An Algorithm for Simulating Data for the Investigation of Maternal-Gene-Environment and Pathway Effects

XIN YOU

Project submitted in partial fulfillment of the requirements for the degree of
Master of Science Mathematics and Statisticsred[1]

Department of Mathematics and Statistics
Faculty of Science
University of Ottawa

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

Abstract

World-wide, there is approximately one case of oral cleft disease for every seven hundred births, making it a fairly common birth defect. Researchers are, therefore interested in discovering how genetic and environmental risk factors contribute to this disease risk. Birth defects may be linked to the mother's genetic variants in conjunction with environmental factors. These factors may contribute to the risk of disease while the baby is in utero. Luckily, pathway-based association analyses have become popular approaches for discovering how genetic variants contribute to diseases, such as cleft palate.

In this project, we develop an approach to simulate data on trios where the probability of disease is increased by maternal genetic effects and maternal gene-environment effects on genes in selected pathways. Development of the simulation algorithm was motivated by a need to compare the performance of pathway-based statistical approaches in detecting maternal gene-environment effects on a cleft palate data set collected from trios. We illustrate the use of the simulation algorithm by investigating the type 1 error rate when there are no true genetic effects and by examining mating type asymmetry when there are true maternal gene-environment effects. Our algorithm will be of use for investigating the power and type 1 error

of pathway-based statistical methods, which are used for detecting maternal genetic effects and maternal gene-environment interactions.

Keywords: Orofacial cleft, maternal effects, simulation, genetic effects, gene–environment interactions, causal variants.

Acknowledgement

I would like to give my deepest thanks to Prof. Kelly Burkett (Faculty of Science, University of Ottawa) and Marie-Helene RoyGagnon (Faculty of Medicine, University of Ottawa) for providing me the opportunity to take on this project. Their supervision though out my time as their student has been an absolute pleasure, they have truly enriched my experience at the University of Ottawa. It has been a great honour to work under their supervision. They have strengthened my enthusiasm for biostatistics and also provided me encouragement when facing difficult situations. They lead me to the door of science and give me strength to walk through it. My critical thinking has greatly enhanced and I am more curious about the world of science then ever before.

I would like to extend my deepest appreciation to Prof. Rafal Kulik (Chair of Mathematics and Statistics, Faculty of Science, University of Ottawa). I remember the first day I came to University of Ottawa, I had trouble understanding lectures. He would explained every topic clearly (even lectures he wasn't conducting), the professor was always patient with my questions as well. I am truly grateful for his help. He has always been a very kind professor and one of the warmest Chair's in our department.

I would also like to thank Prof. Benoit Dionne (Director of Graduate Programs and Associate professor, Faculty of Science, University of Ottawa). He has been very warm towards me, always leaving his door open for students, helping them solve every questions they have. I also remember him leaving some chocolate or candy in his office to help students relax and reduce their stress a little, that was a very thoughtful gesture.

I would like to acknowledge the contribution of Mengche Tsai (Medical Doctor) also one of our team member, and Peter Tea, my colleague. They provided plenty of help during this project, they have always been a good team members and even better friends.

I also want to extend a worm thanks to my mom and dad, Xianghua Kong and Jianming You, for cooking countless delicious meals for me over my lifetime. Day after day, they have given me endless love and support, always believing in me no matter what. They always take my pressures, negative emotions and anxious away with their love. Even though I'm in Canada, they send me love and strength from the other side of the world every day. They made me who I am today.

Lastly, I want thank a very important person in my life, Spencer Payer, for always being there for me no matter what, bring me love and happiness, and being sweetness person I have ever met. My world has become more colourful and lovely because of you.

Contents

Chapter 1

Introduction

Genetic and environmental factors have long been implicated in the disease known as oral cleft. This disease is relatively common across the worlds population, there is one case for every seven hundred births. Individuals with this disease face societal ridicule, making their lives more dicult and causing low self-esteem.[9]. Discovering more accurate methods of determining the probability of oral cleft among the most vulnerable segments of our population, along with discovering new ways of lowering the risk of contracting oral cleft will be highly valuable to humanity.

Researchers would like to determine what factors influences the probability of developing this disease.[15]. Because this disease is present at birth, there are many possible factors that can increase the risk, including the child's genotype, the mother's genotype, the uterine environment which includes substances consumed by the mother, and interactions between genetic and environmental factors. [33].

This project was motivated by an investigation of pathway-based approaches for

discovering maternal genetic and gene-environment risk factors for disease on a cleft palate data set. A challenge in comparing the methods is that it isn't known whether any of the genes investigated have true associations with cleft palate. Therefore, in this project we developed a simulation algorithm that can be used to evaluate methods when there are complex genetic and environmental effects on pathways.

This study collected the genome data from 1000 genome projects and a European sample, we extract 1000 families as the subjects in this study. Each family consisted of a father, a mother and a single child. For each of the family members in a trio. In order to run the simulation, each individuals genotype simulated by using sim1000G.[11]. Simulating the environmental variable by using the Bernoulli distribution[26] with the proportion of this variable in reference data set(eg. 1000Gemoes). Simulation methods are different when simulating the childs phenotype. If data is simulated under the null hypothesis of no association between genetic variants and disease, we assume each child in every family is affected. If data is simulated under an alternative hypothesis, the probability of disease for a child is found according to a logit model depending on the genotypes of the mother, child, and/or the environmental factor. The disease status of the child is then sampled from a Bernoulli distribution with the probability determined from the logit model.

To illustrate the utility of the simulation algorithm, We have two models to test the association, in the first model we assume no genetic effects and no gene-environment effects, we also assume every child is affected in each family, and we only simulate the genes and environmental factor. In the second model we assume that maternal genetic effects or maternal gene-environment interactions exist; In this case,

we simulate the genotype, environmental variables and also the disease status. Since there are multiple SNPs located in a gene, and multiple genes in multiple pathways. We have to randomly choose causal variants. In the second model, 8 scenarios related to selecting the causal variants were created.

The general format of this project will follow a logical flow. Background information regarding genetics and the simulation algorithm will be provided in Chapters 2 and 3. The application of the simulation algorithm will be demonstrated in Chapter4.

Chapter 2

Background

This world is full of variety and diversity. When taking a closer look at the snowflake, we observe that the shape of each snowflake is unique. Or, if we consider human traits like skin color, height, weight, and even our fingerprints, no one is identical. We may wonder why or how these differences arise. For humans, genetic and environmental factors, contribute to our observable traits and diseases. [12].

This project was motivated by a previous comparison of statistical methods for detecting maternal gene-environment effects. [41]. In this chapter, I therefore give genetic and epidemiological background to help the reader understand the project.

To be clear, this chapter does not provide a comprehensive review of human genetics. Instead, we will only focus on the topics related to the statistical methods of detecting genetic effects on families.

2.1 Basic genetic background

In this Chapter, we will introduce some important genetic concepts like DNA, SNPs , and Genotypes. For more descriptions about genetics, the reader is referred to standard textbooks on this topic (e.g., *Refs.*[13, 35]).

2.1.1 DNA

Genetic information is carried by chromosomes within the cells of every individual. In humans, there are 23 unique chromosomes in total, which are made up of *deoxyribonucleic acid* (DNA) and proteins.[13]. DNA carries the genetic information, proteins support different functions, such as DNA's duplication and transmission. The DNA molecule is composed of *nucleotides*, the basic sub–units of DNA. Each nucleotide is composed of one of four nitrogen-containing bases, a sugar called deoxyribose, and a phosphate group. [35]. There are four different bases: $guanine(G)$, $adenine(A)$, $cytosine(C)$, and $thymine(T)$.

2.1.2 Alleles and Genotype

Alleles are the observed variants at a DNA *locus*, also known as loci, which is a marker, a fixed position on a chromosome. A loci is also a variation that can be observed and used to identify individuals or species.[35] A *Genotype* is the combination of alleles at a locus that were inherited from an individuals parents. Therefore, in humans, the genotype is composed of two alleles, with one inherited from the father and the other from the mothers. [20]. Genotypes are often denoted by letters, for example, a locus may have two alleles labeled A and a. The possible genotypes at that locus are shown in Table 2.1.

		Maternal Allele	
Paternal	A	AA	Aa
Allele	a	Aa	aa

Table 2.1: Possible genotypes for a locus with two alleles.

If an individual inherits the same allele from both parents at a locus, the genotype is said to be homozygous. Otherwise, the genotype is heterozygous.[35]. In Table 2.1, the two homozygous genotypes are AA and aa and the heterozygous genotype is Aa. A haplotype is a pair of genes inherited together from a single parent.

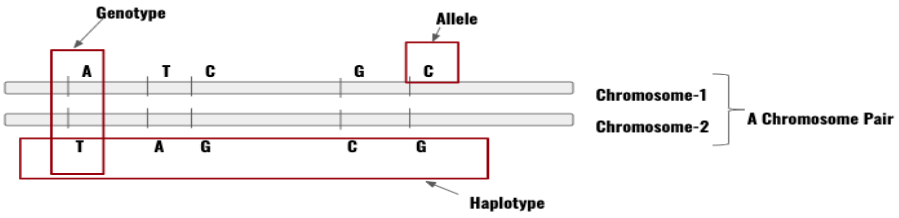


Figure 2.1: The relationship between allele, genotype, haplotype and chromosome.

2.1.3 SNP

Everyone's DNA is different as mutation events create variations in the DNA sequence. *Single nucleotide polymorphisms* ($SNPs$), are the most common genetic variants in DNA.[17]. A SNP is a variant that differs at a single nucleotide in the genome. For example, at a particular site in the genome, some chromosomes might have an A while others have a C. Most of the SNPs in the human genome have no effect on human traits. However, some of them may affect genes function.[40]. Other

SNPs may not directly affect a human trait, but they may be close to unknown variants which do affect traits. Such SNPs are said to be linked.[27]. Therefore, scientist use SNPs as markers to locate genetic variants that are associated with disease. SNPs typically only have 2 alleles, which we often labeled as 0 and 1 for simplicity.

Many SNPs have been identified for numerous disease by association studies. However, the association between a SNP and a disease can result from a causal variant in linkage disequilibrium (LD) with the considered SNP.[10]. We consider a *causal variant* as the variant which is responsible for the association signal at a loci.[16].

2.1.4 Phenotype

A *Phenotype* is an observable characteristic of an individual, that is thought to be at least partially influenced by the individual's genetic makeup.[21]. Common phenotypes include eye color, height, disease history, and even behavior and general disposition.

The association between genotype and phenotype are usually complicated. Most phenotypes are influenced by both genotype and the environment, where environment includes all of the unique circumstances the individual experienced in their life.[12]. In genetic data analysis, environment is defined very broadly and is not limited to the external environment. In family–based studies, environmental factors include a babies experience while in the womb and whether their mother took pre–natal vitamins.[31]. Importantly, genetic and environmental factors often act together to

influence a phenotype; this is called 'gene-environment interaction'. For example, individuals exposed to certain pollutants may be more susceptible to disease due to particular genetic variants that they carry. We refer to this unique circumstances as the "environment interactions".

Several inheritance modes are used for describing a single-gene diseases: autosomal dominant, autosomal recessive, X–linked dominant, X–linked recessive. Inheritance Modes for multiple genes diseases: additive and co-dominant.[6].

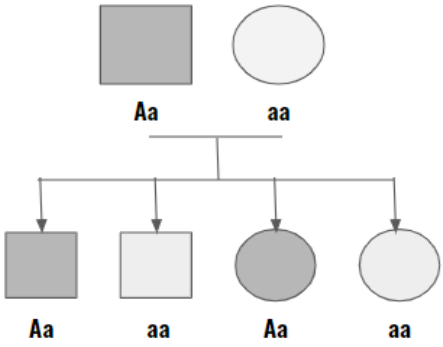


Figure 2.2: Autosomal Dominant: Square represents male and circle represents female. Dark color means affected by disease and light color means non-affected. The first line is father and mother, father is affected and A is the dominant allele he carries, which means only show the phenotype of A no matter the genotype is AA or Aa.

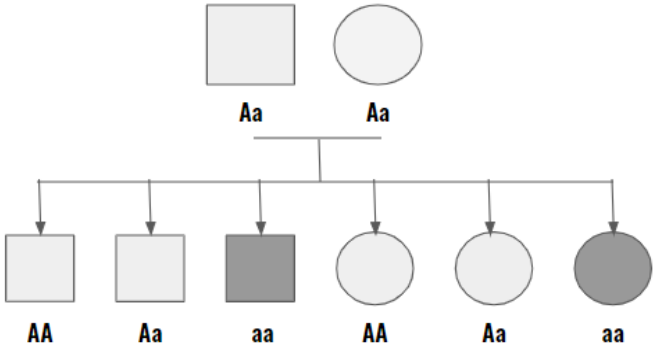


Figure 2.3: Autosomal Recessive: Both parents carries heterozygous genotypes, but only child has the genotype as aa, this will show the disease phenotype.

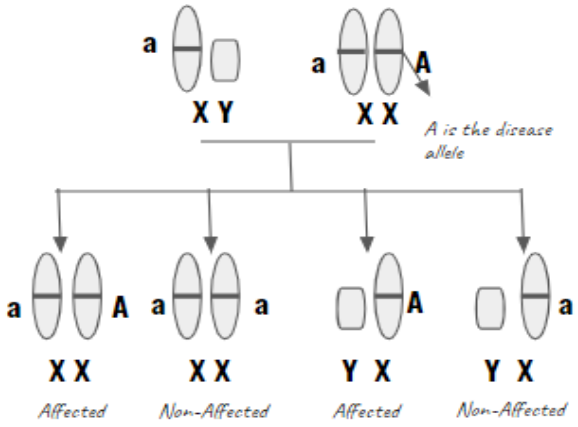


Figure 2.4: X-linked Dominant: 1.Father can only pass Y link to the son. 2. Females have a higher possibility of contracting disease than males, because female have 2 X links from both parents.

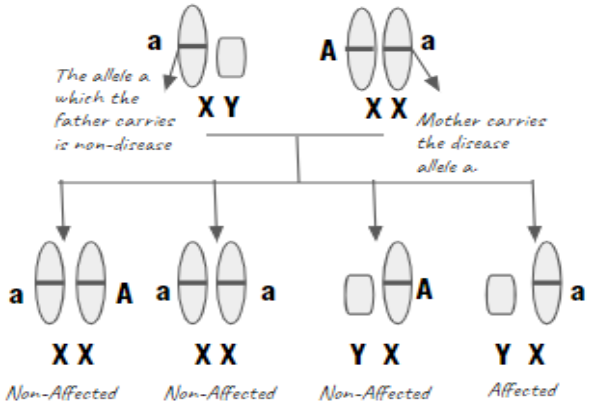


Figure 2.5: X-linked Recessive: 1.Mother must be the carrier of the affected son. 2. Males have a higher possibility of contracting disease than females

Figure 2.6: Picture is from *Ref*.[4]. Additive Inheritance: When dominant alleles are present together they produce double effect. For example, if "R" produces pink as the phenotype of the flowers, and "r" produce white. When both "R" , and "R" appear together, we have a red colour, because the RR produces a double effect.

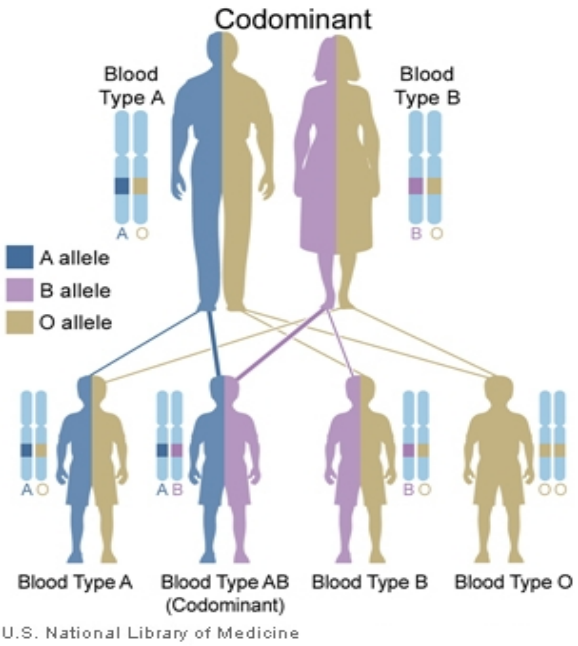


Figure 2.7: Picutre is from Daniel Del Prete. *Ref*.[25]. Co-dominant Inheritance: Co-dominant means both alleles show the full phenotypic expressions together. For example, the second child has allele A from father and allele B from mother. However, allele A and B are both dominant inheritance, which results in the child having blood type AB, because both alleles bring out the full phenotypic expression.

2.1.5 Genes and Pathway

A gene is the name given to a segment of DNA that contains the code needed to produce a protein. The functions of the proteins produced by an organism are diverse. Proteins act along with other proteins to produce some action within a cell. Therefore, we divide proteins and the genes that code for these proteins into *biological pathways.* For example, a *Metabolic pathway* is a series of chemical reactions that occur within a cell, which keeps the cell living, growing and dividing. Since genes code proteins, and proteins act within biological pathways to produce observable phenotypes, pathway-based genetic data analysis has become popular[30]. In such approaches, the association between genetic variants and phenotypes are summarized in different ways over the pathways containing the genes.

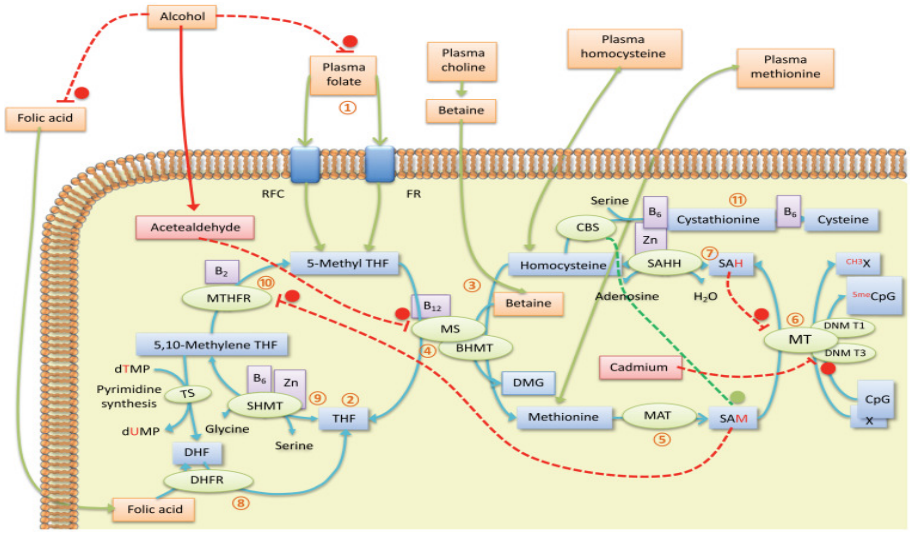


Figure 2.8: Picture is from Text book. See *Ref.*[36]. Schematic illustration of the one-carbon metabolic pathway. There are plenty of genes on a pathway such as MTHFR and DHF. A metabolic pathway can provide nutrition from food consumption, which is highlighted in orange. Red dashes are the enzyme reactions. All the chemical reactions happen on the pathway, which keeps all the organs functioning properly.

2.1.6 Gene environment interactions

Gene−environment interactions (or GxE) occur when the effect of a genetic variant differs depending on environmental factors. For example, Children born with the phenylketonuria genetic variant will have a phenotype that includes a intellectual disability, seizures, behavioral problems, and mental disorders if they eat a normal diet. However, if they eat a specialized diet, they can have a normal phenotypes and better health. In this example, the environmental factor that modifies the risk of phenylketonuria is diet. In epidemiology, gene−environment interactions help us gain a better understanding of some diseases.[22] Figure 2.9 is an example of gene environment interactions, X axis represents genotype, Y axis represents phenotype. There are two lines names "Environment1" and "Environment2" represent different levels of environment intervention. As we can see, the same genotype has different expressions of phenotype under different environmental levels.

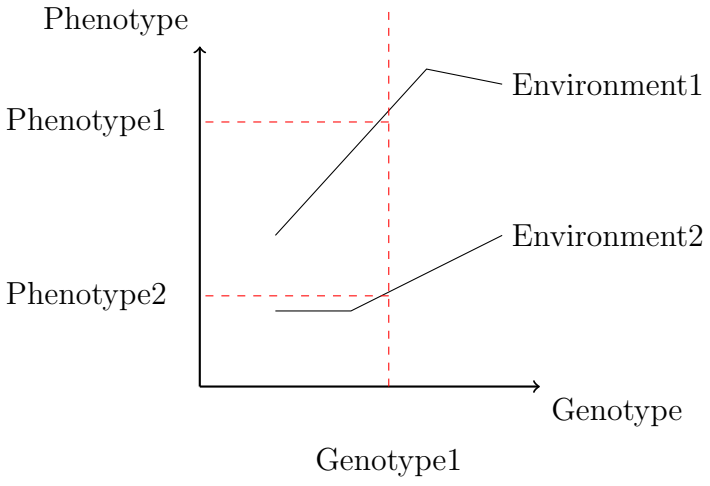


Figure 2.9: Gene-environment interactions

2.1.7 Linkage Disequilibrium

Linkage disequilibrium (LD) refers to the non-random association of independence between the alleles at two or more loci in a general population. The level of linkage disequilibrium between allele A and B is measured by D_{AB}, which is

$$D_{AB} = p_{AB} - p_A p_B$$

where p_{AB} is the probability of the haplotype AB. p_A is the frequency of allele A and p_B is the frequency for allele B. We will use $\mathcal{D}$ to denote the linkage disequilibrium between any two alleles in the following chapters. For more about linkage disequilibrium, the readers referred to Ref.[27].

2.1.8 Hardy Weinberg Equilibrium

Hardy Weinberg Equilibrium describes the relationship between allele and genotypes frequencies. This principle was discovered by Godfrey Harold Hardy and Wilhelm Weinberg in 1908. [41] This principle states that there is a fixed relationship between alleles and genotype, under certain assumptions about the population, including random mating. Consider alleles A, and a, and define the genotype frequencies as

$$P(AA) = p_{11} \quad P(Aa) = p_{12} \quad P(aa) = p_{22} \quad (where p_{11} + p_{12} + p_{22} = 1)$$

$$P(A) = p \quad P(a) = q \quad (wherep + q = 1)$$

Under HWE, we can calculate the frequencies for $P(A)$, $P(a)$, and the genotype frequencies, $P(AA)$, $P(Aa)$, and $P(aa)$, by

$$P(A) = p = p_{11} + \frac{1}{2}p_{12}$$

$$P(a) = q = p_{22} + \frac{1}{2}p_{12}$$

$$P(AA) = p^2, \quad P(Aa) = 2pq, \quad P(aa) = q^2.$$

2.2 Familial inheritance and family–based association studies

In the previous section, we learnt that DNA carries the genetic information, and information about the relationship between genotype and phenotype. In this section, we will discuss how phenotypes are inherited in families. As we know from previous sections, the relationship between genotypes and phenotypes are complex, which means not all individuals carrying the disease-causing genotypes are affected. Therefore, we will focus on both family inheritance and environmental influences.

2.2.1 The Trio study design

In genetic epidemiology, a common study design is to collect genetic and phenotypic data on members of the same family. Usually the family is selected because they show a high rate of disease relative to the general population.[29] The trio study

design is one example of a family-based design and it is the design focused on for this project. In a trio study, we use three family numbers, the *trio*, as a unit to analyze the association between genotypes and phenotypes. *Trios*, consist of two parents and their child, who is normally affected with a disease.[31]. With these designs, we are interested in whether the genetic variants passed to the affected child from his/her parent differs from the variants that were not passed down to the child.

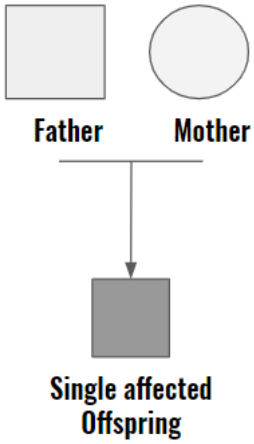
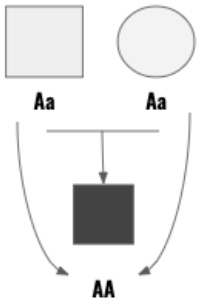


Figure 2.10: A Trio: a trio is composed of 3 individuals, father, mother and an affected child.

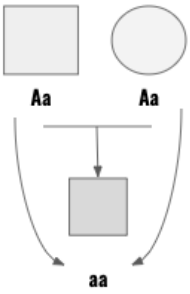
2.2.2 The Classical Transmission Disequilibrium Test (TDT)

Transmission Disequilibrium Test (TDT) is a statistical approach for analyzing trio data. The TDT is used to test for association between a genetic variant and disease. *Refs.*[41, 29].

To test for differential transmission, we consider parents with heterozygous geno-



(a) Transmitted situation: Allele A has been transmitted from both heterozygous parents to the child.



(b) Nontransmitted situation: Allele A has not been transmitted to the child from both parents.

Figure 2.11: Transmitted and Non–transmitted parental alleles under the additive inherent disease mode. The child will only show the disease trait when he/she has alleles A transmitted from both parents.

type only. Each parent has two alleles, one of which is passed to a child and the other which is not. Let T represent the allele that is transmitted and NT the allele that is not transmitted. For example, $p_{1,2} = P(T = 1, NT = 2)$, is the probability that allele1 has been transmitted, while allele2 has not been transmitted to a single affected offspring from one of his or her parents.

Precisely for the TDT, we test the hypotheses

$$H_0 : \frac{p_{A,a}}{p_{A,a} + p_{a,A}} = \frac{1}{2} \quad versus \quad H_1 : \frac{p_{A,a}}{p_{A,a} + p_{a,A}} \neq \frac{1}{2}$$

If the locus with alleles A and a is not associated with the disease, then we do not expect that either allele is preferentially passed to the affected child. Due to the dependent nature of the data, we test this hypothesis using the the standard MacNemar test (2.2.1), which is based on the observed frequencies of transmitted

and non–transmitted alleles from one parent (see Table 2.2). $Refs.$[28, 24].

$$T_{TDT} = \frac{(n_{A,a} - n_{a,A})^2}{n_{A,a} + n_{a,A}} \tag{2.2.1}$$

Table 2.2: Observed frequencies for transmitted and non-transmitted alleles for the transmission disequilibrium test from one heterozygous parent. 2n is the total number of alleles, because each parent has 2 alleles.

Transmitted allele	Non-transmitted A	allele a	Total
A	$n_{A,A}$	$n_{A,a}$	$n_{T=A}$
a	$n_{a,A}$	$n_{a,a}$	$n_{T=a}$
$Total$	$n_{NT=A}$	$n_{NT=a}$	$2n$

2.2.3 Maternal effects

A *Maternal effect* occurs when an individual's phenotype is not only determined by his or her genotype and gene-environment interactions, but also influenced by this individual's mother, which means the causal influence of the maternal genotype or phenotype on the offspring phenotype.[39]. Phenotypic similarities between mothers and their offspring are a result of Maternal effects. An example of this can be found in the offspring of mammals, which will have a higher probability of a healthy birth weight if the mother is in good health and has a body weight that is within a health range during pregnancy. The mother can provide the offspring more nutrition during pregnancy and during the lactation period.[5].

The influence that a mothers gene environment has on her child's phenotype while it is developing in the whom, is known as the *Maternal gene–environment effect*. The reason for this occurrence is that the unborn offspring in humans and

other species requires a environment (a whom) to develop before they can survive on their own. The variants in the child's genome is thought to have environmental and genetic influences. For example, if an individuals mother smoke or drink while she was pregnant, her child might be born with some specic diseases due to a maternal gene environment effect. Phenotype variances with heritable connections have numerous quantitative features, which have been demonstrated to be linked to indirect genetic effects, such as the ones created by the Maternal gene-environment effect.[19]

2.2.4 Bonferroni Correction

The Bonferroni Correction is an adjustment for inflated p–values when several dependent or independent statistical test are being performed at the same time in a same data set. The Bonferroni Correction sets a significant level at $\frac{\alpha}{n}$, where $\alpha = 0.05$ is the significant level, and n is the number of times the hypothesis is being tested.[26]. For example, if the hypothesis is being tested 30 times, then the significance cut–off is $0.05/30 = 0.00167$.

Chapter 3

Algorithm for simulating maternal gene-environment effects on pathways using real human genetic data

In this chapter, we describe our approach for simulating maternal gene-environment effects on pathways using real human genetic data. In chapter 4, we will show the applications of this approach.

3.1 Overview of Algorithm

Our approach assumes that the user has identified biochemical pathways and variants within genes in those pathways. For example, X represents the number of

pathways, Y, the number of genes, and Z is the number of variants. We use population level data (eg. 1000 Genomes) as reference distribution for our simulations; this is referred to as the "reference data set" throughout this chapter.

Null hypothesis: there are no genetic environment interactions between genotypes and phenotypes. Alternative hypothesis : there are genetic environment interactions between genotypes and phenotypes.

We consider the reference data set as a starting point for our simulation design and simulate a new data set as follows:

- For each of the family members in a trio, simulate each individual's genotype. The approach to simulate the genotype will be described in Section 3.2.1
- Simulate the child's phenotype. In this part, simulation methods are different for each project:
  - If data is simulated under the null hypothesis of no association between genetic variants and disease, we assume each child in every family is affected.
  - If data is simulated under an alternative hypothesis, the probability of disease for a child is found according to a logit model depending on the genotypes of the mother, child, and/or the environmental factor. The disease status of the child is then sampled from a Bernoulli distribution with probability determined from the logit model. This model will be described in Section 3.2.3.
- These steps are repeated until the desired number of trios have been simulated.

3.2 Model and implementation details for simulating a data set

3.2.1 Simulating genetic pathway data

Although many approaches exist to simulate genetic data, they typically do not organize the genetic variation into genes and pathways. For this reason, we chose to simulate genetic data based on population genetic data for the Z variants genotyped in the original study. The 1000 Genomes project[1]. has such data available on various human populations. We can download the data from the 1000 Genomes website and extracted the genotypes for the Z SNPs of interest. The genetic data from these individuals are used as our reference distribution for simulating the genetic data. In this way, we can assume that our simulated genetic data would have a similar structure to real data from humans. With the genetic data from the 1000 Genomes restricted to our SNPs and populations of interests, we then use $sim1000G$ to actually simulate the trio data.

Sim1000G is an easy-to-use genetic variant simulator in R, which can simulate genetic data for unrelated individuals and related individual from a family-based design. Sim1000G uses the 1000 Genomes data (in the form of VCF files) to form the reference distribution for simulating new individuals. Sim1000G can capture allele frequency diversity, linkage disequilibrium (LD) patterns, subtle population differences in LD structures without the need for any tuning parameters and fulfill the simulation process with multiple analytic methods for genetic association tests.[11]. We then input VCF files with the data for our SNPs of interest.

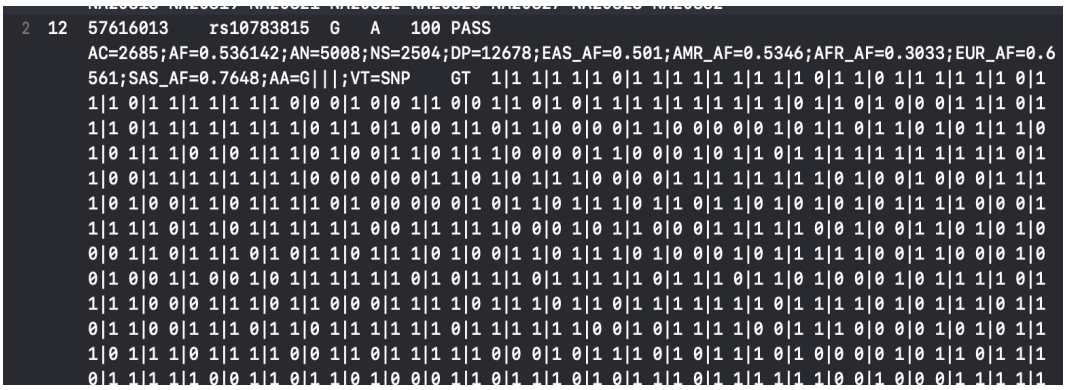


Figure 3.1: An example of VCF file

We simulate genotype data for SNPs selected for the reference data. We use the 1000 Genomes data on these SNPs as our reference distribution for simulating the genetic data for the number of trios we want. Sim1000G requires the data to be simulated separately for each chromosome and also at least 5 SNPs genotyped on each chromosome. Therefore, we need to delete those chromosomes which have less than 5 SNPs. Sim1000G is then used to simulate genotype data through all the SNPs for the trios you selected which are composed of a father, mother and a single child.

3.2.2 Simulating Environmental Variables

To simulate the environmental variable, we chose a variable from the reference data to inspire our probability distribution. In particular, we selected one of the environmental variables in reference data that we are interested in (eg. Env1). We consider it as categorical variable, with 1 representing yes and 0 representing no. Note that in reference data, this variable might be missing for some families, which was indicated by either a 9 or NA. To simulate this variable, we first estimated the

proportions of "no" responses in the reference data using:

$$p_0 = \frac{\#[Env1 = 0]}{\#[(Env1 = 1) + (Env1 = 0)]}$$

Where we only kept the values for Env=1, Env=0, and deleted trios with missing values for Environment1. We then simulate the environmental variables with the observed proportion p_0 and $p_1 = 1 - p_0$, which is the probability of "yes" in the simulations by using Bernoulli distribution. The number of environmental variables being generated equals to the number of trios.

3.2.3 Simulating Phenotype on offspring

Case1: There are no associations between genotype and phenotype. In this case, the distribution of the genotypes in a sample consisting of affected children will be no different than a sample consisting of unaffected children. Therefore, we assume all the trios have single affected offspring. We do not simulate the phenotype for case 1.

Case2: There are associations between genotype and phenotype. When simulating phenotypes used for Alternative models having true gene–disease associations, we must first select the causal variants; these must be selected at the pathway level, genes within pathway, and variants within genes.

- Pathway level.
  - 1 pathway has genes with true association.
  - 2 pathways have genes with true association.

- Genes within pathways.
  - 1 gene within the pathway we have selected has variant(s) with true association.
  - 2 genes within the pathway we have selected have variant(s) with true association.
- Variants within genes.
  - 1 causal variant per gene.
  - 2 causal variants per gene.

We therefore have 8 scenarios involving causal variants. See Figure 3.2 for a visual representation of the causal variant selection process. The 8 scenarios are now given in detail:

1. Randomly choose 1 pathway from the three, randomly choose 1 gene from n genes on that pathway, and randomly choose 1 SNP as the causal variant with a gene within the pathway. We only have 1 causal variant in this scenario.
2. Randomly choose 1 pathway from the three, randomly choose 2 genes from n genes on that pathway, randomly choose 1 SNP as the causal variant from each genes we selected within the pathway. In this scenario, we will have 2 causal variants in total. For example, if the choice is pathway1, the if we observe Gene–1, and Gene–n from the selection, and on each gene, we have SNP–1 and SNP–5. Then, we will have SNP–1 from Gene–1, pathway1, SNP–5 from Gene–n, pathway1, and two causal variants from different genes but the same pathway.

3. Randomly choose 1 pathway from the three, randomly choose 1 gene within that pathway, and randomly choose 2 SNPs as causal variants within that gene within that pathway. In this scenario, we also have 2 causal variants in total.

4. Randomly choose 1 pathway from the three, randomly choose 2 genes within that pathway, and randomly choose 2 SNPs as causal variants within that gene and pathway. There will be 4 causal variants in total for this case.

5. Randomly choose 2 pathways from the three, randomly choose 1 gene from each pathways we selected, and randomly choose 1 SNP from each gene within the pathway. There will be 2 causal variants in total from two different pathways and different genes within that pathway.

6. Randomly choose 2 pathways from the three variables, randomly choose 1 gene within each pathways, and randomly choose 2 SNPs as causal variants within each gene. There will be 4 causal variants in total.

7. Randomly choose 2 pathways from the three, on each pathway, randomly choose 2 genes, and randomly choose 1 SNP from each gene within the pathway. There will be 4 causal variants in tota.

8. Randomly choose 2 pathways, randomly choose 2 genes within each pathway, and randomly choose 2 SNPs within each gene within that pathway. In this scenario, we will have 8 causal variants in total. For example, if the choices are pathway1 and pathway3 from the selection, we have gene–1, gene–2 from pathway1, and gene–3, gene–4 from pathway3. If we have SNP–1, SNP–2 from gene–1, SNP–3, SNP–4 from gene–2, SNP–5, SNP–6 from gene–3, SNP–7 and SNP–8 from gene–4. Then, we will have 8 causal variants, each 2 causal variants

are from different genes, each 2 two genes are from different pathways.

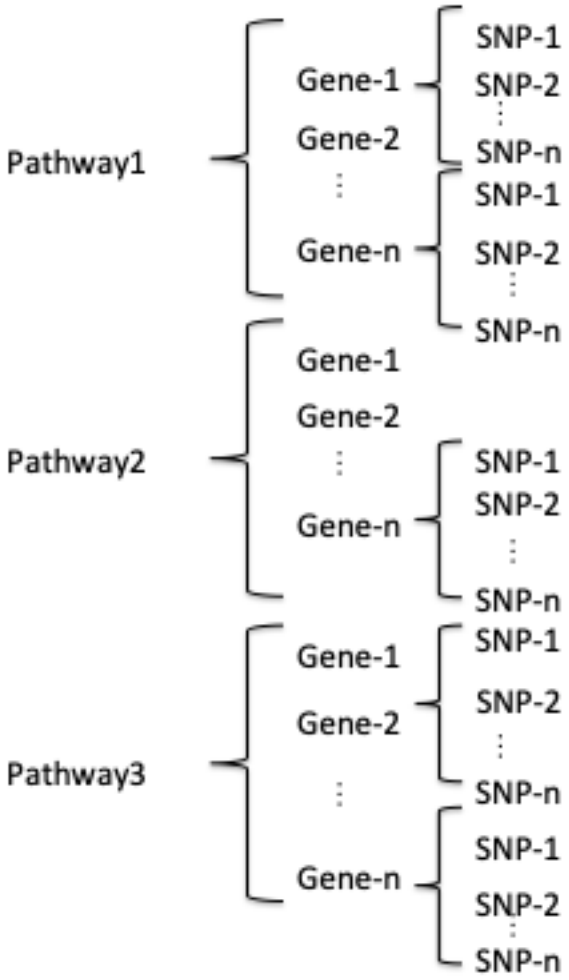


Figure 3.2: Identify Causal Variants

A full version of the function which can return the values for the random selection will be shown on the link in Appendix B. The following is an example of how to apply this technique: I named the function samplegenomes. npathway is instructing the application of this technique, we must choose the number pathways we desire (options are 1 or 2), ngene indicates the application has a decision regarding how

many genes will be chosen on that pathway (options are 1 or 2 as well). Once the program selects a pathway, the genes being chosen will also be derived from that pathway. After the genes are selected, the SNPs being returned will also originate from those genes.

```
> samplegenomes(nsnp = 2,ngene = 2,npathway = 2)
```

Consider our trio-based study of disease outcomes D_i, which is the phenotype we need to simulated for each family, environmental factor E_i, the environmental variables we generated by Bernoulli for each trio, and the mother's genotype in each trio G_i. Our disease model can be formed as 3.2.1 and 3.2.2, also p_i is given by 3.2.4, where OR_i means the Odds Ratio.

$$logit(Pr(D_i = 1|G_i, E_i)) = logit(p_i) \tag{3.2.1}$$

$$logit(p_i) = ln(\frac{p_i}{1-p_i}) = \beta_0 + \beta_G G_i + \beta_E E_i + \beta_{G\times E} G_i \times E_i \tag{3.2.2}$$

$$OR_i = e^{\beta_0+\beta_G G_i+\beta_E E_i+\beta_{G\times E} G_i\times E_i} \tag{3.2.3}$$

$$p_i = \frac{OR_i}{1+OR_i} \tag{3.2.4}$$

We use logit model to achieve the conditional probability of detecting a diseased trio, under the conditions related to the individual's genotype and environmental factors. A probability is generated by each trio, and we only use the mother's genotype in each family since we want to test maternal effects.

The disease status for the offspring in each trio is simulated through Bernoulli distribution, which is $D_i \sim Ber(p_i)$. When $D_i = 1$, we kept this diseased trio, when $D_i = 0$, discard the trio with no disease.

Chapter 4

Application of Data Simulation Algorithm

The motivation for developing the data simulation algorithm is a study investigating whether maternal genetic variants in genes within pre–selected biological pathways interact with environment factors to increase a child's risk of orofacial clefts. In this chapter, I first give some background on the motivating study. I then describe the data simulation strategy.

4.1 Motivating study and question

It is generally known that both genetic and environment factors contribute to the extremely complicated causation of birth defects, like orofacial clefts. Orofacial clefts, are a group of conditions that includes cleft lip (CL), cleft palate (CP), and both together (CLP). These conditions have a probability of 1.8%, making it the most common group of birth defects in the world.[18]. The direct causes of orofacial

clefts amongst most infants are unknown, but we do suspect that genetic factors are involved. Cleft lip and cleft palate are thought to be caused by a combination of genetic effects and maternal effects.[7] The genetic effects can be directly due to the child's genotype, or due to the mother's genotype (maternal effects) which contribute to the child's environment while he or she is in the womb. Environmental factors can also contribute. In the case of birth defects we must consider the mother's environment, such as what the mother eats or drinks and also certain medications being consumed during pregnancy. Finally, certain genetic variants in either the mother or child may modify the effect of environmental risk factors for orofacial clefts (GE interaction).

In the motivating study, data was collected on trios where the child was affected with an orofacial cleft. Genotyping was done for variants in genes in three pathways, Cytosolic Metabolism pathway, Mitochondrial Metabolism pathway, and Cross-membrane Transport pathway. Figure 4.1 provide the gene information on each pathway. Environmental data included variables such as the use of vitamin supplementation during pregnancy. In addition, data was available on both European and Asian samples. Team members were interested in comparing different pathway-based statistical analysis approaches on their ability to detect maternal–gene environment effects. A difficulty with real data is that maternal–gene environmental effects aren't known and so if results differ between methods, it's not possible to determine which one is closest to the truth. Therefore, it's desirable to also compare methods using simulated data where the truth is known. The purpose of this project is therefore to determine an algorithm for simulating genetic and environmental data on trios, where the variants are part of known biological pathways and where the genetic effects can

be due to maternal genotypes and interactions between maternal genotypes and environmental factors. The algorithm for simulating this data is described in chapter 3. In this chapter, we illustrate the applications of the algorithm.

Cytosolic Metabolism	
MTHFR (1:12)	Metabolism: folate utilization
DHFR (272:283)	Metabolism: folate utilization
CTH (13:37)	Metabolism: trans-sulfuration pathway
MTR (38:58)	Metabolism: trans-sulfuration pathway
MTRR (148:236)	Metabolism: trans-sulfuration pathway
BHMT2 (257:261)	Metabolism: trans-sulfuration pathway
BHMT (262:271)	Metabolism: trans-sulfuration pathway
GNMT (284:287)	Metabolism: trans-sulfuration pathway
AHCY (519:521)	Metabolism: trans-sulfuration pathway
CBS (522:538)	Metabolism: trans-sulfuration pathway
DMGDH (237:256)	Metabolism: one carbon interconversion
MTHFD1 (456:468)	Metabolism: one carbon interconversion
MTHFS (469:492)	Metabolism: one carbon interconversion
SHMT1 (493:498)	Metabolism: one carbon interconversion
FTCD (554:568)	Metabolism: one carbon interconversion
TYMS (505:518)	Metabolism: one carbon interconversion
ALDH1L1 (65:147)	Metabolism: one carbon interconversion

(a) The genes located on Cytosolic Metabolism pathway

Mitochondrial Metabolism	
MTHFD2 (59:64)	Metabolism: mitochondrial folate utilization
FPGS (400:406)	Metabolism: mitochondrial folate utilization
SHMT2 (431:433)	Metabolism: mitochondrial folate utilization
MTHFD1L (288:392)	Metabolism: mitochondrial folate utilization
ALDH1L2 (434:455)	Metabolism: mitochondrial folate utilization
Cross-membrane Transport	
FOLH1 (407:418)	Peptidase: peptide cleavage for absorption
FOLR3 (419:424)	Transport: folate receptor (FOLR) family
FOLR1 (425:426)	Transport: folate receptor (FOLR) family
FOLR2 (427:430)	Transport: folate receptor (FOLR) family
SLC19A1 (539:553)	Transport: folate transporter family for transplacental absorption
SLC46A1 (499:504)	Transport: folate transporter family for intestinal membrane absorption
SLC25A32 (393:399)	Transport: folate transporter family for transmitochondrial membrane

(b) The genes located on Mitochondrial Metabolism and Cross-membrane Transport pathways

Figure 4.1: The gene information on each pathway

4.2 Application of simulation pipeline

4.2.1 Application 1 –Simulating Genotype and Environmental variable

To simulate realistic data, we use the data from the motivating study to calibrate our simulation parameters. We focused on the European sample, which had genetic data for 559 SNPs across the three pathways of interest. Several non-genetic variables were available on the trios. We focused on the vitamin supplementation variable. In this section, I will first introduce the application for simulating genotypes, then the application for simulating environmental variables,

Application for Simulating Genotypes

We downloaded the Phase 3 data from the 1000 Genomes website and extracted the genotypes for the 559 SNPs of interest on European-derived populations (British, Finnish, CEPH, Tuscan, Spanish). The 559 SNPs are spread across chromosomes 1, 2, 3, 5, 6, 8, 9, 11, 12, 14, 15, 17, 18, 20, and 21. We then use Sim1000G to simulate the genetic data for 1000 trios. Note that in the orofacial cleft data set, a total of 565 SNPs were genotyped. We removed SNPs from chromosome 20 from our simulation because chromosome 20 had only 3 SNPs genotyped sim1000G requires at least 5. Additionally, we deleted 3 triallellic SNPs that were among the 565, which left us with 14 chromosomes and 565-3-3=559 SNPs in total. Sim1000G is then used to simulate genotype data on these 559 SNPs for 1000 trios composed of a father, mother and a single child.

The genotypes need to be re–coded for subsequent analyses. The genotype of each individual in the orofacial cleft data set is expressed as "0 | 0", "0 | 1", "1 | 0", "1 | 1". After re-coding in sim1000G, the new genotypes are representing as "0", "1","2", where "0" represents genotype "0 | 0", "1" represents "0 | 1" and "1 | 0", and "2" represents "1 | 1". Figure 4.2 shows the procedure of re–coding genotype. Figure 4.3 gives an example of simulated genotypes.

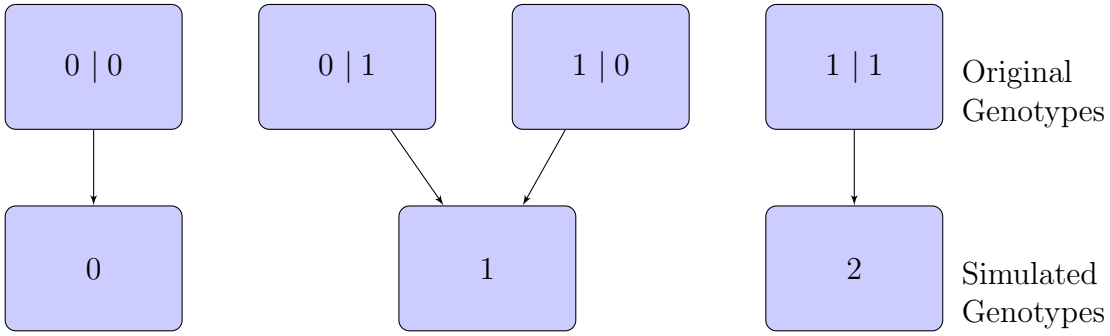


Figure 4.2: Re-coding genotypes

	SNP528	SNP529	SNP530	SNP531	SNP532	SNP533	SNP534	SNP535	SNP536	SNP537
FID-1_1	1	0	2	2	0	0	0	2	2	0
FID-1_2	1	0	2	2	0	0	0	2	2	0
FID-1_3	1	0	2	2	0	0	0	2	2	0
FID-2_1	0	0	0	0	0	0	0	2	1	0
FID-2_2	2	0	1	1	1	1	1	1	1	1
FID-2_3	1	0	0	0	1	1	1	1	0	1
FID-3_1	1	0	0	0	0	0	0	1	1	1
FID-3_2	0	0	1	1	1	1	1	1	1	1
FID-3_3	1	0	1	1	0	0	0	1	1	1
FID-4_1	1	0	0	0	0	0	0	0	0	2
FID-4_2	2	0	0	1	1	1	1	1	1	1
FID-4_3	1	0	0	1	0	0	0	1	1	1
FID-5_1	2	0	0	0	0	1	1	0	0	2
FID-5_2	2	0	0	1	0	0	0	1	1	1
FID-5_3	2	0	0	0	0	1	1	0	0	2
FID-6_1	1	0	0	1	0	0	0	1	1	1
FID-6_2	0	0	0	0	1	1	1	0	0	2
FID-6_3	1	0	0	1	1	1	1	1	1	1
FID-7_1	2	0	1	1	0	1	0	1	1	1
FID-7_2	1	0	0	0	1	1	0	0	0	2
FID-7_3	1	0	1	1	1	1	0	1	1	1

Figure 4.3: Simulated genotypes: Row names are family ID's, where the first index means family number and the second index means personal ID. 1,2 and 3 represents father, mother and child. For example, FID–2_3 means the child in number 2 family. Column names are from SNP1 to SNP559. In this figure, it only captures information from this data, which is family 1 to family 7, SNP528 to SNP 537. The numbers in the center such as 0, 1, and 2 are the simulated genotypes.

Application for Simulating Environmental Variables

For simulating environmental variables, we selected "MVITS", which indicates if the mother took multivitamins or prenatal vitamins in the perinatal period (3 months prior through the 3rd month of pregnancy) from the orofacial cleft data to generate the probability distribution. It is a categorical variable, with 1 representing "yes" and 0 representing "no". Note that in orofacial cleft data, this variable was missing for some families, which was indicated by either a 9 or NA. We estimated the proportions of 'no' responses in the orofacial cleft data using:

$$p_0 = \frac{\#[MVITS = 0]}{\#[(MVITS = 1) + (MVITS = 0)]}$$

We kept the values for MVITS=1, MVITS=0, and deleted families with MVITS=9 and MVITS=NA. The observed proportion was found to be $p_0 = 0.35$. We then simulate the environmental variables with probability $p_0 = 0.35$ and $p_1 = 1 - p_0 = 0.65$, which is the probability of taking multivitamins in the simulations. The environmental variables we generate will have 1000 values since we have 1000 trios.

4.2.2 Application 2 –Using simulated data to estimate type 1 error rate

To illustrate the use of the simulation strategy for estimating the type 1 error rate, we do not need to simulate phenotype data. We assume that all trios have affected children and we use the TDT[28] to test for association. In particular, we use the R package[32] functions colTDT and colGxE to compute our TDT and GE–TDT tests. For each dataset generated, we compute the two test statistics on all 559

SNPs. Because we have 559 SNPs, we are doing multiple comparisons and therefore we expect to see an inflated false positive rate.

To estimate the false positive rate, we simulated 1000 data sets and for each we consider multiple minimums. First, we consider the minimum across all 559 SNPs. Second, we consider the minimum across a selected pathway. Third, we consider the minimum across a selected gene. In this section, I will first introduce what genes and pathways we use and why we use it, then give the results of the simulations.

Identify genes and Pathways

- Identify Genes: Begin with the SNPs located on genes, and genes located on different pathways. In Figure 4.4, we can observe each SNP identified by its RS number and all the SNPs sequenced by their "Base.pair.position", which indicates the location of the SNPs on each gene in the genotype file, this is also in the same order as they are in the Map file. For example, in the Map file from the 1st row to 12th row, the gene called MTHFR is in this sequence. Therefore, we can retrieve the SNPs in the genotype file according to this information. Finally, we chose 4 genes according to their total variant levels, in order to perform the analysis. See Figure 4.5.

- Identify Pathways: Pathway information comes from Figure 4.1 in the first section, we can know what genes are located on which pathways. For example, there are genes called MTHFD2, FPGS, SHMT2, MTHFD1L, and ALDH1L2 which are located on the Mitochon-drial Metabolism pathway. Since we can identify genes from step 1 (Identify Genes), then we know the information of

each pathway, this may be discrete or continuous, depending on the location of the specific gene on the various pathways. Then we can apply it to the genotype matrix in R and perform the analysis. As mentioned earlier, we chose the Cytosolic Metabolism pathway when concentrating on pathway levels, since it has the highest quantity of genes on this pathway.

Chromosome	SNP.identifier	Gene_Name	Base.pair.position
1	rs4846048	MTHFR	11768839
1	rs2184226	MTHFR	11770023
1	rs1476413	MTHFR	11774887
1	rs1801131	MTHFR	11777063
1	rs6541003	MTHFR	11778454
1	rs1801133	MTHFR	11778965
1	rs1572151	MTHFR	11780298
1	rs9651118	MTHFR	11784801
1	rs17367504	MTHFR	11785365
1	rs3737964	MTHFR	11789631
1	rs12404124	MTHFR	11796456
1	rs17376328	MTHFR	11799249
1	rs9804151	CTH	70630356
1	rs7523188	CTH	70631023
1	rs17131272	CTH	70631060
1	rs4650044	CTH	70631985
1	rs4650047	CTH	70639878
1	rs648743	CTH	70648367
1	rs10889869	CTH	70650552
1	rs681475	CTH	70653305
1	rs1145920	CTH	70656428
1	rs6413471	CTH	70659771
1	rs12723350	CTH	70660689
1	rs663649	CTH	70669771
1	rs3767205	CTH	70673923
1	rs515064	CTH	70676656

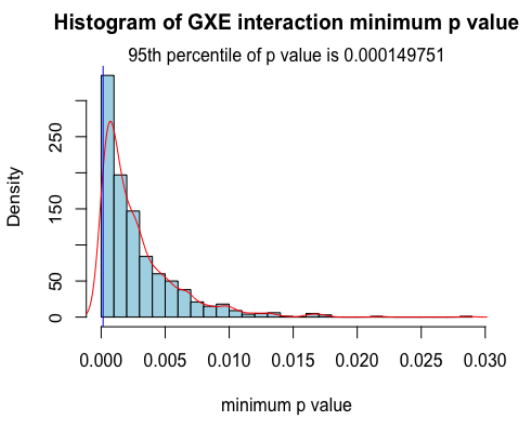
Figure 4.4: Map file

ALDH1L1	ALDH1L2	BHMT	BHMT2	CBS	CTH	DHFR	DMGDH	FOLH1	FOLR1
83	22	9	5	17	25	12	20	11	2
FOLR2	FOLR3	FPGS	FTCD	GNMT	MTHFD1	MTHFD1L	MTHFD2	MTHFR	MTHFS
4	6	7	15	4	13	104	6	12	24
MTR	MTRR	SHMT1	SHMT2	SLC19A1	SLC25A32	SLC46A1	TYMS		
19	88	6	3	15	7	6	14		

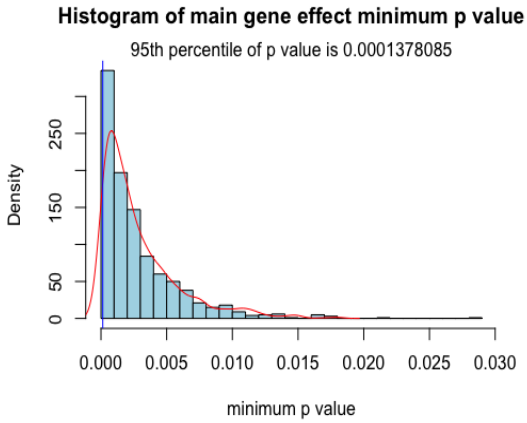
Figure 4.5: Chose 4 Genes which will be used for analyzing at the Gene Level. We chose genes which have the smallest number of variants and the largest number of variants, which will be gene SHMT2 and MTHFD1L. Then we chose 2 in between, a smaller one MTHFR which only has 12 SNPs, and a larger one CTH which has 25 SNPs.

SNP–based associations results

For the SNP–based association results, we examine the distribution of p–values across all SNPs from the gene environment interaction and main gene effect. Figure 4.6 (a) shows the histogram across 1000 simulations of the minimum p values from the GXE test, and (b) shows the histogram of the 1000 minimum p–values from the main gene effects tests. The 95th percentiles of each distributions are 0.0001498 and 0.0001378. That is, to ensure a type 1 error rate of α=0.05, we would reject the null hypothesis if observed p–values were less than approximately 0.00014. Even though there is no true association, the p–values are very small because we are testing 559 SNPs and only storing the minimum. If we instead used a Bonferroni correction, the significance cut–off is $0.05/559 \approx 0.00008945$. The 95th percentiles of the minimum p–values are at least 10 times smaller than the cutoff that we get from our empirical distribution. However, the Bonferroni cutoff assumes independent tests and we know that tests within the same gene will not be independent.



(a) The distribution of 1000 minimum p–values across all 559 SNPs for GXE interaction test



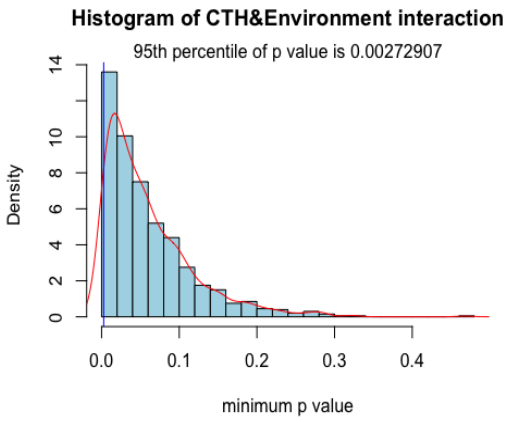
(b) The distribution of 1000 minimum p–values across all 559 SNPs for main gene effect test

Figure 4.6: SNP–based association results

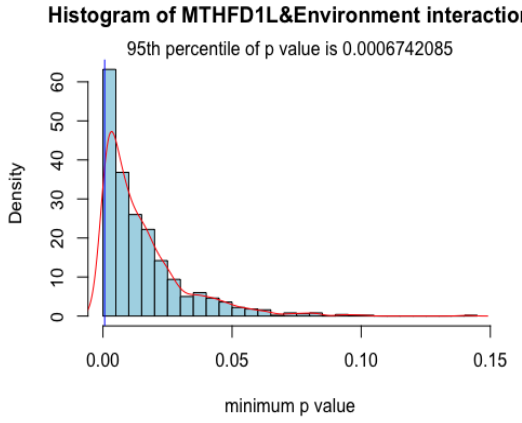
Gene–based Association Results

Figure 4.7 and Figure 4.8 give histograms of minimum p–values when the minimum across all SNPs in our selected genes. Since different genes have a different number of variants, the Bonferroni correction would be different for each gene as well. From Figure 4.7 (a), the 95th percentile of the distribution of minimum p–value for the GXE test for the CTH gene is 0.002729. There are 25 variants in the CTH gene; therefore the Bonferroni correction would be $0.05/25 \approx 0.002$. Again, we see that the p–value cutoff from our empirical distribution is greater than that based on Bonferroni.

From Figure 4.7 (b), the 95th percentile of the minimum p–value for the MTHFD1L gene and environment interaction test is 0.0006742. There are 104 variants in the MTHFD1L gene, therefore the Bonferroni correction would be $0.05/104 \approx 0.00048$.



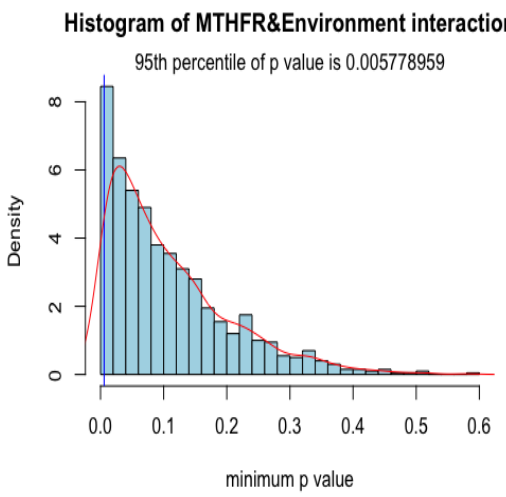
(a) The distribution of 1000 minimum p–values across 25 SNPs in the CTH gene for GXE interactions



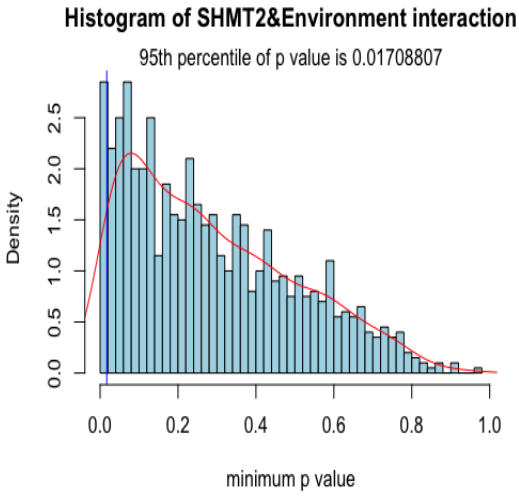
(b) The distribution of 1000 minimum p–values across 104 SNPs in the MTHFD1L gene for GXE interactions

Figure 4.7: Gene–based Association Results

Figure 4.8 gives the results for the GE interaction tests on the MTHFR and SHMT2 genes. From Figure 4.8 (a), the 95th percentile of minimum p values is 0.005779. There are 12 variants in MTHFR, therefore the Bonferroni correction is $0.05/12 \approx 0.004167$. Figure 4.8 (b) gives the 95the percentile of the p value ditribution for SHMT2, which is 0.017088. This value is much higher compared with other genes, this is because there are only 3 variants on SHMT2. The Bonferroni correction would be $0.05/3 \approx 0.016667$ and so both corrections would give similar p value cutoffs.



(a) The distribution of 1000 minimum p–values for MTHFR and environment interactions



(b) The distribution of 1000 minimum p–values for SHMT2 and environment interactions

Figure 4.8: Gene–based Association Results

Pathway–based Association Results

Figure 4.9 gives the results for a pathway–based analysis where we choose the minimum across all SNPs in genes within the selected pathway. The 95th percentile of minimum p–value is 0.0002068. In total there are 366 variants on the pathway. Therefore, the Bonferroni correction is $0.05/366 \approx 0.0001367$.

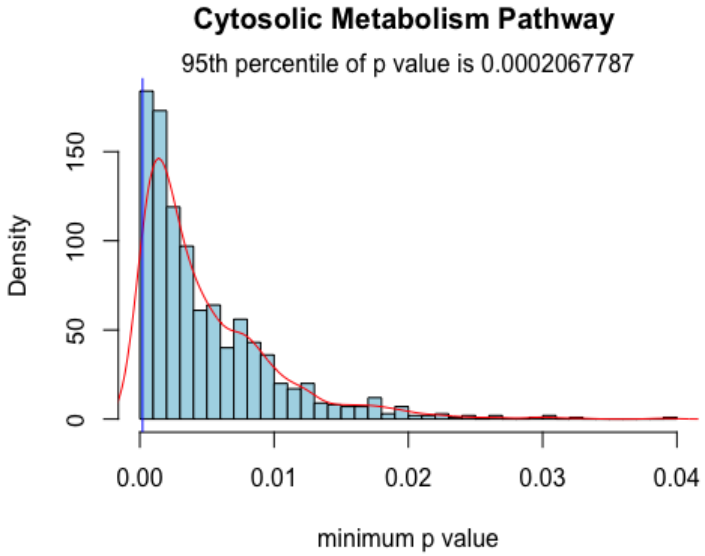


Figure 4.9: Minimum p–values across all SNPs in a pathway for the 1000 simulated data sets

4.2.3 Application 3 –An illustration of mating type asymmetry when there are maternal effects

In this section, we illustrate the use of the data simulation strategy when there are true maternal gene environment effects. We simulate data as described in Chapter 3. For simulating phenotype data, we assume the 8 causal variant scenarios given in Section 3.2.3, case 2.

Under true maternal effects, we expect distortion of the proportions for certain mating types. Normally, if there is no mating asymmetry, we would expect that the proportion of mother's with a particular genotype is equal to the proportion of father's with that same genotype. Let (M,F) denote the Mother's genotype (M) and

the Father's gentoype (F). For example, if there is mating symmetry we have

$$p_{(0,1)} = p_{(1,0)}; p_{(0,2)} = p_{(2,0)}; p_{(1,2)} = p_{(2,1)}.$$

Note that if both parent's have the same genotype, then we trivially have the proportions being equal.

Maternal genetic effects and maternal gene-environment effects are known to cause mating asymmetry in trio studies with affected offspring. Therefore, when we simulate trios with these effects, we should expect to see differences in these proportions.

To illustrate the mating asymmetry caused by true maternal effects, we first simulate the genotype for trios, given the initial values for $\beta_0 = 2.5$, $\beta_G = 2.5$, $\beta_E = -5$, and $\beta_{GE} = -4$, since these values will keep the disease probabilities in a reasonable range. But readers can also try other values for β. Under each scenario described in 3.2.3, simulate p_i for each trio, the disease probability generated by the logit model according to the mother's genotype and environmental factor of each trio. Simulate the phenotype for the child in each trio using a Bernoulli model with the probability of disease p_i . We then discard the trio if the disease status equals to 0 and only keep the affected trios. Continue the simulation for multiple times until we have 1000 affected trios.

Then we check the Father and the Mother's genotype for mating symmetry when E=0 and E=1 (The simulated environmental variable) under each scenarios.

Scenario 1 : 1 SNP 1 Gene 1 Pathway. We have the proportions when environment=0 and environment=1 in Table 4.1. When E=0, we have $p_{(01)} = 0.1916667 \neq 0.2361111 = p_{(10)}$, $p_{(02)} = 0.03055556 \neq 0.03888889 = p_{(20)}$, and $p_{(12)} = 0.1888889 \neq 0.08333333 = p_{(21)}$; When E=1, we have $p_{(01)} = 0.103125 \neq 0.1453125 = p_{(10)}$, $p_{(02)} = 0.0875 \neq 0.01875 = p_{(20)}$, and $p_{(12)} = 0.046875 \neq 0.059375 = p_{(21)}$. In both cases, we observed the distortion of mating asymmetry.

	E=0	E=1
$p_{(01)}$	0.1916667	0.103125
$p_{(10)}$	0.2361111	0.1453125
$p_{(02)}$	0.03055556	0.0875
$p_{(20)}$	0.03888889	0.01875
$p_{(12)}$	0.1888889	0.046875
$p_{(21)}$	0.08333333	0.059375

Table 4.1: The proportions of the Father and Mother's genotype when E=0 and E=1 for Scenario 1

Scenario 2 : 1 SNP 2 Genes 1 Pathway. We have the proportions when environment=0 and environment=1 in Table 4.2. In both cases, when E=0 and E=1, we observed the distortion of mating asymmetry.

	E=0	E=1
$p_{(01)}$	0.2424242	0.1208791
$p_{(10)}$	0.2286501	0.1507064
$p_{(02)}$	0.09366391	0.0266876
$p_{(20)}$	0.03305785	0.0188383
$p_{(12)}$	0.07438017	0.06279435
$p_{(21)}$	0.1515152	0.1051805

Table 4.2: The proportions of the Father and Mother's genotype when E=0 and E=1 for Scenario 2

Scenario 3 : 2 SNPs 1 Gene 1 Pathway. We have the proportions when environment=0 and environment=1 in Table 4.3. In both cases, when E=0 and E=1, we observed the distortion of mating asymmetry.

	E=0	E=1
$p_{(01)}$	0.175	0.1171875
$p_{(10)}$	0.2388889	0.1203125
$p_{(02)}$	0.03888889	0.0390625
$p_{(20)}$	0.04444444	0.0171875
$p_{(12)}$	0.07777778	0.075
$p_{(21)}$	0.1861111	0.06875

Table 4.3: The proportions of the Father and Mother's genotype when E=0 and E=1 for Scenario 3

Scenario 4 : 2 SNPs 2 Genes 1 Pathway. We have the proportions when environment=0 and environment=1 in Table 4.4. When E=0, we have $p_{(01)} = 0.2664756 \neq 0.2722063 = p_{(10)}$, $p_{(02)} = 0.06303725 \neq 0.09169054 = p_{(20)}$, and $p_{(12)} = 0.09455587 \neq 0.06876791 = p_{(21)}$. Same as E = 0, when E=1, we have $p_{(01)} \neq p_{(10)}$, $p_{(02)} \neq p_{(20)}$, and $p_{(12)} \neq p_{(21)}$. In both cases, we observed the distortion of mating asymmetry.

	E=0	E=1
$p_{(01)}$	0.2664756	0.1182796
$p_{(10)}$	0.2722063	0.1198157
$p_{(02)}$	0.06303725	0.04608295
$p_{(20)}$	0.09169054	0.04454685
$p_{(12)}$	0.09455587	0.04454685
$p_{(21)}$	0.06876791	0.07526882

Table 4.4: The proportions of the Father and Mother's genotype when E=0 and E=1 for Scenario 4

Scenario 5 : 1 SNP 1 Gene 2 Pathways. We have the proportions when environment=0 and environment=1 in Table 4.5. In both cases, when E=0 and E=1, we observed the distortion of mating asymmetry.

	E=0	E=1
$p_{(01)}$	0.2386364	0.08950617
$p_{(10)}$	0.2244318	0.1203704
$p_{(02)}$	0.03409091	0.02160494
$p_{(20)}$	0.09375	0.01851852
$p_{(12)}$	0.07954545	0.06481481
$p_{(21)}$	0.1789773	0.03240741

Table 4.5: The proportions of the Father and Mother's genotype when E=0 and E=1 for Scenario 5

Scenario 6 : 2 SNP 1 Gene 2 Pathways. We have the proportions when environment=0 and environment=1 in Table 4.6. In both cases, when E=0 and E=1, we observed the distortion of mating asymmetry, $p_{(01)} \neq p_{(10)}$, $p_{(02)} \neq p_{(20)}$, and $p_{(12)} \neq p_{(21)}$.

	E=0	E=1
$p_{(01)}$	0.2465753	0.103937
$p_{(10)}$	0.1643836	0.1606299
$p_{(02)}$	0.04657534	0.0519685
$p_{(20)}$	0.04109589	0.01574803
$p_{(12)}$	0.1753425	0.06456693
$p_{(21)}$	0.1534247	0.1307087

Table 4.6: The proportions of the Father and Mother's genotype when E=0 and E=1 for Scenario 6

Scenario 7 : 1 SNP 2 Genes 2 Pathways. We have the proportions when environment=0 and environment=1 in Table 4.7. In both cases, when E=0 and E=1, we observed the distortion of mating asymmetry.

	E=0	E=1
$p_{(01)}$	0.2438356	0.1433071
$p_{(10)}$	0.1835616	0.08188976
$p_{(02)}$	0.0739726	0.02362205
$p_{(20)}$	0.03835616	0.04251969
$p_{(12)}$	0.2410959	0.0976378
$p_{(21)}$	0.1068493	0.04094488

Table 4.7: The proportions of the Father and Mother's genotype when E=0 and E=1 for Scenario 7

Scenario 8 : 2 SNP 2 Genes 2 Pathways. We have the proportions when environment=0 and environment=1 in Table 4.8. In both cases, when E=0 and E=1, we observed the distortion of mating asymmetry, $p_{(01)} \neq p_{(10)}$, $p_{(02)} \neq p_{(20)}$, and $p_{(12)} \neq p_{(21)}$.

	E=0	E=1
$p_{(01)}$	0.2672176	0.1632653
$p_{(10)}$	0.1515152	0.100471
$p_{(02)}$	0.02479339	0.02825746
$p_{(20)}$	0.05785124	0.0188383
$p_{(12)}$	0.06887052	0.0989011
$p_{(21)}$	0.1184573	0.03296703

Table 4.8: The proportions of the Father and Mother's genotype when E=0 and E=1 for Scenario 8

Chapter 5

Conclusion

Orofacial clefts are among the most common non–fatal congenital anomalies.[7]. For many years there has been a suspicion that infants affected with oral clef have links to a genetic component which influences them having this disease. Some environmental factors also influence the presence of this disease in infants. Genotypes present within the mother and child are linked to the genetic effects within the child.[23] These Genotypes within the mother affect the unborn child's environment while it is in the womb. What the mother consumed during pregnancy (food, drink, medication, etc) must be examined to evaluate the potential impact it has on birth defects. The mother or unborn child also have specific genetic variants that could influence the environmental risk factors and affect the probability of the child being born with oral clefts.[9]

The trios data collected from a European study of children with oral cleft was previously collected from 1000 genomes. Genotyping was done on genes within three pathways. The pathways were Mitochondrial Metabolism pathway, Cross-membrane

Transport pathway and the Cytosolic Metabolism pathway. The primary environmental factor observed was whether or not the mother consumed vitamins while pregnant. Creating an algorithm which simulates both genetic and environmental data within trios was the primary objective of this project. The genetic affects could be linked to maternal genotypes and associations between these genotypes and environmental influences. Variants are associated with known biological pathways. The simulations created by the algorithm will also examine the hypothesis of no association or an association between genotypes and phenotype.

For the cases where there is no association, we have the results in SNP–based, gene–based and pathway–based associations. The 95th percentile of the minimum p–value across the SNPs, genes and pathways all got cutoff by the Bonferroni Correction. For the association cases, we simulate the phenotype by using 8 scenarios, then we observed the distortions related to mating asymmetry within each scenario.

Similar simulation studies to this point have discovered connections between genetic variants which influence the susceptibility of the child being born with oral cleft. See $Refs.$[15, 23, 33] This project's examination of the environmental factor related to vitamin consumption by the mother during pregnancy and how this factor affects the probability of oral cleft disease within the child is a first of its kind.

Regarding the limitations of this study, I believe being realistic has a strong influence, for example, when simulating under no associations, we assume every child is affected, but we actually don't know whether they are affected or not since we made up their phenotypes. When simulating under association models, we also made

up the values for β, which may cause distorted accuracy of the result. As for the extension of this study, we could investigate the powers or find more precise ways of simulating the disease probabilities.

Appendix I

Some useful commands in $Sim1000G$:

- Inputting data set.

  > `vcf = readVCF( vcf_file, maxNumberOfVariants = 600 , min_maf = 0.01, max_maf = 1)`

- Simulation.

  > `startSimulation(vcf, totalNumberOfIndividuals = 1000)`

- Simulates genotypes for 1 family with n offspring.

  > `newFamilyWithOffspring(family_id, noffspring = 1)`

- Read Genetic Map: Reads a genetic map downloaded from the function `downloadGeneticMap` or reads a genetic map from a specified file. The map must contain a complete chromosome or enough markers to cover the area that will be simulated.

  > `readGeneticMap(chromosome = 22)`

- Retrieve a matrix of simulated genotypes for a specific set of individual IDs.

  > `genotype=retrieveGenotypes(ids)`

Appendix II

The R codes for this project are given in the link as below, `https://drive.google.com/drive/u/1/folders/184YC3XNjiGLHKD5t6mUHdF9gdpopeMhE`.

Bibliography

[1] About IGSR and the 1000 Genomes Project. http://www.internationalgenome.org/data. [Accessed: 10-Apr-2019].

[2] Centre for Advanced Computing Help Document. https://cac.queensu.ca/wiki/index.php/Main_Page. [Accessed: 10-Feb-2019].

[3] Compute Canada. https://www.computecanada.ca/home/. [Accessed: 15-Jan-2019].

[4] Effects of Multiply Alleles, Co-dominance, and Incomplete Dominance on Phenotype. https://genetics-and-dna.weebly.com/effects-of-multiple-alleles-codominance-and-incomplete-dominance-on-phenotype.html. [Accessed: 15-Apr-2017].

[5] *Maternal effects in mammals*. University of Chicago Press, Chicago, [Ill.] ; London, 2009.

[6] G. Alliance and N.Y.M.A.C.G.N.S. Services. *Understanding Genetics: A New York, Mid-Atlantic Guide for Patients and Health Professionals*. NCBI Bookshelf. Genetic Alliance, 2009.

[7] B.H. Bech, C.S. Wu, E.A. Nohr, J. Li, J. Olsen, K.G. Ingstrup, H. Liang, and K. Christensen. Maternal bereavement in the antenatal period and oral cleft in the offspring. *Human Reproduction*, 28(4):1092–1099, 01 2013.

[8] Jenna C Carlson. *Methods for family-based designs in genetic epidemiology studies*. PhD thesis, University of Pittsburgh, 2017.

[9] BR Collett and ML Speltz. A developmental approach to mental health for children and adolescents with orofacial clefts. *Orthodontics & Craniofacial Research*, 10(3):138–148, 2007.

[10] Claire Dandine-Roulland and Hervé Perdry. Where is the causal variant? on the advantage of the family design over the case–control design in genetic association studies. *European Journal of Human Genetics*, 23(10):1357, 2015.

[11] Apostolos Dimitromanolakis, Jingxiong Xu, Agnieszka Krol, and Laurent Briollais. sim1000g: a user-friendly genetic variant simulator in r for unrelated individuals and family-based designs. *BMC bioinformatics*, 20(1):26, 2019.

[12] author Frommlet, Florian. *Phenotypes and genotypes : the search for influential genes*. Computational biology ; v. 18. 2016.

[13] Larry Gonick and Mark Wheelis. *The Cartoon Guide to Genetics (Updated Edition)*. Harper Perennial, 1991.

[14] J.P. Gustafson and R.B. Flavell. *Genomes*. Stadler Genetics Symposia Series. Springer US, 2013.

[15] Margaret A Honein, Sonja A Rasmussen, Jennita Reefhuis, Paul A Romitti, Edward J Lammer, Lixian Sun, and Adolfo Correa. Maternal smoking and envi-

ronmental tobacco smoke exposure and the risk of orofacial clefts. *Epidemiology*, 18(2):226–233, 2007.

[16] Farhad Hormozdiari, Gleb Kichaev, Wen-Yun Yang, Bogdan Pasaniuc, and Eleazar Eskin. Identification of causal genes for complex traits. *Bioinformatics*, 31(12):i206–i213, 2015.

[17] Anke Hüls, Katja Ickstadt, Tamara Schikowski, and Ursula Krämer. Detection of gene-environment interactions in the presence of linkage disequilibrium and noise by using genetic risk scores with internal weights from elastic net regression. *BMC genetics*, 18(1):55, 2017.

[18] K.G. Ingstrup, H. Liang, J. Olsen, E.A. Nohr, B.H. Bech, C.S. Wu, K. Christensen, and J. Li. Maternal bereavement in the antenatal period and oral cleft in the offspring. *Human Reproduction*, 28(4):1092–1099, 2013.

[19] Joseph P Jarvis, Jane Kenney-Hunt, Thomas H Ehrich, L Susan Pletscher, Clay F Semenkovich, and James M Cheverud. Maternal genotype affects adult offspring lipid, obesity, and diabetes phenotypes in lgxsm recombinant inbred strains. *Journal of lipid research*, 46(8):1692–1702, 2005.

[20] Anna Kuchment. Know what's in your genes. *Newsweek*, 150(23), 2007.

[21] S. Malcolm, J. Goodship, and T.H.J. Goodship. *From Genotype to Phenotype*. Human Molecular Genetics. Elsevier Science, 2001.

[22] Michael J. Meaney. Epigenetics and the biological definition of geneenvironment interactions. *Child Development*, 81(1):41–79, 2010.

[23] Timothy A Mousseau and Charles W Fox. The adaptive significance of maternal effects. *Trends in ecology & evolution*, 13(10):403–407, 1998.

[24] Beyoung Yun Park, Jae Woong Sull, Jung Yong Park, Sun Ha Jee, and Terri H Beaty. Differential parental transmission of markers in bcl3 among korean cleft case-parent trios. *Journal of preventive medicine and public health= Yebang Uihakhoe chi*, 42(1):1, 2009.

[25] Daniel Del Prete. codominant inheritance. `https://geneticshbdanieldelprete.weebly.com/codominant-inheritance.html`. [Accessed: 15-Apr-2017].

[26] J. Ranstam. Multiple p-values and bonferroni correction. *Osteoarthritis and Cartilage*, 24(5):763 – 764, 2016.

[27] David E Reich, Michele Cargill, Stacey Bolk, James Ireland, Pardis C Sabeti, Daniel J Richter, Thomas Lavery, Rose Kouyoumjian, Shelli F Farhadian, Ryk Ward, et al. Linkage disequilibrium in the human genome. *Nature*, 411(6834):199, 2001.

[28] Edward A Ruiz-Narváez and Hannia Campos. Transmission disequilibrium test (tdt) for case–control studies. *European journal of human genetics*, 12(2):105, 2004.

[29] Audrey H Schnell and John S Witte. Family-based study designs. *Molecular Epidemiology: Applications in Cancer and Other Human Diseases*, page 19, 2008.

[30] Stefan Schuster, David A Fell, and Thomas Dandekar. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature biotechnology*, 18(3):326, 2000.

[31] Holger Schwender, Qing Li, and Ingo Ruczinski. Preparing case-parent trio data and detecting disease-associated snps, snp interactions, and gene-environment interactions with trio.

[32] Ruczinski I. Schwender H, Li Q. Preparing case-parent trio data and detecting disease-associated snps, snp interactions, and gene-environment interactions with trio.

[33] Gary M Shaw, Cathy R Wasserman, Edward J Lammer, Cynthia D O'Malley, Jeffrey C Murray, Ann M Basart, and Marie M Tolarova. Orofacial clefts, parental cigarette smoking, and transforming growth factor-alpha gene variants. *American journal of human genetics*, 58(3):551, 1996.

[34] C. Smolke. *The Metabolic Pathway Engineering Handbook: Fundamentals*. The Metabolic Pathway Engineering Handbook. CRC Press, 2009.

[35] T Strachan. *Human molecular genetics*. Garland Science, New York, 3rd ed. edition.

[36] T. Tollefsbol. *Handbook of Epigenetics: The New Molecular and Medical Genetics*. Elsevier Science, 2017.

[37] K Van Oers, PJ Drent, G De Jong, and AJ Van Noordwijk. Additive and nonadditive genetic variation in avian personality traits. *Heredity*, 93(5):496, 2004.

[38] Yalu Wen and Qing Lu. Risk prediction models for oral clefts allowing for phenotypic heterogeneity. *Frontiers in genetics*, 6:264; 264–264, 08 2015.

[39] Jason B Wolf and Michael J Wade. What are maternal effects (and what are they not)? *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 364(1520):1107–1115, 04 2009.

[40] Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, et al. Common snps explain a large proportion of the heritability for human height. *Nature genetics*, 42(7):565, 2010.

[41] 1966 Ziegler, Andreas. *A statistical approach to genetic epidemiology.* Wiley-VCH, Weinheim, 2nd ed. edition.