origin by environment interactions. The results for the maternal imprinting by environment and paternal imprinting by environment models are shown in Figures 4.3 and 4.4, respectively.

From these results, we see that the relative bias for the interaction effect, V, under maternal/paternal imprinting by environment interactions is low in both stratified and unstratified models, not exceeding 1.5%. The Type 1 error rate is also low in both approaches, with it not exceeding 7% in the stratified approach and 6% in the unstratified approach. However, when examining power (i.e., true effect 1.5 or 1.6), we find that the unstratified model has substantially higher power than the stratified model. For example, with $2,000$ case-triads simulated with $V = 1.5$, a 30% environmental variable frequency rate, and a maternal imprinting by environment interaction, the unstratified model had a power of 88%, while the stratified model had a power of 35%. We find similar observations with $1,000$ case-triads and control-triads under the same conditions, with a power of 69% and 26% for the unstratified and stratified models, respectively (the reduction in power is expected, as the number of informative case-triads is lower). This highlights a major advantage of **TriL-LIEM**; it enables users to easily use the unstratified approach, which we show has better power than the stratified approach when modelling specific parent-of-origin by environment interactions.

Figure 4.3: Comparison between rejection rates and relative biases of the non-stratified (unhatched) and stratified (hatched) models for maternal imprinting by environment interactions. Data was simulated with other genetic risks set to 1, no mating asymmetry, and analyzed using a HWE mating asymmetry model with $\mathrm{C}, \mathrm{M}, I_{\mathrm{M}}, \mathrm{E} : I_{\mathrm{M}}$ as genetic effects (alongside $\mathrm{E} : \mathrm{C}, \mathrm{E} : \mathrm{M}$ under the stratified model). The red line represents the desired Type 1 error rate of 5% when there are no true effects.

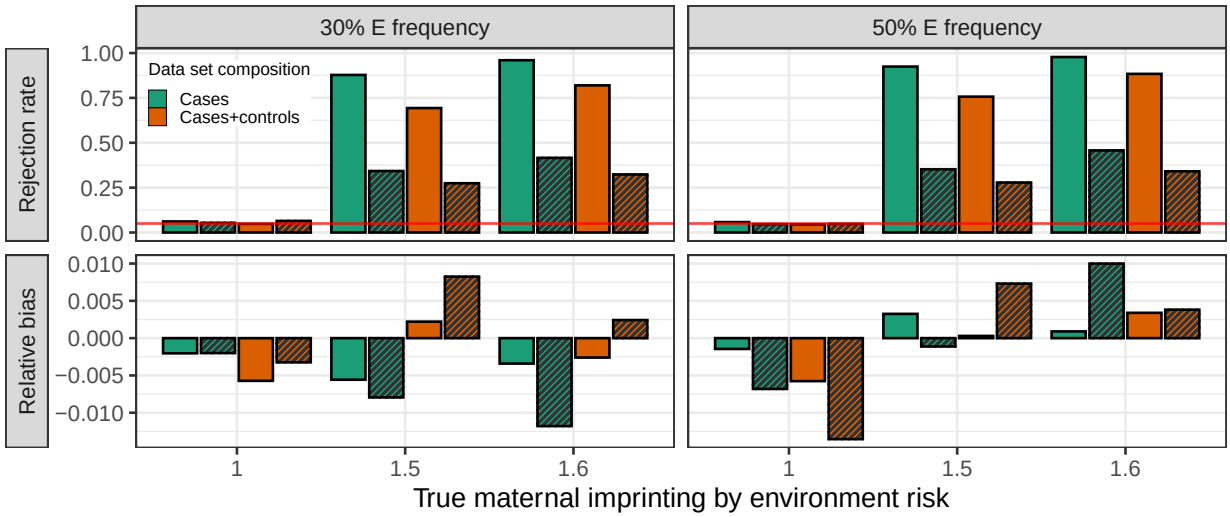
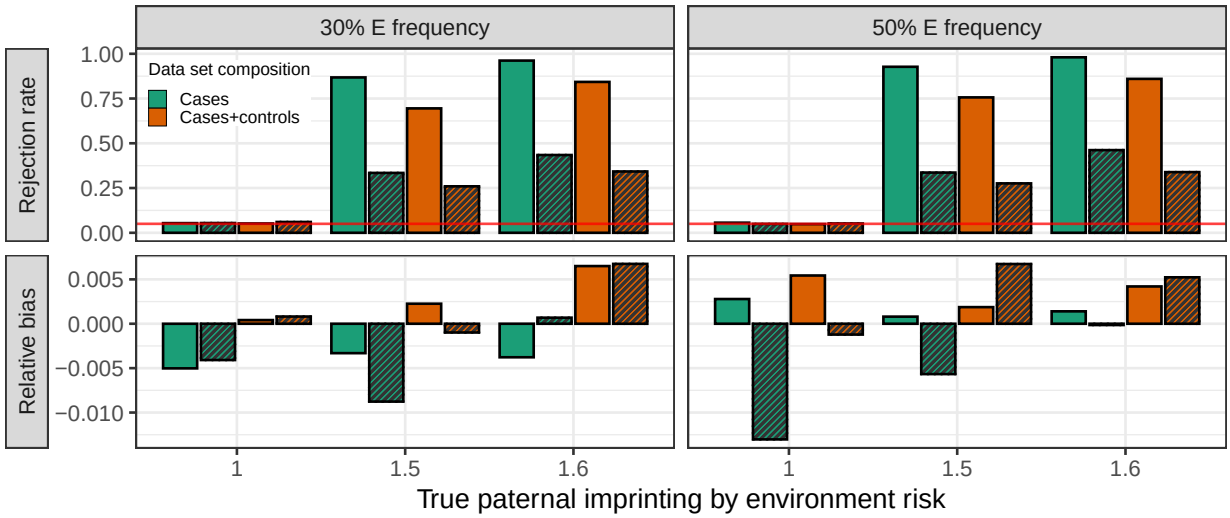


Figure 4.4: Comparison between rejection rates and relative biases of the non-stratified (unhatched) and stratified (hatched) models for paternal imprinting by environment interactions. Data was simulated with other genetic risks set to 1, no mating asymmetry, and analyzed using a HWE mating asymmetry model with $\mathrm{C}, \mathrm{M}, I_{\mathrm{F}}, \mathrm{E} : I_{\mathrm{F}}$ as genetic effects (alongside $\mathrm{E} : \mathrm{C}, \mathrm{E} : \mathrm{F}$ under the stratified model). The red line represents the desired Type 1 error rate of 5% when there are no true effects.



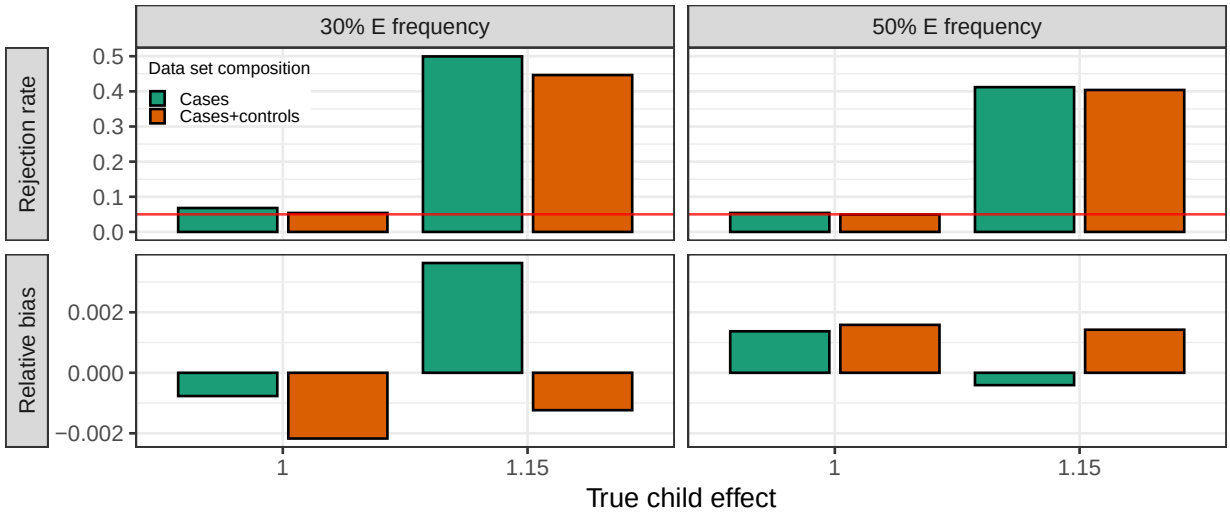
Haplin uses a modified version of the stratified approach when modelling parent-of-origin effects, which has better power than the basic stratified approach [6]. However, it does not allow users to test $\mathrm{E} : I_{\mathrm{M}}$ or $\mathrm{E} : I_{\mathrm{F}}$ interactions individually, instead comparing the ratios of maternal and paternal imprinting in the E strata. Child effects also cannot be tested for simultaneously, as this approach requires I_{M} and I_{F} to both be present. This highlights another advantage of the unstratified approach, which is its flexibility.

While testing for environmental interactions, we may also concurrently include other genetic effects in the model and perform Wald tests. For example, we can test for child effects using a non-stratified model with maternal imprinting by environment interactions (using $\mathrm{C}, \mathrm{M}, I_{\mathrm{M}}, \mathrm{E} : I_{\mathrm{M}}$ as effects); the results are shown in Figure 4.5. We see that the relative error for C in this approach is low, never exceeding 0.3%. In significance tests, we find that the Type 1 error rate is between 5% and 8%, which slightly exceeds our desired Type 1 error rate of 5%. For the power, we observe that for data simulated with $R = 1.15$, power is roughly 47% when the environmental variable frequency is 30%, and 40% when the frequency is 50%. While certainly not ideal, results such as these could help inform model selection. The Type 1 and Type 2 error rates can be improved by removing non-significant terms from the model, such as the $\mathrm{E} : I_{\mathrm{M}}$ interaction or the I_{M} term when appropriate. This can be done via a likelihood ratio test, which is not currently implemented in **TriLLIEM** (though it is planned as future work).

4.1.3.3 Imprinting Effects

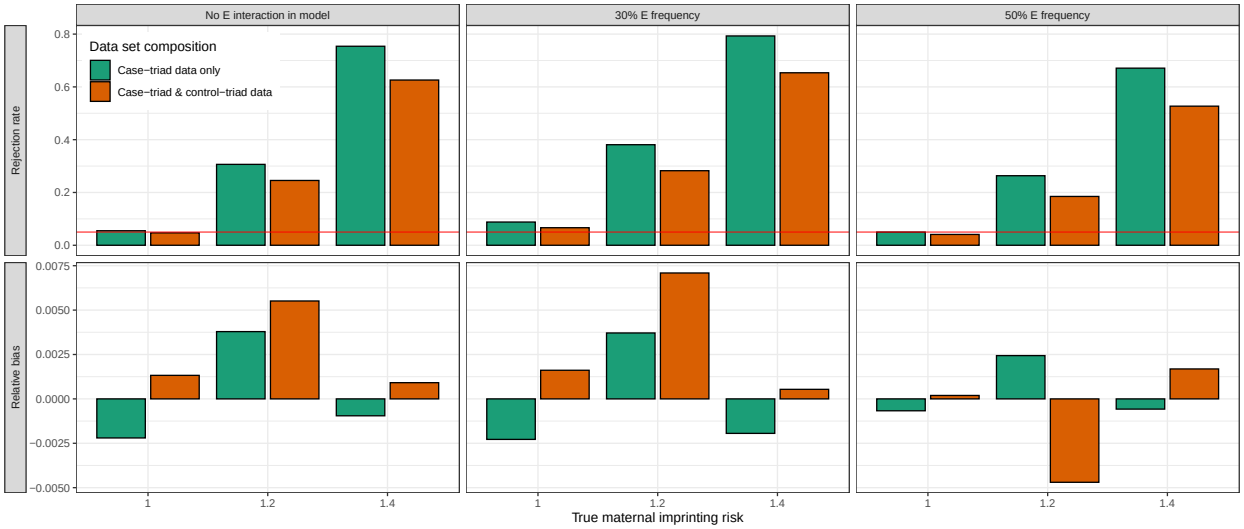
Examining the ability of the log-linear model to correctly identify imprinting effects, we find that the model's power for detecting these effects is fair. Similar to our previous assessment for child effects, we display results for modelling true imprinting effects when an imprinting by environment interaction effect is included in the model in Figure 4.6. Relative bias is low, being consistently $< 1\%$, and power increases as the magnitude of I_{M}'s true value

Figure 4.5: Rejection rates and relative biases for C in the non-stratified model for maternal imprinting by environment interactions. Data was simulated with all genetic risks aside from C set to 1, no mating asymmetry, and analyzed using a HWE mating asymmetry model with $\mathrm{C}, \mathrm{M}, I_{\mathrm{M}}, \mathrm{E} : I_{\mathrm{F}}$ as genetic effects. The red line represents the desired Type 1 error rate of 5% when there are no true effects.



increases, with case-triad data with 30% environmental variable frequency having a power of 36% when $I_{\mathrm{M}} = 1.2$ and 79% when $I_{\mathrm{M}} = 1.4$. From the Type 1 error rates, we see that at a 30% environmental variable frequency, the false rejection rate for I_{M} is 9% with $2,000$ case-triads and 8% with $1,000$ case-triads and $1,000$ control-triads. This false rejection rate is concerning, though again, it can be reduced by removing non-significant terms from the model, such as C or M when applicable. From these results, we find that the model power for detecting maternal imprinting is high even when $\mathrm{E} : I_{\mathrm{M}}$ is included in the model; the rejection rates for the data with $I_{\mathrm{M}} \neq 1$ are similar between the model with no E interactions and the model with $\mathrm{E} : I_{\mathrm{M}}$ and 30% environmental variable frequency. The power shown in Figure 4.6 is actually slightly higher in the model with the added $\mathrm{E} : I_{\mathrm{M}}$ parameter when using data with 30% environmental variable frequency, which is unexpected since a model with fewer parameters should have more power, though this could be due to simulation error.

Figure 4.6: Rejection rates and relative biases for I_{M} in our simulation study on true maternal imprinting. Modelled genetic effects are $\mathrm{C}, \mathrm{M}, I_{\mathrm{M}}$, and $\mathrm{E} : I_{\mathrm{M}}$, with a HWE mating type on data simulated without mating asymmetry.



4.2 Population stratification simulation

In another simulation study, we examined the confounding effects of population stratification on the estimation of gene-environment interactions. We simulated a population in which there were 2 hidden sub-populations. These sub-populations had several differing demographic parameters, which we specify in Table 4.2. For our population stratification simulations, we use the notation MAF_i for minor allele frequency, $\varepsilon_{i,1}$ and $\varepsilon_{i,0}$ for proportion of case- (1) or control- (0) triads with the environmental variable, respectively, α_i for disease prevalence, and ω_i for proportion of the entire population from the ith sub-population (or equivalently, probability of sampling an individual from this sub-population).

Table 4.2: Parameters across the two sub-populations in our simulation study on population stratification.

Parameter	MAF_i	$\varepsilon_{i,1}$	$\varepsilon_{i,0}$	α_i	ω_i
Sub-population 1 (i = 1)	0.2	0.3	0.3	0.1	0.4
Sub-population 2 (i = 2)	0.4	0.5	0.5	0.2	0.6

A total of $1,000$ case-triads and $1,000$ control-triads are simulated in each data set. To

determine the proportion of case/control trios having the environmental variable, we assumed that the disease status, D, was conditionally independent of the environmental variable, E, given population status. Since there are no true genetic effects in this simulation study, this is a reasonable assumption, and it allows us to compute the probability of an exposed case in sub-population $X = i$ in Equation (4.2.1) as

$$\begin{aligned}\Pr(X = i, \mathrm{E} = 1 \mid D = 1) &= \frac{\Pr(D = 1, E = 1 \mid X = i)\Pr(X = i)}{\Pr(D = 1)} \\ &= \frac{\Pr(D = 1 \mid X = i)\Pr(E = 1 \mid X = i)\Pr(X = i)}{\Pr(D = 1)} \\ &= \frac{\alpha_i \varepsilon_{i,1} \omega_i}{\alpha_1\omega_1 + \alpha_2\omega_2}. \end{aligned} \tag{4.2.1}$$

Likewise, we repeat for the other combinations of D and E in Equation (4.2.2) to obtain

$$\Pr(X = i, \mathrm{E} = 0 \mid D = 1) = \frac{\alpha_i(1 - \varepsilon_{i,1})\omega_i}{\alpha_1\omega_1 + \alpha_2\omega_2}, \tag{4.2.2a}$$

$$\Pr(X = i, \mathrm{E} = 1 \mid D = 0) = \frac{(1 - \alpha_i)\varepsilon_{i,0}\omega_i}{(1 - \alpha_1)\omega_1 + (1 - \alpha_2)\omega_2}, \tag{4.2.2b}$$

$$\Pr(X = i, \mathrm{E} = 0 \mid D = 0) = \frac{(1 - \alpha_i)(1 - \varepsilon_{i,0})\omega_i}{(1 - \alpha_1)\omega_1 + (1 - \alpha_2)\omega_2}. \tag{4.2.2c}$$

In each sub-population, we assumed the minor allele did not affect one's disease risk (i.e. all effects were 1), set all mating asymmetry parameters to 1, and assumed HWE within each sub-population and Mendelian inheritance (Tables 2.1 and 2.2). We sampled $1,000$ case-triads, assuming a multinomial distribution with probabilities in Equations (4.2.1) and (4.2.2) to get the number of cases from sub-population 1 with $\mathrm{E} = 1$, population 1 with $\mathrm{E} = 0$, etc., and repeated this for the controls. Then within each of these 8 groups, we simulated the number of triads in each genotype category, again from a multinomial distribution with probabilities from $\Pr(\mathrm{M}, \mathrm{F}, \mathrm{C}|\mathrm{D}, \mathrm{E}, X = i) = \Pr(\mathrm{M}, \mathrm{F}|X = i) \cdot \Pr(\mathrm{C}|\mathrm{M}, \mathrm{F})$, and then combined the counts in each genotype category from the two sub-populations to form our final data set. $2,000$ data sets were simulated in this manner. We then ran **TriLLIEM** using

the maternal environment interaction model and both imprinting by environment interaction models, with C and M always being included. This was repeated assuming the HWE, MS, and MaS mating type models. With this setup, we can examine how hidden population substructures affect the log-linear model's performance when environmental interactions and control trios are present.

The results with both case- and control-triads are shown in Tables 4.3 and 4.4. From these tables, we see that, as expected, population stratification does indeed lead to an increase in Type 1 error for C and M across all environmental interaction models, as well as larger relative biases in the range of 6%–22%. For instance, the false positive rates for C and M are greater than 20% for nearly all models, except C in a paternal imprinting by environmental effect model, where instead, the paternal imprinting term is falsely significant in over 20% of the simulations. However, examining the environmental interaction terms, we see that the false rejection rates do not exceed 7%, and the relative bias is roughly 4%. This indicates that even under population stratification, the estimates and significance tests for environmental interactions are still valid, even though other genetic effects are inflated.

We repeated this procedure but with only $2,000$ case trios and no control trios to verify that the case trio model is indeed robust to population stratification. We show the results of the case-only model in Tables 4.5 and 4.6. As expected, with only cases, for the main effects, the relative bias shrinks to $< 0.5\%$ and the rejection rate is consistently near 0.05. In addition, the interaction effects are also estimated with little bias and an appropriate rejection rate.

Table 4.3: Average relative biases across 2,000 datasets for various models in our simulation study on population stratification, where the data has both cases and controls.

C	M	E : M	E : I_{M}	E : I_{F}	I_{M}	I_{F}	Mating type
0.08137	0.09690	-0.03367					HWE
0.16134	0.15991		-0.03268		-0.16046		HWE
-0.03861	0.15952			-0.03684		0.22777	HWE
0.07813	0.09288	-0.03178					MS
0.15258	0.15118		-0.03059		-0.15466		MS
-0.03844	0.15079			-0.03594		0.21797	MS
0.06040	0.15136	-0.04313					MaS
0.12591	0.17841		-0.03446		-0.12695		MaS
-0.03170	0.17826			-0.04367		0.18582	MaS

Table 4.4: Rejection rates at $\alpha < 0.05$ across 2,000 datasets for various models in our simulation study on population stratification, where the data includes both cases and controls.

C	M	E : M	E : I_{M}	E : I_{F}	I_{M}	I_{F}	Mating type
0.2775	0.2345	0.0670					HWE
0.4825	0.4675		0.0670		0.1915		HWE
0.0570	0.4670			0.055		0.257	HWE
0.2625	0.2175	0.0630					MS
0.4485	0.4285		0.0665		0.1835		MS
0.0660	0.4295			0.057		0.242	MS
0.1650	0.3035	0.0645					MaS
0.2945	0.4805		0.0610		0.1325		MaS
0.0600	0.4795			0.066		0.176	MaS

Table 4.5: Average relative biases across $2,000$ datasets for various models in our simulation study on population stratification, where the data only has cases.

C	M	E : M	E : I_{M}	E : I_{F}	I_{M}	I_{F}	Mating type
-0.00010	-0.00245	0.00110					HWE
0.00229	0.00044		-0.00032		-0.00489		HWE
-0.00229	0.00021			-0.00068		0.00441	HWE
-0.00008	-0.00253	0.00123					MS
0.00246	0.00054		-0.00026		-0.00521		MS
-0.00243	0.00034			-0.00073		0.00479	MS

Table 4.6: Rejection rates at $\alpha < 0.05$ across $2,000$ datasets for various models in our simulation study on population stratification, where the data only has cases.

C	M	E : M	E : I_{M}	E : I_{F}	I_{M}	I_{F}	Mating type
0.0510	0.0475	0.052					HWE
0.0435	0.0450		0.0505		0.0500		HWE
0.0485	0.0450			0.0425		0.0445	HWE
0.0540	0.0510	0.056					MS
0.0510	0.0505		0.0565		0.0555		MS
0.0565	0.0500			0.0475		0.0530	MS

Chapter 5

Discussion & Conclusion

In this thesis, we introduced our software package, **TriLLIEM**, for fitting the Weinberg et al. [27] log-linear model, compared it to existing software, and used **TriLLIEM** to assess the log-linear model in circumstances that have not been thoroughly studied before. We designed **TriLLIEM** to fill the gaps in software packages, thereby providing users with an accessible method of fitting log-linear models to genotype data on triads.

In Chapter 2, we introduced the genetic background required for modelling genotype data on trios. We described how the categories corresponding to maternal, paternal, and child genotype values could be modelled with a multinomial distribution with hyperparameters corresponding to genetic and environmental effects. This also motivated the log-linear model, which allows the probabilities of the multinomial distribution to be captured by a Poisson generalized linear model. We then discussed the genetic factors that could contribute to disease risk, such as maternal effects, imprinting, and gene-environment interactions, and we showed how they are modelled.

In Chapter 3, we outlined the features of existing software packages for multinomial (**EMIM**) and log-linear (**Haplin**) modelling. We provided details on the use of **TriLLIEM** to simulate and model triad data. Using simulated data with various effect sizes, we com-

pared estimates between **EMIM** and **Haplin** to ensure agreement when it was expected. The differences in functionality between **EMIM**, **Haplin**, and **TriLLIEM** are summarized in Table 5.1.

Table 5.1: Comparison of features between **Haplin**, **EMIM**, and **TriLLIEM**.

Functionality	**Haplin**	**EMIM**	**TriLLIEM**
Availability	R package, available on CRAN	Standalone FORTRAN package, compiled from source or Windows/Linux binaries	R package, available via GitHub
Underlying probability model	Poisson log-linear model	Multinomial model	Poisson log-linear model
Mating type models	HWE only	HWE, MS, MaS	HWE, MS, MaS
Main genetic effects	• Maternal effects • Child effects	• Maternal effects • Child effects	• Maternal effects • Child effects
Parent-of-origin effects	• Parent-of-origin effects (modified parameterization)	• Maternal imprinting effect • Paternal imprinting effect	• Maternal imprinting effect • Paternal imprinting effect
Gene-environment interactions	Stratified approach implemented in the software	Not implemented, user may code the stratified approach themselves	Both stratified and non-stratified approaches are implemented

The main analysis in this thesis was in Chapter 4, where we conducted two simulation studies using **TriLLIEM**. The first was to examine the log-linear model's performance when modelling single populations simulated under many different scenarios (e.g. mating asymmetry, true genetic effects), and to compare the stratified and non-stratified approaches to environmental interactions. The second was to examine the log-linear model's Type 1 error when modelling gene-environment interactions when population stratification is present. These specific simulations were chosen as they act as extensions to the findings from other researchers. For instance, imprinting effects were considered in the original

Weinberg model [27], and further studied in Ainsworth et al. [1], but work on imprinting by environment interactions is limited to the rigid parameterization used in **Haplin** [6]. A goal we had with this work was to exploit the flexibility of generalized linear models to explore gene-environment interactions.

From our first simulation study, we confirmed that mating asymmetry results in a biased maternal effect estimator when the HWE or MS mating type model is erroneously assumed. Furthermore, we observed spurious significance for other parameters (e.g. child effects when modelling maternal imprinting by environment interactions) when control triads were present and the data were simulated under mating asymmetry. We also showed that the stratified approach results in a substantial loss of power for imprinting by environment interactions compared to the non-stratified approach.

From our second simulation study, we found that Type 1 error rates for the main genetic effects were inflated under all mating type models when data were simulated under population stratification with both case- and control-triads when estimating gene-environment interactions. These interaction terms, however, had Type 1 error rates in the range of 5.5% to 6.7%, which is only slightly inflated from the desired 5% rate, and in some cases is within simulation error. We also found that when the data included only case-triads, population stratification did not affect the Type 1 error rates of any of the modelled genetic factors. These findings corroborated those found by Shin et al. [21].

TriLLIEM can now be used to repeat the analyses in several other studies that could only use **Haplin** or **EMIM**. For instance, in Rasevic et al. [19], the authors used **Haplin** to examine risk of orofacial clefts based on environmental factors like smoking and folic acid supplementation. This analysis used **Haplin** to identify maternal and imprinting effects, however this required that they used the HWE mating type model, which may not have held in the population under study, and they were forced to use the stratified approach for assessing maternal gene-environment and parent-of-origin by environment interactions.

TriLLIEM could be used to repeat this analysis under other mating type models, and permits the selection of specific interaction terms to be included to increase statistical power. Similarly, for the other congenital diseases mentioned in Chapter 2 (spina bifida, maternal phenylketonuria, etc.), **TriLLIEM** can be used to identify and quantify risk from various genetic effects. Most importantly, **TriLLIEM** is accessible and easy to use, making the research process more efficient and more open to researchers from non-statistical backgrounds. We believe that this improved accessibility will lead to more frequent and accurate discoveries in the field of genetic epidemiology.

There are several future goals we have in mind for improving **TriLLIEM**. First, we will allow users to include data with missing counts, and use the EM algorithm to estimate the missing counts to perform the appropriate significance tests, as is possible in **Haplin** [7]. For instance, there could be genotype failure during data collection that results in the father's genotype being unknown. Since this is already a feature in **Haplin** and **EMIM**, allowing sporadic missing genotype data in **TriLLIEM** will increase its utility. We will also implement significance tests beyond the Wald test for individual parameters in the model, such as the likelihood ratio test for comparing nested models. Although the likelihood is provided by the `glm()` function, which is used for estimation of genetic effects, when the EM algorithm is required (imprinting, missing genotype data), the likelihood returned by the `glm()` function in R is not correct. Several optimizations could be made in **TriLLIEM**, such as its implementation of the EM algorithm, which runs significantly slower than **Haplin**'s, and pales in speed compared to **EMIM**'s direct search. This can be improved by refactoring some of the relevant code to use a compiled C++ library rather than relying purely on R. Lastly, we hope to add more features that make **TriLLIEM** an effective option for users from non-statistical backgrounds, such as functions that automate model selection or provide clear interpretations of the statistical tests and results that **TriLLIEM** outputs.

Appendix A

Detailed Conditions and Results for the Simulation Studies

Table A.1: Conditions used for simulating data in our simulation study. This table is a detailed breakdown of the 72 different conditions outlined in Table 4.1.

ID	R	S	V	C_i	I_{M}	I_{F}	ε	Interaction
1	1.00	1	1.0	0.85	1.0	1.0	0.3	I_{M}
2	1.15	1	1.0	0.85	1.0	1.0	0.3	I_{M}
3	1.00	1	1.5	0.85	1.0	1.0	0.3	I_{M}
4	1.00	1	1.6	0.85	1.0	1.0	0.3	I_{M}
5	1.00	1	1.0	1.00	1.0	1.0	0.3	I_{M}
6	1.15	1	1.0	1.00	1.0	1.0	0.3	I_{M}
7	1.00	1	1.5	1.00	1.0	1.0	0.3	I_{M}
8	1.00	1	1.6	1.00	1.0	1.0	0.3	I_{M}
9	1.00	1	1.0	1.15	1.0	1.0	0.3	I_{M}
10	1.15	1	1.0	1.15	1.0	1.0	0.3	I_{M}
11	1.00	1	1.5	1.15	1.0	1.0	0.3	I_{M}

12	1.00	1	1.6	1.15	1.0	1.0	0.3	I_{M}
13	1.00	1	1.0	0.85	1.2	1.0	0.3	I_{M}
14	1.00	1	1.0	1.00	1.2	1.0	0.3	I_{M}
15	1.00	1	1.0	1.15	1.2	1.0	0.3	I_{M}
16	1.00	1	1.0	0.85	1.4	1.0	0.3	I_{M}
17	1.00	1	1.0	1.00	1.4	1.0	0.3	I_{M}
18	1.00	1	1.0	1.15	1.4	1.0	0.3	I_{M}
19	1.00	1	1.0	0.85	1.0	1.0	0.3	I_{F}
20	1.15	1	1.0	0.85	1.0	1.0	0.3	I_{F}
21	1.00	1	1.5	0.85	1.0	1.0	0.3	I_{F}
22	1.00	1	1.6	0.85	1.0	1.0	0.3	I_{F}
23	1.00	1	1.0	1.00	1.0	1.0	0.3	I_{F}
24	1.15	1	1.0	1.00	1.0	1.0	0.3	I_{F}
25	1.00	1	1.5	1.00	1.0	1.0	0.3	I_{F}
26	1.00	1	1.6	1.00	1.0	1.0	0.3	I_{F}
27	1.00	1	1.0	1.15	1.0	1.0	0.3	I_{F}
28	1.15	1	1.0	1.15	1.0	1.0	0.3	I_{F}
29	1.00	1	1.5	1.15	1.0	1.0	0.3	I_{F}
30	1.00	1	1.6	1.15	1.0	1.0	0.3	I_{F}
31	1.00	1	1.0	0.85	1.0	1.2	0.3	I_{F}
32	1.00	1	1.0	1.00	1.0	1.2	0.3	I_{F}
33	1.00	1	1.0	1.15	1.0	1.2	0.3	I_{F}
34	1.00	1	1.0	0.85	1.0	1.4	0.3	I_{F}
35	1.00	1	1.0	1.00	1.0	1.4	0.3	I_{F}
36	1.00	1	1.0	1.15	1.0	1.4	0.3	I_{F}
37	1.00	1	1.0	0.85	1.0	1.0	0.5	I_{M}
38	1.15	1	1.0	0.85	1.0	1.0	0.5	I_{M}

39	1.00	1	1.5	0.85	1.0	1.0	0.5	I_{M}
40	1.00	1	1.6	0.85	1.0	1.0	0.5	I_{M}
41	1.00	1	1.0	1.00	1.0	1.0	0.5	I_{M}
42	1.15	1	1.0	1.00	1.0	1.0	0.5	I_{M}
43	1.00	1	1.5	1.00	1.0	1.0	0.5	I_{M}
44	1.00	1	1.6	1.00	1.0	1.0	0.5	I_{M}
45	1.00	1	1.0	1.15	1.0	1.0	0.5	I_{M}
46	1.15	1	1.0	1.15	1.0	1.0	0.5	I_{M}
47	1.00	1	1.5	1.15	1.0	1.0	0.5	I_{M}
48	1.00	1	1.6	1.15	1.0	1.0	0.5	I_{M}
49	1.00	1	1.0	0.85	1.2	1.0	0.5	I_{M}
50	1.00	1	1.0	1.00	1.2	1.0	0.5	I_{M}
51	1.00	1	1.0	1.15	1.2	1.0	0.5	I_{M}
52	1.00	1	1.0	0.85	1.4	1.0	0.5	I_{M}
53	1.00	1	1.0	1.00	1.4	1.0	0.5	I_{M}
54	1.00	1	1.0	1.15	1.4	1.0	0.5	I_{M}
55	1.00	1	1.0	0.85	1.0	1.0	0.5	I_{F}
56	1.15	1	1.0	0.85	1.0	1.0	0.5	I_{F}
57	1.00	1	1.5	0.85	1.0	1.0	0.5	I_{F}
58	1.00	1	1.6	0.85	1.0	1.0	0.5	I_{F}
59	1.00	1	1.0	1.00	1.0	1.0	0.5	I_{F}
60	1.15	1	1.0	1.00	1.0	1.0	0.5	I_{F}
61	1.00	1	1.5	1.00	1.0	1.0	0.5	I_{F}
62	1.00	1	1.6	1.00	1.0	1.0	0.5	I_{F}
63	1.00	1	1.0	1.15	1.0	1.0	0.5	I_{F}
64	1.15	1	1.0	1.15	1.0	1.0	0.5	I_{F}
65	1.00	1	1.5	1.15	1.0	1.0	0.5	I_{F}

66	1.00	1	1.6	1.15	1.0	1.0	0.5	I_{F}
67	1.00	1	1.0	0.85	1.0	1.2	0.5	I_{F}
68	1.00	1	1.0	1.00	1.0	1.2	0.5	I_{F}
69	1.00	1	1.0	1.15	1.0	1.2	0.5	I_{F}
70	1.00	1	1.0	0.85	1.0	1.4	0.5	I_{F}
71	1.00	1	1.0	1.00	1.0	1.4	0.5	I_{F}
72	1.00	1	1.0	1.15	1.0	1.4	0.5	I_{F}

Table A.2: Relative biases for simulated data sets without controls where E = 0.3, where the interaction parameter is I_M. S and NS stand for the stratified and non-stratified approaches for gene-environment interactions, respectively. The MaS column represents the value of the mating asymmetry coefficients used during simulation. To see the conditions scenario conditions, please use the ID column to refer to the appropriate row of Table A.1. Results from data simulated with MaS coefficients of 1.15 are omitted for brevity.

ID	C	M	E : I_M	I_M	MT	E approach	True effect	MaS
1	0.00157	-0.21201		-0.00090	HWE	no E	None	0.85
1	0.00059	-0.21250	-0.00715	0.00104	HWE	S	None	0.85
1	0.00170	-0.21193	-0.00332	-0.00057	HWE	NS	None	0.85
1	0.00158	-0.21385		-0.00092	MS	no E	None	0.85
1	0.00056	-0.21439	-0.00728	0.00108	MS	S	None	0.85
1	0.00173	-0.21384	-0.00349	-0.00051	MS	NS	None	0.85
2	0.00023	-0.21012		-0.00009	HWE	no E	C	0.85
2	0.00005	-0.21057	-0.00064	0.00037	HWE	S	C	0.85
2	0.00035	-0.21008	-0.00033	-0.00056	HWE	NS	C	0.85
2	-0.00024	-0.21220		0.00006	MS	no E	C	0.85
2	-0.00028	-0.21266	-0.00013	0.00046	MS	S	C	0.85
2	-0.00011	-0.21226	-0.00058	-0.00026	MS	NS	C	0.85
3	0.00156	-0.20933		0.13400	HWE	no E	E : I_M	0.85
3	0.00063	-0.20909	-0.00213	-0.00227	HWE	S	E : I_M	0.85
3	0.00184	-0.20912	-0.00052	-0.00329	HWE	NS	E : I_M	0.85
3	0.00168	-0.21027		0.13356	MS	no E	E : I_M	0.85
3	0.00063	-0.21108	-0.00081	-0.00222	MS	S	E : I_M	0.85
3	0.00188	-0.21062	0.00154	-0.00374	MS	NS	E : I_M	0.85
4	0.00311	-0.20749		0.15162	HWE	no E	E : I_M	0.85

Table A.2: *(continued)*

ID	C	M	E : I_{M}	I_{M}	MT	E approach	True effect	MaS
4	0.00209	-0.20824	-0.01110	-0.00481	HWE	S	E : I_{M}	0.85
4	0.00323	-0.20742	-0.00696	-0.00665	HWE	NS	E : I_{M}	0.85
4	0.00319	-0.20818		0.15106	MS	no E	E : I_{M}	0.85
4	0.00209	-0.20993	-0.01121	-0.00486	MS	S	E : I_{M}	0.85
4	0.00329	-0.20852	-0.00567	-0.00727	MS	NS	E : I_{M}	0.85
5	-0.00092	0.00209		-0.00220	HWE	no E	None	1.00
5	-0.00176	0.00240	-0.00200	-0.00142	HWE	S	None	1.00
5	-0.00077	0.00224	-0.00204	-0.00228	HWE	NS	None	1.00
5	-0.00095	0.00212		-0.00218	MS	no E	None	1.00
5	-0.00183	0.00237	-0.00246	-0.00133	MS	S	None	1.00
5	-0.00080	0.00224	-0.00207	-0.00223	MS	NS	None	1.00
6	0.00349	0.00105		-0.00460	HWE	no E	C	1.00
6	0.00470	0.00155	0.00509	-0.00631	HWE	S	C	1.00
6	0.00363	0.00116	0.00091	-0.00537	HWE	NS	C	1.00
6	0.00343	0.00102		-0.00448	MS	no E	C	1.00
6	0.00462	0.00147	0.00534	-0.00610	MS	S	C	1.00
6	0.00357	0.00111	0.00095	-0.00524	MS	NS	C	1.00
7	0.00070	-0.00238		0.14066	HWE	no E	E : I_{M}	1.00
7	0.00034	-0.00249	-0.00795	0.00616	HWE	S	E : I_{M}	1.00
7	0.00120	-0.00188	-0.00558	0.00483	HWE	NS	E : I_{M}	1.00
7	0.00070	-0.00239		0.14077	MS	no E	E : I_{M}	1.00
7	0.00037	-0.00243	-0.00704	0.00608	MS	S	E : I_{M}	1.00

Table A.2: *(continued)*

ID	C	M	E : I_{M}	I_{M}	MT	E approach	True effect	MaS
7	0.00118	-0.00202	-0.00528	0.00495	MS	NS	E : I_{M}	1.00
8	-0.00165	-0.00066		0.16310	HWE	no E	E : I_{M}	1.00
8	-0.00225	-0.00321	-0.01179	0.00641	HWE	S	E : I_{M}	1.00
8	-0.00142	-0.00044	-0.00341	0.00342	HWE	NS	E : I_{M}	1.00
8	-0.00158	-0.00053		0.16291	MS	no E	E : I_{M}	1.00
8	-0.00214	-0.00308	-0.01163	0.00621	MS	S	E : I_{M}	1.00
8	-0.00131	-0.00020	-0.00339	0.00313	MS	NS	E : I_{M}	1.00
13	-0.00195	-0.21477		0.00484	HWE	no E	I_{M}	0.85
13	-0.00275	-0.21580	-0.00694	0.00711	HWE	S	I_{M}	0.85
13	-0.00183	-0.21470	-0.00314	0.00515	HWE	NS	I_{M}	0.85
13	-0.00185	-0.21546		0.00433	MS	no E	I_{M}	0.85
13	-0.00256	-0.21644	-0.00620	0.00639	MS	S	I_{M}	0.85
13	-0.00171	-0.21547	-0.00291	0.00463	MS	NS	I_{M}	0.85
14	-0.00094	-0.00139		0.00379	HWE	no E	I_{M}	1.00
14	-0.00098	-0.00354	-0.00659	0.00576	HWE	S	I_{M}	1.00
14	-0.00083	-0.00128	-0.00124	0.00371	HWE	NS	I_{M}	1.00
14	-0.00093	-0.00141		0.00387	MS	no E	I_{M}	1.00
14	-0.00098	-0.00360	-0.00692	0.00596	MS	S	I_{M}	1.00
14	-0.00080	-0.00123	-0.00132	0.00374	MS	NS	I_{M}	1.00
16	-0.00296	-0.21261		0.00199	HWE	no E	I_{M}	0.85
16	-0.00250	-0.21136	0.00592	0.00056	HWE	S	I_{M}	0.85
16	-0.00291	-0.21261	-0.00090	0.00199	HWE	NS	I_{M}	0.85

Table A.2: *(continued)*

ID	C	M	E : I_{M}	I_{M}	MT	E approach	True effect	MaS
16	-0.00281	-0.21280		0.00147	MS	no E	I_{M}	0.85
16	-0.00232	-0.21151	0.00669	-0.00006	MS	S	I_{M}	0.85
16	-0.00273	-0.21283	-0.00079	0.00142	MS	NS	I_{M}	0.85
17	0.00156	-0.00021		-0.00095	HWE	no E	I_{M}	1.00
17	-0.00088	-0.00110	-0.00557	0.00097	HWE	S	I_{M}	1.00
17	0.00169	-0.00008	0.00207	-0.00194	HWE	NS	I_{M}	1.00
17	0.00157	-0.00022		-0.00100	MS	no E	I_{M}	1.00
17	-0.00088	-0.00103	-0.00586	0.00103	MS	S	I_{M}	1.00
17	0.00171	-0.00008	0.00173	-0.00191	MS	NS	I_{M}	1.00

Table A.3: Rejection rates at $\alpha < 0.05$ for simulated data sets without controls where E = 0.3, where the environmental interaction parameter is I_{M}. S and NS stand for the stratified and non-stratified approaches for gene-environment interactions, respectively. The MaS column represents the value of the mating asymmetry coefficients used during simulation. To see the scenario conditions, please use the ID column to refer to the appropriate row of Table A.1. Results from data simulated with MaS coefficients of 1.15 are omitted for brevity.

ID	C	M	E : I_{M}	I_{M}	MT	E approach	True effect	MaS
1	0.0450	0.8370		0.0465	HWE	no E	None	0.85
1	0.0915	0.7575	0.0600	0.0990	HWE	S	None	0.85
1	0.0680	0.8680	0.0490	0.0875	HWE	NS	None	0.85
1	0.0430	0.8430		0.0455	MS	no E	None	0.85
1	0.0910	0.7635	0.0610	0.0960	MS	S	None	0.85
1	0.0615	0.8735	0.0525	0.0870	MS	NS	None	0.85
2	0.4355	0.8290		0.0420	HWE	no E	C	0.85
2	0.4020	0.7500	0.0600	0.0940	HWE	S	C	0.85
2	0.4890	0.8595	0.0515	0.0840	HWE	NS	C	0.85
2	0.4320	0.8295		0.0410	MS	no E	C	0.85
2	0.4010	0.7560	0.0600	0.0960	MS	S	C	0.85
2	0.4880	0.8635	0.0540	0.0805	MS	NS	C	0.85
3	0.0490	0.8225		0.1705	HWE	no E	E : I_{M}	0.85
3	0.0800	0.7375	0.3545	0.0905	HWE	S	E : I_{M}	0.85
3	0.0710	0.8580	0.8775	0.0850	HWE	NS	E : I_{M}	0.85
3	0.0495	0.8250		0.1705	MS	no E	E : I_{M}	0.85
3	0.0775	0.7360	0.3525	0.0895	MS	S	E : I_{M}	0.85
3	0.0710	0.8590	0.8800	0.0840	MS	NS	E : I_{M}	0.85
4	0.0460	0.8105		0.1975	HWE	no E	E : I_{M}	0.85

Table A.3: *(continued)*

ID	C	M	E : I_{M}	I_{M}	MT	E approach	True effect	MaS
4	0.0755	0.7460	0.4195	0.0935	HWE	S	E : I_{M}	0.85
4	0.0675	0.8470	0.9435	0.0855	HWE	NS	E : I_{M}	0.85
4	0.0455	0.8130		0.1950	MS	no E	E : I_{M}	0.85
4	0.0725	0.7440	0.4150	0.0945	MS	S	E : I_{M}	0.85
4	0.0665	0.8475	0.9440	0.0845	MS	NS	E : I_{M}	0.85
5	0.0480	0.0460		0.0550	HWE	no E	None	1.00
5	0.0750	0.0770	0.0540	0.0955	HWE	S	None	1.00
5	0.0680	0.0685	0.0615	0.0880	HWE	NS	None	1.00
5	0.0480	0.0455		0.0535	MS	no E	None	1.00
5	0.0755	0.0765	0.0515	0.0935	MS	S	None	1.00
5	0.0675	0.0665	0.0625	0.0885	MS	NS	None	1.00
6	0.4335	0.0520		0.0570	HWE	no E	C	1.00
6	0.4045	0.0790	0.0705	0.1100	HWE	S	C	1.00
6	0.4995	0.0665	0.0545	0.0900	HWE	NS	C	1.00
6	0.4330	0.0525		0.0580	MS	no E	C	1.00
6	0.4050	0.0790	0.0670	0.1105	MS	S	C	1.00
6	0.4965	0.0685	0.0540	0.0890	MS	NS	C	1.00
7	0.0435	0.0505		0.1760	HWE	no E	E : I_{M}	1.00
7	0.0730	0.0720	0.3425	0.0965	HWE	S	E : I_{M}	1.00
7	0.0610	0.0680	0.8780	0.0795	HWE	NS	E : I_{M}	1.00
7	0.0445	0.0525		0.1755	MS	no E	E : I_{M}	1.00
7	0.0725	0.0725	0.3395	0.0965	MS	S	E : I_{M}	1.00

Table A.3: *(continued)*

ID	C	M	E : I_{M}	I_{M}	MT	E approach	True effect	MaS
7	0.0580	0.0690	0.8760	0.0800	MS	NS	E : I_{M}	1.00
8	0.0460	0.0545		0.2220	HWE	no E	E : I_{M}	1.00
8	0.0715	0.0785	0.4160	0.0955	HWE	S	E : I_{M}	1.00
8	0.0655	0.0705	0.9600	0.0800	HWE	NS	E : I_{M}	1.00
8	0.0435	0.0525		0.2180	MS	no E	E : I_{M}	1.00
8	0.0720	0.0790	0.4115	0.0930	MS	S	E : I_{M}	1.00
8	0.0685	0.0715	0.9600	0.0785	MS	NS	E : I_{M}	1.00
13	0.0420	0.8400		0.3050	HWE	no E	I_{M}	0.85
13	0.0695	0.7665	0.0605	0.3435	HWE	S	I_{M}	0.85
13	0.0610	0.8710	0.0450	0.3865	HWE	NS	I_{M}	0.85
13	0.0410	0.8430		0.3030	MS	no E	I_{M}	0.85
13	0.0685	0.7660	0.0575	0.3410	MS	S	I_{M}	0.85
13	0.0605	0.8690	0.0500	0.3830	MS	NS	I_{M}	0.85
14	0.0535	0.0545		0.3065	HWE	no E	I_{M}	1.00
14	0.0765	0.0840	0.0650	0.3350	HWE	S	I_{M}	1.00
14	0.0720	0.0765	0.0520	0.3810	HWE	NS	I_{M}	1.00
14	0.0520	0.0530		0.3025	MS	no E	I_{M}	1.00
14	0.0740	0.0860	0.0640	0.3305	MS	S	I_{M}	1.00
14	0.0715	0.0775	0.0485	0.3750	MS	NS	I_{M}	1.00
16	0.0545	0.8260		0.7490	HWE	no E	I_{M}	0.85
16	0.0820	0.7455	0.0680	0.7150	HWE	S	I_{M}	0.85
16	0.0770	0.8680	0.0605	0.7940	HWE	NS	I_{M}	0.85

Table A.3: *(continued)*

ID	C	M	E : I_{M}	I_{M}	MT	E approach	True effect	MaS
16	0.0535	0.8305		0.7450	MS	no E	I_{M}	0.85
16	0.0830	0.7420	0.0695	0.7155	MS	S	I_{M}	0.85
16	0.0775	0.8665	0.0585	0.7910	MS	NS	I_{M}	0.85
17	0.0465	0.0425		0.7540	HWE	no E	I_{M}	1.00
17	0.0755	0.0780	0.0635	0.7145	HWE	S	I_{M}	1.00
17	0.0635	0.0635	0.0475	0.7930	HWE	NS	I_{M}	1.00
17	0.0465	0.0420		0.7515	MS	no E	I_{M}	1.00
17	0.0740	0.0735	0.0585	0.7120	MS	S	I_{M}	1.00
17	0.0640	0.0605	0.0485	0.7885	MS	NS	I_{M}	1.00

Table A.4: Relative biases for simulated data sets with controls where E = 0.3, where the environmental interaction paramter is I_{M}. S and NS stand for the stratified and non-stratified approaches for gene-environment interactions, respectively. The MaS column represents the value of the mating asymmetry coefficients used during simulation. To see the scenario conditions, please use the ID column to refer to the appropriate row of Table A.1. Results from data simulated with MaS coefficients of 1.15 are omitted for brevity.

ID	C	M	E : I_{M}	I_{M}	MT model	E approach	True effect	MaS
1	0.11489	-0.11880		-0.13176	HWE	no E	None	0.85
1	0.11685	-0.11951	0.00974	-0.13359	HWE	S	None	0.85
1	0.11490	-0.11882	0.00044	-0.13280	HWE	NS	None	0.85
1	0.11511	-0.11906		-0.13202	MS	no E	None	0.85
1	0.11702	-0.11981	0.00980	-0.13380	MS	S	None	0.85
1	0.11511	-0.11915	0.00040	-0.13300	MS	NS	None	0.85
1	-0.00015	0.00029		-0.00144	MaS	no E	None	0.85
1	0.00137	0.00000	0.00883	-0.00323	MaS	S	None	0.85
1	-0.00020	0.00024	-0.00132	-0.00165	MaS	NS	None	0.85
2	0.11611	-0.11922		-0.13167	HWE	no E	C	0.85
2	0.11655	-0.12048	0.00115	-0.13137	HWE	S	C	0.85
2	0.11625	-0.11915	-0.00120	-0.13243	HWE	NS	C	0.85
2	0.11634	-0.11955		-0.13189	MS	no E	C	0.85
2	0.11677	-0.12089	0.00076	-0.13156	MS	S	C	0.85
2	0.11652	-0.11949	-0.00109	-0.13269	MS	NS	C	0.85
2	-0.00231	-0.00155		0.00552	MaS	no E	C	0.85
2	-0.00179	-0.00307	0.00109	0.00580	MaS	S	C	0.85
2	-0.00222	-0.00143	-0.00123	0.00497	MaS	NS	C	0.85
3	0.11418	-0.11915		-0.01109	HWE	no E	E : I_{M}	0.85

Table A.4: *(continued)*

ID	C	M	E : I_{M}	I_{M}	MT model	E approach	True effect	MaS
3	0.11330	-0.12198	-0.00667	-0.12693	HWE	S	E : I_{M}	0.85
3	0.11383	-0.11948	-0.00156	-0.12974	HWE	NS	E : I_{M}	0.85
3	0.11437	-0.11941		-0.01138	MS	no E	E : I_{M}	0.85
3	0.11360	-0.12238	-0.00598	-0.12724	MS	S	E : I_{M}	0.85
3	0.11397	-0.11988	-0.00133	-0.12986	MS	NS	E : I_{M}	0.85
3	-0.00209	-0.00231		0.14127	MaS	no E	E : I_{M}	0.85
3	-0.00135	-0.00324	0.00573	0.00358	MaS	S	E : I_{M}	0.85
3	-0.00228	-0.00258	-0.00087	0.00423	MaS	NS	E : I_{M}	0.85
4	0.11819	-0.11742		0.00402	HWE	no E	E : I_{M}	0.85
4	0.11856	-0.11880	0.00081	-0.13473	HWE	S	E : I_{M}	0.85
4	0.11800	-0.11762	-0.00152	-0.13549	HWE	NS	E : I_{M}	0.85
4	0.11834	-0.11756		0.00383	MS	no E	E : I_{M}	0.85
4	0.11876	-0.11916	0.00094	-0.13488	MS	S	E : I_{M}	0.85
4	0.11811	-0.11783	-0.00118	-0.13566	MS	NS	E : I_{M}	0.85
4	0.00183	-0.00126		0.15879	MaS	no E	E : I_{M}	0.85
4	0.00351	-0.00002	0.01201	-0.00520	MaS	S	E : I_{M}	0.85
4	0.00192	-0.00110	-0.00137	-0.00258	MaS	NS	E : I_{M}	0.85
5	-0.00233	0.00151		0.00132	HWE	no E	None	1.00
5	-0.00372	0.00199	-0.00324	0.00283	HWE	S	None	1.00
5	-0.00217	0.00166	-0.00572	0.00161	HWE	NS	None	1.00
5	-0.00226	0.00159		0.00119	MS	no E	None	1.00
5	-0.00365	0.00207	-0.00312	0.00271	MS	S	None	1.00

Table A.4: *(continued)*

ID	C	M	E : I_{M}	I_{M}	MT model	E approach	True effect	MaS
5	-0.00210	0.00174	-0.00573	0.00149	MS	NS	None	1.00
5	-0.00119	0.00091		-0.00015	MaS	no E	None	1.00
5	-0.00245	0.00140	-0.00264	0.00123	MaS	S	None	1.00
5	-0.00115	0.00104	-0.00358	-0.00004	MaS	NS	None	1.00
6	-0.00133	-0.00011		-0.00082	HWE	no E	C	1.00
6	-0.00308	-0.00353	-0.01423	0.00403	HWE	S	C	1.00
6	-0.00124	-0.00003	-0.00195	-0.00124	HWE	NS	C	1.00
6	-0.00132	-0.00009		-0.00080	MS	no E	C	1.00
6	-0.00308	-0.00358	-0.01444	0.00408	MS	S	C	1.00
6	-0.00118	-0.00001	-0.00188	-0.00132	MS	NS	C	1.00
6	-0.00161	0.00061		-0.00029	MaS	no E	C	1.00
6	-0.00331	-0.00261	-0.01386	0.00438	MaS	S	C	1.00
6	-0.00153	0.00066	-0.00248	-0.00032	MaS	NS	C	1.00
7	-0.00006	-0.00058		0.13454	HWE	no E	E : I_{M}	1.00
7	0.00055	-0.00093	0.00825	-0.00258	HWE	S	E : I_{M}	1.00
7	0.00003	-0.00049	0.00221	-0.00249	HWE	NS	E : I_{M}	1.00
7	-0.00007	-0.00059		0.13465	MS	no E	E : I_{M}	1.00
7	0.00055	-0.00095	0.00809	-0.00250	MS	S	E : I_{M}	1.00
7	0.00010	-0.00046	0.00226	-0.00257	MS	NS	E : I_{M}	1.00
7	0.00065	-0.00095		0.13373	MaS	no E	E : I_{M}	1.00
7	0.00134	-0.00121	0.00862	-0.00340	MaS	S	E : I_{M}	1.00
7	0.00071	-0.00096	0.00246	-0.00293	MaS	NS	E : I_{M}	1.00

Table A.4: *(continued)*

ID	C	M	E : I_{M}	I_{M}	MT model	E approach	True effect	MaS
8	-0.00348	-0.00198		0.16407	HWE	no E	E : I_{M}	1.00
8	-0.00450	-0.00117	0.00242	0.00435	HWE	S	E : I_{M}	1.00
8	-0.00352	-0.00202	-0.00260	0.00416	HWE	NS	E : I_{M}	1.00
8	-0.00345	-0.00183		0.16395	MS	no E	E : I_{M}	1.00
8	-0.00449	-0.00117	0.00147	0.00435	MS	S	E : I_{M}	1.00
8	-0.00341	-0.00168	-0.00259	0.00374	MS	NS	E : I_{M}	1.00
8	-0.00271	-0.00230		0.16303	MaS	no E	E : I_{M}	1.00
8	-0.00378	-0.00141	0.00152	0.00356	MaS	S	E : I_{M}	1.00
8	-0.00276	-0.00222	-0.00104	0.00285	MaS	NS	E : I_{M}	1.00
13	0.11876	-0.11802		-0.13423	HWE	no E	I_{M}	0.85
13	0.11999	-0.11934	0.00238	-0.13394	HWE	S	I_{M}	0.85
13	0.11883	-0.11799	-0.00293	-0.13447	HWE	NS	I_{M}	0.85
13	0.11887	-0.11817		-0.13445	MS	no E	I_{M}	0.85
13	0.12010	-0.11953	0.00243	-0.13423	MS	S	I_{M}	0.85
13	0.11895	-0.11819	-0.00263	-0.13474	MS	NS	I_{M}	0.85
13	0.00098	-0.00141		0.00145	MaS	no E	I_{M}	0.85
13	0.00182	-0.00232	0.00179	0.00203	MaS	S	I_{M}	0.85
13	0.00098	-0.00139	-0.00459	0.00204	MaS	NS	I_{M}	0.85
14	-0.00226	-0.00004		0.00551	HWE	no E	I_{M}	1.00
14	-0.00371	-0.00127	-0.01156	0.00999	HWE	S	I_{M}	1.00
14	-0.00216	0.00005	-0.00835	0.00709	HWE	NS	I_{M}	1.00
14	-0.00222	0.00002		0.00542	MS	no E	I_{M}	1.00

Table A.4: *(continued)*

ID	C	M	E : I_{M}	I_{M}	MT model	E approach	True effect	MaS
14	-0.00365	-0.00116	-0.01180	0.00994	MS	S	I_{M}	1.00
14	-0.00208	0.00018	-0.00856	0.00695	MS	NS	I_{M}	1.00
14	-0.00174	-0.00034		0.00511	MaS	no E	I_{M}	1.00
14	-0.00308	-0.00157	-0.01208	0.00965	MaS	S	I_{M}	1.00
14	-0.00168	-0.00025	-0.00785	0.00681	MaS	NS	I_{M}	1.00
16	0.12028	-0.11463		-0.14026	HWE	no E	I_{M}	0.85
16	0.11701	-0.11410	0.00634	-0.14077	HWE	S	I_{M}	0.85
16	0.12032	-0.11462	0.00744	-0.14275	HWE	NS	I_{M}	0.85
16	0.12023	-0.11465		-0.14031	MS	no E	I_{M}	0.85
16	0.11694	-0.11412	0.00658	-0.14091	MS	S	I_{M}	0.85
16	0.12031	-0.11464	0.00780	-0.14294	MS	NS	I_{M}	0.85
16	0.00234	-0.00085		-0.00388	MaS	no E	I_{M}	0.85
16	-0.00066	-0.00047	0.00523	-0.00406	MaS	S	I_{M}	0.85
16	0.00230	-0.00083	0.01061	-0.00732	MaS	NS	I_{M}	0.85
17	-0.00125	0.00046		0.00091	HWE	no E	I_{M}	1.00
17	0.00008	-0.00057	0.00746	0.00026	HWE	S	I_{M}	1.00
17	-0.00126	0.00045	-0.00044	0.00054	HWE	NS	I_{M}	1.00
17	-0.00125	0.00025		0.00131	MS	no E	I_{M}	1.00
17	0.00017	-0.00072	0.00764	0.00054	MS	S	I_{M}	1.00
17	-0.00119	0.00024	-0.00046	0.00084	MS	NS	I_{M}	1.00
17	-0.00046	-0.00032		0.00045	MaS	no E	I_{M}	1.00
17	0.00046	-0.00063	0.00589	0.00042	MaS	S	I_{M}	1.00

Table A.4: *(continued)*

ID	C	M	E : I_{M}	I_{M}	MT model	E approach	True effect	MaS
17	-0.00051	-0.00046	-0.00306	0.00126	MaS	NS	I_{M}	1.00

Table A.5: Rejection rates at $\alpha < 0.05$ for simulated data sets with controls where E = 0.3, where the environmental interaction paramter is I_{M}. S and NS stand for the stratified and non-stratified approaches for gene-environment interactions, respectively. The MaS column represents the value of the mating asymmetry coefficients used during simulation. To see the scenario conditions, please use the ID column to refer to the appropriate row of Table A.1. Results from data simulated with MaS coefficients of 1.15 are omitted for brevity.

ID	C	M	E : I_{M}	I_{M}	MT	E approach	True effect	MaS
1	0.2815	0.3100		0.1520	HWE	no E	None	0.85
1	0.2465	0.2725	0.0585	0.1765	HWE	S	None	0.85
1	0.3095	0.3440	0.0515	0.1840	HWE	NS	None	0.85
1	0.2825	0.3140		0.1505	MS	no E	None	0.85
1	0.2495	0.2725	0.0580	0.1760	MS	S	None	0.85
1	0.3085	0.3405	0.0545	0.1825	MS	NS	None	0.85
1	0.0560	0.0505		0.0505	MaS	no E	None	0.85
1	0.0785	0.0535	0.0645	0.0875	MaS	S	None	0.85
1	0.0690	0.0570	0.0555	0.0885	MaS	NS	None	0.85
2	0.8745	0.3090		0.1635	HWE	no E	C	0.85
2	0.7865	0.2815	0.0560	0.1795	HWE	S	C	0.85
2	0.8945	0.3405	0.0435	0.1975	HWE	NS	C	0.85
2	0.8745	0.3085		0.1635	MS	no E	C	0.85
2	0.7850	0.2820	0.0555	0.1810	MS	S	C	0.85
2	0.8935	0.3405	0.0430	0.1985	MS	NS	C	0.85
2	0.3770	0.0480		0.0520	MaS	no E	C	0.85
2	0.3235	0.0620	0.0545	0.0935	MaS	S	C	0.85
2	0.4120	0.0560	0.0465	0.0835	MaS	NS	C	0.85
3	0.2785	0.3115		0.0530	HWE	no E	E : I_{M}	0.85

Table A.5: *(continued)*

ID	C	M	E : I_{M}	I_{M}	MT	E approach	True effect	MaS
3	0.2385	0.2685	0.2460	0.1725	HWE	S	E : I_{M}	0.85
3	0.3010	0.3490	0.6720	0.1880	HWE	NS	E : I_{M}	0.85
3	0.2785	0.3105		0.0540	MS	no E	E : I_{M}	0.85
3	0.2435	0.2665	0.2485	0.1730	MS	S	E : I_{M}	0.85
3	0.3035	0.3530	0.6710	0.1860	MS	NS	E : I_{M}	0.85
3	0.0515	0.0550		0.1405	MaS	no E	E : I_{M}	0.85
3	0.0680	0.0630	0.2445	0.0860	MaS	S	E : I_{M}	0.85
3	0.0625	0.0585	0.5870	0.0870	MaS	NS	E : I_{M}	0.85
4	0.2905	0.3055		0.0520	HWE	no E	E : I_{M}	0.85
4	0.2575	0.2625	0.3190	0.1780	HWE	S	E : I_{M}	0.85
4	0.3145	0.3305	0.8025	0.1895	HWE	NS	E : I_{M}	0.85
4	0.2890	0.3055		0.0520	MS	no E	E : I_{M}	0.85
4	0.2585	0.2720	0.3180	0.1785	MS	S	E : I_{M}	0.85
4	0.3120	0.3320	0.8030	0.1875	MS	NS	E : I_{M}	0.85
4	0.0495	0.0490		0.1585	MaS	no E	E : I_{M}	0.85
4	0.0700	0.0565	0.3135	0.0880	MaS	S	E : I_{M}	0.85
4	0.0605	0.0525	0.7315	0.0800	MaS	NS	E : I_{M}	0.85
5	0.0455	0.0505		0.0470	HWE	no E	None	1.00
5	0.0600	0.0515	0.0645	0.0790	HWE	S	None	1.00
5	0.0540	0.0620	0.0485	0.0665	HWE	NS	None	1.00
5	0.0460	0.0505		0.0460	MS	no E	None	1.00
5	0.0605	0.0490	0.0645	0.0775	MS	S	None	1.00

Table A.5: *(continued)*

ID	C	M	E : I_{M}	I_{M}	MT	E approach	True effect	MaS
5	0.0530	0.0630	0.0485	0.0665	MS	NS	None	1.00
5	0.0520	0.0505		0.0520	MaS	no E	None	1.00
5	0.0685	0.0490	0.0670	0.0795	MaS	S	None	1.00
5	0.0645	0.0550	0.0500	0.0780	MaS	NS	None	1.00
6	0.4155	0.0515		0.0595	HWE	no E	C	1.00
6	0.3415	0.0625	0.0615	0.0905	HWE	S	C	1.00
6	0.4465	0.0630	0.0575	0.0935	HWE	NS	C	1.00
6	0.4150	0.0505		0.0570	MS	no E	C	1.00
6	0.3395	0.0610	0.0635	0.0925	MS	S	C	1.00
6	0.4475	0.0600	0.0545	0.0955	MS	NS	C	1.00
6	0.3645	0.0540		0.0615	MaS	no E	C	1.00
6	0.3175	0.0570	0.0625	0.0895	MaS	S	C	1.00
6	0.3925	0.0625	0.0505	0.0910	MaS	NS	C	1.00
7	0.0505	0.0475		0.1430	HWE	no E	E : I_{M}	1.00
7	0.0585	0.0655	0.2745	0.0790	HWE	S	E : I_{M}	1.00
7	0.0610	0.0610	0.6935	0.0695	HWE	NS	E : I_{M}	1.00
7	0.0505	0.0470		0.1425	MS	no E	E : I_{M}	1.00
7	0.0585	0.0650	0.2725	0.0785	MS	S	E : I_{M}	1.00
7	0.0610	0.0585	0.6940	0.0650	MS	NS	E : I_{M}	1.00
7	0.0490	0.0475		0.1285	MaS	no E	E : I_{M}	1.00
7	0.0595	0.0585	0.2595	0.0805	MaS	S	E : I_{M}	1.00
7	0.0570	0.0520	0.6030	0.0745	MaS	NS	E : I_{M}	1.00

Table A.5: *(continued)*

ID	C	M	E : I_{M}	I_{M}	MT	E approach	True effect	MaS
8	0.0480	0.0580		0.1735	HWE	no E	E : I_{M}	1.00
8	0.0640	0.0690	0.3235	0.0955	HWE	S	E : I_{M}	1.00
8	0.0600	0.0665	0.8200	0.0840	HWE	NS	E : I_{M}	1.00
8	0.0480	0.0570		0.1685	MS	no E	E : I_{M}	1.00
8	0.0640	0.0695	0.3205	0.0950	MS	S	E : I_{M}	1.00
8	0.0625	0.0660	0.8185	0.0830	MS	NS	E : I_{M}	1.00
8	0.0590	0.0520		0.1655	MaS	no E	E : I_{M}	1.00
8	0.0690	0.0640	0.3060	0.1005	MaS	S	E : I_{M}	1.00
8	0.0670	0.0580	0.7465	0.0965	MaS	NS	E : I_{M}	1.00
13	0.2920	0.3090		0.0605	HWE	no E	I_{M}	0.85
13	0.2610	0.2620	0.0575	0.0995	HWE	S	I_{M}	0.85
13	0.3220	0.3455	0.0545	0.0885	HWE	NS	I_{M}	0.85
13	0.2925	0.3085		0.0605	MS	no E	I_{M}	0.85
13	0.2590	0.2645	0.0575	0.0995	MS	S	I_{M}	0.85
13	0.3225	0.3445	0.0535	0.0880	MS	NS	I_{M}	0.85
13	0.0450	0.0485		0.2150	MaS	no E	I_{M}	0.85
13	0.0595	0.0640	0.0625	0.2355	MaS	S	I_{M}	0.85
13	0.0585	0.0570	0.0510	0.2490	MaS	NS	I_{M}	0.85
14	0.0610	0.0545		0.2455	HWE	no E	I_{M}	1.00
14	0.0775	0.0675	0.0605	0.2540	HWE	S	I_{M}	1.00
14	0.0740	0.0645	0.0460	0.2825	HWE	NS	I_{M}	1.00
14	0.0595	0.0530		0.2450	MS	no E	I_{M}	1.00

Table A.5: *(continued)*

ID	C	M	E : I_{M}	I_{M}	MT	E approach	True effect	MaS
14	0.0785	0.0665	0.0605	0.2535	MS	S	I_{M}	1.00
14	0.0735	0.0630	0.0450	0.2825	MS	NS	I_{M}	1.00
14	0.0600	0.0530		0.2270	MaS	no E	I_{M}	1.00
14	0.0750	0.0600	0.0590	0.2540	MaS	S	I_{M}	1.00
14	0.0735	0.0620	0.0505	0.2800	MaS	NS	I_{M}	1.00
16	0.2950	0.2875		0.2285	HWE	no E	I_{M}	0.85
16	0.2595	0.2605	0.0620	0.2470	HWE	S	I_{M}	0.85
16	0.3210	0.3285	0.0520	0.2610	HWE	NS	I_{M}	0.85
16	0.2965	0.2875		0.2260	MS	no E	I_{M}	0.85
16	0.2600	0.2625	0.0605	0.2480	MS	S	I_{M}	0.85
16	0.3210	0.3285	0.0520	0.2625	MS	NS	I_{M}	0.85
16	0.0450	0.0490		0.5670	MaS	no E	I_{M}	0.85
16	0.0610	0.0600	0.0590	0.5280	MaS	S	I_{M}	0.85
16	0.0540	0.0545	0.0540	0.6005	MaS	NS	I_{M}	0.85
17	0.0555	0.0465		0.6260	HWE	no E	I_{M}	1.00
17	0.0630	0.0640	0.0655	0.5675	HWE	S	I_{M}	1.00
17	0.0655	0.0585	0.0530	0.6535	HWE	NS	I_{M}	1.00
17	0.0560	0.0480		0.6210	MS	no E	I_{M}	1.00
17	0.0635	0.0665	0.0665	0.5720	MS	S	I_{M}	1.00
17	0.0650	0.0595	0.0540	0.6515	MS	NS	I_{M}	1.00
17	0.0515	0.0415		0.5930	MaS	no E	I_{M}	1.00
17	0.0640	0.0595	0.0645	0.5460	MaS	S	I_{M}	1.00

Table A.5: *(continued)*

ID	C	M	E : I_{M}	I_{M}	MT	E approach	True effect	MaS
17	0.0590	0.0495	0.0495	0.6260	MaS	NS	I_{M}	1.00

Bibliography

[1] H. F. Ainsworth, J. Unwin, D. L. Jamison, and H. 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, 2011.

[2] M. Bourgey, J. Healy, P. Saint-Onge, H. Massé, D. Sinnett, and M.-H. Roy-Gagnon. Genome-wide detection and characterization of mating asymmetry in human populations. *Genetic Epidemiology*, 35(6):526–535, 2011.

[3] L. R. Cardon and L. J. Palmer. Population stratification and spurious allelic association. *The Lancet*, 361(9357):598–604, 2003.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, Sept. 1977.

[5] M. A. Garland, B. Sun, S. Zhang, K. Reynolds, Y. Ji, and C. J. Zhou. Role of epigenetics and miRNAs in orofacial clefts. *Birth Defects Research*, 112(19):1635–1659, Sept. 2020.

[6] M. Gjerdevik, Ø. A. Haaland, J. Romanowska, R. T. Lie, A. Jugessur, and H. K. Gjessing. Parent-of-origin-environment interactions in case-parent triads with or without independent controls. *Annals of Human Genetics*, 82(2):60–73, 2018.

[7] H. K. Gjessing and R. T. Lie. Case-parent triads: estimating single- and double-dose effects of fetal and maternal disease gene haplotypes. *Annals of Human Genetics*, 70(Pt 3):382–396, May 2006.

[8] J. Healy, M. Bourgey, C. Richer, D. Sinnett, and M.-H. 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(1):369534, 2010.

[9] J. N. Hellwege, J. M. Keaton, A. Giri, X. Gao, D. R. Velez Edwards, and T. L. Edwards. Population stratification in genetic association studies. *Current Protocols in Human Genetics*, 95(1):1.22.1–1.22.23, 2017.

[10] C. A. Hobbs, M. A. Cleves, S. L. MacLeod, S. W. Erickson, X. Tang, J. Li, M. Li, T. Nick, S. Malik, and the National Birth Defects Prevention Study. Conotruncal heart defects and common variants in maternal and fetal genes in folate, homocysteine, and transsulfuration pathways. *Birth Defects Research Part A: Clinical and Molecular Teratology*, 100(2):116–126, 2014.

[11] R. Howey and H. J. Cordell. PREMIM and EMIM: tools for estimation of maternal, imprinting and interaction effects using multinomial modelling. *BMC Bioinformatics*, 13(1), June 2012.

[12] J. Hudson. *Maternal gene-environment effects: An evaluation of statistical approaches to detect effects and an investigation of the effect of violations of model assumptions*. MSc. thesis, University of Ottawa, Ottawa, ON, Sept. 2019. Available at `http://dx.doi.org/10.20381/ruor-23880`.

[13] A. Jugessur, Skare, J. R. Harris, R. T. Lie, and H. K. Gjessing. Using offspring-parent triads to study complex traits: A tutorial based on orofacial clefts. *Norsk Epidemiologi*, 21(2), Apr. 2012.

[14] J. B. Lang. Obtaining the observed information matrix for the poisson log linear model with incomplete data. *Biometrika*, 79(2):405–407, 1992.

[15] P. Lee, D. Ridout, J. Walter, and F. Cockburn. Maternal phenylketonuria: report from the united kingdom registry 1978–97. *Archives of Disease in Childhood*, 90(2):143–146, 2005.

[16] T. A. Louis. Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):226–233, 1982.

[17] P. Portin. The birth and development of the DNA theory of inheritance: sixty years since the discovery of the structure of DNA. *J. Genet.*, 93(1):293–302, Apr. 2014.

[18] R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.

[19] N. Rasevic, J. Bastasic, M. Rubini, M. R. Rakesh, K. M. Burkett, D. Ray, P. A. Mossey, B. Peterlin, M. F. J. Khan, A. Ravaei, L. Autelitano, M. C. Meazzini, J. Little, and M.-H. Roy-Gagnon. Maternal and parent-of-origin gene–environment effects on the etiology of orofacial clefting. *Genes*, 16(2), 2025.

[20] S.A.G.E. Statistical Analysis for Genetic Epidemiology, 2021. Release 6.4.2: http://darwin.cwru.edu.

[21] J.-H. Shin, B. McNeney, and J. Graham. On the use of allelic transmission rates for assessing gene-by-environment interaction in case-parent trios. *Annals of Human Genetics*, 74(5):439–451, 2010.

[22] T. Strachan and A. P. Read. *Human molecular genetics 2*. BIOS Scientific, 2nd ed. edition, 1999.

[23] D. M. Umbach and C. R. Weinberg. The use of case-parent triads to study joint effects of genotype and exposure. *The American Journal of Human Genetics*, 66(1):251–261, Jan. 2000.

[24] J. Vermunt. LEM: A general program for the analysis of categorical data. 01 1997.

[25] C. Weinberg and D. Umbach. A hybrid design for studying genetic influences on risk of diseases with onset early in life. *The American Journal of Human Genetics*, 77(4):627–636, Oct. 2005.

[26] C. R. Weinberg. Methods for detection of parent-of-origin effects in genetic studies of case-parents triads. *The American Journal of Human Genetics*, 65(1):229–235, July 1999.

[27] 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. *American Journal of Human Genetics*, 62(4):969–978, Apr. 1998.

[28] A. J. Wilcox, C. R. Weinberg, and R. T. Lie. Distinguishing the effects of maternal and offspring genes through studies of "case-parent triads". *American Journal of Epidemiology*, 148(9):893–901, 11 1998.