Accounting for the phenomenon of Transmission Ratio Distortion in family-based genetic association studies

Lam Opal Huang

Doctor of Philosophy

Department of Epidemiology, Biostatistics and Occupational Health

McGill University

Montréal, Québec

2015-11-02

A thesis submitted to McGill University in partial fulfillment of the requirements

of the degree of Doctor of Philosophy

DEDICATION

This thesis is dedicated to my parents, Alice and Edward,

and my brother, Jonathan.

ACKNOWLEDGEMENT

It has been a long and arduous journey for the quest of knowledge and character strength that I have taken on as a challenge, with the trust in the support from those who are around me, sharing my deepest struggles and doubts, and cheering me on. Countless many whom I am indebted to for the accomplishment that I have achieved so far, whom in spirit and in action, spurred me onwards in moments of great challenge.

I would like to thank my supervisor Dr. Aurélie Labbe, who has devoted time and effort to guide me in the development of my research skills by sharing her insight and knowledge, debated with me to challenge the angle I have taken on, and encouraged me when I was bewildered by the multitude of investigative approaches. Her guidance has undoubtedly taught me the way of thinking and approaching problems with incisive clarity, and of writing for an audience with succinct vocabulary. I appreciate her provision of a research project for my application for the CIHR scholarship I received in my early years of PhD, and of the additional funding in later years. She also generously provided me with computer equipment which is crucial for supporting my research work.

I would also like to thank my co-supervisor Dr. Claire Infante-Rivard, who met with me when I first came to McGill, and was in search of a research topic. She generously provided me with different options from her projects, and encouraged me to make an independent decision on my own, trusting that my interest in the science could carry on my endeavor in the research of this topic. Little did I know at that point that how critical such seemingly small decision would sustain the weight of my thesis among other things. She always brought new insights, shed light on my work from different angles, and revealed more comprehensive understanding of the value of my research. Her suggestions have embellished my research work for the taste of a wider audience.

Apart from the research experience, both of my supervisors have been very supportive when I was facing personal issues. It was surely a challenging journey for me, but certainly for my supervisors as well. Nevertheless, they continued on with their commitment and carried me until the very end. For this, I am truly grateful.

I would like to thank the numerous teachers I have taken courses with in my first few years of PhD, from the Biostatistics faculty: Drs. Robert Platt and James Hanley, from the Statistics faculty: Drs. Russell Steele, David Stephens, David Wolfson and Dr. Alain Vandal, whom has left McGill years ago. Their passion in teaching and helpfulness in reaching out to students have really encouraged me when I was in my first year at McGill, and felt the crushing pressure of comprehensive exams. They also inspired me with the respect they showed for every question coming from their students, and the encouragement they brought for finding my own path of learning.

I am indebted to my colleagues in the Biostatistics department, Willy, Esther, Amy, and Mireille, in the Epidemiology department, Samantha, Lisa and Daphne, and my friends, Rebekah and Christy, for their tremendous support along the way. They humbled me, and taught me what friendship means, rather than competitiveness which is sometimes valued more than personal relationships.

I am also thankful towards the staff at the Student Affair office, Katherine, Andre Yves, Deirdre, and Suzanne, whom has already retired. They had answered numerous questions, and always went an extra step to make my life easier.

I would like to thank my mentor, Dr. Thomas Choy, whom has seen me through different stages of my life spanning more than a decade, facing increasing level of challenges throughout these stages because of the career decisions I have made. He never painted a rosy picture and gave me empty praises, rather presented hard facts, and let me make my own decisions weighing the risks. Even though he has retired, his constant presence during the most difficult times in my life will always be remembered.

I reserved the last mentioning to my family, my parents and my brother. They have nurtured me every step of my life until this very day. The person who I am today, and the achievements I have accomplished so far, I owe all of these to them. They challenged me when I had to make hard decisions, celebrated with me when I overcame my challenges, encouraged me when I fell, comforted me when I felt fighting a lonesome battle. At times that I felt I was stretched to my limit, a word, a hug and a smile from them encouraged me to just take one more step. They have walked alongside with me through this journey, closer than anyone, and to them I dedicate this thesis.

ABSTRACT

Transmission Ratio Distortion (TRD) is a genetic phenomenon where one of the two alleles from either parent is transmitted to the offspring with a probability different than the expected 0.5. This leads to a departure from the Mendelian inheritance ratio. There have been many animal studies reporting TRD on gene regions of known functions. These findings have triggered interest in identifying TRD loci in humans. However, human studies are relatively few, and TRD remains largely unexplored in the field of statistical genetics.

We argue that TRD is in fact an important phenomenon which lies at the intersection of three different genetic fields: developmental, population and statistical genetics. In developmental genetics, where embryonic growth is being studied, understanding TRD mechanisms can contribute to the identification of the biological processes leading to differential survival. From a population genetics perspective, TRD could give rise to rare variants which, due to the many counter-balancing evolutionary forces are maintained at a low frequency. This leads to a change in genetic diversity in the population where TRD occurs.

Since TRD involves allele transmission from parents to offspring, it can only be studied using family-based designs, and in this work, we concentrate on family trios (parents and affected offspring). Results from such studies are commonly analyzed using a log-linear model. Here, we extend this model by using the transmission probability of minor allele from parents to child as an offset. We adjust for two types of TRD: non-sex-of-parent-specific TRD (NST), and sex-of-parent-specific TRD (ST).

By conducting simulations of case-parent-trio populations, we show that either NST or ST can confound the relative risk estimates for child genotype. For ST, it further confounds the imprinting effect estimates. This leads to the inflation of Type 1 error, loss in power, and poor performance in sensitivity and specificity. We also show that spurious results due to TRD can be eliminated and correct inference restored. One limitation with this approach is the availability of the transmission probability of the minor allele which may exist or not in publically available dataset or needs to be estimated in appropriately selected control trios.

Studying TRD is worthwhile because of the close evolutionary history it might share with that of rare variants, and the confounding effect it has on imprinting effect. Both of these phenomena may help uncovering of the "missing heritability" components from past GWAS. In a more applied and immediate perspective, correct adjustment of TRD could increase consistency in findings from association studies.

ABRÉGÉ

Le “Transmission Ratio Distortion” (TRD) est un phénomène génétique au cours duquel l'un des deux allèles d’un parent est transmis à sa descendance avec une probabilité différente de celle attendue, c’est à dire 0.5. Cela conduit donc à une déviation du rapport de l'hérédité mendélienne. La présence de TRD au niveau de gènes ou de régions génétiques fonctionnelles a été rapportée dans de nombreuses études animales. Ces résultats ont suscité un intérêt pour l'identification de loci affectés par le TRD chez les humains. Cependant, les études humaines sont relativement peu nombreuses, et le TRD reste largement inexploré dans le domaine de la statistique génétique.

Nous pensons que le TRD est un phénomène important qui se trouve à l'intersection de trois domaines génétiques différents: le développement, la génétique de population et la statistique génétique. En ce qui concerne la génétique du développement, qui étudie la croissance embryonnaire, la compréhension des mécanismes du TRD peut contribuer à l'identification des processus biologiques conduisant à une différence de survie. Du point de vue de la génétique des populations, le TRD pourrait donner lieu à l’apparition de variants rares qui, en raison des nombreuses forces évolutives de ré-équilibrage, sont maintenus à basse fréquence. Cela conduit à un changement dans la diversité génétique de la population affectée par le TRD.

Puisque le TRD implique une transmission des allèles des parents à leurs enfants, il ne peut être étudié qu’à l’aide de concepts basés sur les études familiales. Dans ce travail, nous nous concentrons sur des familles-trios (parents et descendant atteint). Les données de telles études sont généralement analysées à l’aide d’un modèle log-linéaire. Ici, nous généralisons ce modèle en utilisant la probabilité de transmission de l'allèle mineur des parents à l'enfant en tant qu’``offset''. Nous ajustons le modèle pour deux types de TRD: le TRD non sexe-spécifique (NST), et TRD le sexe-spécifique (ST).

En effectuant des simulations de populations cas-parent-trio, nous montrons que le NST ou le ST peuvent agir comme facteur confondant au niveau des estimations du risque relatif par rapport au génotype de l'enfant. Le ST peut de plus agir comme facteur confondant pour les effets d'impression (imprinting). Cela engendre une inflation des erreurs de type 1, une perte de puissance, et de mauvaises performances en terme de sensibilité et de spécificité. Nous montrons également

que des résultats erronés obtenus à cause du TRD peuvent être éliminés et les paramètres du modèle corrigés. Une limitation de cette approche est la connaissance de la probabilité de transmission de l'allèle mineur, information qui peut être disponible dans des ensembles de données publiques, ou qui doit être estimée à l'aide de trios-contrôles sélectionnés de manière appropriée.

Etudier le TRD est important en raison de l'histoire évolutive que le locus du TRD pourrait partager avec ceux de variants rares, et de l'effet confondant que ce phénomène peut avoir sur l'effet d'impression. Ceci peut nous amener à découvrir une partie de l'héritabilité manquante, phénomène identifié dans les GWAs. Dans une perspective plus appliquée et à court terme, une prise en compte du TRD dans les analyses pourrait accroître la cohérence des résultats des études d'association.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AUC	Area Under Curve
CEPG	Conditional on Exchangeable Parental Genotype
CPG	Conditional on Parental Genotype
DSL	Disease Susceptibility Locus
DSP	Discordant Sib-Pair
EM Algorithm	Expectation-Maximization Algorithm
FBAT	Family-Based Association Test
GRR	Genotype Relative Risk
GWAS	Genome-wide Association Studies
HGP	Human Genome Project
HWE	Hardy-Weinberg Equilibrium
IBD	Identical-By-Descent
IUGR	Intra-Uterine Growth Restriction
LD	Linkage Disequilibrium
LL-LRT	Loglinear Likelihood Ratio Test
LRT	Likelihood Ratio Test
MAF	Minor Allele Frequency
MFC	Mother-Father-Child
MLE	Maximum Likelihood Estimator
MST	Maternal ST
NCP	Non-Centrality Parameter
NST	Non-Sex-of-parent-specific TRD
OR	Odds Ratio

PAT	Parental Asymmetry Test
PCA	Principle Component Analysis
PDT	Pedigree Disequilibrium Test
POET	Parent-of-Origin Effects Test
PO-LRT	Parent-of-Origin Likelihood Ratio Test
PST	Paternal ST
ROC Curve	Receiver Operating Characteristic Curve
RR	Relative Risk
SNP	Single Nucleotide Polymorphisms
ST	Sex-of-parent-specific TRD
TAT	Transmission Asymmetry Test
TDT	Transmission Disequilibrium Test
TDTI	TDT for Imprinting
TRD	Transmission Ratio Distortion
WGS	Whole Genome Sequencing

PREFACE: CONTRIBUTION OF AUTHORS

The contribution to this thesis is described as follows:

Chapter 1: This chapter was drafted by Lam Opal Huang (LOH). Aurélie Labbe (AL) and Claire Infante-Rivard (CIR) provided review comments on all the revisions of the Chapter.

Chapter 2: This chapter was drafted by LOH. AL and CIR provided review comments on all the revisions of the Chapter.

Chapter 3: The outline of the content was conceived by LOH. LOH drafted the manuscript and developed the original simulation studies. AL augmented the section on study design and method, and developed the corresponding figures. AL and CIR provided guidance and review comments on all the revisions of the manuscript. This manuscript was published in a peer-reviewed journal [1].

Chapter 4: The research question for this manuscript was conceived by LOH. LOH developed, implemented and applied the method for simulation studies and real data analysis, and wrote the R software package 'TRD'. AL contributed to a revision of the statistical model. LOH drafted the manuscript. AL and CIR provided guidance and review comments on all revisions of the manuscript.

Chapter 5: The research question for this manuscript was conceived by LOH. LOH developed, implemented and applied the method for simulation studies. LOH drafted the manuscript. AL and CIR provided guidance and review comments on all revisions of the manuscript.

Chapter 6: This chapter was drafted by LOH. AL and CIR provided review comments on all the revisions of the Chapter.

PREFACE: STATEMENT OF ORIGINALITY

This thesis is comprised of two introductory chapters (Chapters 1 and 2), three manuscripts (Chapters 3-5), and a discussion (Chapter 6). Each of the manuscripts includes a preamble for the overall content.

Chapter 1: This chapter is the introduction of the thesis, laying out the context for Transmission Ratio Distortion (TRD), its role among recent development in genetics, the motivation for and limitations of our proposed methods.

Chapter 2: This chapter is the background of the thesis. It includes 1) a short summary of TRD, 2) a literature review on family-based studies and statistical tests, 3) a literature review on likelihood-based models for family-based studies, comparing competing models, and 4) a literature review on methods for testing imprinting effect.

Chapter 3: This chapter is a literature review on the role of TRD in the fields of developmental, population, and statistical genetics. It is an original review, the first of its kind in the current literature, comparing the study designs, methods and results across all included studies. It also includes original material from a simulation study.

Chapter 4: This chapter investigates the impact of TRD on family-based association studies, with the development of a new method which extends a loglinear model, adjusting for the effect of TRD. The performance of the extended model is assessed using a set of simulation studies and the method is applied to a real dataset.

Chapter 5: This chapter presents an original investigation on the impact of a specific type of TRD on family-based association studies. We developed a method extending a loglinear model that adjusts for the effect of this type of TRD. Performance of the extended model is assessed by simulation studies.

Chapter 6: This chapter is the summary and discussion section of the thesis. It summarizes the importance of TRD in human genetics studies, and discusses the limitations, benefits, initiatives and challenges for future development.

PREFACE: FINANCIAL SUPPORT

The research work of this thesis is supported by:

1. McGill Provost Graduate Scholarship
2. Canadian Institute of Health Research Doctoral Scholarship in Population genetics, Genetic epidemiology and Complex diseases
3. Canadian Institutes of Health Research Operating Grant: PI - Dr. Aurélie Labbe (MOP-93723)
4. Fonds de recherche du Québec Subvention d'établissement: Dr. Aurélie Labbe (20057)
5. Medical Special Allocation: Dr. Aurélie Labbe (217123)

PREFACE: ETHICS APPROVAL

The manuscript 2 (Chapter 4) in this thesis includes analyses of previously collected data from human subjects. Ethics approval for the collection of the data was obtained by the original studies, available upon request.

Chapter 1

Introduction and objectives

1.1 The Human Genome Project

The Human Genome Project (HGP) was initiated in the 1990s resulting in advancement in mapping and sequencing the human genome. This was made possible by combining the disciplines of molecular cell biology and classical genetics, with the contribution of computational sciences. There were five main domains in the HGP which were using genomics to understand the structure of genome, understanding the biology of genome and its relationship to human diseases, and using all the former to advance the science of medicine and to improve the effectiveness of health care. Audacious strategies were planned to advance the technology and analytical methodology with the goal of correctly interpreting sequencing and other results.

1.2 Genome-wide Association Studies and missing heritability

With the rise of the popularity of the Genome-Wide Association Studies (GWAS), much progress has been made in the field of genomics in the last decade, to identify relationship of genome to human diseases. These studies have resulted in the identification of genes for several hundred traits and diseases. However, amid the apparent success of GWAS, many loci discovered could not be replicated consistently. Furthermore, they only account for a small percentage of the heritability for most complex diseases. This led geneticists to re-examine the supposed hypothesis of "common disease - common variant". Rare variants were then suggested to be the link to the "missing heritability" in high penetrance diseases, which usually cluster in families. Furthermore, scientists found another layer of genetic information to explain the "missing heritability": the epigenomic coding on the DNA sequence, which regulates gene expression of the human genome and is likely influential in determining the severity of disease.

1.3 Transmission Ratio Distortion in the fields of genetics

Transmission Ratio Distortion (TRD), which we investigated in this thesis, is a biological phenomenon where one of the alleles on a locus from either parents is over-transmitted to the next generation, violating the Mendelian transmission ratio. Different TRD mechanisms can interrupt

either the gametic or embryonic development processes, and these are explained in Chapter 3. TRD interestingly lies at the intersection of three different but related genetic fields: developmental, population and statistical genetics. This provides great incentives to investigate the role of TRD in human genetics.

1.4 Overview of manuscript 1 (Chapter 3)

TRD has been well-studied in plants and animals. However, human studies on TRD have been relatively few. We consolidated 26 such studies in the last two decades; TRD loci are involved in a whole range of disease conditions, such as various forms of cancer, neurological conditions, and others. TRD loci are also implicated in imprinting. However, the link between disease etiology and TRD mechanisms has not been established, except for embryo viability. In our review, we also included some mouse studies documented in the last decade to underscore the different methods used to study TRD as well as to compare with human study results. The studies listed in Chapter 3 include different types of designs, statistical models and tests that can be used to identify TRD loci under the influence of various forms of TRD. Representative study designs and tests are reported and used to develop working examples and figures. Note that TRD can only be studied in family-based study design instead of case-control study design because it affects the transmission of alleles from one generation to the next.

Since TRD in the parental transmission of disease allele leads to a deviation from the Mendelian ratio in the offspring generation. If TRD persists over many generations, it is possible for the over-transmitted allele to reach fixation in the population where TRD occurs, and hence lead to a slow disappearance of the disadvantaged allele. However, there are many evolutionary forces in place to regulate and maintain the disadvantaged allele at a low frequency in the gene pool and hence, resulting in rare variants. Examples of such mechanisms are mutations, recombination, genetic drift, and the presence of an immunogenetic advantage for survival in later adulthood. These mechanisms are further explained in Chapter 3. Understanding the role of TRD in the evolutionary context can provide a more comprehensive perspective of population genetics. We speculate that many of the rare variants observed in the current gene pool of various populations might indeed have a TRD origin. It is likely that identifying TRD loci could assist the discovery of rare variants

and their role in many complex diseases in regards to the "missing heritability" from classical GWAS.

1.5 Overview on manuscript 2 (Chapter 4)

TRD occurs in the diseased and the non-diseased. The presence of TRD will then lead to the over-transmission and hence, over-representation of disease allele in the offspring generation, in both diseased (case) and non-diseased (control) populations. Conventional family-based association studies recruit cases to assess over-representation of disease allele in the case populations. If this over-representation significantly deviates from the null (Mendelian inheritance ratio), an association between disease susceptible locus (DSL) and disease is then established. Since TRD and the true association between DSL and disease outcome both lead to deviation from Mendelian ratio, the measured association may be confounded when TRD occurs. In order to correctly interpret the results, we have to adjust for the effect of TRD in the measured association signal. The model and its extension we used for this purpose are developed in Chapter 4.

1.6 Overview on manuscript 3 (Chapter 5)

The second layer of genetic information in our DNA sequence is the epigenomic coding, which regulates the transcription activities of mRNA from our genome blueprint. A well-known example of epigenomic coding is imprinting, where paternally- and maternally-inherited alleles can lead to different levels of gene expression at a neighbouring disease gene in the offspring. Cases recruited from a population that exhibits an imprinting effect influencing a particular gene will have a higher proportion of disease allele in offspring inherited from the parent who induces a higher expression level. It is believed that more than 1% of all mammalian genes exhibit imprinting effect. Imprinting could potentially account for some of the "missing heritability" in genetic studies.

In Chapter 5, we investigated a special form of TRD, called sex-of-parent-specific TRD (ST). ST occurs when one parent consistently over-transmits an allele, while the other transmits under the Mendelian inheritance ratio. With ST resulting from over-transmission of disease allele from one parent, the case population recruited, if representative, will have an over-representation of the disease allele from this parent. ST effect then confounds with an imprinting effect because they show the same results in the case population. It is then crucial to address ST when studying

imprinting genes in order to reduce spurious findings. The model and its extension used for adjusting ST are developed in Chapter 5.

1.7 Overview on background information and conclusion (Chapters 2 and 6)

To provide some background information, we included in Chapter 2 a brief description of TRD, the fundamentals of family-based association analyses, some likelihood-based approaches, and recently developed methods for detecting imprinting effect in family studies. All this background knowledge laid out the context where TRD is examined and studied for the purpose of our investigation in this thesis. Chapter 6 concluded our goals for the study on TRD and proposed future initiatives following this line of research.

1.8 Motivation

We intended to investigate each of these aspects of TRD in family-based association studies in order to 1) highlight and document the study of TRD in terms of study design, methodology, and the link to known disease loci in current literature, 2) to quantify the consequences of TRD on statistical measures which leads to the possibility of spurious association results, and 3) to develop a methodology to correct for potential confounding due to TRD in association studies estimating offspring genotype relative risk and imprinting effects. The extent of TRD in human is still largely unknown. The published articles on human studies we documented in Chapter 3 are few. One of the reasons is because we cannot easily manipulate the parental mating genotype in human as we do in mouse strains. There is also no established evidence in current literature of a link between disease etiology and mechanisms of TRD, except for fetal survival. The role of TRD in human genome and its impact on complex diseases are indeed under-studied.

1.9 Limitations

In our methods described in Chapters 4 and 5, we used an offset in a log-linear model to adjust for TRD. This offset is computed as the transmission ratio of disease allele from parent to child in control-parent-trios. We tested this method in a real dataset of case- and control-trios with the measured phenotype as intrauterine growth restriction (IUGR). We were able to find loci exhibiting TRD and adjust for it. To generalize this method, we assume that 1) this transmission ratio is available from independent samples of control-trios from major consortia, such as the

HapMap project, and 2) the control-trios are recruited from the same population as the case-trios. However, control-trios are not conventionally collected in most studies of genetic diseases, due to the lack of incentives. Therefore, initiatives to recruit samples with such information and requirement are rare, especially for sex-of-parent-specific allele transmission ratio. Nevertheless we envision the research results from this thesis could generate interest and lead to an increasing awareness of TRD and its significance in human studies.

Chapter 2

Background

2.1 Overview of Transmission Ratio Distortion

TRD is the genetic phenomenon where either or both of the parents over-transmit one of their alleles to the child, leading to a departure from the Mendelian inheritance ratio. TRD can manifest itself in a non-sex-of-parent-specific (Figure 2.1) or sex-of-parent-specific (Figure 2.2) manner. There are different types of TRD which result from disruption in the gametic or embryonic development stages. Examples of TRD include germline selection, meiotic drive, gametic competition, imprint resetting error, and embryo lethality. The biological mechanisms behind these TRD processes are further explained in Figure 3.1 of Chapter 3 in more detail.

TRD lies at the intersection of three different genetic fields: developmental genetics which studies the role of genes in controlling the development of an organism, population genetics which deals with the genetic diversity in human populations due to evolutionary forces, and statistical genetics which studies the relationship between genes and human health. Studying TRD in developmental genetics can identify biological processes responsible for differential survival of zygote or embryo. In population genetics, it provides additional information on evolutionary forces that affect the diversity of the current gene pool. It can also contribute to the discovery of rare variants responsible for high penetrance diseases clustered in high risk families. Finally, in statistical genetics it leads to the correct interpretation of, on the one hand, the association or linkage signals between disease and genes and, on the other hand, of the interplay between epigenetic and DSL genetic effects.

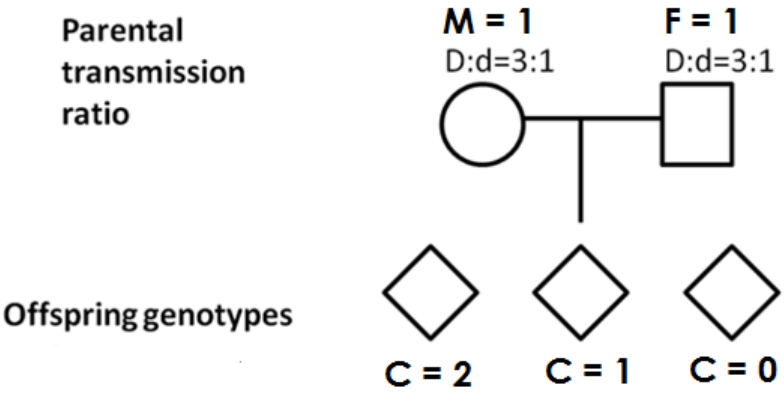
Even though TRD has been relatively well-studied in animals and plants, its prevalence in humans remains largely unknown. We searched for human studies on TRD in the last two decades. Some of the TRD loci overlap with known loci responsible for a whole range of diseases, such as cancers, Type 1 and Type 2 diabetes, developmental abnormalities, etc. Study designs and statistical methods used in these studies vary, depending on the nature of the TRD process being investigated. For example, non-sex-of-parent-specific TRD (NST) such as germline selection can be detected using trios with offspring unselected for phenotype or control-trios, by applying the Transmission

Disequilibrium Test (TDT). Embryo lethality due to epimutation can be assessed using two-generation families, comparing expected versus observed offspring genotype ratio by the Pearson's Chi-square test. Grandparental origin of TRD such as due to imprint resetting error can be detected using 3-generation families with multivariate logistic regression predicting the grandparental source of inherited allele using variables such as sex of offspring, cross (in mouse) and their interaction.

When linkage/association is being assessed between a DSL and disease status, the presence of TRD can be a confounding factor. When a parent over-transmits the disease allele due to TRD, TRD is in the same direction as the linkage/association signal, and hence, it inflates the true signal. When a parent under-transmits the disease allele due to TRD, TRD is in the opposite direction of the linkage/association signal, and hence, it attenuates the true signal. Therefore, if TRD is present but not accounted for, it can lead to false positives or false negatives and consequently, spurious conclusions. This highlights the importance of developing a statistical method which adjusts for TRD and provides correct interpretation of the linkage/association signal.

The impact of TRD at the organismal level can lead to consequences in terms of genetic diversity. When selective pressure on the disadvantaged allele occurs consistently over generations, it can cause the allele to become extinct. On the other hand, the over-transmitted allele can then reach fixation, and reduce the allelic diversity in the gene pool. This has perhaps led to the slow disappearance of TRD loci on the genome. However, different evolutionary forces can sometimes maintain the disadvantaged allele at a low frequency, such as mutations, recombinations, genetic drift and the presence of an immunogenetic advantage for survival in later adulthood. Some of these TRD loci result in rare variants. Rare variants are currently under intense research investigation, and identifying TRD loci could help in the discovery of these variants. TRD is human populations is largely under-explored, yet it holds potential to shed light on many areas of genetics ultimately contributing to our knowledge of the relationship between genes and human health.

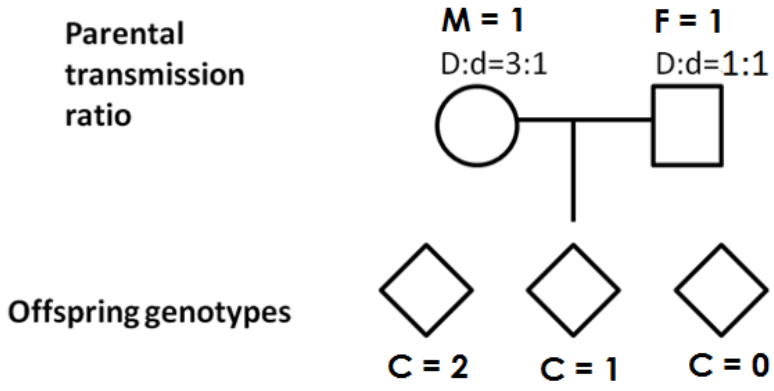
Figure 2.1 Non-sex-of-parent-specific TRD (NST)



Expected proportion under Mendelian inheritance (D:d = 1:1)	1/4	1/2	1/4
Observed proportion under NST inheritance (D:d =3:1)	9/16	6/16	1/16

*** Genotype notation of mother (M), father (F) and child (C) uses the additive model, which counts the number of minor allele that the individual carries.**

Figure 2.2 Sex-of-parent-specific TRD (ST): Maternal ST (MST)



Expected proportion under Mendelian inheritance (D:d = 1:1)	1/4	1/2	1/4
Observed proportion under MST inheritance (D:d = 3:1)	3/8	4/8	1/8

2.2 Family-based association studies

2.2.1 The evolving role of family-based studies

Family-based study design were prevalent in the latter half of the twentieth century for identifying genes associated with rare Mendelian diseases, in closely linked regions on the genome, usually with some preliminary biological evidence. Examples of such are cystic fibrosis [2] and Huntington's disease [3]. Study designs can range from sib-pairs (discordant or concordant), case-parent trios, relatives, and more complex pedigrees. For complex and more frequent diseases, amid some challenges, these studies also allowed the identification of some important genes involved in the etiology. For example, BRCA1 and BRCA2, known to predispose individual carriers to breast cancer, were discovered by linkage study [4].

As case-control GWAS became available due to the advancement in genotyping technology, they quickly replaced family-based studies, allowing coverage at low cost of millions of single nucleotide polymorphisms (SNP) on the genome in large samples. The feasibility of these studies provided information to identify multiple genes associated with complex diseases, such as coronary heart disease [5, 6], Crohn's disease [7-9], numerous forms of cancer [10-12], Type 1 [13, 14] and Type 2 diabetes [15-17], schizophrenia [18-21], and bipolar disorder [18, 22, 23].

Regardless of these successes, common variants identified in GWAS have usually revealed only small risk increment for common diseases. This led to the suggestion that rare variants which have high penetrance in affected families are accountable for the "missing heritability" [24-27]. This leads to a renewal of interest in conducting family-based studies to identify these rare variants, especially with the availability of whole genome sequencing (WGS) technology [28-30] where billion base-pairs on the genome can in principle be sequenced.

2.2.2 Population-based association study

Population-based association studies usually utilizes affected subjects and compare them with unrelated controls from the same genetic population; this case-control design serves as the basis of GWAS. Both cases and controls are genotyped for a large number of SNPs across the genome. Association with disease is then estimated at each locus and the disease status (affected or

unaffected) of the individuals, usually with a Cochran-Armitage trend test or logistic regression. A departure from independence between disease and marker is taken as evidence to the presence of association [31]. However, the control sample which is presumed from the same population as the cases may be different in ways that are difficult to measure and account for. This leads to population stratification and has possibly contributed to the poor success rate in replication of the findings [32, 33]. Population stratification seen in case-controls studies can be corrected using a number of different methods among which principle component analysis (PCA) [34] or Bayesian outlier method [35] are used. These methods may be complex and of limited use in studies of candidate genes where only a limited number of SNPs have been genotyped.

2.2.3 Family-based association study

Family-based association study design uses related subjects. The controls in these study designs are inherently matched to the cases in terms of population structure which guards against population stratification. Ideally, every member of the study unit is genotyped at each potential DSL. However, some statistical methods have the flexibility to accommodate for missing data. Linkage and/or association with case-parent trios is commonly assessed by the TDT. Other tests can be applied depending on the design or the genetic models, which are illustrated in the next section. When disease is associated with DSL, the disease allele is transmitted more (or less) often than expected under the null, indicating a departure from the Mendelian inheritance ratio.

2.2.4 TDT and family-based association tests

TDT is the simplest version of family-based association test, as well as the most commonly used [36]. It is a type of McNemar Test which uses only heterozygous parents. The original design using the TDT is a case-parent trio study where transmitted and non-transmitted disease alleles from heterozygous parents to the child are counted. The non-transmitted alleles from the parents are used to form the ethnically matched control to the case child. Therefore, it is robust against population stratification.

The McNemar Test table for TDT is shown in Table 2.1. Both parents with heterozygous genotype are non-discriminately used in computing the counts. Considering a bi-allelic locus, the number of

heterozygous parents who transmitted disease allele D to child is counted as b_1, and the number who transmitted the non-diseased allele d to the child is counted as c_1. The χ^2_{case} statistic is $\frac{(b_1-c_1)^2}{(b_1+c_1)}$, and has 1 degree of freedom (df). The corresponding TDT Chi-square statistic for control-trios is χ^2_{ctrl}. In a simulation study included in Chapter 3, we showed that χ^2_{ctrl} by itself can be used to test for the presence of TRD.

Table 2.1: TDT on case- and control-trios

	Case trios		Control trios	
	Non-transmitted allele		Non-transmitted allele	
Transmitted allele	D	d	D	d
D	a_1	b_1	a_2	b_2
d	c_1	d_1	c_2	d_2
TDT statistics	$\chi^2_{case} = \frac{(b_1 - c_1)^2}{(b_1 + c_1)}$		$\chi^2_{ctrl} = \frac{(b_2 - c_2)^2}{(b_2 + c_2)}$	

In addressing the phenomenon of segregation distortion (a type of TRD), which confounds with linkage and association signals, Spielman et al. [36] suggested the use of both case- and control-trios. He proposed a Chi-square test statistic, which we called χ^2_{CC}, also with 1df. It uses the heterozygous counts of both case- and control-trios to detect an excess or deficit in transmission counts of the minor allele compared to the major allele, between case- and control-trios. The test statistic is shown in Table 2.2. This Chi-Square statistic measures the significance of true association and linkage signal given the possible presence of TRD.

Table 2.2: Pearson's Chi-square test on case- and control-trios

	Transmitted allele in heterozygous parents		
	D	d	Row total
Case trios	b_1	c_1	n_1
Control trios	b_2	c_2	n_2
Column total	n_b	n_c	n
Pearson's Chi-square test statistic	$\chi^2_{CC} = \frac{n(b_1c_2 - c_1b_2)^2}{(n_1n_2n_bn_c)}$		

The dual-null hypothesis of TDT in GWAS or candidate gene studies with no previous linkage signal is no linkage nor association. Therefore, it has power only when both linkage and association are present. Rejecting the null hypothesis implies linkage disequilibrium (LD) between disease and DSL, which means that association is due to lack of recombination, not population stratification. The null distribution of the TDT statistic is a central Chi-square statistic with 1 df, and TDT is non-parametric. It makes no assumption on underlying genetic model or distribution of disease in the population, and hence is robust against misspecification of disease model or trait distribution.

Later developed methods generalized the TDT and accommodated nuclear families with multiple affected and unaffected offspring, such as the Family-based association test (FBAT) [37, 38]. A natural basis for association statistics is the covariance between trait and genotype. The definition of the FBAT statistics includes factor X (counts of copies of minor allele in child), Y (trait), T (coding of trait derived from Y), and P (genotype of parents). We define T as Y-μ, where Y is the phenotypic variable and μ is a fixed, pre-specified value that depends on the nature of the sample and phenotype. The covariance statistic used in the FBAT statistic is:

$$U = \sum T\,(X - E[X|P]) \qquad (2.2.1)$$

The FBAT test statistic under the null hypothesis is a central Chi-square statistic with 1 df, which is:

$$S = \frac{U^2}{var(U)} \tag{2.2.2}$$

Changing the way T is defined, one can include unaffected offspring, fit alternative traits or multiple traits, whereas changing the way X is defined, one can test alternative genetic models (recessive, dominant), and multiple alleles at a locus. Therefore, FBAT is widely applicable to many test situations. FBAT can be generalized to address arbitrary pedigree, missing parents/founders or haplotypes, or extended to handle complex phenotypes, arbitrary genetic models, and multiallelic markers. FBAT and TDT are the same under the condition that both parents are genotyped, T = 1 when affected, 0 otherwise, and X is the number of disease alleles.

Another alternative association test is the Pedigree Disequilibrium Test (PDT), which is specifically designed for analysis of LD in general pedigrees [39]. This method builds on informative pedigrees which have at least 1 informative trio or 1 informative discordant sib-pair (DSP). An informative trio has at least 1 affected child and 1 heterozygous parent, whereas an informative DSP has at least 1 affected and 1 unaffected sibling and may or may not have parental genotype data. In an informative trio, define X_T = count (D is transmitted) – count (D is not transmitted), where D is the minor allele, and in an informative DSP, define X_S = count (D in affected sib) – count (D is in unaffected sib). A summary statistic for a pedigree with n_T informative trios and n_S informative DSP is then:

$$D_i = \frac{1}{n_{Ti}+n_{Si}}\left[\sum_j^{n_{Ti}} X_{Tij} + \sum_j^{n_{Si}} X_{Sij}\right] \tag{2.2.3}$$

where i is the ith pedigree, and j is the jth trio or sib-pair within an independent pedigree. The PDT statistic is:

$$T = \frac{\sum_i^N D_i}{\sqrt{\sum_i^N D_i^2}} \tag{2.2.4}$$

where N is the total number of unrelated pedigrees. This T statistic is asymptotically normal with mean 1 and variance 0 under the null hypothesis of no LD. If we use the same data (trios only), both TDT and PDT will be asymptotically equivalent under the null hypothesis.

2.2.5 Advantages of family-based studies

In family-based studies, a significant finding usually implies both linkage and association, not population stratification. It is true that recruiting cases and unrelated controls is usually much easier than family members, especially for late-onset diseases. Regardless of the difficulties in ascertaining and genotyping multiple family members, it was shown that in rare diseases, trios-design achieves greater power than case-control design with the same number of study unit: 3 individuals for a trio, and 2 individuals for a case-control-pair [38]. Furthermore, a family-based study has additional advantages because with the proper analysis it can provide more genetic information than the case-control study; for example, an imprinting effect can be tested. Case-control studies have fallen short of accounting for high penetrance rare diseases because they usually have low power to detect rare variants (less than 1%). Rare variants are usually clustered in families, and so far have been best addressed in family-based studies. Both rare variants [24, 26, 27] and imprinting as an epigenetic effect [40-43] have been considered with greater interest because of their potential role in the "missing heritability" from the classical GWAS. This has raised interest in family-based studies in recent years. With respect to our own interest, Transmission Ratio Distortion (TRD) can only be studied using a family-based study design, where information on parental transmission of allele to the offspring is available.

2.3 Likelihood-based approaches in family-based association studies: loglinear, logistic and conditional logistic models

2.3.1 Likelihood methods for family-based studies using case-parent trios

There are a few likelihood approaches which involve testing for association using the likelihood ratio test (LRT) or the score test. For our investigation on TRD, we have considered the loglinear [44], the logistic [45] and the conditional logistic regression models [46, 47]. These family-based study approaches use the conventional case-parent-trios design and therefore are robust against

population stratification. They have features in handling complex test scenarios either in the original framework or through later extensions, which will be explained in detail in section 2.3.5.

2.3.2 Weinberg et al. (1998) loglinear model

Weinberg et al.'s [44] loglinear model is based on the multinomial likelihood of a 15-category genotype combinations, indexed by the mother (M), father (F) and child genotype (C) counting the number of copies of minor alleles in these individuals. The general form of the count probability for genotype category MFC in this model can be written as:

$$P[MFC|D] = \frac{P[D|MFC]P[C|MF]P[MF]}{P[D]}$$

where *P[D|MFC]* is the penetrance function of disease given the trio genotype MFC, *P[C|MF]* is the inheritance probability of child genotype given parental genotype, *P[MF]* is the mating type frequency, and *P[D]* = d is the disease prevalence.

This loglinear model estimates two relative risk (RR) parameters for child genotype (1 or 2 copies of minor allele), and two for maternal genotype. The genotype type is coded 0 as homozygous wild-type, 1 as heterozygous, and 2 as homozygous mutant. The loglinear model is presented as:

$$log\{E\,[n_{MFC}|D]\} = \gamma_j + log(2)I_{[MFC=111]} + \beta_1 I_{[C=1]} + \beta_2 I_{[C=2]} + \alpha_1 I_{[M=1]} + \alpha_2 I_{[M=2]} \quad (2.3.1)$$

where j indicates the j-th mating type stratum MF, ranging from 1 to 6, based on 6 unique exchangeable parental mating types as a result of the assumption on mating symmetry. RR for child genotype 1 and 2 are $R_1 = exp(\beta_1)$ and $R_2 = exp(\beta_2)$, respectively. RR for maternal genotype 1 and 2 are $S_1 = exp(\alpha_1)$ and $S_2 = exp(\alpha_2)$, respectively. This model also provides a likelihood ratio Chi-square statistic to test for significance of association between marker and disease. The LRT with child-only effect under a log-additive relative risk model is asymptotically equivalent to the TDT.

Weinberg et al. [44] have shown a possible extension of this loglinear model to further accommodate parent-of-origin effects by adding two imprinting variables for maternally (I_M) and paternally (I_F) inherited disease allele for a heterozygous child. The category MFC = 111 can then

be further divided into MFC = 111M (child disease allele inherited from mother) and MFC = 111F (child disease allele inherited from father). The augmented model based on the 16 MFC categories shown in Table 2.3 can be written as:

$$log\{E\,[n_{MFC}|D]\} = \gamma_j + log(2)I_{[MFC=111]} + \beta_1 I_{[C=1]} + \beta_2 I_{[C=2]} + \alpha_1 I_{[M=1]} + \alpha_2 I_{[M=2]} + \varepsilon_F I_F + \varepsilon_M I_M \quad (2.3.2)$$

Table 2.3: Components of original loglinear model with child, maternal and imprinting effect: equation (2.3.2)

Mating Type (MT)	MFC Genotype	Mating Type frequency $P[MF]$	Probability of transmission $P[C\|MF]$	Genotype frequency $P[MFC]$	Penetrance probability $P[D\|MFC]$	Conditional Genotype frequency $P[MFC\|D]$
1	222	p^4	1	μ_1	$f_0\ R_2\ S_2\ I_M\ I_F$	$f_0\ R_2\ S_2\ I_M\ I_F\ \mu_1/d$
2	212	$2p^3(1-p)$	1/2	μ_2	$f_0\ R_2\ S_2\ I_M\ I_F$	$f_0\ R_2\ S_2\ I_M\ I_F\ \mu_2/d$
	211	$2p^3(1-p)$	1/2	μ_2	$f_0\ R_1\ S_2\ I_M$	$f_0\ R_1\ S_2\ I_M\ \mu_2/d$
	122	$2p^3(1-p)$	1/2	μ_2	$f_0\ R_2\ S_1\ I_M\ I_F$	$f_0\ R_2\ S_1\ I_M\ I_F\ \mu_2/d$
	121	$2p^3(1-p)$	1/2	μ_2	$f_0\ R_1\ S_1\ I_F$	$f_0\ R_1\ S_1\ I_F\ \mu_2/d$
3	201	$p^2(1-p)^2$	1	μ_3	$f_0\ R_1\ S_2\ I_M$	$f_0\ R_1\ S_2\ I_M\ \mu_3/d$
	021	$p^2(1-p)^2$	1	μ_3	$f_0\ R_1\ I_F$	$f_0\ R_1\ I_F\ \mu_3/d$
4	112	$4p^2(1-p)^2$	1/4	μ_4	$f_0\ R_2\ S_1\ I_M\ I_F$	$f_0\ R_2\ S_1\ I_M\ I_F\ \mu_4/d$
	111M	$4p^2(1-p)^2$	1/4	μ_4	$f_0\ R_1\ S_1\ I_M$	$f_0\ R_1\ S_1\ I_M\ \mu_4/d$
	111F	$4p^2(1-p)^2$	1/4	μ_4	$f_0\ R_1\ S_1\ I_F$	$f_0\ R_1\ S_1\ I_F\ \mu_4/d$
	110	$4p^2(1-p)^2$	1/4	μ_4	$f_0\ S_1$	$f_0\ S_1\ \mu_4/d$
5	101	$2p(1-p)^3$	1/2	μ_5	$f_0\ R_1\ S\ I_M$	$f_0\ R_1\ S_1\ I_M\ \mu_5/d$
	100	$2p(1-p)^3$	1/2	μ_5	$f_0\ S_1$	$f_0\ S_1\ \mu_5/d$
	011	$2p(1-p)^3$	1/2	μ_5	$f_0\ R_1\ I_F$	$f_0\ R_1\ I_F\ \mu_5/d$
	010	$2p(1-p)^3$	1/2	μ_5	f_0	$f_0\ \mu_5/d$
6	000	$(1-p)^4$	1	μ_6	f_0	$f_0\ \mu_6/d$

This model is slightly different from Weinberg et al. [44] which uses a 15 MFC category model. Both of these two models encounter a multicollinearity problem because:

$$C = I_{[C=1]} + 2I_{[C=2]} = I_F + I_M$$

and in its full form is not statistically identifiable. The Expectation Maximization (EM) algorithm can be used with this model to address missing parental origin information for the triply heterozygous genotype category (MFC = 111). Simulation results showed that convergence can be achieved.

2.3.3 Weinberg (1999) logistic model

Weinberg later proposed a logistic regression model, termed parent-of-origin likelihood ratio test (PO-LRT) [45]. It proposed to tackle the problem of missing parental origin information, by using trios with parents carrying unequal copies of variant allele, so that the parental origin of disease allele in the child is known. This logistic model is written as:

$$log\left[\frac{P[M>F|MT,C]}{P[M<F|MT,C]}\right] = \alpha\, I_{[C=1]} + \beta\, I_{[M+F>1]} + \gamma\left(I_{\{M+F=1]} - I_{[M+F>2]}\right) \qquad (2.3.3)$$

where the numerator of the logit is the probability of mother carrying more copies of minor allele than father, given mating type (MF) and child genotype (C), and the denominator is the probability of father carrying more copies of minor allele than mother, given the same conditions. A different parameterization method was used in comparison with Weinberg et al. [44] as shown in Table 2.4, where R_2, R_1 and I_M are equal to $R_2I_MI_F$, R_1I_F and I_M/I_F in Weinberg et al. [44], respectively. Equation (2.3.3) uses only mating type strata 2, 3 and 5 from Table 2.4.

To calculate the conditional probability of P[M>F|MT, C, D] in equation (2.3.3) using values in Table 2.4, for example with mating type (MT) 2, child genotype 2 and M>F, we have:

$$P[M > F|MT, C, D] = \frac{P[M = 2, F = 1, C = 2|D]}{P[M = 2, F = 1, C = 2|D] + P[M = 1, F = 2, C = 2|D]}$$

$$P[M > F|MT, C, D] = \frac{f_0 R_2 S_2 \mu_2/d}{f_0 R_2 S_2 \mu_2/d + f_0 R_2 S_1 \mu_2/d} = \frac{S_2}{S_2 + S_1}$$

where D indicates disease status. Similarly, for the MT 2, child genotype 2 and M<F, we have:

$$P[M < F|MT, C, D] = \frac{f_0 R_2 S_1 \mu_2/d}{f_0 R_2 S_2 \mu_2/d + f_0 R_2 S_1 \mu_2/d} = \frac{S_1}{S_2 + S_1}$$

Table 2.4: Components of logistic model with maternal and imprinting effect: equation (2.3.3)

Mating Type (MT)	MFC Genotype	Mating Type frequency $P[MF]$	Probability of transmission $P[C\|MF]$	Genotype frequency $P[MFC]$	Penetrance probability $P[D\|MFC]$	Conditional Genotype frequency $P[MFC\|D]$
1	222	p^4	1	μ_1	$f_0 R_2 S_2$	$f_0 R_2 S_2 \mu_1/d$
2	212	$2p^3(1-p)$	1/2	μ_2	$f_0 R_2 S_2$	$f_0 R_2 S_2 \mu_2/d$
	211	$2p^3(1-p)$	1/2	μ_2	$f_0 R_1 S_2 I_M$	$f_0 R_1 S_2 I_M \mu_2/d$
	122	$2p^3(1-p)$	1/2	μ_2	$f_0 R_2 S_1$	$f_0 R_2 S_1 \mu_2/d$
	121	$2p^3(1-p)$	1/2	μ_2	$f_0 R_1 S_1$	$f_0 R_1 S_1 \mu_2/d$
3	201	$p^2(1-p)^2$	1	μ_3	$f_0 R_1 S_2 I_M$	$f_0 R_1 S_2 I_M \mu_3/d$
	021	$p^2(1-p)^2$	1	μ_3	$f_0 R_1$	$f_0 R_1 \mu_3/d$
4	112	$4p^2(1-p)^2$	1/4	μ_4	$f_0 R_2 S_1$	$f_0 R_2 S_1 \mu_4/d$
	111M	$4p^2(1-p)^2$	1/4	μ_4	$f_0 R_1 S_1 I_M$	$f_0 R_1 S_1 I_M \mu_4/d$
	111F	$4p^2(1-p)^2$	1/4	μ_4	$f_0 R_1 S_1$	$f_0 R_1 S_1 \mu_4/d$
	110	$4p^2(1-p)^2$	1/4	μ_4	$f_0 S_1$	$f_0 S_1 \mu_4/d$
5	101	$2p(1-p)^3$	1/2	μ_5	$f_0 R_1 S_1 I_M$	$f_0 R_1 S_1 I_M \mu_5/d$
	100	$2p(1-p)^3$	1/2	μ_5	$f_0 S_1$	$f_0 S_1 \mu_5/d$
	011	$2p(1-p)^3$	1/2	μ_5	$f_0 R_1$	$f_0 R_1 \mu_5/d$
	010	$2p(1-p)^3$	1/2	μ_5	f_0	$f_0 \mu_5/d$
6	000	$(1-p)^4$	1	μ_6	f_0	$f_0 \mu_6/d$

In another example for MT 2, child genotype 1 and M>F, we have:

$$P[M > F|MT, C, D] = \frac{P[M = 2, F = 1, C = 1|D]}{P[M = 2, F = 1, C = 1|D] + P[M = 1, F = 2, C = 1|D]}$$

$$P[M > F|MT, C, D] = \frac{f_0 R_1 S_2 I_M \mu_2/d}{f_0 R_1 S_2 I_M \mu_2/d + f_0 R_1 S_1 \mu_2/d} = \frac{S_2 I_M}{S_2 I_M + S_1}$$

Similarly, for MT 2, child genotype 1 and M<F, we have:

$$P[M < F|MT, C, D] = \frac{f_0 R_1 S_1 \mu_2/d}{f_0 R_1 S_2 I_M \mu_2/d + f_0 R_1 S_1 \mu_2/d} = \frac{S_1}{S_2 I_M + S_1}$$

With the same approach, we obtained equation (2.3.3) for all MFC categories in strata 2, 3, and 5, as shown in Table 2.5.

Table 2.5: Parents carrying unequal variant allele counts

Stratum	MFC genotype	P[M>F|MT, C, D]	P[M<F|MT, C, D]	P[M>F|MT, C, D]/ P[M<F|MT, C, D]
2	212 and 122	$S_2/(S_1+S_2)$	$S_1/(S_1+S_2)$	S_2/S_1
2	211 and 121	$S_2 I_M/(S_2 I_M + S_1)$	$S_1/(S_2 I_M+ S_1)$	$S_2 I_M / S_1$
3	201 and 021	$S_2 I_M/(S_2 I_M +1)$	$1/(S_2 I_M +1)$	$S_2 I_M$
5	101 and 011	$S_1 I_M/(S_1 I_M +1)$	$1/(S_1 I_M +1)$	$S_1 I_M$
5	100 and 010	$S_1/(S_1+1)$	$1/(S_1+1)$	S_1

The predictors in this model (equation (2.3.3)) are more difficult to interpret. The imprinting variable I_M is uniquely present in numerator of rows 2 to 4 in Table 2.5, where the common condition is C = 1. Therefore, $I_M = exp(\alpha)$, where α is the regression parameter for indicator variable $I_{[C=1]}$. The maternal variable S_2 is only present in numerator of rows 1 to 3, where M+F = 2 or 3. The common condition of these 3 rows are therefore M+F>1. Therefore, $S_2 = exp(\beta)$ where β is the regression parameter of the indicator variable $I_{[M+F>1]}$. Then, the maternal variable S_1 is only in the denominator of rows1 and 2, and in the numerator of rows 4 and 5, and is not present in row 3. Therefore, a positive indicator variable for rows 4 and 5 is M+F = 1, and a negative indicator variable for row 1 and 2 is M+F = 3, or M+F > 2, because negative in the log scale corresponds to division in the original scale. Therefore, $S_1 = exp(\gamma)$ is in the numerator when M+F = 1, and in the denominator when M+F > 2, and γ is the regression parameter of the difference between indicator variables $I_{[M+F=1]}$ and $I_{[M+F>2]}$. The model is fitted without an intercept,

to ensure the predictors estimate exactly the maternal genotype 1 and 2 effects, and parent-of-origin (imprinting) effect without a reference level. This model does not give an estimate for child genotype effect.

2.3.4 Cordell et al. (2002, 2004) conditional logistic model

Cordell et al. also proposed a case-parent trio approach, but used the untransmitted allele from the parents to generate pseudo-controls and fitted them in a conditional logistic regression model [46, 47]. For example, each case-parent trio contains 4 parental haplotypes. Taking 1 haplotype out of each parent, 4 possible phased child genotype can be created, one of which is the actual case child and the remaining 3 are pseudo-controls. This is the conditional on parental genotype (CPG) approach. If we assume parental genotypes are exchangeable, we will have 4 more pseudo-controls, and this becomes the conditional on exchangeable parental genotype (CEPG) model. However, not all pseudo-controls formed from either CPG or CEPG approaches are useful in the sense that only the ones with deducible parental-origin and/or phase information can be retained for fitting the model. This is due to the restriction on the model being fitted, whether it depends on parental-origin and/or phase. In such cases, pseudo-controls for which parental-origin and/or phase cannot be determined are discarded. For the model that does not depend on either parental-origin or phase, there could be only 1 pseudo-control, using the left-over alleles once the case alleles are removed from the parental genotype pair.

Notation of trio genotypes for Cordell's approach [46, 47] is g_c, g_m, g_f, for child, mother and father, respectively, is a notation equivalent to that used in Weinberg et al. [44, 45] as C, M and F. We retained the notation by Cordell et al. here for easier reference to the original papers [46, 47]. The general form of conditional probability of g_c for each trio contributing to the CPG conditional likelihood can be written as:

$$P\left[g_c \middle| g_m, g_f, D, \xi\right] = \frac{P[D|g_c, g_m, g_f]}{\sum_{g_c^* \epsilon G_\xi} P[D|g_c^*, g_m, g_f]} \tag{2.3.4}$$

where g_c, g_m, g_f are as defined previously, D is the disease status of the child, ξ is the event where parental-origin and/or phase can be deduced depending on the model being fitted, and $g_c^* \epsilon G_\xi$ are

all the g_c that met the condition defined by ξ. For CEPG conditional likelihood, each trio's contribution to the conditional likelihood can be written as:

$$P\left[g_c \middle| g_m, g_f, D, \xi\right] = \frac{P[D|g_c, g_m, g_f]}{\sum_{g_c^*, g_m^*, g_f^* \epsilon G_\xi} P[D|g_c^*, g_m^*, g_f^*]} \quad (2.3.5)$$

where $g_c^*, g_m^*, g_f^* \ \epsilon G_\xi$ are all the (g_c, g_m, g_f) combinations that met the condition defined by ξ under exchangeable parental genotypes. The proofs for equations (2.3.4) and (2.3.5) are shown in the appendix of Cordell et al. [46].

As shown in Self et al. [48] and Schaid [49], the conditional probability in equation (2.3.4) is equivalent to that used in the conditional logistic regression with a case of (phased) genotype g_c matched to a number of pseudocontrols of (phased) genotype g_c^* where $g_c^* \epsilon G_\xi$. The likelihood for the whole dataset is the product of the conditional probability across all N case-parent trios. Note that the conditional likelihoods in equations (2.3.4) and (2.3.5) are without the nuisance parameters, $P[g_c|g_m, g_f, \xi]$ and $P[g_m, g_f, \xi]$ [46].

This conditional logistic regression approach provides a natural and flexible framework to incorporate epistasis (gene-gene interaction), gene-environment interaction, and parent-of-origin effect, and can handle multi-allelic loci, multiple linked loci, and multiple linked loci in a multiple unlinked region, without the need to adjust for nuisance parameters [46]. The more restricted model assuming parental allelic exchangeability generates 4 additional pseudo-controls and increases power when studying parent-of-origin effect. However, simulation using this method shows that there is limited power to distinguish parent-of-origin effect from mother-fetal genotype interaction.

2.3.5 Comparison of Weinberg and Cordell approaches

Cordell and Weinberg's approaches make no assumption about Hardy-Weinberg Equilibrium (HWE) or random mating. We consider Cordell's conditional logistic regression model [46, 47] a robust and competitive alternative to Weinberg's loglinear [44] and logistic [45] models. Cordell incorporated many features in her proposed method [46, 47], which cover a myriad of study

designs, inclusion of genetic and non-genetic factors, several types of genotype parameterization, and the use of phenotypic data. Multiple extensions of Weinberg et al. [50-54] have also made the approach more appealing in facing the challenges of complex test scenarios, such as quantitative trait, missing data, multiple offspring, and multi-allelic locus.

The conditional logistic approach [46, 47] does not require the fitting of the nuisance parameters, $P[g_c|g_m, g_f, \xi]$ and $P[g_m, g_f, \xi]$, which is an advantage. However, this approach does not make full use of missing data. When there is a missing parent, only one pseudo-control can be generated. When inference is not possible, trios are discarded, which leads to reduced power because some data is lost. On the other hand, Kistner et al. [51, 55] extended Weinberg's approach [44, 45] by using an EM algorithm to retrieve the missing information, and hence make use of the incomplete trios.

An extension of Weinberg's approach by Gjessing et al. [50] is the ability to handle multi-allelic loci as well as multiple linked and unlinked loci with unknown phase, but it requires specialized software HAPLIN. Cordell's approach [46, 47] also requires the specialized program PSEUDOCC to generate pseudo-controls. The features of Weinberg et al. [44] and Cordell et al. [46, 47] approaches are enlisted in Table 2.6.

2.3.6 Application to study on TRD

In order to extend existing method for handling TRD, we made use of the multinomial probability of Weinberg et al. loglinear model [44], and separate out the component of transmission probability of child genotype given parental mating type, *P[C|MF]*. The details of this extension are shown in our Chapters 4 and 5. As for Cordell's conditional logistic regression [46, 47], an extension is possible but less convenient because the nuisance parameters of the term $P\left[g_c\middle|g_m, g_f, \xi\right]$ is canceled in the calculation of the conditional likelihood $P\left[g_c\middle|g_m, g_f, D, \xi\right]$ under Mendelian inheritance [46]. When there is TRD, the resulting conditional likelihood does not simply depend on the penetrance function $P\left[D\middle|g_c, g_m, g_f\right]$ alone, but also a function of $P\left[g_c\middle|g_m, g_f, \xi\right]$ which complicates the maximization procedure of the regression parameters in a standard conditional logistic regression framework. Weinberg's PO-LRT [45] although provides

estimate for parent-of-origin effect, cannot be used for the extension for TRD because there is no natural component in the model readily available for such purpose.

Table 2.6. Comparing Weinberg and Cordell approaches

Author	Weinberg (extensions)	Cordell (extensions)
Model	Log-linear model [44]	Conditional logistic model [46, 47]
Study design	Case-trios	Case-trios/matched pseudo-controls
Assumption on HWE	No	No
Assumption on random mating	No	No
Estimation of nuisance parameters	Yes	No
Handle maternal-fetal genotype interaction	Yes (Sinsheimer et al. [54])	Yes
Handle Maternal effect	Yes	Yes
Handle Parent-of-origin	Yes	Yes
Handle Multiple offspring	Yes (Kistner et al. [56])	Yes
Handle Multi-allelic locus	Yes (Gjessing et al. [50])	Yes
Handle Multiple linked/unlinked loci with unknown phase	Yes (Gjessing et al. [50], Shi et al. [53])	Yes
Handle missing data	Yes (Kistner et al. [51, 55, 56]) Does not discard trios with ambiguous parent-origin and unknown phase	No Discard trios with ambiguous parent-origin and unknown phase
Handle Gene-environment interaction	Yes (Kistner et al. [52])	Yes
Handle Gene-gene interaction	No	Yes
Handle Quantitative trait	Yes (Kistner et al. [51, 52, 55, 56])	Yes (Wheeler et al.[57])
Specialized software	HAPLIN for multi-allelic or multiple haplotype with unknown phase (Gjessing et al. [50])	PSEUDOCC (in stata) to generate pseudo-controls (Clayton [58])

2.4 Current methods on testing for parent-of-origin (imprinting) effect

2.4.1 Overview of the current literature

Imprinting is the phenomenon when the disease allele inherited by the offspring from the father induces a different level of gene expression at a neighbouring disease gene, compared to disease allele inherited from the mother, which determines the amount of transcription activities at the DSL. The differential level of expression changes the penetrance of disease in child depending on the parental-origin of the inherited disease allele, and hence the RR of the child genotype. It is believed that more than 1% of the mammalian genes are subject to imprinting. Few lines of methodology in current literature that have been developed to study parent-of-origin effect in association studies. These include extensions of the TDT, the Parental Asymmetry Test (PAT), the loglinear, logistic, and conditional logistic models. We now examine the basic principle of these approaches for binary traits.

2.4.2 Extensions of Transmission Disequilibrium Test (TDT)

Zhou et al. have developed the parent-of-origin effects test (POET), based on a McNemar test, to detect the presence of imprinting effect for case-parent trios [59]. Assuming an additive genotype model counting the number of copies of the disease allele (noted D), there are a total of 15 mother-father-child (MFC) genotype categories with exchangeable parental mating types. These 15 categories can be divided into 3 groups: 1) mother and father carry an equal number of disease allele, 2) mother carries more disease allele than father, and 3) father carries more disease allele than mother. The corresponding counts are: $N_{M=F} = N_{222}+N_{112}+N_{111}+N_{110}+N_{000}$, $N_{M>F} = N_{212}+N_{211}+N_{201}+N_{101}+N_{100}$, $N_{F>M} = N_{122}+N_{121}+N_{021}+N_{011}+N_{010}$, respectively, where N_{MFC} is the number of trios with maternal (M), paternal (F) and child (C) genotype combination.

When there is imprinting, it is more likely for the affected child in the sample to have inherited the disease allele from the parent who induces a higher expression level at a neighbouring disease gene. Under the null hypothesis of no imprinting, counts in groups 2 and 3 should be equal. Therefore, the POET can be defined as a McNemar test in the form of:

$$POET = \frac{N_{F>M} - N_{M>F}}{\sqrt{N_{F>M} + N_{M>F}}} \quad (2.4.1)$$

which follows a standard normal distribution under the null hypothesis of no imprinting. A significant p-value indicates the presence of an imprinting effect. This test can be performed even when the marker is not necessarily the DSL. Weinberg [45] has previously noted that when both parents are heterozygous, the transmissions of disease allele from mother and father are not statistically independent. In POET, the MF = 11 category is excluded in computing the test statistic.

A TDT-imprinting (TDTI) was proposed in [60], to test for linkage/association in the presence of maternal or paternal imprinting. This test is a combination of the POET and the regular TDT statistics. Hu et al. [60] re-define the TDT statistic as the square-root of the original TDT statistic by Spielman et al. [36]. This TDT test statistic, when both parents are included, distributed as standard normal under the null hypothesis and can be written as:

$$TDT_b = \frac{u^T N - v^T N}{\sqrt{u^T N + v^T N}} \quad (2.4.2)$$

where $u = (u_j)_{j=1}^{15}$ is a vector of indicator variables representing the categories of the event that the disease allele is transmitted, $v = (v_j)_{j=1}^{15}$ represents the event that disease allele is not transmitted, from either or both heterozygous parents, and $N = (N_j)_{j=1}^{15}$ is the vector of the number of trios which belongs to category j (or MFC). Similar indicator vectors $u_f = (u_{fj})_{j=1}^{15}$ and $v_f = (v_{fj})_{j=1}^{15}$ are defined for heterozygous fathers who transmit and do not transmit the disease allele to child, respectively, and u_m and v_m for heterozygous mothers. TDT statistics separately for heterozygous mothers and fathers are defined as:

$$TDT_m = \frac{u_m^T N - v_m^T N}{\sqrt{u_m^T N + v_m^T N}} \quad (2.4.3)$$

and

$$TDT_f = \frac{u_f^T N - v_f^T N}{\sqrt{u_f^T N + v_f^T N}} \quad (2.4.4)$$

both of which are distributed as standard normal. The combined TDTI statistic is then a combination of the original TDT, and TDTs for mother and father, with the significance of POET statistic as an indicator, which determines inclusion of any of these three statistics. This TDTI statistic can be written as:

$$TDTI = TDT_m I_{[POET<-\frac{z_\alpha}{2}]} + TDT_f I_{[POET>\frac{z_\alpha}{2}]} + TDT_b\, I_{[|POET|\leq\frac{z_\alpha}{2}]} \quad (2.4.5)$$

where z_α is the two-sided significance level for the POET test for imprinting. Under the null hypothesis of no imprinting, POET and the TDT_m, TDT_f and TDT_b are asymptotically independent, therefore, TDTI is asymptotically standard normal. This TDTI statistic is shown to be more powerful than TDT when parent-of-origin effect is significant, while less powerful when it is not significant [60].

When there is only one parent available (either mother or father), the corresponding 1-POET test for imprinting can be written as [61]:

$$1POET = \frac{w(N_{M<C}-N_{M>C})-(1-w)(N_{F<C}-N_{F>C})}{\sqrt{w^2(N_{M\neq C})+(1-w)^2(N_{F\neq C})-(n_m+n_f)^{-1}(N_{M<C}-N_{M>C})(N_{F<C}-N_{F>C})}} \quad (2.4.6)$$

where $w = \frac{n_f}{n_f+n_m}$, and n_f is the number of case-father pairs, and n_m is the number of case-mother pairs. The corresponding 1-TDTI test for linkage/association in the presence of imprinting is [61]:

$$1TDTI = \frac{w(N_{M<C}-N_{M>C})+(1-w)(N_{F<C}-N_{F>C})}{\sqrt{w^2(N_{M\neq C})+(1-w)^2(N_{F\neq C})+(n_m+n_f)^{-1}(N_{M<C}-N_{M>C})(N_{F<C}-N_{F>C})}} \quad (2.4.7)$$

which is also distributed as a standard normal under the null hypothesis of no imprinting. Xia [62] extended the TDTI to test for imprinting effect in complete and incomplete families with one or multiple children (C-TDTI). Xia [63] later extended the test to address quantitative traits (Q-C-TDTI).

2.4.3 Extensions of Parental Asymmetry Test (PAT)

The PAT also uses case-parent-trios to detect parent-of-origin effect. Only the categories that have heterozygous child, and different maternal and paternal genotypes are used. If we set $N_{F>M}$ to be the counts of trios with father carrying more disease allele than mother, and $N_{M>F}$ to be the counts of mother carrying more disease allele than father, the PAT statistic can be written as:

$$PAT = \frac{N_{F>M,C=1} - N_{M>F,C=1}}{\sqrt{N_{F>M,C=1} + N_{M>F,C=1}}} \tag{2.4.8}$$

which is distributed as a standard normal under the null hypothesis of no imprinting. Note that the PAT proposed by Weinberg [45] is the square of this statistic and follows a Chi-square (1) distribution.

When only one parent is available along with an arbitrary number of children, the 1-PAT was proposed by Zhou et al. [64] to address the study design and test for imprinting in the presence of linkage/association. This statistic can be written as:

$$1PAT = \frac{w\left(N_{M<C,C=1} - N_{M>C,C=1}\right) + (1-w)\left(N_{F<C,C=1} - N_{F>C,C=1}\right)}{\sqrt{w^2(N_{M\neq C}) + (1-w)^2(N_{F\neq C}) + (n_m + n_f)^{-1}\left(N_{M<C,C=1} - N_{M>C,C=1}\right)\left(N_{F<C,C=1} - N_{F>C,C=1}\right)}} \tag{2.4.9}$$

which is also distributed as a standard normal. Similar to 1-TDT, $w = \frac{n_f}{n_f + n_m}$, and n_f is the number of case-father pairs, and n_m is the number of case-mother pairs. It can also be extended to include multiple affected offsprings in one family. In the same paper, Zhou et al. [64] proposed C-PAT, which combines PAT and 1-PAT including complete and incomplete nuclear families, respectively, in a single test for imprinting effect (full mathematical details can be referred to in Zhou et al. [64]).

Becker [65] proposed an extension to PAT for nuclear families using haplotype, which is termed HAP-PAT. The corresponding statistic for HAP-PAT is a McNemar test as the TDT, which can be written as:

$$HAP - PAT = \frac{n-1}{n} \sum_i \frac{(t_{i1} - t_{i2})^2}{t_{i1} + t_{i2}} \tag{2.4.10}$$

where n is the total number of nuclear families in the sample, t_{i1} is the count of i-th haplotype a child inherited from the father, and t_{i2} is the count of ith-haplotype inherited from the mother, where i-th haplotype $h_i \in H$, for i ranges from 1 to n, representing n possible haplotypes. This HAP-PAT test for imprinting in the presence of association. With the same study design, Zhou [66] developed the HAP-1-PAT, by using multiple tightly linked markers for families with only one parent available. The test statistic can be written as:

$$HAP\ 1PAT = \frac{n-1}{n}\sum_{i=1}^{n}\frac{[w(t_{Mi1}-t_{Mi2})+(1-w)(t_{Fi1}-t_{Fi2})]^2}{w^2(t_{Mi1}+t_{Mi2})+(1-w)^2(t_{Fi1}+t_{Fi2})} \quad (2.4.11)$$

where the weight $w = \frac{n_f}{n_{f+n_m}}$, with n_f and n_m as previously defined. The count t_{Mi1} is number of heterozygous child inheriting haplotype h_i from the father, and t_{Mi2} is the number of heterozygous child inheriting haplotype h_i from the mother, both in case-mother families. The counts t_{Fi1} and t_{Fi2} are the corresponding counts for the case-father families. Zhou [66] further extended it to include families with either both parents or one parent by the HAP-C-PAT test.

Zhou et al. [67] extended the PAT to include general pedigrees in a method named PPAT, which uses all informative family trios from pedigrees. The PPAT statistic can be written as:

$$PPAT = \frac{\sum_{j=1}^{N}\sum_{i=1}^{n_i} S_{ij}}{\sum_{j=1}^{N}\left[\sum_{i=1}^{n_i} S_{ij}\right]^2} \quad (2.4.12)$$

where S_{ij} is the PAT statistic for ith trio in jth pedigree.

Zhou et al. [68] also proposed to make use of control children in families when testing for imprinting to increase statistical power in detecting imprinting effect in the presence of association. The PATu and 1-PATu are developed to include families with both parents and one parent respectively. The C-PATu was then developed to combine the complete and incomplete families in one analysis, with weighted contribution from case and control families based on disease prevalence [68]. The extended PATs with inclusion of control-families are shown to have greater power than using case-families alone, and are robust to population stratification. Furthermore, it was shown that misspecification of population prevalence of disease can reduce the power of C-PATu, but will not invalidate it.

2.4.4 Transmission Asymmetry Test (TAT)

Weinberg [45] constructed the TAT in the spirit of TDT, but for detecting parent-of-origin effect. TAT is essentially the same as TDT except that case-trios where both parents are heterozygous are excluded in the analysis. Taking only heterozygous father married to homozygous mother and heterozygous mother married to homozygous father, the TAT tests for equal transmission of the disease and non-disease alleles. The resulting test is a 1-df Chi-square McNemar test, as the TDT. Weinberg [45] used simulated data to show that the power of TAT is poor.

2.4.5 Loglinear model by Weinberg et al. (1998)

Weinberg et al. [44] proposed a loglinear model, with details described in Chapters 2.3.2, as a competing model with TAT, to detect parent-of-origin effect. Covariates entered into the model are child and maternal genotypes, and paternal and maternal imprinting variables. Then, a likelihood ratio test is performed against the background null model with no covariates, which is termed loglinear likelihood ratio test (LL-LRT). This LL-LRT is shown to have better performance in terms of power than TAT [45]. The loglinear model in Weinberg et al. [44] results in a LRT which tests for both association and parent-of-origin effect.

2.4.6 Logistic model by Weinberg (1999)

In later study, Weinberg [45] proposed a logistic model framework that only uses mating types with unequal copies of disease allele in the father and mother. The resulting model, PO-LRT gives an estimate for imprinting effect, and maternal effect with one or two copies of disease allele in the mother, as shown in Chapter 2.3.3. Weinberg [45] noted that when the investigator is certain that there is no maternal effect, then samples used in PO-LRT are further reduced to trios containing only heterozygous children. The result is the PAT, which is shown to have better power than PO-LRT.

2.4.7 Conditional logistic model by Cordell et al. (2002, 2004)

A separate line of research uses conditional logistic regression to test for imprinting effect, where three pseudo-controls are generated by the untransmitted alleles from parents to an affected child

[46, 47] (see Chapter 2.3.4). This approach incorporates a wide-range of solutions to address the relevant statistical issues, including parent-of-origin effect, and many others mentioned previously. This conditional logistic model conditions on parental genotypes and child being diseased, and does not include nuisance parameter such as the mating type frequencies. The study shows that when the condition is relaxed to exchangeable parental genotype, power to detect parent-of-origin effect is increased. However, this method discards trios with ambiguous parent-of-origin or unknown phase information, which leads to a 3-9% loss of trios [46]. A simulation study also reveals that this method has an inflation of Type 1 error. Based on simulation, Cordell et al. [46] stated that their method shows limited power in differentiating parent-of-origin effect and mother-child interaction effect.

2.4.8 Application to scenarios with sex-of-parent-specific TRD (ST)

Our goal to study imprinting effect is the situation when ST confounds with this signal. The approaches described above to extend TDT and PAT [59-69] have exhausted all the possible development to the existing TDT and PAT methods, with increasing mathematical complexity and decreasing practicality as different study designs, availability of genotype or haplotype data, and missing data problem are added to the scenarios. These methods do not have a readily available component for adjusting Non-Mendelian transmission. Furthermore, the existing framework of these tests cannot easily incorporate covariates such as child, maternal, and maternal-fetal genotype interaction effects, which are in close relation with the imprinting effect and are sometimes being studied together. Therefore, this line of developed methods does not fit our current and future research goals.

As it is described more fully in Chapter 5, a sex-of-parent-specific offset, which is a slight modification to the non-sex-of-parent-specific offset proposed in Chapter 4, can be used to address ST in the loglinear model. It is intuitive and simple to implement with essentially no change to the original test framework (model and study design). Logistic model by Weinberg does not offer such property [45]. Similar extension of the conditional logistic model [46, 47] might be possible, but involves a more complicated likelihood for maximization as explained in Chapter 2.3.6.

Chapter 3

Transmission ratio distortion:

Review of concept and implications for genetic association studies

3.1 Preamble

This chapter constitutes the basis of TRD in the context of three separate but related genetic fields: developmental, statistical and population genetics. We defined TRD in statistical term, and underscored the importance of TRD in these three fields. From a developmental genetics perspective, knowledge of TRD can provide additional information on the relationship between genes and growth of organism, and eventually increase the understanding of zygotic and embryonic development of humans. TRD is also important from a population genetics perspective because it contributes as part of the evolutionary forces in determining the genetic diversity of the human genome in different populations. Alleles under TRD are sometimes maintained at a low frequency due to various evolutionary forces such as recombination, mutation, drift and the presence of an immunogenetic advantages in later adulthood. The result of which is the rise of rare variants. There has not been many human studies in identifying TRD loci in the last two decades. With the number of TRD studies available, many different study designs have been proposed, each with various statistical tests or models. We described several TRD mechanisms, which require corresponding study design and statistical model to detect and quantify the TRD signal.

In the 26 TRD studies we investigated, four gene regions (*SUPT3H-MIRN586-RUNX2, IGF2/INS, DMPK,* and *H19*) were replicated across multiple studies in exhibiting TRD. Given the limited number of studies that were included, we considered this as ample evidence for the existence of TRD. Most cited studies used family-based study design with population unselected for phenotypes from major consortia such as Framingham Heart Studies, HapMap project and Centre d'Etude du Polymorphisme Humain (CEPH). However, these loci found that exhibit TRD are mapped to known gene regions for various types of diseases even though study populations are not ascertained for their phenotypes. Other studies used families of affected and unaffected individuals, or carriers of disease allele to assess the excess in transmission of disease allele with respect to the non-disease allele.

The most interesting aspect from our point of view is the role of TRD in the midst of exciting discoveries of new loci responsible for various disease condition and traits. Even though the extent of TRD is relatively unknown in human, we noted that the implication of TRD on genetic linkage and association studies cannot be simply ignored.

The presence of TRD can lead to spurious conclusion on newly discovered disease loci, if not accounted for. TRD is an often overlooked phenomenon in human genetic studies. This chapter has brought into light the importance of TRD in three different genetic fields. It also highlighted the current progress on study designs and methods developed for detecting TRD in the field of statistical genetics, which serves as a precursor to further development of models to adjust for TRD in the presence of true linkage/association signals.

Manuscript 1: Transmission ratio distortion: review of concept and implications for genetic association studies

Lam Opal Huang,[1,$] Aurélie Labbe,[1,2,$] Claire Infante-Rivard[1,$,*]

1 Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal Quebec, Canada, H3A 1A3

2 Douglas Mental Health University Institute, Montreal, Quebec, Canada, H4H 1R3

$ All authors contributed equally to this work.

*Address for correspondence: Claire Infante-Rivard. claire.infante-rivard@mcgill.ca.

Tel : 514-398-4231; Fax: 514-398-7435

Keywords: Transmission Ratio Distortion, Transmission Disequilibrium Test, Imprinting, Mendelian ratios, Association studies

3.2 Abstract

Transmission ratio distortion (TRD) occurs when one of the two alleles from either parent is preferentially transmitted to the offspring. This leads to a statistical departure from the Mendelian law of inheritance, which states that each of the two parental alleles is transmitted to offspring with a probability of 0.5. A number of mechanisms are thought to induce TRD such as meiotic drive, gametic competition, and embryo lethality. TRD has been extensively studied in animals, but the prevalence of TRD in humans remains largely unknown. Nevertheless, understanding the TRD phenomenon and taking it into consideration in many aspects of human genetics has potential benefits that have not been sufficiently emphasized in current literature. In this review, we discuss the importance of TRD in three distinct but related fields of genetics: developmental genetics which studies the genetic abnormalities in zygotic and embryonic development, statistical genetics/genetic epidemiology which utilizes population study designs and statistical models to interpret the role of genes in human health, and population genetics which is concerned with genetic diversity in populations in an evolutionary context. From the perspective of developmental genetics, studying TRD leads to the identification of the processes and mechanisms for differential survival observed in embryos. As a result, it is a genetic force which affects allele frequency at the population, as well as, at the organismal level. Therefore, it has implications on genetic diversity of the population over time. From the perspective of genetic epidemiology, the TRD influence on a marker locus is a confounding factor which has to be adequately dealt with to correctly interpret linkage or association study results. These aspects are developed in this review. In addition to these theoretical notions, a brief summary of the empirical evidence of the TRD phenomenon in human and mouse studies is provided. The objective of our paper is to show the potentially important role of TRD in many areas of genetics, and to create an incentive for future research.

3.3 Introduction

Transmission ratio distortion (TRD) is observed when one of the two alleles from either parent is preferentially transmitted to the offspring, leading to a statistical departure from the Mendelian inheritance ratio of 0.5 [70]. When observed in affected offspring, as conventionally measured by the transmission disequilibrium test (TDT) [36], this departure is interpreted as suggesting the presence of linkage and association between the allele and the offspring condition. Allelic transmission from parents to affected offspring has been used in genetic association studies as one way to provide validation for case-control results because, contrary to case-controls results, transmission results are not affected by population structure bias. However, the TRD phenomenon has also been empirically observed in apparently unaffected offspring [71-77], although the extent of TRD in the human genome is not well known. The presence of this departure from the expected Mendelian transmission has an impact on the interpretation of results from linkage and association studies in affected individuals because it occurs in the general population [72, 73, 75, 78].

TRD could potentially inflate or attenuate the linkage or association signal in identical-by-descent (IBD) or TDT-like test results, respectively. Two or more alleles are said to be IBD if they are identical copies of the same ancestral allele. An over-sharing of alleles IBD between related affected individuals at a specific marker indicates linkage between this marker and the disease susceptibility locus. A TDT assesses over-transmission of a minor allele with respect to the major allele of certain marker locus in case (affected) trios. If the result of TDT is significant, it suggests an association and linkage between the marker locus and the disease locus. Therefore, TRD on a marker locus that causes excess or deficit in allele sharing and transmission, which acts independently from the disease status, can lead to false positives or negatives in IBD sharing and TDT results. In fact, this TRD phenomenon is observed in the general population which includes both affected and unaffected individuals. Therefore, a linkage or association signal observed in affected individuals does not necessarily guarantee true linkage and association between marker and disease loci. Moreover, the presence of TRD leads to significant power loss in such studies. All these aspects have not been sufficiently emphasized in current literature and will be addressed in this review. A listing of TRD studies and their results has been included in Table 3.1 and 3.2, which will be discussed more extensively later in the paper.

Many biological mechanisms governing the passage from gametic formation to embryonic development can contribute to TRD [70, 71, 75, 78-82]. These mechanisms lead to differential survival in gametes, zygotes, and embryos and have implications on developmental genetics. Moreover, when TRD repeatedly occurs over many generations, the frequency of the allele that is favored and the alleles at close loci begin to shift upwards in the population [83]; as a consequence, the disadvantaged allele at the TRD locus gradually becomes rare in the population. We set up a simulation study to trace the marker allele frequency over time for a locus under TRD influence [84]. We found that even under a strong departure from Mendelian law of inheritance, it can take more than 10 generations for the advantaged allele to reach complete fixation, i.e. the allele frequency becomes 1 (results available from authors). This simulation set up will be discussed in more detail under the section on population genetics perspective. This observation has implications for population genetics because it reduces the diversity of the population gene pool over generations as disadvantaged alleles are eliminated through time [83, 85, 86].

A review of TRD will therefore lead us to address aspects related to both developmental and population genetics, in addition to statistical interpretation of genetic association studies in its presence. We begin this review by discussing the possible TRD mechanisms; thereafter, specific methods to detect TRD in different study designs are presented as they are related to the underlying biological/developmental processes. We then report results from studies evaluating TRD in the literature of human (Table 3.1) and mouse (Table 3.2) studies. The importance of TRD as a confounding factor in linkage and association studies is discussed next. Finally, we briefly address TRD from the perspective of population genetics linking it with current strategies to uncover rare variants.

3.4 TRD mechanisms

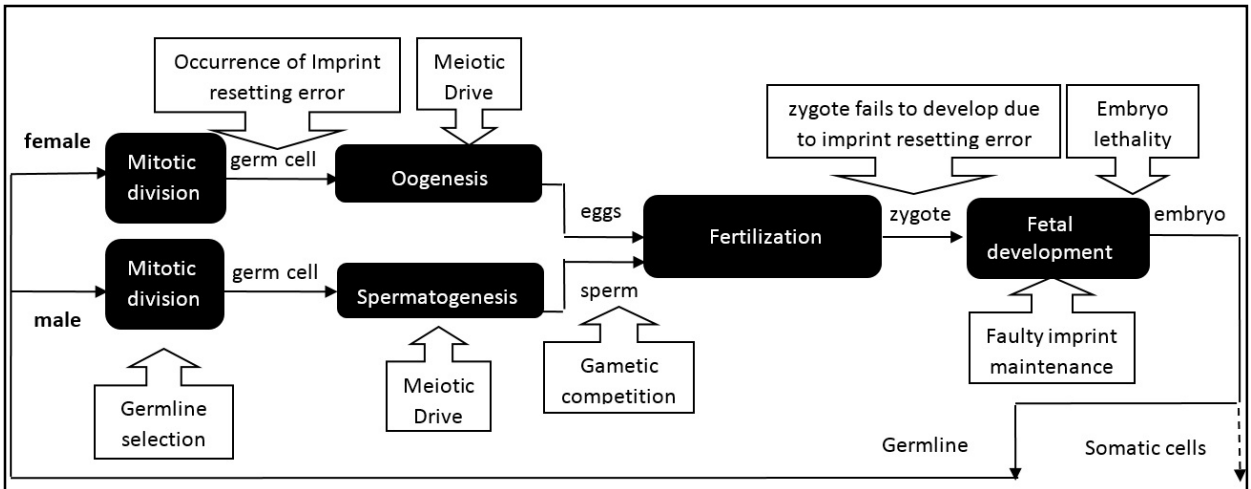
TRD has been identified and modeled in humans [71-73, 75, 77, 87-90], mouse [91-102], drosophila [103-105], and lesser kestrel [106]. It is a result of disruptive mechanisms during the gametic or embryonic development stages (Figure 3.1). These TRD mechanisms include germline selection during mitosis of germ cells [79], meiotic drive during female and male meiosis [70], gametic competition of sperm to achieve fertilization [75], embryo lethality due to deleterious genotype or mother-fetal incompatibility [75], as well as imprint resetting error in parental germ

cells when the parents are still embryos in the body of the grandparents, or faulty imprint maintenance at fertilization or in early embryonic development stage of the offspring [80, 89, 90].

Except for the two imprint regulation processes mentioned above, TRD at a marker locus can be observed from a sample of unaffected offspring and their parents' genotypes. In this situation, a deviation from the Mendelian 1:1 ratio of allelic transmission is observed. On the other hand, imprint resetting error and faulty imprint maintenance both lead to a more complex form of TRD, in which the deviation of Mendelian ratios is attributed to parent-of-origin distortion. Genomic imprinting occurs when certain genes are expressed in a parent-of-origin specific manner, through an inheritance process independent of Mendelian inheritance. For example, the imprinted allele from the mother is silenced such that only the non-imprinted allele inherited from the father is activated, and likewise for the imprinted allele from the father and the corresponding non-imprinted allele from the mother.

Before meiosis happens in parents, imprint resetting occurs in parental germ cells when they are still embryos in the body of the grandparents, and parental imprints are erased and re-established according to the sex of the parents [107]. The father's two imprints from the paternal grandparents are both reset to paternal imprints, while the mother's imprints from the maternal grandparents are reset to maternal imprints, such that the four sister chromatids resulting from meiosis in either parent could all have the same imprint. This reprogramming ensures that every sperm cell contains a paternal imprint and that egg cell contains a maternal imprint. When a sperm unites with an egg to form a functional zygote, there is one paternal and one maternal imprint, which is essential for survival. If the erasure process fails, for example in the female, a proportion of eggs would contain a paternal imprint. An egg having a faulty paternal imprint unites with a sperm carrying a paternal imprint will form a zygote with two paternal imprints, which is incompatible with survival [80]. Unsuccessful imprint resetting in males leads to the same consequence. Under such circumstances, several authors have suggested that if the normal function of the imprinted gene is necessary for successful fertilization or embryo survival, imprinting resetting errors may point to grandparental-origin-TRD [80, 90, 108]. This will be discussed in detail in the next section.

Figure 3.1: Underlying biological mechanisms behind TRD



(1) **Germline selection** - Germ cell life cycle begins when a mature embryo is formed. The germ cells first start division through mitosis. During mitosis, mechanisms such as mutation, recombination and gene conversion, collectively called germline selection mechanisms cause cells with certain genotypes to be produced at a higher proportion than others. Hence, germ cells entering the next stage, meiosis, have an imbalanced genotype ratio.

(2) **Meiotic drive** - Female meiosis is called oogenesis, and male spermatogenesis. Since oogenesis is asymmetric by nature, only one of the four chromatids becomes a functional gamete, and the others become polar bodies and are eliminated. The chromatid of the haplotype with structural advantage in facilitating the orientation and replication during meiosis tends to be transmitted more. This mechanism is called meiotic drive. Although rare, meiotic drive can occur in male eukaryotes as well. There is another type of meiotic drive called sex chromosome drive that occurs during spermatogenesis, which leads to unequal production of X- or Y-bearing gametes.

(3) **Gametic competition** - In some male organisms, sperms survived through meiotic drive tend to compete with each other to achieve fertilization. This is called gametic selection. Well-studied models of gametic selection include t-haplotype system in mouse and segregation distorter in drosophila.

(4) **Imprinting errors** - Imprint resetting occurs during the postimplantation stage, where parental imprints are erased and re-established. When an error occurs during imprint resetting, the resulting embryo may be incompatible for survival. Faulty imprint maintenance during embryonic development can also lead to the death of embryos.

(5) **Embryo lethality** - After the embryo is formed, there are other mechanisms of selection termed embryo lethality. One example of embryo lethality is the Rh+ system where mother and fetal blood types are incompatible. During delivery when the placenta ruptures, upon the blending of maternal blood with fetal blood stream, the fetus dies.

3.5 TRD inference: study designs and methods

3.5.1 Overview

Since TRD involves a deviation from the Mendelian 1:1 ratio of allelic transmission from parents to offspring, it can only be observed in family-based studies. However, the ascertainment of these families differs depending on the goal of the study. If the intention of the study is to search for association or linkage between a marker and a specific disease, families will have been ascertained based on the disease of interest. Therefore, the presence of TRD becomes a confounding signal and can be falsely interpreted as a linkage or association signal. On the other hand, if the search for TRD loci is unrelated to a specific disease but rather the primary research goal, families with offspring unselected for phenotype or disease should be genotyped. Under these study conditions, an observed deviation from Mendelian inheritance may be attributed to one of the underlying biological mechanisms of TRD described in the previous section, or to some others that remain unknown.

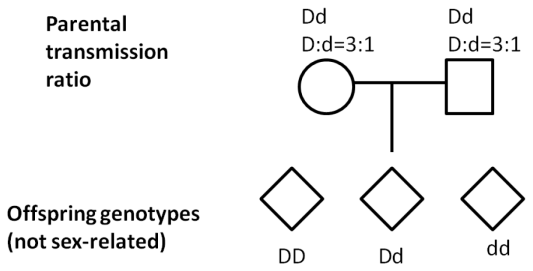
Depending on these biological mechanisms, TRD can be observed in different family structures unselected for phenotype. Choice of family structure includes i) two-generation families (parents and offspring) for the general case of TRD, to assess transmission from parents to offspring or from parent to female (male) offspring for sex-of-offspring specific TRD, ii) larger families, to study over-sharing of alleles identical by descent (IBD) between “affected” sib pairs which are defined to be the "survived" offspring, and iii) three-generation families (grandparents, parents and child) for grandparental-origin-dependent TRD. The variety of these designs targeting specific underlying biological processes suggests that different statistical analyses are appropriate in each of their corresponding contexts. These different scenarios are reviewed in detail in the following sections.

3.5.2 Detecting TRD in trios with offspring unselected for phenotype

The over-transmission of an allele from heterozygous parents to offspring is conventionally measured by the TDT in a sample of trios (parents and their offspring) [36]. Figure 3.2 illustrates this most general form of TRD, where allelic transmission disequilibrium occurs in a non-sex-specific manner. In this example, observed offspring genotypes do not follow the Mendelian ratio,

leading to a departure from the expected genotype distribution. This type of TRD can be identified in trios unselected for phenotype using TDT, which is a McNemar test assessing the null hypothesis that the transmission of one allele is the same as the transmission of the alternative allele at a marker locus in heterozygous parents [36].

Figure 3.2: General case of TRD observed in trios, using a TDT approach. Consider a TRD locus with 2 alleles D and d, where the allelic transmission ratio from parent to unaffected offspring is D:d=3:1. This figure illustrates all possible offspring genotypes regardless of their sex, arised from a pair of heterozygote parents.



Expected proportion under Mendelian inheritance (D:d = 1:1)	1/4	1/2	1/4
Observed proportion in a sample of families with heterozygote parents	9/16	6/16	1/16

TDT in a sample of n=90 families with heterozygote parents

	D non transmitted	**d non transmitted**	**Total**
D transmitted	a = 0	b = 120	120
d transmitted	c = 60	d = 0	60
Total	60	120	2n=180

The TDT tests for the null hypothesis of Mendelian allelic transmission D:d=1:1

Null hypothesis $H_0 : \frac{b}{b+c} = \frac{c}{b+c} = 0.5$ $\chi^2 = \frac{(b-c)^2}{b+c} = 20$ P-value is 7.7×10^{-6}

Over-transmission of a marker allele from parents to offspring can also occur in a sex-of-parent-specific manner. An over-transmission from mother to offspring not observed in father can be explained by female meiotic drive, whereas an over-transmission from father to offspring not observed in mother can be explained by male meiotic drive, which is rare, or by gametic competition (see Figure 3.1). Examples of these TRD mechanisms were seen in two human studies included in Table 3.1 [81, 82]. In principle, these TRD mechanisms can be uncovered using the

TDT with trios, stratifying the transmission counts according to maternal or paternal origin, where the over-transmission from heterozygote mothers or fathers is tested using a TDT, as shown previously in Figure 3.2.

However, when both parents are heterozygous, TDT on mothers vs. TDT on fathers is no longer a valid test due to lack of statistical independence between paternal and maternal transmissions [45]. Other tests have been proposed in determining parent-of-origin effect, such as Transmission Asymmetry Test (TAT), Likelihood Ratio Test (LRT), and Parental Asymmetry Test (PAT). However, these tests require the absence of prenatal maternally-mediated effect defined as the effect of maternal genotype on phenotype of child. TAT omits counts when both parents are heterozygous and therefore ensures independence of parental transmission. However, prenatal maternally-mediated effect can cause differential weighting of the paternal and maternal transmission in TAT, and may give spurious parent-of-origin effect. The LRT from a log-linear model can take into account of both prenatal maternally-mediated and parent-of-origin effect. However, this test might not be valid if the allele tested is a marker in proximity of a neighboring disease susceptibility locus instead of a candidate gene itself, due to possibility of recombination during the formation of gametes where parent-of-origin might be interchanged.

Another approach was proposed with the Parent-Of-Origin Likelihood Ratio Test (PO-LRT); its aim is to determine parent-of-origin effect by stratifying population according to parental mating type and child genotype. This stratification removes the dependence on the parental inheritance, the inherited copies of allele in child, and possible gametic recombination, so that within strata counts depend only on prenatal maternally-mediated effect and parent-of-origin effect. When there is assumed to be no prenatal maternally-mediated effect, PO-LRT is reduced to PAT, which uses only heterozygous cases (child who inherited 1 copy of disease allele) where parents transmit different alleles to the child, while the other trios are no longer informative because both parents transmit the same allele. Therefore, for the scenario where diseases are subject to prenatal maternally-mediated effects and the investigated locus is possibly a marker in proximity of a disease locus, PO-LRT remains the only valid testing procedure [45].

3.5.3 Detecting TRD in extended families with offspring unselected for phenotype

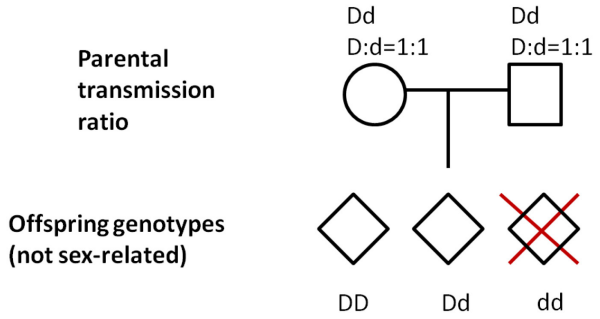
A deviation from Mendelian inheritance cannot always be attributed to a biological process occurring in parents. After the embryo is formed, there are other mechanisms of selection which are collectively termed embryo lethality (Figure 3.1). In this case, embryos with a specific genotype are eliminated, leading to an imbalance in the offspring genotypic ratios as illustrated in Figure 3.3A. Another form of embryo lethality involves an epimutation instead of a DNA mutation, where methylation on imprinted genes which control gene expression is disturbed. This could result in spontaneous abortion [109]. Note that, embryo lethality is different from the previous example of germline selection, meiotic drive, and gametic competition, where the advantaged alleles are transmitted at a higher proportion while the disadvantaged genotype is still observable in the offspring generation. A TDT approach using both parents as described above can be used.

An alternative analytical strategy with larger families is to use non-parametric linkage analysis, which looks at over-sharing of alleles identical by descent (IBD) between "affected" related pairs. In this specific case, all offspring in the extended families are labeled "affected", which essentially means "having survived", and the objective is to determine regions in the genome linked to the phenotype defined as "being alive in the last generation" [73]. The over-sharing of alleles IBD between sib pairs at an embryo-lethality TRD locus is illustrated in Figure 3.3B. Note that over-sharing of alleles IBD in related pairs can be observed only in families with heterozygote parents at the TRD locus. In the example of Figure 3.3, homozygote dd individuals could not have survived and homozygote DD parents could not produce dd embryos implying that a deviation from Mendelian ratio cannot be detected unless both parents are heterozygote. This constraint has some consequences in the statistical analysis, as IBD sharing between sib pairs cannot be detected with doubly heterozygote parents. In this case, multipoint linkage analysis, where IBD status is estimated from neighboring markers, should be performed. This analytical strategy was used by Paterson et al. [73] in the Framingham Heart Study cohort, but no loci met the genome-wide criteria for linkage. Note that embryo lethality can also be sex-specific, which induces a sex-of-offspring specific TRD. The analytical strategy is the same as above, except that linkage analysis is performed only in female (respectively male) offspring, i.e., between sisters (brothers) in the example of Figure 3.3B. However, since one looks at over-transmission (Figure 3.3A) or over-sharing (Figure 3.3B) of a marker allele while embryos with the faulty genotype could not have

survived, it is impossible to determine whether the observed TRD occurred in the parents or at the embryonic stage.

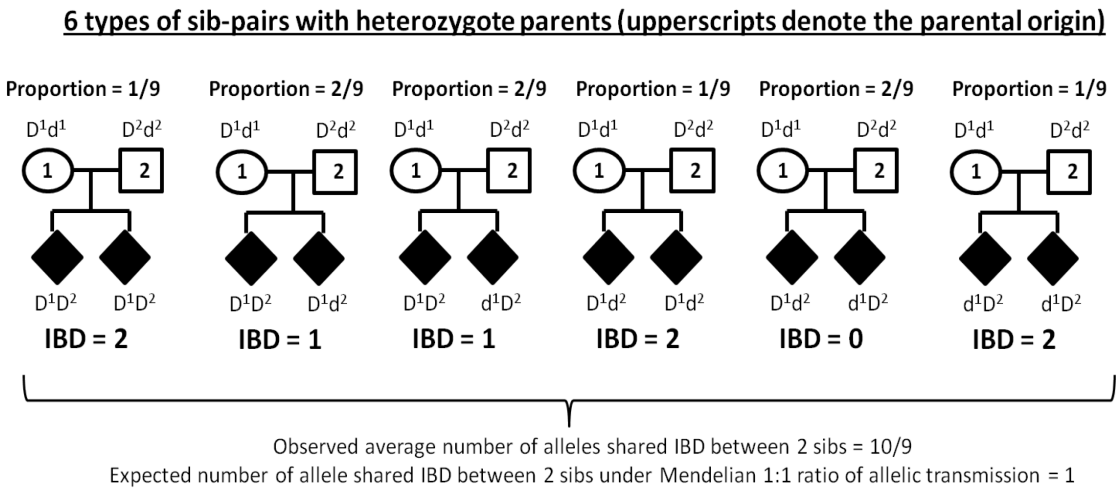
Figure 3.3: TRD caused by embryo lethality. We assume here that the mutant allele is d and that lethality is autosomal recessive. As a result, dd genotype is eliminated before birth. (A) Deviation from genotypic Mendelian ratios in offspring, observed in families with heterozygote parents. (B) Illustration of the IBD sharing between sib pairs in families with heterozygote parents when there is TRD. Note that in practice, the parental origin of the genotype in these samples needs to be inferred using neighboring markers.

(A)



Observed proportion in embryos before lethality	1/4	1/2	1/4
Observed proportion in live borns	1/3	2/3	0

(B)



3.5.4 Grandparental origin TRD: imprinting errors

In the two types of TRD described above, the deviation in allelic transmission from Mendelian ratio is inferred based on what is observed in the offspring genotypes (see Figure 3.2 and Figure 3.3A). Another form of TRD can occur which is induced by an imbalance in the grandparental origin of the offspring's genotypes. Under Mendelian inheritance in humans, each individual contains the genetic information transmitted by his/her four grandparents, with an expected transmission ratio of 1:1:1:1. However, a deviation from this ratio, which is also a form of TRD, can be explained by a possible imprint resetting errors in the parent's germline, or erroneous maintenance of parental imprints in early embryonic development stage. Figure 3.4.1 illustrates an example of a three-generation family with correct imprint resetting and maintenance. In this example, we assume that the genetic locus is maternally imprinted, which means that only paternal alleles are activated in offspring. As we see in Figure 3.4.1, imprint marks have been correctly reset in grandparents A, B, C and D, so that each egg cell contains a maternal imprint and each sperm cell contains a paternal imprint. As a result, both individuals in the second generation inherit a correctly imprinted allele from their mother and a correctly non-imprinted allele from their father. The same resetting process successfully occurs in the germline of the second generation individuals (father and mother) before meiosis. Then, when the egg from the mother is fertilized by the sperm of the father, each of them transmits a correctly imprinted allele to the offspring. As seen in Figure 3.4.1, there is no deviation from the Mendelian ratio in either the offspring genotypic ratios, nor in the allelic origin of parents and grandparents.

Figure 3.4.2 illustrates the scenario where an imprint resetting error occurred on allele 2 of the mother, which is incompatible with embryonic survival. This leads to the deviation from Mendelian inheritance ratio in the allelic origin of the grandparents. Interestingly, this also leads to a deviation from the Mendelian ratio in the offspring, which seems to suggest that this phenomenon could be captured by using the TDT approach in trios described above. For comparison, Figure 3.4.3 illustrates a similar scenario, but the imprint resetting error occurred on allele 1 of the father. Similarly, the allele which failed to reset correctly is under-transmitted. A deviation from Mendelian ratio of the alleles from grandparents can be observed in the offspring. This observation is the basis of the statistical analyses aiming to uncover TRD induced by imprinting errors.

Figure 3.4.1: Example of a three-generation family including 4 grandparents, 2 parents and offspring. We consider a marker with 2 alleles, denoted as1 and 2. Grandparents are denoted as A, B, C and D and superscripts at each genotype indicate the grandparent origin. In this example, correct imprint resetting occurs in the germline before the production of eggs and sperm cells. We assume here that the marker is maternally imprinted and imprinted marks are represented by a red triangle.

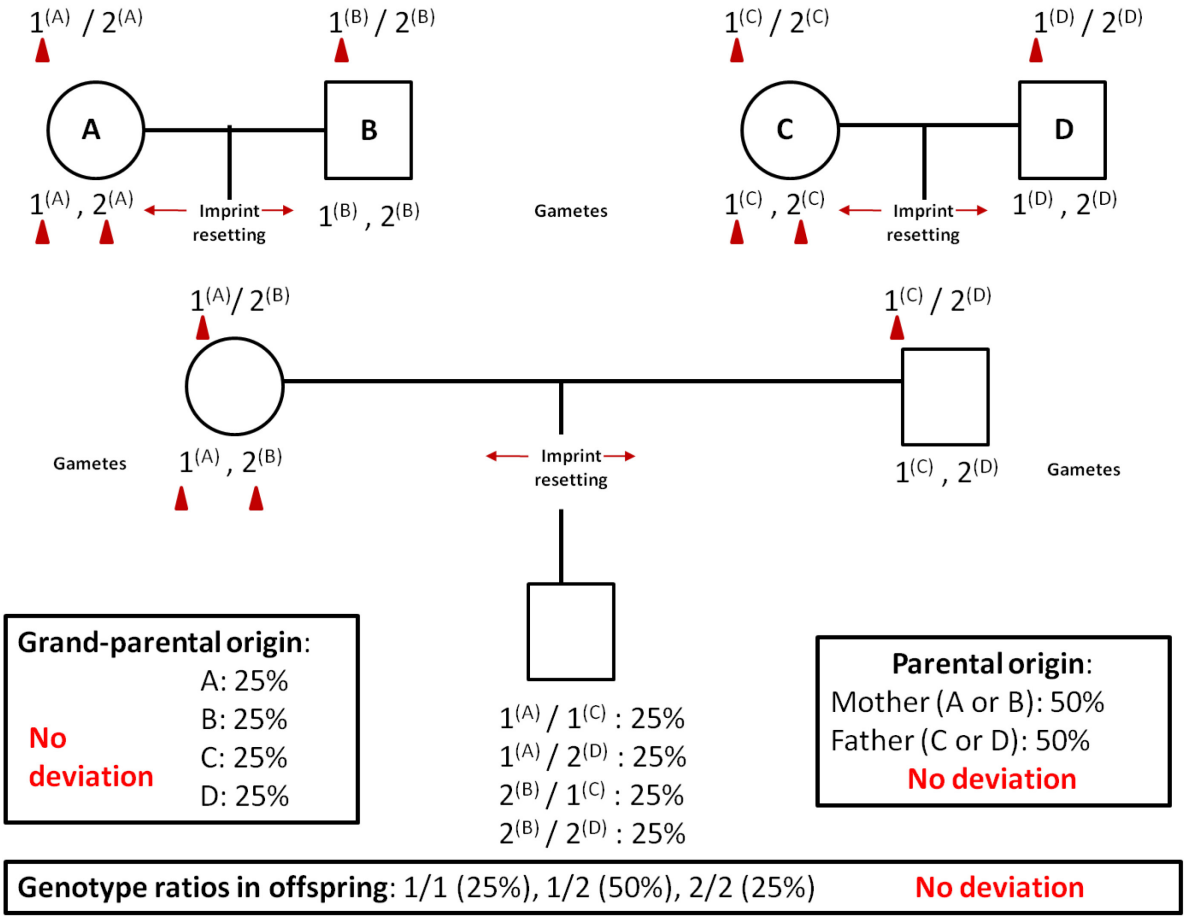


Figure 3.4.2: Example of a three-generation family with imprint resetting error at allele 2 in mother. Same scenario as in Figure 3.4.1, an imprint resetting error occurred in the mother, which is incompatible for embryonic survival.

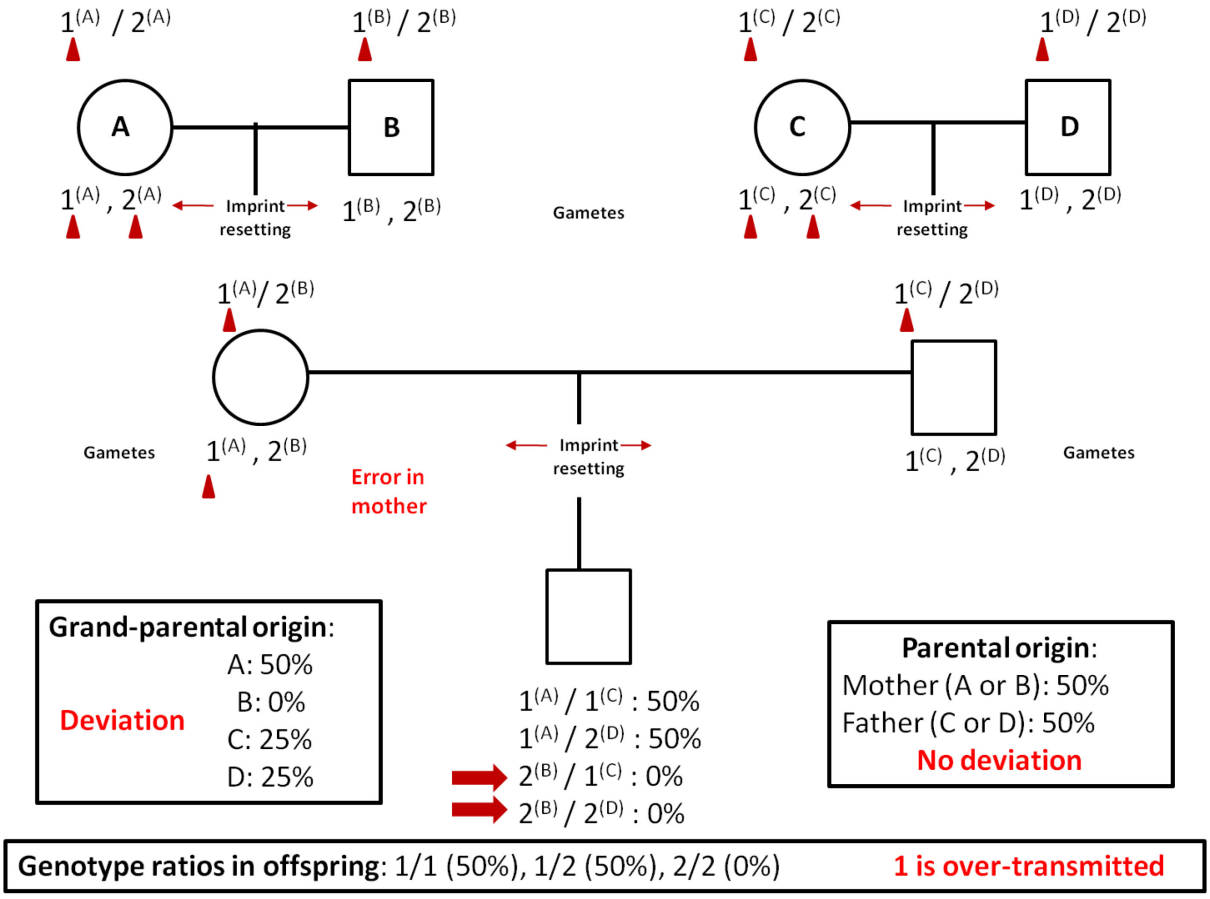
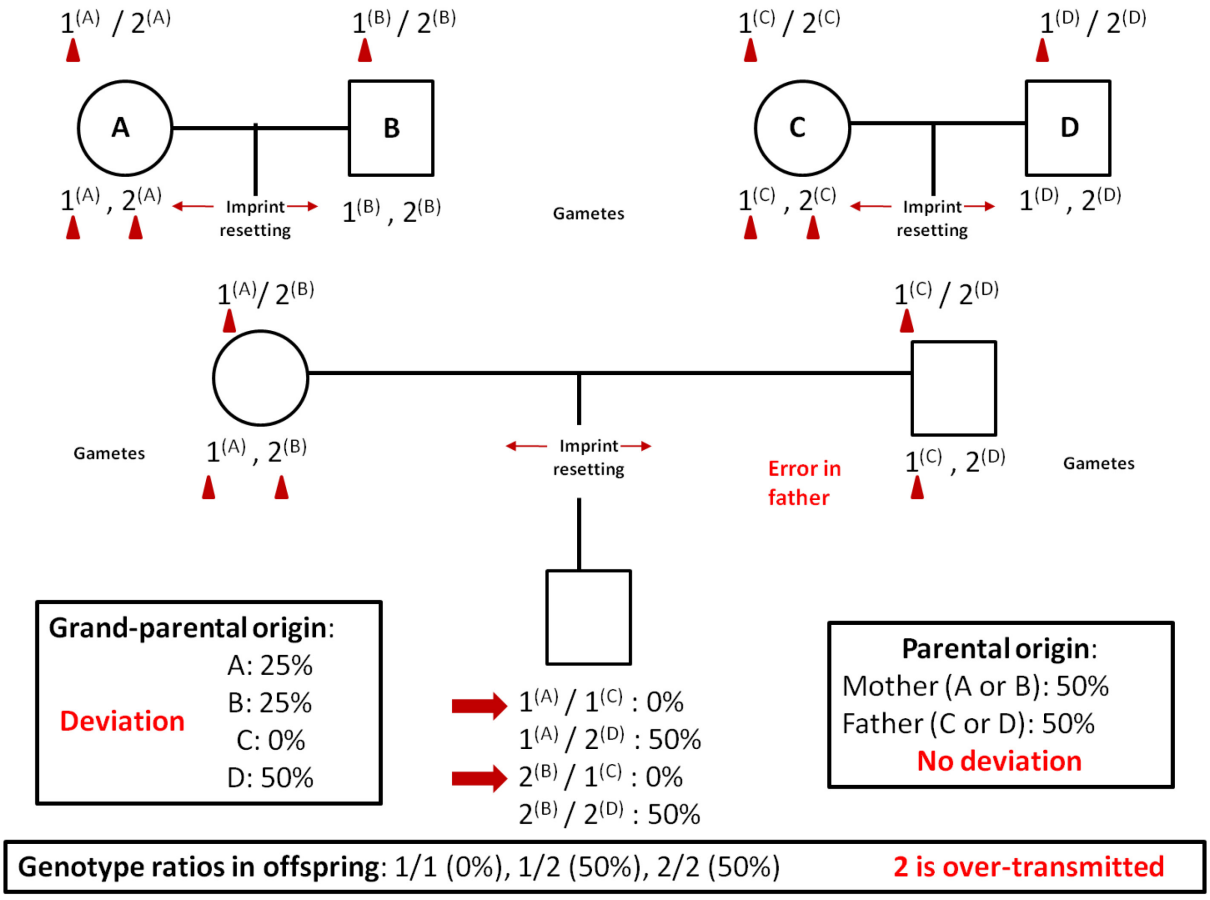


Figure 3.4.3: Example of a three-generation family with imprint resetting error at allele 1 in father. Same example as in Figure 3.4.1, an imprint resetting error occurred in the father, which is incompatible for embryonic survival.



Two analytical strategies have been proposed in the literature to determine the grandparental origin of TRD. First, a simple binomial test can be used by determining if the proportions of grandpaternal alleles and grandmaternal alleles are equal in the offspring's genotypes for a given marker. In practice, TRD is estimated by the proportion of grandmaternal alleles transmitted to the offspring [71, 80, 90]. The method of maximum likelihood [110] can be used to estimate TRD in the presence of missing genotypes, by using neighboring flanking markers as well as map distances [108]. In cases where embryo lethality due to imprinting error occurs in a sex-of-offspring specific manner, TRD can also be estimated by using a logistic regression model predicting grandparental source (dichotomous outcome), where variables such as sex of offspring and mating type of parents are included in the model [80]. In Yang et al.'s paper [80], grandparental-origin TRD locus was inferred on the basis of genotypes of the closest microsatellite markers. For non-informative markers, it was inferred on the basis of the grandparental origin of the flanking markers.

3.6 TRD empirical findings in previous literature

Several studies using some of the designs and methods reviewed above successfully uncovered numerous TRD loci in human (Table 3.1) and mouse models (Table 3.2). Mouse studies have been an incentive for much of the research on TRD in humans. The preferential transmission of the t-haplotype on the segregation distorter gene of the t-complex region on Chromosome 17 is a well-studied TRD example, and it has puzzled scientists for decades [86]. TRD influence on sperm motility due to t-complex transmission distortion has been reported [91, 92, 100]. Two studies have also shown TRD locus on Chromosome 7 that affects imprinting [99, 102], the latter being associated with a Robertsonian translocation. Another study has investigated the phenomenon of embryo lethality due to TRD [101] on the Ovum mutant (*Om*) gene. Developmental disabilities have been associated with TRD loci on Chromosomes 2 [97] and 7 [102]. A few studies have found a TRD influence on loci associated with diseases such as Cystic Fibrosis [94] and limb-girdle muscular dystrophy Type 2A [98]. TRD have also been detected on the *SPAM1* gene, which is believed to influence reduced transcript sharing of spermatids during male meiosis [96]. Casellas et al. [93] used Bayesian binomial model in search of TRD loci in the mouse and has found multiple loci on Chromsomes 1, 2, 3, 5, 12, 13, and 14, although none was mapped to specific gene regions.

Many of the reported TRD loci in human studies play a role in tumour suppression and have been found in genes associated with colon cancer, leukemia, bladder cancer, intestinal adenoma, node-positive breast cancer and other cancers [73, 90, 111, 112]. A number of TRD loci are within gene regions responsible for imprinting [80, 90, 109, 111], such as *DNMT1* on chromosome 12 and *H19* on 11p15.5, leading to loss of imprint and embryonic lethality. Abnormal methylation during imprint resetting on (i) imprinting centre (IC) genes which regulate the expression of imprinted gene, such as ICs *H19* and *KNCQ1OT1*, and on (ii) imprinted gene *CDKN1C*, has been linked to embryo lethality mechanisms which result in spontaneous abortion [109]. The result on imprint region *H19* was also a replication of a previous study [90]. A more recent study also found a region on Chromosome 1 that is responsible for infertility and recurrent pregnancy loss, but is not mapped to any specific SNP [113].

Many loci with observed TRD in humans are also linked to autoimmunity functions, located on the Major Histocompatibility complex (MHC) region on chromosome 6 [88, 90], the absence of which can progress to autoimmune diseases such as Type 1 diabetes, rheumatoid arthritis, or other diseases. It is also worth noting that the TRD finding on *INS/IGF2* gene region has been replicated in three studies [90, 111, 112]. Two studies have also uncovered TRD on the short arm of human chromosome 6 in the region of the transcription factor-encoding genes *SUPT3H* and *RUNX2*, as well as the microRNA locus MIRN586, with one SNP (rs12199720) included in both studies showing statistically significant results [74, 77]. This is interesting as *RUNX2* in particular is involved neoplastic development in hematopoietic lineages [114]. There are also many TRD loci that are linked to abnormal development in neurogenesis, neuronal differentiation, and other cognitive function in central and peripheral nervous system [71, 73, 82, 90, 111, 115-117].

In assessing the quality of the TRD findings in their study, Paterson et al. [73] speculate about the SNPs found to have excessive transmission of major alleles; previous studies have shown that when these alleles have a low Minor Allele Frequency (MAF), it may indicate genotyping error. However, no such observations apply to SNPs with excessive transmission of minor alleles. The 8 SNPs in this study (included in Table 3.1) that were found to have excessive transmission of minor alleles were shown to have good genotyping quality as well as significant TDT p-value.

Meyer et al. [77] also discussed the possibility of genotyping error. Among the three datasets they analyzed, results from Framingham Heart Study and Hutterite of European ancestry remain inconclusive. However, for the Austism Genetic Resource Exchange (AGRE) dataset, many signals extended across multiple SNPs, which is unlikely a result of genotyping error. In Naumova et al. [90], in order to eliminate the possibility of genotyping error, datasets from different labs were used to validate the results. Paterson and Petronis [115], Hanchard et al. [88] have raised the possibility of genotyping error, but did not specifically address the issue.

Mitchell et al. [118] investigated some studies that used TDT-derived association statistics, and found that genotyping error can lead to false inflation of such statistics if, for example, a number of homozygous parents are miscalled as heterozygous. However, genotyping error is more of a concern with genome-wide scan because a large number of SNPs are genotyped at the same time. Quality control normally needs to be in place to filter out SNPs inconsistent with Hardy-Weinberg Equilibrium or having low MAF, which should be applied with caution because these features are expected for loci exhibiting TRD. The majority of the studies we included, with the exception of Meyer et al. [77], Paterson et al. [72, 119], Paterson and Petronis [115], were candidate gene analyses, and therefore less prone to genotyping error. Furthermore, we have seen replications of a number of gene regions exhibiting TRD across multiple studies.

Paterson and Petronis [115], found evidence showing the association between some loci on chromosome 10 and schizophrenia, as well as bipolar disorder. There are several loci on chromosome 19q13 showing evidence of TRD which are associated with the severity of cystic fibrosis phenotype and endophenotype [120]. Three papers have shown multiple gene regions under TRD influence that are associated with Type 1 and Type 2 long-QT syndrome [121] and human muscular dystrophy [81, 122], with the gene *DMPK* replicated in the last two studies. One paper exclusively studied the *SMN1* gene, which is associated with human spinal muscular atrophy, and found significant evidence of TRD [123].There are a few other TRD loci that are linked to blood coagulation and insulin regulation [76, 111, 112]. Three studies found TRD influence in regions on Chromosome 2 [124] and 10 [125, 126], that are linked to Inflammatory Bowel Diseases; the latter two studies each identified the gene *DLG5*. There is one study that found a TRD region on Chromosome 17 that is linked to bone deficiency which express itself as Split-

hand/foot malformation (SHFM), and SHFM with long bone deficiency (SHFLD) which is a congenital disorder characterized by severe malformation of the distal limbs [127].

In Table 3.1 and 3.2, effect sizes are shown as reported in the studies. They were estimated using a variety of measures, such as TRD ratio, odds ratio, relative risk, NPL score, and grandmaternal allele transmission ratio. In this paper, we define TRD ratio to be the proportion of the preferred allele transmission counts among all transmission counts from parents to offspring at a specific locus. For example, if it is three times more likely to transmit advantaged over disadvantaged allele, the TRD ratio is 3/(3+1) = 0.75. The TRD ratios of the advantaged allele over all alleles found in most studies are within the range of 0.3 to 0.6. There are a few exceptions [82, 109], which show a more extreme skew. Grandmaternal allele transmission ratios for the two grandparental-origin-dependent TRD studies are between 0.4 and 0.65 [80, 90], which also represent mild distortions. Two analyses of TRD from HapMap data are not included in the table, as the list of genes and SNP reported is quite extensive. The first report [87] shows more extreme skewness in the ratio than the ones in Table 3.1, up to greater than 0.9 in both YRI and CEU population with p-values less than 10^{-4}. A later analysis searching for TRD from approximately 630,000 HapMap SNPs [128] reports 1,205 transmission outliers (based on Fisher's exact test) in 224 candidate genes, although results have not been adjusted with the Bonferroni correction. However, results from the permutation tests reached significance level. In this study, as well as the previous ones, genes with TRD signals were found on a substantial number of biological pathways, including in particular the protein phosphorylation pathway.

Table 3.1: Transmission Ratio Distortion findings in current literature of human studies

First Author	Study population (analytical method)	Gene	Gene Location	Effect size	p-value	Function of genes
Klopocki (2012)	Pedigrees with affected members (simple ratio)	*BHLHA9*	17p13.3	r^a=0.3(12/40)^f r^a=0.7(30/42)^m	-	Split-hand/foot malformation (SHFM), SHFM and long bone

						deficiency (SHFLD)
Meyer (2012)	Trios from: three generation Framingham Heart Study; Hutterite families from South Dakota; families from Autism Genetic Resource Exchange project (TDT)	(partial results- see Table 2 in article for full results) SNP: rs12199720 (MAF=0.45) *SUPT3H-MIRN586-RUNX2*	6p21.1	r^a =0.593 [f]	1.77e-05[mt-]	Human growth, osteoblastic differentiation, skeletal morphogenesis, height, cleft palate, neoplastic development in hematopoietic lineages
		SNP: rs748001 (MAF=0.36)	Chr 10	r^a =0.585	4.55e-08 [mt-]	-
Honeywell (2012)	One five-generation family of carriers and non-carriers of pericentric chromosome inversion (ratio of miscarriages)	-	Region between 1p36.21 & 1q42.13	r^g=0.15(13/88) r^a=0.71(15/21)	-	Infertility, recurrent pregnancy loss, higher risk for congenital abnormalities in offspring
Shoubridge (2012)	39 multi-generation families with affected and unaffected individuals (Pearson's $\chi 2$ test)	*ARX*	Xp21.3	r^a =0.6(149/247)[m]	0.002	Non-syndromic intellectual disability, infantile spasms or serious brain malformations
Liu (2012)	HapMap phase 3 trios	*ATG16L1* (SNP:rs3792106, MAF=0.48)	2q37.1	r^a =15/38[m]	0.19	Inflammatory Bowel Diseases
				r^a =21/32[f]	0.077	
				r^a =13/40[mo]	0.027	

	(Pearson's $\chi 2$ test)			r^{a} =23/40^{fa}	0.34	
		LRP2 (rs6733122)	Chr2	r^{a}= 0.65(228/353)	0.029$^{mo;mt+}$	Donnai-Barrow syndrome (DBS) and facio-oculoacoustico-renal syndrome (FOAR)
		ZNF133 (rs926716)	Chr20	r^{a}= 0.37(176/473)	0.018 $^{mo;mt+}$	osteoblastoma
Henckaerts (2010)	DZ twin pairs with 1 died in uterus (ratio of genotype)	*DLG5*	10q23	alive: $r^{cc:ct:tt}$= 0.78:0.2:0.02 (32:8:1) dead: $r^{cc:ct:tt}$= 0.8:0.16:0.04 (56:11:3)	-	Inflammatory Bowel Diseases
Santos (2009)	Hap Map YRI and CEPH trios on Chr 6 (TDT)	*SUPT3H-MIRN586-RUNX2* rs6899845 (MAF=0.457)	6p21.1	r^{a}= 0.94(16/17)^{fa} r^{a}= 0.64(7/11)^{ma}	3.0e-04$^{fa,mt+}$ 0.0233$^{ma,mt+}$	Human growth, osteoblastic differentiation, skeletal morphogenesis, height, cleft palate, neoplastic development in hematopoietic lineages
		rs2677101 (MAF=0.45)	6p21.1	r^{a}= 0.94(17/18)^{fa}	2.0e-04 $^{mt+}$	
Paterson (2009)	Multi-generation families unselected for	Intergenic *NBPF8, HFE2*	1q21.1	OR=0.58	7.7e-06 $^{mt-}$	Cognitive development and tumour suppressor,

	phenotype in Framingham Heart Study (NPL, TDT)					iron metabolism
		TMEM37 intron I	2q14.2	OR=0.47	1.0e-06 [mt-]	Resistance to pathogens
		SAG intron 6	2q37.1	OR=0.49	7.4e-10 [mt-]	Night blindness in Oguchi disease
		MEGF10 Intron 6	5q33	OR=0.35	8.2e-07 [mt-]	Brain functions
		SPOCK1 intron 2	5q31	OR=0.36	1.4e-06 [mt-]	Unknown
		C9orf3 intron 5,	9q22.32	OR=0.45	2.4-07 [mt-]	Lipid, apolipoprotein
		DBC1, CDK5RAP2, MEGF9	9q32-q33, 9q33.2, 9q32-q33.2	OR=1.36	3.7e-06 [mt-]	Bladder cancer, neuronal differentiation, central and peripheral nervous system
		CTDP1 intron 4	18q23	OR=0.75	1.8e-06 [mt-]	Congenital cataract, facial dysmorphism, peripheral neuropathy
Bettencourt (2008)	102 Sib-pairs with parents of normal families (Pearson's $\chi 2$ test)	*ATXN3*	14q32.1	r^a =0.569 r^a =0.581[fa] r^a =0.557[mo]	0.013 0.04 >0.05	Machado-Joseph disease (MJD), also known as Spinocerebellar ataxia type 3 (SCA3)
Sazenova (2008)	Tissues from 84 spontaneous	*H19*	11p15.5	r^a =0	-	Imprinting centre control

	abortions from women					synthesis of IGF2
		CDKN1C	11p15.5	r^a =0	-	Tumour suppressor
		KNCQ1OT1	11p15.5	r^g =0.095	-	Imprinting centre control activation of imprinted genes including CDKN1C
Yang (2008)	Three-generation mouse families and CEPH families (Binomial exact test, logistic regression)	*DNMT1* (D12Nds2 ^{ms})	Chr12	r^a = 0.17 ^{cohort1}	0.068^{mt-} 0.016 $^{mt-}$	Imprinted region
Becker (2007)	37 Nuclear family with affected twins or siblings, and discordant sibs (HAP-PAT test)	rs1982073 (*TGFβ1*, MAF=0.445)	19q13	r^a = 0.33 ^{cohort2}	0.000145^{mt+}	Control severity phenotype and endophenotype of cystic fibrosis (CF)
		rs1800469 (*TGFβ1*, MAF=0.359)	19q13	-	-	
		D19S112 (DMPK)	19q13	-	0.0304^{mt+}	
De Rango (2007)	Concordant and discordant cousin-pairs with centenarian parents (likelihood ratio test)	*TNFb* and *TNFa*	6p21.3	-	0.007^{d} $^{,mt-}$, 0.06^{c} $^{,mt-}$	Tumour necrosis (death)
		HSP70.1	6p21.3	-		Graft-vs-host-disease
		SIRT3	11p15.5	-	0.015$^{d,mt-}$, 0.0396$^{c,mt-}$	Node-positive breast cancer
		HRAS1	11p15.5	-		Oncogene

		IGF2	11p15.5	-		Intestinal adenoma (tumour)
		INS	11p15.5	-		Hyperinsulinism (above normal insulin level)
		TH	11p15.5	-		Neuropathology
		-	14q32	r^a=0.469; 0.539^f; 0.401^m	-	Imprinted region in human
Friedrichs (2006)	Case-trios and control-trios (multivariate logistic regression)	*DLG5* (rs1248696) MAF=0.042	10	OR^e=1.75 OR = 1.52 OR = 2.49^m OR = 1.01^f	0.025 0.021 $^{mt-}$ <0.001 $^{mt-}$ 0.979 $^{mt-}$	Inflammatory bowel disease (IBD)
Imboden (2006)	Nuclear family with carrier parents of Type 1 and Type 2 long-QT syndrome (Pearson's $\chi 2$ test)	*KCNQ1*	11p15.5	r^a =0.57 r^a =0.59 ^{f} r^a =0.54 ^{m}	<0.001 $^{mt+}$ <0.001 $^{mt+}$ >0.05 $^{mt+}$	Long-QT syndrome Type 1
		KCNH2	7q36.1	r^a =0.57 r^a =0.60 ^{f} r^a =0.53 ^{m} $r^{f \text{ vs } m}$=0.57	0.001 $^{mt+}$ <0.001 $^{mt+}$ >0.05 $^{mt+}$ 0.02 $^{mt+}$	Long-QT syndrome Type 2
Dean (2006)	335 preimplantation embryo selected on heterozygosity of parents (Binomial exact test)	*DMPK*	19q13.3	r^a=0.59 r^a =0.6^{mo} r^a =0.59^{fa} r^a =0.55^m r^a =0.65^f	0.0004^{mt-} 0.0055^{mt-} 0.03^{mt-} 0.2^{mt-} 0.0001^{mt-}	Human muscular dystrophy
Hanchard (2006)	Trios unselected for phenotype (Pearson's $\chi 2$ test)	*CLIC-2230* (in central *MHC*)	6p21.3	r^a=0.6(70/116)	0.025 $^{mt-}$	Autoimmunity, regulation of cellular processes
Botta (2005)	Trios of fetus with	*SMN1*	5q13.2	r^a= 0.45(284/628)	0.016	Spinal muscular

	heterozygous carrier parents(Pearson 's $\chi 2$ test)					atrophy (SMA)
Infante-Rivard (2005)	Case-trios and control-trios of unaffected newborns (TDT)	*MTHFR*	1p36.3	RR=0.73	<0.005 [mt-]	occlusive vascular disease, neural tube defects, colon cancer and acute leukemia
		Factor V Leiden	1q23	RR=0.38	<0.002 [mt-]	blood coagulation cascade, hemorrhagic diathesis, thrombophilia
		Factor II (prothrombin)	11p11	RR=0.24	<0.001 [mt-]	blood coagulation cascade, maintain vascular integrity during development and postnatal life, thrombosis and dyspro-thrombinemia
Paterson (2003)	Two-generation Framingham Heart Study families (multipoint NPL LOD score)	-	Chr2 cM200	NPL=1.9 [m]	0.0011 [mt-]	-
		-	Chr4 cM168	NPL=1.86 [m]	0.0013 [mt-]	-
		-	Chr10 cM14	NPL=2.05	7.5e-04 [mt-]	

		-	Chr17 cM65	NPL=1.82[f], 0.59 [m]	0.0017 [mt-], 0.037 [mt-]	-
		-	Chr17 cM86	NPL=0.61[f], 1.77 [m]	0.0420 [mt-], 0.0016 [mt-]	-
		-	Chr20 cM96	NPL=1.10	0.0087 [mt-]	-
		-	Chr22 cM41	NPL=1.75	0.0016 [mt-]	-
Naumova (2001)	Three-generation CEPH families (Exact binomial test)	*IGF2*	11p15.5	Tgm[b]=0.62[m] Tgm[b]=0.50[f]	-	Intestinal adenoma (tumour)
		H19	11p15.5			Loss of imprinting of IGF2
		MASH2 (ASCL2)	11p15.5			Neuronal precursor for central and peripheral nervous system
		IGFR2 (FCGR2B)	6q25-q27	Tgm[b]=0.6[m] Tgm[b]=0.59[f]	-	Autoimmune disease
Paterson (1999)	Two and three-generation CEPH families (Multipoint NPL)	-	10p11-p15	NPL=1.84	0.04 [mt-]	Chromosome 10 was known to be associated with schizophrenia, bipolar affective disorder, obesity, Type 1 diabetes and alcoholism
Eaves (1999)	Nuclear family with children	*IGF2*	11p15.5	r[a]=0.54	0.002	Intestinal adenoma (tumour)

	unselected for disease (TDT)	*INS*	11p15.5			Hyperinsulinism (above normal insulin level)
Magee (1998)	Pedigree of affected and unaffected sib-pairs with parents	*DMPK*	19q13.3	r^a=0.63 r^a=0.583^{fa} r^a=0.687^{mo}	0.007 - 0.009	Myotonic dystrophy (DM)
Naumova (1998)	Three-generation CEPH families (Exact Binomial Test)	*DXS1068*	Xp11.4	Tgm^b =0.62 Tgm^b =0.52	0.0032^{mt+} 0.628 $^{mt+}$	Duchenne muscular dystrophy, Cognitive functions, Type1 Diabetes
Riess (1997)	Nuclear family of affected and unaffected offspring (Pearson's χ2 test)	*SCA1*	6p23	r^a =0.85	<0.05	spinocerebellar ataxia Type 1
		SCA3	14q24.3 –q31	r^a =0.62 r^a =0.73^{mo}	<0.05 <0.01	spinocerebellar ataxia Type 3

Table 3.2: Transmission Ratio Distortion findings in current literature of mouse studies

First Author	Study population (analytical method)	Gene	Gene Location	Effect size	p-value	Function of gene
Bauer (2012)	Wild type and mutant strains of mouse testis (Pearson's χ2 test)	*NME3* (distorter locus), t-complex *SMOK1* (responder locus)	Chr 17	NME^{mu}:0.352 vs control:0.27	0.0095	Sperm motility
				NME^{wt}:0.59 vs control:0.443	0.0006	
Casellas (2012)	Mouse crosses	rs3663003^{do,} ^{MAF*}	1	PM=0.358	4.3$^{PO,\ mt+}$	-

	(Bayesian Binomial Model)	rs3694780 [do, MAF*]	3	PM=0.330	3.9 [PO, mt+]	
		rs3698001 [do, MAF*]	12	PM=0.312	4.1 [PO, mt+]	
		rs3678616 [do, MAF*]	13	PM=0.331	4.5 [PO, mt+]	
		D14Mit44 [do, MAF*]	14	PM=0.562	29976,7 [PO, mt+]	
		rs13476816[ad, MAF*]	2	PM= -0.318	1.3 [PO, mt+]	
		rs6289734 [ad, MAF*]	3	PM= -0.193	21.7 [PO, mt+]	
		rs13482595 [ad, MAF*]	5	PM= -0.163	1.8 [PO, mt+]	
Eversley (2010)	Two-generation mouse families (Pearson's $\chi 2$ test)	rs8260829 [MAF*]	Chr 7	r^a=0.591	0.005 [mt-]	imprinted genes influencing fetal and placental growth, neurological disorder
		rs4228380 [MAF*]	Chr 10	r^a=0.317	3.0e-08 [mt-]	-
		rs3707772 [MAF*]	Chr 11	r^a=0.353	8.0e-06 [mt-]	-
Veron (2009)	Mouse sperm cells (Pearson's $\chi 2$ test)	*Tcd 1-4* (distorter locus) t-complex *SMOK1* (responder locus)	Chr 17	Tcd^{mu}:0.766	1.27e-14	Sperm motility
				Tcd^{wt}:0.555	0.789	
Haston (2007)	Cystic fibrosis Mouse	*D5Mit239*	Chr 5	$r^{wt/wt:wt/mu:mu/mu}$= 0.21:0.41:0.38[f, ncf]	5.7e-15	Cystic fibrosis
		D5Mit239	Chr 5	$r^{wt/wt:wt/mu:mu/mu}$=	0.035	

	crosses (Pearson's $\chi 2$ test)			0.12:0.6:0.27[m, cf]		
		DXMit16	Chr X	$r^{wt/wt:wt/mu:mu/mu}$= 0.31:0.27:0.42 [f, ncf]	3.0e-35	
Bauer (2007)	Wild type and mutant strains of mouse testis (Pearson's $\chi 2$ test)	*FGD2* (distorter locus) t-complex Tcr (responder locus)	Chr 17	$Fgd2^{mu}$:0.35 $Fgd2^{wt}$:0.47	- 0.01	Sperm motility
Schulz (2006)	Normal vs Robertsonian translocation crosses of mouse (Pearson's $\chi 2$ test)	(2.8) Robertsonian translocation	Chr 2	r^a=0.44 r^a=0.44[m] r^a=0.45[f]	0.0013 0.0093 0.0515	Developmental disabilities and mental retardation
Martin-DeLeon (2005)	Transgene vs wild-type crosses of mouse (Pearson's $\chi 2$ test)	*SPAM1*	Chr 6	r^a=0.67 (2/3)	<0.001	Transcript sharing of spermatids
Wu (2005)	Two-generation mouse families (Pearson's $\chi 2$ test)	*Ovum mutant (Om)*	Chr 11	r^a=0.561(198/353)	<0.05	Embryo lethality
Underkoffler (2005)	Normal vs Robertsonian translocation crosses of mouse (McNemar Test)	(7.18) Robertsonian translocation	Chr 7 and Chr 18	r^a=0.54[fa] r^a=0.46[mo]	0.02 0.02	Imprinting
Taveau (2004)	Wild type and mutant	*CAPN3*	Chr 2	$r^{wt/wt:wt/mu:mu/mu}$= 0.17:0.50:0.33	<0.01	Limb-girdle muscular

	crosses of mouse (Pearson's $\chi 2$ test)					dystrophy Type 2A

a, TRD ratio of transmission of minor allele vs all alleles

ad, additive model

b, Transmission ratio in grandmaternal alleles

c, concordant cousin pairs (De Rango 2008)

cc, CC genotype of the SNP

cf, Cystic Fibrosis lethal genotype

ct, CT genotype of the SNP

d, discordant cousin pairs (De Rango 2008)

do, dominant model

e, maternal vs. paternal transmission of risk allele, deviation from Mendelian ratio indicates parent-of-origin effect

f, female offspring

fa, father

g, ratio of miscarriage (due to embryo lethality)

m, male offspring

MAF, Minor allele frequency

MAF*, MAF for the SNP is not available

mo, mother

ms, microsatellite

mt+, adjusted for multiple testing

mt-, not adjusted for multiple testing

mu, mutant breed

ncf, non-Cystic Fibrosis lethal genotype

wt, wild-type breed

NPL, non-parametric linkage score

OR, odds ratio of transmitting the major allele

PM, posterior mean

PO, posterior odds: >100, decisive evidence, 10<PO<31.62, strong evidence, 3.16<PO<10, substantial evidence

RR, relative risk for the newborn genotype using the gene-dosage model

tt , TT genotype of the SNP

-, not available

3.7 TRD as a confounding signal in association or linkage analysis

The presence of TRD at a marker locus in the general population can influence the results of a linkage or association analysis in the affected population, by over- or under- estimating the true signal [72, 73, 75, 78]. As a result, it would be necessary to detect TRD as a confounding parameter in studies searching for disease. If TRD occurs at a distal locus from the disease susceptibility locus (DSL) and is not in LD with the DSL, a linkage or association signal would be detected, leading to a false positive signal. On the other hand, if TRD occurs on a locus in the vicinity of the DSL and is in LD with it, it would inflate or attenuate the linkage or association signal, potentially leading to a false positive or false negative signal.

Greenwood et al. [78] simulated linkage between a marker and disease loci, in a population of affected brother pairs. In this study, the marker locus is designed to be under influence of both TRD and linkage. The authors used a TRD ratio, defined as the ratio of a grandparental allele transmitted from the mother to a male child vs. all grandparental allele transmission, which is different from our definition of TRD used in this paper. The impact of these conditions was examined on three parameters: the TRD ratio on the X-linked marker locus, the relative risk of disease recurrence in an individual given an affected brother compared to the population prevalence, and the expected IBD sharing of alleles at the X-linked marker locus. Since the marker locus is X-linked, the maximum IBD sharing between affected brother pairs is 1. It was shown that

as TRD increases while relative risk remains the same, the expected allele sharing biased away from 0.5, giving a false positive signal. The results indicate that IBD sharing patterns for affected sib pairs are strongly affected by TRD and that the estimated statistical significance of a sib-pair linkage study may be extremely biased.

The same study also showed that the presence of TRD leads to significant power loss. Assuming a baseline of expected sharing due to TRD, the null and alternative hypotheses together are testing for additional expected sharing due to linkage. Therefore, expected sharing under the alternative hypothesis is always greater than or equal to that of the null hypothesis because of the additional sharing. When the baseline TRD ratio increases, expected allele sharing under both hypotheses increases as well. However, the difference between the expected sharing of null and alternative hypotheses decreases as they converge to a maximum sharing of 1. This then makes it more difficult to differentiate a true signal from a false one at higher values of TRD ratio. As such, Type 2 error increases and power decreases accordingly.

On the other hand, Spielman et al. [36] proposed to use a mixture of case trios (affected offspring with parents) and control trios (unaffected offspring with parents) to differentiate true linkage or association signals from false positives due to TRD by applying a TDT to both types of trios. The study concluded that 1) a statistically significant TDT in case trios suggests evidence of either linkage and association or TRD or both, 2) a statistically significant TDT in control trios suggests evidence of TRD or both TRD and linkage/association, 3) a statistically significant TDT in case trios but not in control trios suggests evidence of true linkage and association, and 4) when a statistically significant TDT is observed in both case trios and control trios, a significant Pearson Chi-square statistic of case trios vs. control trios transmission counts suggests evidence of true linkage and association.

To verify Spielman et al.'s [36] findings, we set up a simulation study for the 4 scenarios described in Table 3.3. The disease allele frequency (p) in the population was set between 0.01 and 0.05 indicating a rare to moderately rare disease frequency. The marker allele frequency (q) was set at 0.1 as a minor allele. The underlying TRD influence on the marker locus had a ratio between 0.6 and 0.9 for the minor allele, exploring mild to extreme skew of transmission. The recombination fraction between disease and marker loci (θ) was specified as 0.1 in the scenarios 3 and 4 when

there was linkage and association between disease and marker loci, or otherwise is set to 0.5 (scenario 1 and 2). A pre-specified linkage disequilibrium (LD) parameter (δ) was adjusted for each disease allele frequency being tested, to ensure positive haplotype frequencies, which depend on disease and marker allele frequencies. Therefore, LD was set to be slightly less than the minimum of p (1-q) and q (1-p) when there was linkage and association (scenario 3 and 4), and set to 0 otherwise (scenario 1 and 2). We simulated random mating in a population of 600,000 trios (parents and child) with the above specified parameters. Assuming a recessive mode of inheritance at the disease loci, we sampled 500 case trios and 500 control trios from the simulated population. We then applied the TDT at the marker, for both the case and control trios. As suggested by Spielman et al. [36], we further applied the Pearson's $\chi 2$ test to assess the excess transmission of minor allele over major allele in case trios vs control trios. This procedure was repeated 500 times, and the results of the test statistics were averaged over these 500 simulations. The p-values are computed accordingly using each of the averaged test statistics over 500 simulations. Our results support the proposal of study design, statistical method, and conclusions suggested by Spielman et al. [36], as shown in Table 3.3. This simulation study was repeated for a dominant mode of inheritance, and the same results were obtained.

Greenwood and Morgan [78] suggested that if TRD is suspected during the planning stage of a study, the planned sample size of the study needs to be increased by only a small amount to maintain the desired power to detect linkage. For example, it was shown in simulations that with an original sample size of 30 sib-pairs, when the TRD ratio increases from 0.5 to 0.62, then adding 11% sib-pairs will approximately guarantee the original power. When the TRD ratio is at 0.7, the sample size needs to be increased by one third to achieve the same desired power. Similarly, Evans et al. [129] carried out simulations to estimate the sample size required for various power level and Type 1 error level to detect transmission distortion in genome-wide studies using trios unselected for phenotype. They found that when distortion is small (TRD=0.51), one needs hundreds of thousand trios to achieve 80% power. However, for moderate value of TRD (0.7), only hundreds of trios are needed. They also showed that the number of trios decreases when the parental heterozygote frequency increases.

Table 3.3: Simulation results for 4 scenarios each averaged over 500 simulations based on TDT & Pearson's Chi-square test*

Presence of linkage and association	Presence of TRD	Significance of TDT in case-trios	Significance of TDT in control-trios	Significance of Pearson's Chi-square test of case-trios vs. control-trios transmission counts
No	No	No	No	No
No	Yes	Yes	Yes	No
Yes	No	Yes	No	Yes
Yes	Yes	Yes	Yes	Yes

*Methods referenced in Spielman et al.(1993)

3.8 TRD from a population genetics perspective

The impact of TRD at the organismal level could become manifest at the population level as the human genome evolves over time. Therefore, TRD is also a main study objective in a population genetics context because this genetic force leads to changes in the diversity of the population gene pool over generations. By using the formulae in Chevin and Hospital [83], we set up a simulation study designed to trace the marker allele frequency and LD between marker and disease loci over generations. First we defined marker allele frequency to be the MAF at the marker locus. Disease allele frequency was set to be rare. Recombination fraction and LD were specified accordingly to indicate linkage and association. In equation 1 of Chevin and Hospital [83], the change in marker allele frequency in i-th generation is a function of TRD ratio and marker allele frequency for the (i-1)-th generation, and as such, the marker allele frequency increases over each generation. LD in the i-th generation is a function of TRD ratio, recombination fraction, LD of the (i-1)-th generation, and marker allele frequency at the (i-1)-th generation as seen in equation 8 of Chevin and Hospital [83]. We simulated this change in LD and marker allele frequency for many generations and over time, LD decays and marker allele frequency eventually reached fixation with a frequency of 1 in the population. For a TRD ratio of 0.9, fixation can be reached in about 10 generations. As for a TRD ratio of 0.6, it can take up to 80 generations to reach fixation, depending on the strength of linkage and association between marker and disease loci. These

changes in genetic diversity over time culminate to an equilibrium state of involved parameters in the population, namely the MAF and haplotype frequency at TRD and neighboring loci, and LD between marker and disease loci [84].

As we have seen, TRD can be detected within two or three generations by observing transmission patterns from parents and grandparents to offspring. If TRD is persistent through many generations, a gradual shift in the allele frequency at the TRD locus would be observed. Over time, the advantaged allele(s) could become fixed in the population, while the alternatives are completely eliminated. This may provide an explanation as to why studies have been able to discover only a small number of TRD loci, because alleles at some of these TRD loci have already become monomorphic. Therefore, no genetic variation could be detected in the population on these "disappeared" TRD loci. However, through observation on some other identified TRD loci, disadvantaged alleles still appear to exist at a low frequency and remain polymorphic as rare variants. This raises questions as to why TRD did not sweep the advantaged allele into fixation. Several authors have tried to answer this question by suggesting theories on sources of counter-balancing forces which keep the allele in polymorphic state, such as recombination which breaks up linkage between distortion driver and responder genes [130], mutation and genetic drift acting in the opposite direction of the TRD [131], and an immunogenetic advantage for survival in later adulthood regardless of low fertility of the disadvantaged genotypes [132].

The existence of these rare variants provides us with great insight into the understanding of TRD and the importance of corresponding gene functions at these loci. Rare disease variants are currently the focus of genome-wide association studies in search of missing heritability in complex disorders [133]. It has been hypothesized that rare disease variants could be more functional than common variants and have high penetrance [134-136]. This suggests a potentially similar role for disadvantaged TRD rare variants when their gene functions determine survival. Since there is usually low power to detect rare variants using a standard genome-wide genotyping platform with feasible sample size, there are intense ongoing research efforts to address this issue [137, 138]. These efforts should lead to a better understanding of TRD and its contribution to the rare variant phenomenon itself.

3.9 Conclusion

In conclusion, TRD is a complex and understudied area with challenges such as access to very large and error-free genotype databases with unselected phenotypes. Recent sequencing studies have included unaffected subjects as well as affected subjects. Moreover, with the change of focus back to family-based studies, these data may be conveniently used to the study of TRD. As discussed, TRD is a phenomenon with potential impact on practical aspects of human genetics such as correct interpretation of association study results, as well as more theoretical ones, such as frequency of variants and related population genetics issues. This review aimed at underscoring the importance and interest of TRD in human genetics.

Ethical standards: The study complies with the current laws of Canada.

Conflict of interest: The authors declare that they have no conflict of interests.

Chapter 4

Adjusting for Transmission Ratio Distortion in the analysis of case-parent trios using a loglinear model

4.1 Preamble

Transmission Ratio Distortion (TRD) has been captured statistically in various family-based study designs using control-trios, or child unselected for phenotypes as enlisted in Tables 3.1 and 3.2 of Chapter 3, for human and mouse studies, respectively. The detection of deviation from Mendelian inheritance in apparently unaffected individuals indicates the potential presence of TRD. Methods to detect TRD include the TDT, the Binomial Exact Test, the Pearson's Chi-square test, the multipoint non-parametric linkage test, the Mann-Whitney U test and the multivariate logistic model. However, all of these methods only provide a p-value for the significance of the TRD signal, without any mean to adjust for it. We utilized the loglinear model framework developed by Weinberg et al. [44] and extend it to adjust for non-sex-of-parent-specific TRD (NST). This loglinear model not only provides a LRT p-value which measures association signal, but also offers RR estimates for child genotype 1 or 2. Assumptions of this model include Mendelian transmission and mating symmetry, but not HWE or random mating.

We proposed to take an existing component, *P[C|MF]*, in the multinomial conditional probability of the loglinear model, and replace it with a category-specific offset based on the transmission probability of minor allele (t) . This probability is computed from control-trios. Our simulation showed that without adjusting for the presence of TRD, there is an inflation of RR and Type 1 error, and significant power loss. We also applied the extended model to a real dataset on IUGR dataset and identified 2 loci that were influenced by TRD, and recovered the correct significance level. Note that our method depends on the fact that control-trios are available for the computation of t. However, this might not always be feasible. Sensitivity analysis was conducted to test the effect of misspecification of t to the estimation of model parameters. The results showed that the validity of our conclusion is very sensitive to the misspecification. However, for a sample size of 500 control-trios, the 95%CI for the estimated t lies within ± 0.07 of true value of t, which does not lead to significant inflation in the parameter estimates.

Manuscript 2: Adjusting for Transmission Ratio Distortion in the analysis of case-parent trios using a loglinear model

Lam Opal Huang [1*], Claire Infante-Rivard [1], Aurélie Labbe [1-3]

[1] Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal Quebec, Canada, H3A 1A2

[2] Douglas Mental Health University Institute, Montréal, Quebec, Canada, H4H 1R3

[3] Department of Psychiatry, McGill University, Montréal Quebec, Canada, H3A 1A1

* Correspondence author

Contact information for correspondence author:

1020 Pine Avenue West, Montreal, Quebec H3A 1A2
Tel.: 514-398-6258

Fax: 514-398-4503

opal.huang@mail.mcgill.ca

Conflict of Interest

This work was supported in part by a research grant from Canadian Institutes of Health Research Operating Grant: PI: Dr. Aurélie Labbe; MOP-93723. The authors declare no conflict of interests.

4.2 Abstract

Transmission of the two parental alleles to offspring not following the Mendelian ratio has been termed Transmission Ratio Distortion (TRD). It is the result of mechanisms occurring during gametic and embryonic developmental stages. TRD has been well-studied in animal and plant models, but remains largely unknown in human studies. The Transmission Disequilibrium Test (TDT) was first proposed to test for association and linkage by estimating departure from the expected allele transmission proportions in families composed of an affected offspring and the two parents (case-trios); adjusting for possible TRD using control trios was recommended. However, the TDT does not provide parameter estimates for different genetic models. A loglinear model for association studies was later proposed providing relative risk (RR) estimates of disease for the child and maternal effects. This model assumes Mendelian transmission. Results from our simulation study showed that case-trios RR estimates using the loglinear model are biased in the presence of TRD. Power and Type 1 error are also compromised. In this paper, we propose an extended loglinear model including a separate component for TRD. Under this extended model, RR estimates, power and Type 1 error are correctly restored. We then applied this model to a real dataset on intrauterine growth restriction, and showed consistent results with a previously used approach that adjusted for TRD using control-trios. Our findings suggested the need to adjust for TRD to avoid spurious results in association studies. Documenting TRD in the population is therefore essential for the correct interpretation of genetic association studies.

4.3 Introduction

Transmission Ratio Distortion (TRD) occurs when the transmission of the two alleles from a heterozygous parent to the offspring violates the Mendelian law. TRD results from disruptive mechanisms occurring during the gametic or embryonic developmental stages [1], including germline selection [79], meiotic drive [70], gametic competition [75], embryo lethality [75], and imprint resetting error [80, 90]. The presence of TRD can lead to spurious conclusions in association studies.

Studies in animal models have contributed to our understanding of TRD using backcrosses [139] or F_2 crosses [93]. A recent study uses a Bayesian framework to model TRD in boars and piglets and was shown to achieve appealing statistical performance [140]. In humans, individuals unselected for phenotype have been studied to detect TRD in the general population, such as in the Framingham Heart study [73, 77], the Centre d'Etude du Polymorphisme Humain [80, 90], and the HapMap project [87].

In some studies both case and control populations were analyzed separately to detect a difference in transmission [117, 125]. For example, Spielman et al. [36] analyzed both case- and control-trios using the TDT. True association was assessed using a Pearson's Chi-square test. Deng and Chen [141] proposed a TDT statistic that is the sum of TDT statistics for case- and control-trios. Previously, we suggested a modified TDT statistics where the two diagonal counts in McNemar test are multiplied by t and (1-t), respectively, where t is the transmission ratio of the minor allele in control-trios [142].

Other statistical measures have also been proposed to study affected offspring, such as the Binomial exact test [80, 81], the Pearson's Chi-square test [116, 121], the multipoint non-parametric linkage (NPL) test [72, 115], the Mann-Whitney U test [111], and the multivariate logistic model [80]. These methods and TDT-type analyses only give statistical significance of linkage and association, but do not estimate the disease relative risk. Newer methods were proposed to address these limitations.

The family-based association test (FBAT) [143, 144] and likelihood methods that use case-trios to construct conditional logistic [46], unconditional logistic [45], and loglinear models [44, 50-52, 54] have also been used in family-based studies. In particular, Weinberg et al. proposed a loglinear model to detect an association between a marker and disease [44]. This model estimates a relative risk of disease for the offspring, and assumes Mendelian transmission. It has a probability component that can be easily extended to adjust for TRD. Our proposed method uses the transmission ratio of a minor allele in control trios, ideally obtained from an external dataset such as HapMap, to account for TRD through an offset parameter in the model. This transmission ratio likely varies across different populations because of the unique evolutionary history each population carries. However, the HapMap project offers control data on populations with different ethnicities and hence, can address this issue.

This extended loglinear model was validated through extensive simulation studies. It was also applied to an intrauterine growth restriction (IUGR) case-control study augmented with a case- and control-trio study [76, 145], investigating the role of thrombophilic genes in IUGR. The current literature in support of the association between thrombophilia and IUGR is inconsistent. We explored the possible role of TRD in these inconsistencies.

4.4 Material and Methods

We investigated the association between a bi-allelic disease susceptibility locus (DSL) and a disease. Assuming an additive model, we defined genotype by the number of copies of the minor allele. Therefore, homozygous wild-type individuals were coded as genotype 0, heterozygous as genotype 1, and homozygous mutant as genotype 2.

4.4.1 Loglinear model by Weinberg et al. (1998)

The loglinear model proposed by Weinberg et al. [44] assumes Mendelian transmission and mating symmetry, but makes no assumption about the Hardy-Weinberg Equilibrium (HWE). For the purpose of this paper, we considered the simpler form of this model where only parameters associated with the child genotypes are included.

In this model, the response variable is the number of trios (counts) for each of the 15 mother-father-child (MFC) genotype categories, as described in Table 4.1. These 15 categories can be subdivided into 6 mating types defined by the paired parental genotypes. Covariates entering the model include two indicator variables for child inheriting one or two disease alleles and five additional ones corresponding to the first five mating types. The model which includes an intercept and an offset parameter, is described as:

$$log\{E\ [n_{MFC}|D]\} = \rho_6 + \sum_{j=1}^{5} \rho_j I_{[S=j]} + log(2) I_{[MFC=111]} + \beta_1 I_{[C=1]} + \beta_2 I_{[C=2]} \qquad (4.1)$$

where M, F, and C represent the mother, father and child genotypes, respectively; n_{MFC} is the number of trios with genotypes MFC, and D is the disease status of the child. The $\rho_j + \rho_6$ terms (i = 1 to 5) are the regression coefficients for the first 5 parental mating types in Table 4.1; ρ_6 is the intercept corresponding to the 6th mating type MF=00; β_1 and β_2 are the regression coefficients for child genotype 1 and 2, respectively such that $\beta_1 = log\ (R_1)$ and $\beta_2 = log\ (R_2)$. R_1 and R_2 are the corresponding relative risks with respect to genotype 0. This model, which we call model 1, operates under the assumption of Mendelian transmission. The complete derivation of this model is shown in Chapter 4.7.1.

4.4.2 Loglinear model with adjustment for TRD

Without the assumption of Mendelian transmission at the DSL, model 1 can be generalized into:

$$log\{E\ [n_{MFC}|D]\} = \xi_6 + \sum_{j=1}^{5} \xi_j\ I_{[S=j]} + log\ \tau_{MFC} + \beta_1 I_{[C=1]} + \beta_2 I_{[C=2]} \qquad (4.2)$$

where τ_{MFC} is the transmission offset $P[C|MF]$, $\xi_j + \xi_6$ terms (i = 1 to 5) are the regression coefficients for the first 5 parental mating types in Table 4.1, and ξ_6 is the intercept corresponding to the 6th mating type. The coefficients β_1 and β_2 are as defined in model 1. This model, which accounts for TRD, is denoted as model 2 in the remaining of the paper, with derivation shown in Chapter 4.7.1.

The offset τ_{MFC} depends on the TRD ratio t, defined as the transmission probability of a minor allele from a heterozygous parent to the child. This leads to a different offset in each MFC genotype category, which corrects for TRD in that specific trio combination. The TRD parameter t can take

on values either greater than or less than 0.5. The value $t = 0.5$ corresponds to Mendelian transmission, in which case models 1 and 2 are equivalent (see Chapter 4.7.1).

We fitted both loglinear models (1) and (2) to obtain estimates of RR for child genotype 1 and 2, and their corresponding p-values using Z-tests. To assess significance of the association between the disease phenotype and the DSL, a Likelihood Ratio Test (LRT) was used. We refer to Chapter 4.7.2 for more details about the distribution of the LRT under the null and alternative hypotheses.

Table 4.1: Relative Risk, stratum frequency, and probability of transmission (TRD or Mendelian) for case-parent trios

Stratum	MFC Genotype	Stratum frequency under HWE	Probability of transmission (τ_{MFC})		Relative Risk
			TRD	Mendelian	
1	222	p^4	1	1	R_2
2	212	$2p^3(1-p)$	t	1/2	R_2
	211	$2p^3(1-p)$	1-t	1/2	R_1
	122	$2p^3(1-p)$	t	1/2	R_2
	121	$2p^3(1-p)$	1-t	1/2	R_1
3	201	$p^2(1-p)^2$	1	1	R_1
	21	$p^2(1-p)^2$	1	1	R_1
4	112	$4p^2(1-p)^2$	t^2	1/4	R_2
	111	$4p^2(1-p)^2$	2t(1-t)	1/2	R_1
	110	$4p^2(1-p)^2$	$(1-t)^2$	1/4	1
5	101	$2p(1-p)^3$	t	1/2	R_1
	100	$2p(1-p)^3$	1-t	1/2	1
	11	$2p(1-p)^3$	t	1/2	R_1
	10	$2p(1-p)^3$	1-t	1/2	1
6	0	$(1-p)^4$	1	1	1

4.4.3 Simulation study

In order to assess the performance of model 2 with respect to model 1, a simulation study was set up to generate different TRD scenarios. RR parameters, RR p-values, LRT p-values, Type 1 error, and power were compared between the 2 models, where the true t value was used in model 2. A sensitivity analysis was also carried out to test the impact on RR estimates and power when an incorrect parameter t is used in model 2.

4.4.3.1 Simulation setup

We considered a causal locus under study with no recombination. A random population of 100,000 trios was generated, from which 500 case trios were sampled. Parental genotypes at the DSL were generated under HWE assuming a minor allele frequency (MAF) = 0.1. A TRD parameter t was specified which varied between 0.1 and 0.9. Offspring were assigned to diseased or non-diseased phenotypes with penetrance factors f_0, f_1 and f_2 for homozygous wild type, heterozygous and homozygous mutant genotypes, respectively, and only the case-trios were sampled. Such penetrance values varied depending on the scenario studied. The simulation was repeated 100 times and averaged RR estimates, p-values of the averaged Z statistics for RR and p-values of the averaged LRT statistics are reported.

4.4.3.2 Measuring impact of TRD on association statistics

We compared the RR (95%CI) values and LRT p-values of both models under two main scenarios: (1) a common disease associated with a low penetrance disease allele at $f_0=0.1, f_1=0.11, f_2=0.15$, and (2) a rare disease with penetrance factors at $f_0=0.1, f_1=0.5, f_2=0.5$. In scenario (2), a dominant genotype model was assumed, and the estimated RR (noted as $R_{1/2}$) is for individuals carrying at least one disease allele, compared to individuals having genotype 0. To measure the inflation in RR and LRT p-values in model 1 when there is TRD but it is not modeled, we computed the log (base 10) ratio of RR and LRT p-values in model 1 with respect to model 2. We also varied f_2 fixing f_1=0.11 and varied f_1 fixing f_2=0.15 to describe the inflation of LRT p-values with respect to penetrance factors. To assess the inflation of Type 1 error, we set the penetrance factors to $f_0 = f_1 = f_2 = 0.1$ assuming no association while varying t from 0.1 to 0.9. Using sample sizes of 100, 300 and 500, we computed the Type 1 error of detecting a false association signal. Finally, we

evaluated the power of both models to detect a true association signal in the presence of TRD. In this case, we set $f_0 = 0.1, f_1 = 0.2, f_2 = 0.3$, varying t from 0.1 to 0.9 in the simulation, with sample sizes of 100, 300 and 500. Critical value for declaring significance was $\alpha = 0.05$.

4.4.3.3 Sensitivity analysis

The assumption in the simulation study was that the true value of t in Table 4.1 is known. In reality one might have an approximate idea of where t lies, or one can compute t in control-trios if they are available as part of the study or from major consortia such as the HapMap project. We performed a sensitivity analysis to examine the consequences of a misspecification of the TRD parameter t on the RR estimates and the power to detect true association. Therefore, we simulated three scenarios with true association signal, $f_0 = 0.1, f_1 = 0.2, f_2 = 0.3$, and true transmission ratio of the minor allele as $t=0.3$, $t=0.5$ and $t=0.7$. For each scenario, model 2 was fitted with the offset τ_{MFC} calculated using a selected t varying between 0.1 and 0.9. We then evaluated the log (base 10) ratio of RR obtained from model 2 using true t values vs selected t values that adjust for TRD. Power was also evaluated.

4.4.4 Application of models 1 and 2 to a real dataset

We applied our model to an intrauterine growth restriction (IUGR) case-control study augmented with a case- and control- trio study [146, 147], of which data were collected from a Canadian hospital between 1998 and 2000. The original study was intended to study the relationship between thrombophilia and IUGR. IUGR in this dataset is defined as birth weight less than the 10th percentile according to gestational age and sex, based on the national standards. The sample we used includes 493 case-trios and 472 control-trios with approximately 25% being black.

We examined six thrombophilic genes: Coagulation Factor XIII, A1 polypeptide (*F13A1* [MIM 134570]), Serpin peptidase inhibitor clade E member 1/Plasminogen activator inhibitor type 1 (*SERPINE1/PAI-1* [MIM 173360]), Methylenetetrahydrofolate reductase variant A1298C (*MTHFR A1298C* [MIM 607093]), Methylenetetrahydrofolate reductase variant C677T (*MTHFR C677T* [MIM 607093]), Coagulation Factor V (*F5* [MIM 612309]), and Coagulation Factor II (*F2* [MIM 176930]). The number of complete case-trios for *F13A1, PAI-1, MTHFR A1298C, MTHFR*

C677T, F5, and *F2* were 208, 176, 243, 246, 240, and 258, respectively. The number of complete control-trios for the same genes were 222, 153, 231, 217, 239, and 243, respectively.

We computed the MAF using all complete trios and t using control-trios. We compared our extended loglinear model 2 with another method proposed by Infante-Rivard and Weinberg [76] to assess and quantify the extent of TRD in the same IUGR population with the use of control-trios, specifically for *F5*. The difference between our model 2 and the model used in Infante-Rivard and Weinberg [76] is that the former inserts t as an offset in the loglinear model fitted with case-trios only, while the latter uses both case- and control-trios (12 strata) adding an interaction term between child genotype and case status to estimate RR in cases. Our approach has the advantage of not requiring the collection of control-trios sample. However, the model proposed by Infante-Rivard and Weinberg [76] remained a reliable validation for our results because it does not depend on the selected value of t.

4.5 Results

4.5.1 Simulation Study

4.5.1.1 Inflation of RR estimates

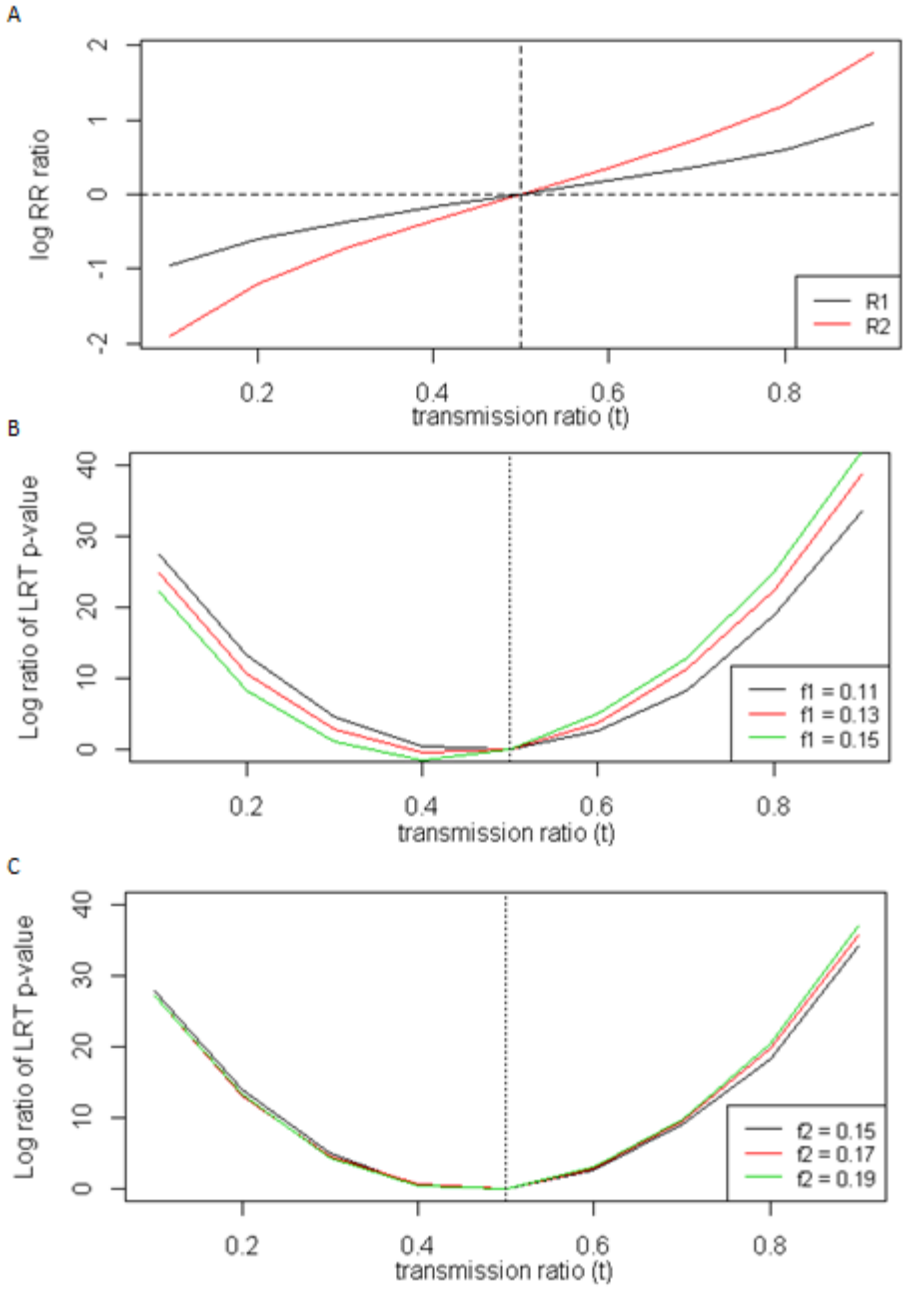
When the transmission ratio was Mendelian, i.e. $t=0.5$, models 1 and 2 yielded the exact same RR estimates and 95%CI as expected (Table 4.2), and were close to the ratios of the underlying penetrance factors f_1/f_0 and f_2/f_0. When testing $t=0.3$ where the disease allele is under-transmitted, the RR for model 1 was attenuated excluding 1 in the 95% CI, whereas RR estimates, p-value and LRT p-value were restored in model 2. Similarly, for $t=0.7$, the RR for model 1 were inflated whereas this false inflation in RR estimates, p-values and LRT p-values was removed under model 2. As seen in Figure 4.1A, the RR inflation ratio increased and decreased exponentially with respect to t, implying that even small deviation from $t = 0.5$ can lead to a substantial RR inflation. The slope of RR ratio for R_2 was double that of R_1, showing that inflation due to TRD affected R_2 much more severely than R_1.

Figure 4.1: Inflation on RR and LRT p-values from models 1 and 2

(A) Log ratio (base 10) of relative risk R_1 and R_2 for model 1 to model 2

(B) Log ratio (base 10) of LRT p-values for model 1 to model 2 when $f_2 = 0.15$

(C) Log ratio (base 10) of LRT p-values for model 1 to model 2 when $f_1 = 0.11$



4.5.1.2 Inflation of p-values

We see in Table 4.2 that when TRD is not adjusted for, the significance of the LRT p-value was inflated in either direction of deviation from $t = 0.5$. In Figure 4.1B, we observed that for $t< 0.5$, smaller f_1 leads to greater inflation, whereas it was the opposite for $t> 0.5$. This is because when $t < 0.5$, the false association signal is in the opposite direction of the disease effect, whereas when $t > 0.5$, they are in the same direction. However, in Figure 4.1C, this effect seemed to be less pronounced for varying f_2, as some of the lines were crossing each other as t changed. This is because R_2 had a relatively larger confidence interval than R_1. LRT p-value was less sensitive to changes in R_2.

Table 4.2: Relative Risk with 95% CI and p-values, and Likelihood Ratio Test p-values of models 1 and 2 when $t = 0.3, 0.5$ and 0.7 with population parameters:
(1) $p= 0.1, f_0 = 0.1, f_1 = 0.11, f_2 = 0.15$ for low penetrance common disease, and
(2) $p = 0.01, f_0 = 0.1, f_1 = 0.5, f_2 = 0.5$ for high penetrance rare disease.

Low penetrance common disease						
t	Model	R_1 (95%CI)	p-value	R_2 (95%CI)	p-value	LRT p-value
0.3	1	0.47 (0.33,0.65)	6.00E-06	0.25 (0.06,1.08)	0.07	2.85E-06
	2	1.09 (0.78,1.51)	0.59	1.34 (0.30,5.84)	0.51	0.28
0.5	1	1.10 (0.81,1.51)	0.53	1.40 (0.51,3.89)	0.43	0.26
	2	1.10 (0.81,1.51)	0.53	1.40 (0.51,3.89)	0.43	0.26
0.7	1	2.52 (1.78,3.57)	2.00E-07	8.01 (3.18,20.17)	8.27E-06	6.57E-10
	2	1.08 (0.76,1.53)	0.7	1.47 (0.58,3.70)	0.42	0.25
High penetrance rare disease						
t	Model	$R_{1/2}$ (95%CI)	p-value			LRT p-value
0.3	1	2.44 (1.20,4.94)	0.014			0.025
	2	5.71 (2.82,11.57)	1.29E-06			8.62E-07
0.5	1	5.58 (2.55,12.21)	1.55E-05			6.55E-07
	2	5.58 (2.55,12.21)	1.55E-05			6.55E-07
0.7	1	13.73 (4.99,37.79)	1.57E-07			2.62E-13
	2	5.87 (2.13,16.16)	0.000504			2.23E-05

Note: Models for high penetrance rare disease were fitted assuming a dominant genotype model and $R_{1/2}$ represents the RR of cases carrying 1 or 2 copies of disease allele.

4.5.1.3 Inflation of Type 1 error

Figure 4.2A shows the theoretical Type 1 error by computing LRT using a Non-Central Chi-square distribution, with a non-centrality parameter (NCP) calculated based on equation 4.8 and 4.9 derived in Chapter 4.7.2. Figure 4.2B shows the empirical Type 1 error we observed by fitting the loglinear model. The empirical results shown in Figure 4.2B are similar to our theoretical results in Figure 4.2A. Type 1 error of the TRD-adjusted model 2 remained the same across all t values (i.e. close to 0.05), and were exactly the same for all sample sizes. Therefore, NCP for model 2 does not depend on sample size or t, which means that this model is robust to the effect of TRD when the null hypothesis is true. In Figure 4.2A and 4.2B, we see that Type 1 error for the unadjusted model 1 increased as t deviated from 0.5 which led to a false inflation of the association signals.

4.5.1.4 Power loss

Relatively consistent results were obtained between theoretical power (Figure 4.3A) and empirical power (Figure 4.3B). Power for sample size n = 100 was poor in both Figures 4.3A and 3B, which was true even TRD was absent. We also noticed that model 2 gave relatively stable power for the most part in the range of t, while model 1 power suffered from the effect of TRD. However, when t was lower than 0.2 or greater than 0.5, model 1 power was greater than that of model 2. This is because a strong TRD actually inflates the power of detecting an association signal in either direction. Power for model 2 decreased slightly when $t > 0.7$, which suggested that the TRD offset overcompensates the inflation in power. However, a TRD ratio as large as 0.9 is rare, but even when $t = 0.8$, the power was still maintained around 0.8 for sample sizes of 300 and 500.

Documented results for TRD studies of human and mouse shown in Chapter 3 (Tables 3.1 and 3.2) mostly show minor allele transmission ratio between 0.3 and 0.8. The power for our model to detect association was still adequate for a t between 0.2 and 0.8, with a sample size greater than 300 case-trios, from a randomly generated population with MAF around 0.1. For rare variants, which is conventionally defined as frequency < 1% in the population, sample size has to be in the

thousands in order to achieve a similar level of power. A dominant model can also be used when mother-father-child genotype category counts with child genotype 2 are small.

Figure 4.2: Type 1 error plot of models 1 and 2 for sample size 100, 300, and 500 when there is no association between disease and DSL where $f_0 = f_1 = f_2 = 0.1$.

(A) Theoretical results from equations 4.8 and 4.9 in Chapter 4.7.2

(B) Empirical results from simulation

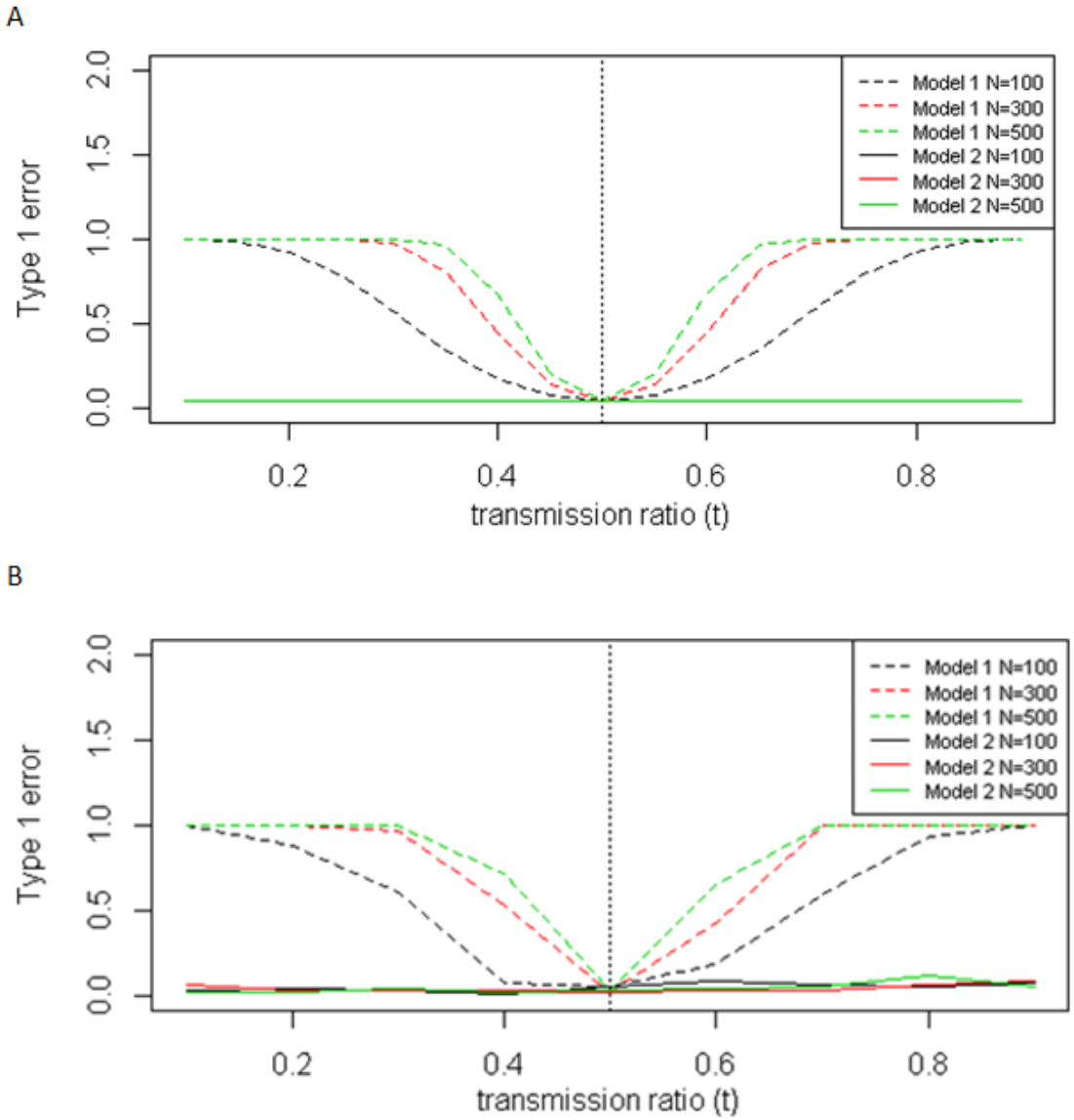
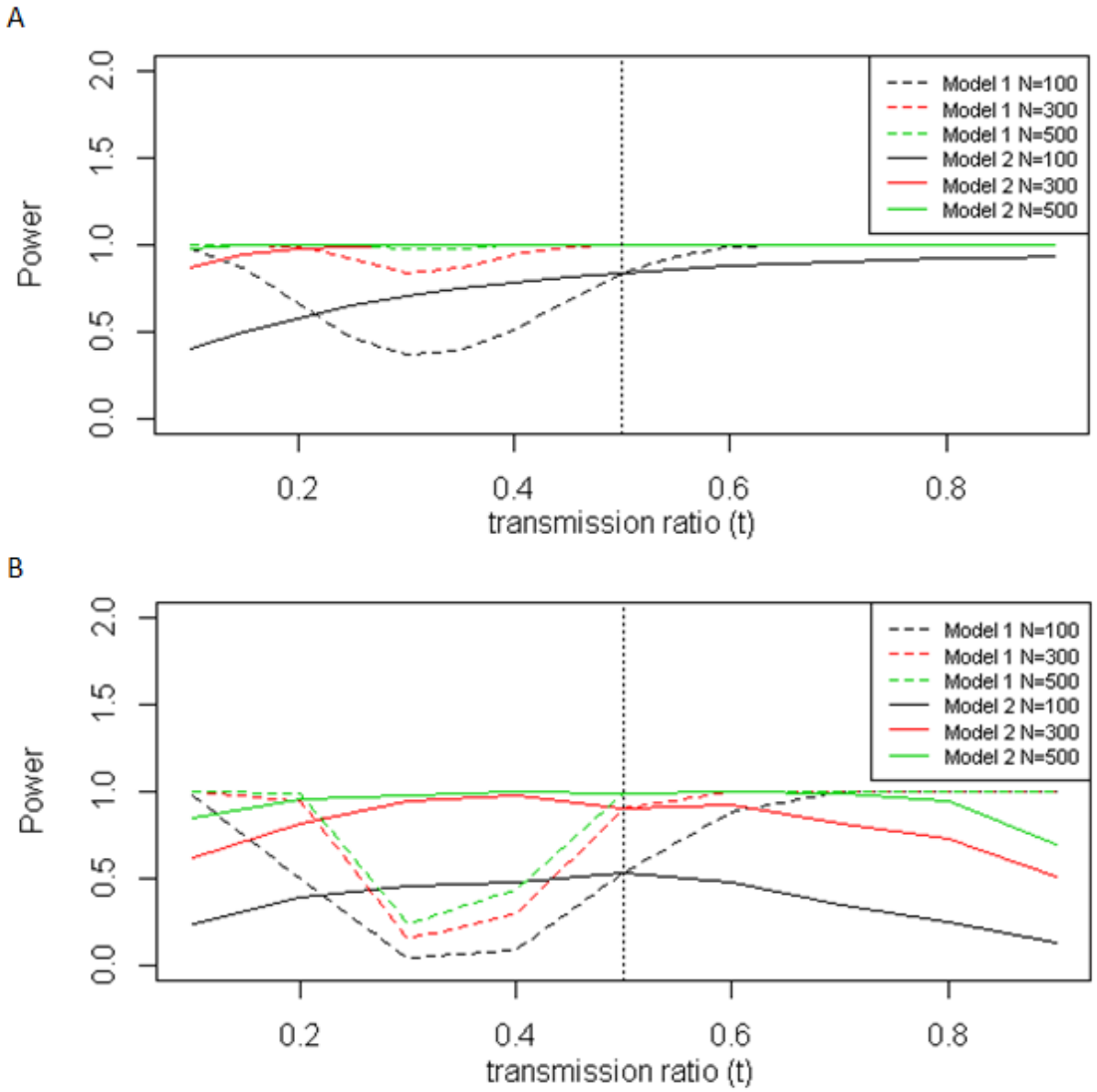


Figure 4.3: Power plot of models 1 and 2 for sample size 100, 300, and 500 when there is true association between disease and DSL where $f_0 = 0.1, f_1 = 0.2, f_2 = 0.3$.

(A) Theoretical results from equations 4.8 and 4.9 in Chapter 4.7.2

(B) Empirical results from simulation



4.5.1.5 Sensitivity analysis: Inflation in RR estimates

We observed that using an under-estimated t value less than the true t in model 2 led to some inflation in the RR (log ratio greater than 0), while an over-estimated t (greater than the true t value) led to attenuation (log ratio less than 0) in the RR, as seen in both Figures 4.4A and 4.4B for R_1 and R_2, respectively. We also noted that the inflation curve of the log RR ratio was linear, which means that the inflation and attenuation are exponential in nature for both R_1 and R_2. When the

difference between the true and selected t was ± 0.1, the inflation ratio lied between $10^{0.25} = 1.78$ and $10^{-0.25} = 0.56$ for R_1. When the difference was greater than ± 0.1, the inflation ratio became more pronounced. The slope of the log RR ratio curve for R_2 in Figure 4.4B was twice that of R_1 in Figure 4.4A. Therefore, the inflation or attenuation in R_2 was more severe than in R_1. Results from our model 2 were highly sensitive, on an exponential scale, to a correct input of t value.

4.5.1.6 Sensitivity analysis: Attenuation and inflation in power

In Figure 4.4 (C), (D) and (E), for $t = 0.3$ and 0.5, the power to detect true association was completely restored when the selected t was equal to the true t. However, setting the selected and true at $t = 0.7$, the power for detecting true association was not completely restored. This was consistent with what we observed in the previous section of power analysis. We also observed that there was a decrease in power when (1) true $t = 0.3$ but the selected t was between 0.3 and 0.6, (2) when true $t = 0.5$ and the selected t was between 0.5 and 0.8, and (3) when true $t = 0.7$, while the selected t was between 0.7 and 0.9. This is due to the partial cancellation of the true signal by the selected t. From these observations, we see that power was also highly sensitive to correct t, even when selected t was slightly greater than the true t.

4.5.1.7 Accuracy of estimated t from control-trios populations

We estimated the mean and standard deviation of the empirical t over 100 iterations in a simulated control-trios population with sample size 500. The 95% CI of the estimated t approximately lies within ± 0.07 of true t value. Increasing the sample size beyond this point did not significantly change the 95%CI. This uncertainty in the estimation of t cannot be built into the likelihood under the current model framework because it is included in the model as an offset, not a variable. Approaches that could account for this uncertainty would likely suffer a price in statistical power.

4.5.2 Application to a case-control, case- and control-parent trio study of IUGR newborn carried out in a Canadian hospital

The MAF calculated from all complete trios in our sample was 23.8% for *F13A1*, 46.4% for *SERPINE1/PAI-1*, 27.1% for *MTHFR A1298C*, 28.9% for *MTHFR C677T*, 2.92% for *F5*, and 1.68% for *F2* (Table 4.3). Except for *MTHFR A1298C*, all MAF were close to the expected range from the literature [146-151]. Discrepancies were likely due to the fact that the samples were genetically heterogeneous with approximately 25% being black.

4.5.2.1 Application to 6 IUGR genes

Applying models 1 and 2 to the IUGR dataset [145], we see in Table 4.3 that *F13A1, SERPINE1/PAI-1* and *MTHFR C677T* all had transmission ratios around 0.5. *MTHFR A1298C* had slightly lower transmission of the disease allele with $t = 0.45$. However, *F5* and *F2* had transmission deviate significantly from the Mendelian ratio with $t = 0.36$ and 0.11. Genotype relative risks from the loglinear model showed no significant association for *F13A1, SERPINE1/PAI-1*, *MTHFR A1298C* and *MTHFR C677T* variants (Table 4.3), similar to previous reports [145, 152]. Due to the small number of cases with 2 copies of *F5* and *F2*, these two genes were analyzed under a dominant model. We see that for *F5*, RR, RR p-value and LRT p-value changed from insignificant (model 1) to significant (model 2), suggesting a deleterious effect of the minor allele. For *F2*, we observed the opposite trend (Table 4.3).The change in significance of the *F5* statistics means that the minor allele is under-transmitted, and operates in the opposite direction of the effect on disease. The change in significance of the *F2* statistics shows that TRD acts in the same direction as the effect of the minor allele on disease. The change in risk after adjustment for TRD was coherent with the expected effects from these variants given that they are known to affect placental circulation and thus potentially fetal growth.

Figure 4.4: Log ratio of Relative Risk, and power with selected t (ranging from 0.1 to 0.9) vs true t in model 2

(A) Log ratio of Relative Risk R_1

(B) Log ratio of Relative Risk R_2

(C) Power of model 2 when true $t = 0.3$

(D) Power of model 2 when true $t = 0.5$

(E) Power of model 2 when true $t = 0.7$

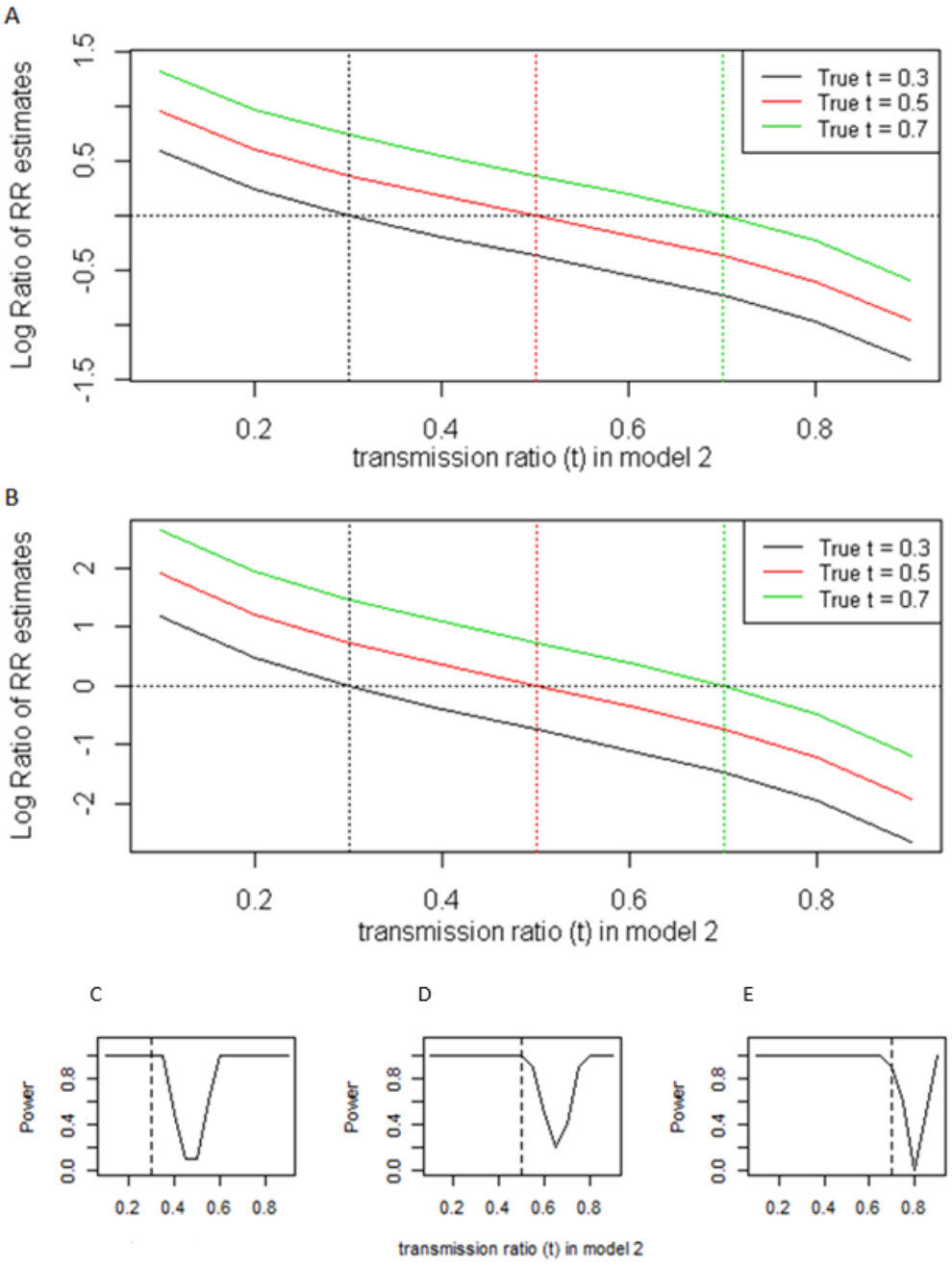


Table 4.3: RR estimates, LRT p-value of adjusted model 2 and unadjusted model 1 for 6 thrombopilic genes, with MAF, transmission ratio (t) and number of genotype 2 cases (G2).

GRR Model									
Gene	Model	MAF	t	G2	R_1 (95%CI)	R_1 p-value	R_2 (95%CI)	R_2 p-value	LRT p-value
F13A1	1	0.24	0.54	16	0.97 (0.66,1.43)	0.89	1.41 (0.68,2.94)	0.354	0.57
	2				0.82 (0.56,1.21)	0.32	1.01 (0.48,2.1)	0.98	0.55
SERPINE1/ PAI-1	1	0.46	0.49	42	0.80 (0.49,1.30)	0.37	0.97 (0.52,1.82)	0.93	0.53
	2				0.83 (0.51,1.35)	0.46	1.06 (0.57,1.98)	0.86	0.53
MTHFR A1298C	1	0.27	0.45	18	0.84 (0.60,1.19)	0.34	0.78 (0.40,1.52)	0.46	0.58
	2				1.04 (0.74,1.47)	0.82	1.18 (0.60,2.31)	0.63	0.89
MTHFR C677T	1	0.29	0.50	19	0.95 (0.67,1.35)	0.8	0.75 (0.39,1.43)	0.38	0.67
	2				0.94 (0.67,1.34)	0.75	0.73 (0.38,1.40)	0.34	0.65
Dominant Model									
Gene	Model	MAF	t	G2	$R_{1/2}$ (95%CI)	p-value			LRT p-value
F5	1	0.03	0.36	2	1.29 (0.57,2.93)	0.54			0.53
	2				2.35 (1.039,5.33)	0.04			0.039
F2	1	0.017	0.11	0	0.31 (0.11,0.85)	0.023			0.014
	2				2.5 (0.91,6.82)	0.074			0.1

Note: F5 and F2 genes have been analyzed under a dominant model.

4.5.2.2 Comparison with TRD analysis in Infante-Rivard (2005) on Coagulation factor V gene

Infante-Rivard and Weinberg [76] found in their study that both *F5* and *F2* exhibited evidence of TRD, as well as *MTHFR A1298C* but to a lesser extent, which is consistent with our estimation from control-trios (Table 4.3). Pursuing the analysis of results for *F5*, the authors used 6 more strata from control-trios together with an interaction term between child genotype and case status. A gene-dosage model ($R_2=R_1^2$) was used implicitly to adjust for TRD; the RR for cases was estimated to be 3.59. We also fitted the augmented loglinear model 2 using a gene-dosage model, and obtained a RR estimate of 2.88 with 95% CI (1.3072, 6.3476). This result is in the range of the estimate from Infante-Rivard and Weinberg [76]. Of note, the number of trios included in these two analyses was different as Infante-Rivard and Weinberg [76] used the LEM software with a built-in EM algorithm for missing data whereas here we only used complete trios. This shows that results from our extended loglinear model 2, which adjusts for TRD were comparable to those from the augmented model proposed in Infante-Rivard and Weinberg [76].

4.6 Discussion

Studies using animal models can potentially provide new insights in handling the phenomenon of TRD. Unlike human studies, the genetic make-up of the animals can be fine-tuned to achieve the desirable study design. In their study, Casella et al. [140] reported many SNPs with TRD that are associated with biological processes involved in embryo viability, confirming previous findings [1].

TRD is much less studied in humans than in animals or plants. In fact, in most genetic association studies in the current literature TRD remains largely unaccounted for. We previously reviewed a number of human studies on TRD [70, 73, 75, 87, 88, 90] and discussed the various methods and study designs in detecting TRD [1].

Here we extend a model used for family-based association studies by accounting for TRD. Our simulation study showed that when TRD is unaccounted for as in model 1, the RR is inflated or attenuated exponentially. Power and Type 1 error also suffered greatly. These results support the need to adjust for TRD. Using a real dataset where the *F5* gene was studied as a determinant of

IUGR, we validated our model in comparison with an approach using control trios [76]. However, we noted that the accuracy of our results depended on the correct TRD offset used in model 2. If we conduct a study with less well-known DSL and diseases, it is unlikely that we will have information on the TRD factor. However, by leveraging on studies such as the HapMap project [87], it may be possible to obtain such information for many DSL.

The extended loglinear model we proposed uses the transmission ratio of minor alleles estimated from control-trios without using actual control-trio data directly in fitting the model. There are other approaches in the current literature which utilize control-trios data directly in model fitting [153, 154]. Since genetic materials from fathers is less likely to be available in practice, authors also have suggested the use of case-mother and control-mother duos [155-157] or supplementing case-trios with control-mother duos [158], via logistic regression [157] multinomial likelihood model [155] or retrospective likelihood approach [156]. These methods have the advantage of testing for violation in Mendelian assumption, but require more genotyping and complex modeling.

The software developed by van Den Oord and Vermunt [159] that was used by Weinberg et al. [160] to fit a loglinear model is LEM, based on the programming language PASCAL. It does not readily have a component for including a TRD offset. However, we implemented the TRD offset method used in this paper in an R package (named TRD) available on the Comprehensive R Archive Network (CRAN).

Currently, there is no comprehensive knowledge on the extent of TRD in the human genome. As TRD can potentially inflate or attenuate an association signal, with such large sets of SNPs being tested, results can be severely biased leading to spurious conclusions. Since TRD over generations leads to reduced mutational diversity in the genome, many of these TRD loci contain rare variants which are currently intensively researched. When transmission counts are small, even a slight distortion could lead to major impact on the outcome of the studies. Given what we observed in our simulation study, there is a need to sequence a control population to identify and quantify the extent of TRD in the human genome. Incorporating this information in the analysis of genetic association studies could provide more accurate and valid estimates. Therefore, we suggest that knowledge of TRD in genomic databases is essential to determine the relevance of genes in various diseases.

4.7 Appendix

4.7.1 Derivation of model 1 (without TRD offset) and 2 (with TRD offset)

4.7.1.1 Derivation of the general model

Let M, F, and C represent the mother, father and child genotypes respectively. The 15 MFC genotype categories are described in Table 4.1. We also let n_{MFC} represent the number of trios with genotypes MFC, and let D represent the disease status of the child. The probability of each MFC cell in Table 4.1 can be written as:

$$P[MFC|D] = E\left[\frac{n_{MFC}}{n}|D\right] = \frac{P[D|MFC]P[C|MF]P[MF]}{P[D]} \quad (4.3)$$

where

P[D|MFC] = Probability that the child is affected given a trio genotype MFC

P[C|MF] = Probability that the child genotype is C given parental genotypes MF

P[MF] = Probability of mating type MF for the parents

P[D] = Disease prevalence

Since we assume that there is no maternal or imprinting effect on the disease status of the child, we can write *P[D|MFC] = P[D|C]*, which means that the disease status of the child depends solely on the child's genotype. Furthermore, we re-write:

$$P[D|C] = P[D|C=0]\frac{P[D|C]}{P[D|C=0]} = f_0 R_c \quad (4.4)$$

where f_0 is the penetrance factor for child genotype 0 and R_c is the RR of child genotype C, and C can be 1 or 2.

Therefore, equation 4.3 can be written as:

$$log\left\{E\left[\frac{n_{MFC}}{n}|D\right]\right\} = log\ P[D|C] + log\ P[C|MF] + log\ P[MF] - log\ P[D]$$

Using the notations $P[C|MF] = \tau_{MFC}$, $P[MF] = \mu_{MF}$, and $P[D] = d$ (see Table 4.1), and using equation 4.4 for *P[D|C],* we obtain:

$$log\{E\ [n_{MFC}|D]\} = log(f_0 R_c) + log\ \tau_{MFC} + log\ \mu_{MF} + log\ n - log\ d$$

$$= log\left(\frac{f_0 n}{d}\right) + log\ \tau_{MFC} + log\ \mu_{MF} + \beta_c \qquad (4.5)$$

where $log(R_c) = \beta_c$.

Model 1 described in this paper corresponds to the scenario where t = 0.5 is substituted into τ_{MFC} (Mendelian transmission). Model 2 corresponds to the scenario where t is not restricted to 0.5, and can take on values between 0 and 1, excluding 0 and 1.

4.7.1.2 Statistical equation for model 1

In order to fit the model described in equation 4.5, we use different grouping schemes for model 1 and model 2. For Weinberg's model (model 1), the terms $log(\tau_{MFC})$ and $log\ (\mu_{MF})$ are grouped together, which we temporarily term φ_{MF} plus an offset term, $log(2)I_{[MFC=111]}$, which only appears for MFC category 111 (seen in last column of Table 4.4). This is because it is the same within each stratum, except for stratum 4, where the sum $log(\tau_{MFC}) + log(\mu_{MF})$ in Table 4.4 (last column) for MFC=111 is 2 times of MFC=112 and 110.

Therefore, to derive the statistical equation for model 1, equation 4.5 can be re-written as

$$log\{E\ [n_{MFC}|D]\} = log\left(\frac{f_0 n}{d}\right) + \sum_{MF=mf} \varphi_{MF}\ I_{[MF=mf]} + log(2)I_{[MFC=111]} + \beta_1 I_{[C=1]} + \beta_2 I_{[C=2]}$$

We can then absorb the constant term $\frac{f_0 n}{d}$ into the summation of φ_{MF} terms and have

$$log\{E\ [n_{MFC}|D]\} = \sum_{MF=mf} log\left[\left(\frac{f_0 n}{d}\right) exp(\varphi_{MF})\right] I_{[MF=mf]} + log(2) I_{[MFC=111]} + \beta_1 I_{[C=1]} + \beta_2 I_{[C=2]}$$

Table 4.4: Stratum frequency, probability of transmission (Mendelian) for case-parent trios

Stratum	MF genotype	C genotype	Stratum frequency (μ_{MF})	Probability of transmission (τ_{MFC})	$log(\mu_{MF}) + log(\tau_{MFC}) = \varphi_{MFC} + log(2) I_{[MFC=111]}$
1	22	2	p^4	1	$log[p^4]+0$
2	21 or 12	1 or 2	$2p^3(1-p)$	1/2	$log[p^3(1-p)]+0$
3	20 or 02	1	$p^2(1-p)^2$	1	$log[p^2(1-p)^2]+0$
4	11	2	$4p^2(1-p)^2$	1/4	$log[p^2(1-p)^2]+0$
	11	1	$4p^2(1-p)^2$	1/2	$log[p^2(1-p)^2]+log2$
	11	0	$4p^2(1-p)^2$	1/4	$log[p^2(1-p)^2]+0$
5	10 or 01	0 or 1	$2p(1-p)^3$	1/2	$log[p(1-p)^3]+0$
6	00	0	$(1-p)^4$	1	$log[(1-p)^4]+0$

By noting γ_{MF} as the first term of the above equation, model 1 can be written as:

$$log\{E\ [n_{MFC}|D]\} = \sum_{MF=mf} \gamma_{MF} I_{[MF=mf]} + log(2) I_{[MFC=111]} + \beta_1 I_{[C=1]} + \beta_2 I_{[C=2]}$$

Since there are 6 strata (S) of MF mating types, by fitting the model with an intercept, we finally obtain:

$$log\{E\ [n_{MFC}|D]\} = \rho_6 + \sum_{j=1}^{5} \rho_j I_{[S=j]} + log(2) I_{[MFC=111]} + \beta_1 I_{[C=1]} + \beta_2 I_{[C=2]} \qquad (4.6)$$

where $\gamma_6 = \rho_6$ and $\gamma_j = \rho_6 + \rho_j$ for j = 1 to 5.

4.7.1.3 Statistical equation for model 2

For model 2, we do not group the terms $log(\tau_{MFC})$ and $log\ (\mu_{MF})$ together, but assign $log(\tau_{MFC})$ as an offset given a specific value of t (Table 4.1), and estimate $log\ (\mu_{MF})$. Therefore, equation 4.5 can be re-written as:

$$log\{E\ [n_{MFC}|D]\} = log\left(\frac{f_0 n}{d}\right) + \sum_{MF=mf} log\ \mu_{MF}\ I_{[MF=mf]} + log\ \tau_{MFC} + \beta_1 I_{[C=1]} + \beta_2 I_{[C=2]}$$

$$= \sum_{MF=mf} log\left(\frac{f_0 n}{d}\right) \mu_{MF}\ I_{[MF=mf]} + log\ \tau_{MFC} + \beta_1 I_{[C=1]} + \beta_2 I_{[C=2]}$$

By noting $log\left(\frac{f_0 n}{d}\right)\mu_{MF}$ as α_{MF}, model 2 can be written as:

$$log\{E\ [n_{MFC}|D]\} = \sum_{MF=mf} \alpha_{MF}\ I_{[MF=mf]} + log\ \tau_{MFC} + \beta_1 I_{[C=1]} + \beta_2 I_{[C=2]}$$

By fitting the model with an intercept, we finally obtain:

$$log\{E\ [n_{MFC}|D]\} = \xi_6 + \sum_{j=1}^{5} \xi_j\ I_{[S=j]} + log\ \tau_{MFC} + \beta_1 I_{[C=1]} + \beta_2 I_{[C=2]} \quad (4.7)$$

where $\alpha_6 = \xi_6$ and $\alpha_j = \xi_6 + \xi_j$ for j = 1 to 5 and *S* = stratum.

Therefore, final statistical formula for model 1 is written in equation (4.6) and for model 2 in equation (4.7).

4.7.2: Non-Central Chi-square Likelihood for model 1 (without TRD offset) and model 2 (with TRD offset)

To perform the Likelihood Ratio Test (LRT) in assessing significance of association between the disease phenotype and DSL, we set up a null model for both model 1 and 2 with null hypothesis $H_0 : \beta_1 = \beta_2 = 0$. The corresponding LRT test statistic, which is the difference in deviance between null and full model, has an asymptotic Chi-Square distribution with 2 degrees of freedom accounting for the two extra terms R_1 and R_2. Agresti [161] showed that when the null hypothesis

is not true for a loglinear model, the resulting LRT is a chi-square statistic with a non-centrality parameter (NCP):

$$\lambda = 2n \sum_{MFC} \pi_{MFC}(M_a) log \left(\frac{\pi_{MFC}(M_a)}{\pi_{MFC}(M_0)}\right)$$

where $\pi_{MFC}(M_a)$ is the true probability of each cell with MFC combination, and $\pi_{MFC}(M_0)$ is the value under the null hypothesis. We also denoted the degree of freedom as υ, which is 2 in our LRT because there are 2 extra variables R_1 and R_2 in the alternative model than the corresponding null model.

To calculate Type 1 error and power comparable to our theoretical values, we need to have the exact likelihood. Our likelihood for the alternative hypothesis is shown in equation 4.3 and re-written as:

$$\pi_{MFC}(M_a) = \frac{f_0 R_c \tau_{MFC} \mu_{MF}}{d}$$

where $f_0 R_c$, τ_{MFC}, μ_{MF} and d are defined as in equation 4.4 and 4.5.

In the presence of TRD, we know that even when the null hypothesis is true, the LRT still has a non-Central Chi-square distribution. The null model is different for models 1 and 2 because TRD is being adjusted in the offset of model 2 but not in model 1. Under the null hypothesis, *P[D|MFC] = P[D]*, and hence, $f_0 R_c / d = 1$. The likelihoods for models 1 and 2 under null hypothesis are then, respectively:

$$\pi_{MFC}(M_{01}) = \mu_{MF}\, \tau_{MFC}\ [0.5]$$

and

$$\pi_{MFC}(M_{02}) = \mu_{MF}\ \tau_{MFC}[t]$$

Under the alternative hypothesis, NCP for model 1 is:

$$\lambda_1 = 2n \sum_{MFC} \frac{f_0 R_c \tau_{MFC} \mu_{MF}[t]}{d}\ log\left(\frac{f_0 R_c \tau_{MFC}[t]}{\tau_{MFC}[0.5]\ d}\right) \quad (4.8)$$

and the NCP for model 2 is:

$$\lambda_2 = 2n \sum_{MFC} \frac{f_0 R_c \tau_{MFC} \mu_{MF}[t]}{d}\ log\left(\frac{f_0 R_c}{d}\right) \quad (4.9)$$

Note that when t is not equal to 0.5, even though there is no association signal, the LRT is still a NCP chi-square statistic. The NCP for model 1 is 0 when t = 0.5 (Mendelian transmission) and $\frac{f_0 R_c}{d}$=1 (no association). Therefore, null hypothesis for model 1 requires both Mendelian transmission and no association between disease and DSL. However, since TRD has already been adjusted for in model 2, the NCP is 0 when $\frac{f_0 R_c}{d}$=1 (no association).

Web resources

R package 'TRD', http://cran.r-project.org/web/packages/TRD/index.html

Online Mendelian Inheritance in Man (OMIM), http://www.omim.org

HUGO Gene Nomenclature Committee (HGNC), http://www.genenames.org

Chapter 5

Modeling sex-of-parent-specific Transmission Ratio Distortion and imprinting effect in loglinear model using case-trios

5.1 Preamble

We have examined non-sex-of-parent-specific TRD (NST) in Chapter 4, and the implication of its presence in invalidating association study results if not accounted for. We proposed an offset of transmission probability of minor alleles, which arises from a natural component in the loglinear model framework developed by Weinberg et al. [44]. This offset is shown to be successful in restoring the correct RR, Type 1 error, and power. However, TRD might also occur in sex-of-parent-specific manner, which we call sex-of-parent-specific TRD (ST). This ST can be sub-categorized into maternal ST (MST) and paternal ST (PST), which refer to non-Mendelian transmission in only mother and only father, respectively.

ST is problematic because it mimics another mechanism, the imprinting effect, which is believed to influence more than 1% of all mammalian genes. Imprinting occurs when disease allele inherited from the father induces a different expression level at a neighbouring disease gene than that inherited from the mother. It leads to an over-representation of disease allele in the child from the parent who induces a higher expression level. On the other hand, ST can also lead to an over-representation of disease allele in the child from one parent, when that parent has a higher transmission ratio of the minor allele. Therefore, ST not only affects the RR estimates for child effect, but also confounds the imprinting effect. In this chapter, we will show the results of ST on RR of child and imprinting effects, type 1 error, sensitivity and specificity, and illustrate the effectiveness of applying the sex-of-parent-specific transmission offset to the loglinear model in restoring the correct measures.

Manuscript 3: Modeling sex-of-parent-specific Transmission Ratio Distortion and imprinting effect in loglinear model using case-trios

Lam Opal Huang [1*], Claire Infante-Rivard [1], Aurélie Labbe [1-3]

[1] Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal Quebec, Canada, H3A 1A2

[2] Douglas Mental Health University Institute, Montréal, Quebec, Canada, H4H 1R3

[3] Department of Psychiatry, McGill University, Montréal Quebec, Canada, H3A 1A1

* Correspondence author

Contact information for correspondence author:

1020 Pine Avenue West, Montreal, Quebec H3A 1A2
Tel.: 514-398-6258

Fax: 514-398-4503

opal.huang@mail.mcgill.ca

Conflict of Interest

The authors declare no conflict of interests.

5.2 Abstract

Transmission Ratio Distortion (TRD) is a phenomenon where parental transmission of disease allele to the child does not follow the Mendelian inheritance ratio. TRD can occur in a sex-of-parent-specific or non-sex-of-parent-specific manner. In our previous paper, the loglinear model proposed by Weinberg et al. based on case-trios study design was extended to address non-sex-of-parent-specific TRD (NST). An offset computed from the transmission probability of the minor allele in control-trios is used to adjust for TRD. It was shown that adjusting the model with the offset can remove the inflation in RR and Type 1 error introduced by NST. The loglinear model in Weinberg et al. was then further extended to estimate an imprinting parameter. It is believed that more than 1% of all mammalian genes are imprinted. In the presence of imprinting, child inheriting disease allele from the parent who induces a higher expression level at a neighbouring disease gene is over-represented in the sample. As we know that TRD mechanisms such as meiotic drive and gametic competition also occur in sex-of-parent-specific manner. Therefore, sex-of-parent-specific TRD (ST) can lead to over-representation of maternal or paternal alleles in the affected child in a similar fashion. As a result, ST confounds with the imprinting effect when present in the sample. We proposed to specify a sex-of-parent-specific transmission offset in adjusting the loglinear model to account for ST. We found that the extended model restores the correct RR estimates for child and imprinting effects, adjusts for inflation in Type 1 error, and improves performance on sensitivity and specificity compared to the original model without TRD offset. We conclude that in order to correctly interpret association signal and imprinting effect, adjustment for ST is necessary to ensure valid conclusions.

5.3 Background

Transmission Ratio Distortion (TRD) is the genetic phenomenon where one of the two alleles from a parent is favorably transmitted to the child, hence violating the 1:1 Mendelian inheritance law [70]. There are many forms of TRD that arise from a range of biological mechanisms during the gametic and embryonic developmental stages [1, 71, 79, 81]. In a previous paper, we have examined and modeled the simplest form of TRD, where transmission probability of minor allele in both parents are the same. Here we will call this non-sex-of-parent-specific TRD (NST) (Chapter 4). However, TRD can occur in sex-of-parent-specific manner [89, 95, 101, 104, 124, 162]. Biological mechanisms involved in this type of TRD disrupts cell processes such as gametic formation during meiosis [101], and zygote production during fertilization [163]. We call this type of TRD sex-of-parent-specific TRD (ST).

5.3.1 Meiotic Drive

During female meiosis, a germ cell is divided into 4 cells each containing a sister chromatid. Only one of these cells becomes a gamete (egg) while the other 3 become polar bodies and are eventually eliminated [70]. Since this process is asymmetric, when one sister chromatid has a structural advantage over the others, it tends to have a survival advantage. This process, which happens predominantly in female, is called meiotic drive [70]. In male meiosis, all 4 cells result in functional gametes and therefore, the process is not affected by this survival advantage [70]. However, male meiotic drive also exists in species such as sciara, but is rare [70].

5.3.2 Gametic competition

Gametic competition occurs at the fertilization stage, where some sperms outperform others in reaching successful fertilization [75] and leads to over-transmission of corresponding alleles in the winners. Classical gametic selection systems include the mouse t-haplotype and segregation distorter in drosophila [70]. Gametic competition occurs only in males, and therefore, are paternal-specific.

5.3.3 Impact of TRD on association studies

One of the common study design to study association between disease and genetic markers is the case-trio family-based study design. These family-based association studies are robust to population stratification because the transmitted allele from parents to child is perfectly matched with age, sex, and ethnicity of the non-transmitted allele from the same child [36]. Control-trios which are composed of both parents and their unaffected offspring, have been previously used in controlling for TRD [36]. TRD can lead to over- or under-transmission of the disease allele in the cases, but also in the general population represented by controls [71-77]. Our recent work has shown that when the effect of NST is ignored in case-trio studies, the association signal measured can be inflated or attenuated, leading to spurious results (Chapter 4). Therefore, we concluded that NST can confound the true association signal.

5.3.4 Loglinear model and child effect

Weinberg et al. [44] proposed a loglinear model to measure the magnitude of association (relative risk) between a disease susceptibility locus (DSL) and the expression of disease in the child. The simplest loglinear model consists of two variables, one for child of genotype 1 and one for child of genotype 2, where the former is defined as the heterozygous genotype and the latter as the homozygous mutant genotype. The homozygous wild-type genotype 0 serves as the baseline. The parameter estimates of these 2 variables measure the relative risk (RR) of child inheriting 1 or 2 copies of disease allele from the parents. Furthermore, the deviance of the full loglinear model against the null can be used in a likelihood ratio test to measure the significance of the association signal. Since this loglinear model can estimate genotype relative risk (GRR), as well as test for significance of association, it is advantageous over a test such as the transmission disequilibrium test (TDT) which only offers a p-value for the significance of the association signal [36]. Therefore, we chose to use the loglinear model to illustrate the impact of TRD in our results.

5.3.5 Loglinear model and child effect with NST offset

Recently, we proposed an extension to the Weinberg et al log-linear model (Chapter 4) by taking into account the TRD probability; this involves including in the model an offset parameter

computed using the minor allele transmission probability estimated from control-trios. This TRD offset which is different at each of the 15 mother-father-child (MFC) genotype categories, adjusts for the effect of NST. We showed that this offset can restore the true RR, significance of association, and compensate for the inflated Type 1 error and power loss for the likelihood ratio test.

5.3.6 Imprinting (parent-of-origin) effect

Imprinting effect, which also known as parent-of-origin effect, expresses itself when a disease allele is activated when inherited from one of the parents, but not from the other [76, 90, 164, 165]. According to Mendelian inheritance, when a disease is paternally (maternally) imprinted, the corresponding disease allele is silenced, while one inherited from the mother (father) is activated. However, in complex diseases, the silencing and activation of the imprinted disease allele is not absolute [41]. For example, the disease allele inherited from the mother to the child may not be fully penetrant, and the one inherited from the father may not be fully silenced. Nevertheless, an imprinting effect can be statistically measured as the ratio of probability of the maternal vs paternal expression at a neighbouring disease gene [45].

5.3.7 Joint modeling of child genotype and imprinting effect

The imprinting and child effects can be modeled as multiplicative factors which form the combined penetrance function [44, 45], and is additive in log scale. In this paper, we will consider both child and imprinting effects in a model similar to the loglinear model previously proposed by Weinberg et al. [44]. Details are explained in Chapter 5.4.

5.3.8 Relationships between sex-of-parent-specific TRD and imprinting

We denote t_m and t_f to be the transmission probabilities of the minor allele from mother and father, respectively, to child in control-trio populations. Let the genotype of a heterozygous child inheriting the minor allele from the mother be Dd, and from the father be dD. As seen in the example illustrated by Figure 5.1, the proportions of heterozygous children inheriting minor allele (D) from the mother and the father in the population are equal under NST. However, when there

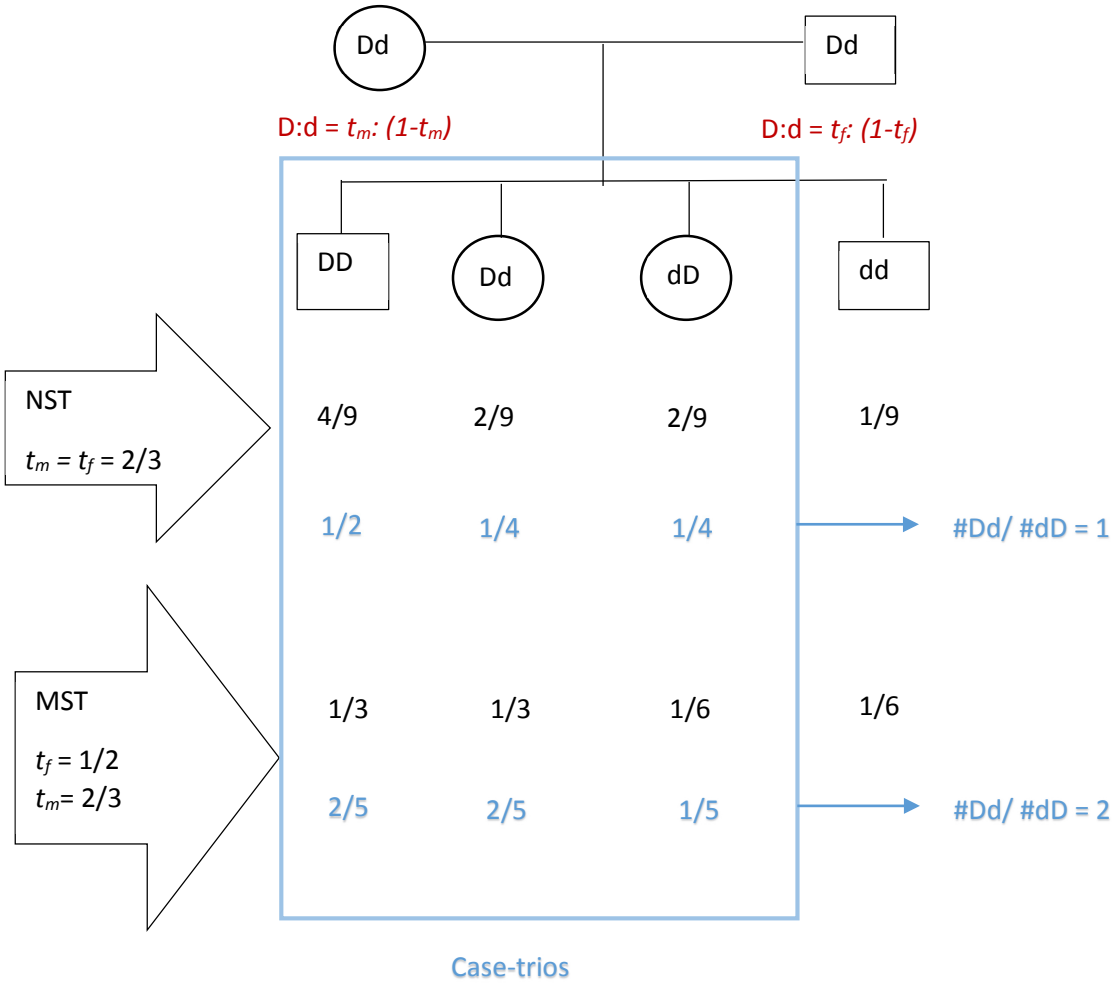
is a maternal sex-of-parent-specific TRD (MST), where mothers over-transmit the minor allele at a 2:1 ratio for example, the ratio of Dd to dD cases is also 2:1 in the population.

Another example is shown in Figure 5.2, where maternal expression is higher than paternal expression at a ratio of 2:1. The proportion of diseased children with Dd genotype is therefore, twice as high as with the dD genotype, under NST. If there is MST instead, in the presence of imprinting, the ratio of Dd to dD cases could rise to 4:1 in the population. Therefore, we see that when ST and imprinting effects both exist, and act in the same direction, the imprinting effect is inflated (Figure 5.2). On the other hand, if the two effects act in opposite directions, the imprinting effect is attenuated. As a result, a ST signal can confound the significance of an imprinting effect. This confounding effect is the main focus of this paper. We intend to adjust for this ST factor in a loglinear model, and evaluate the inflation in RR estimates for child and imprinting effects, Type 1 error, and performance on sensitivity and specificity.

5.4 Material and Methods

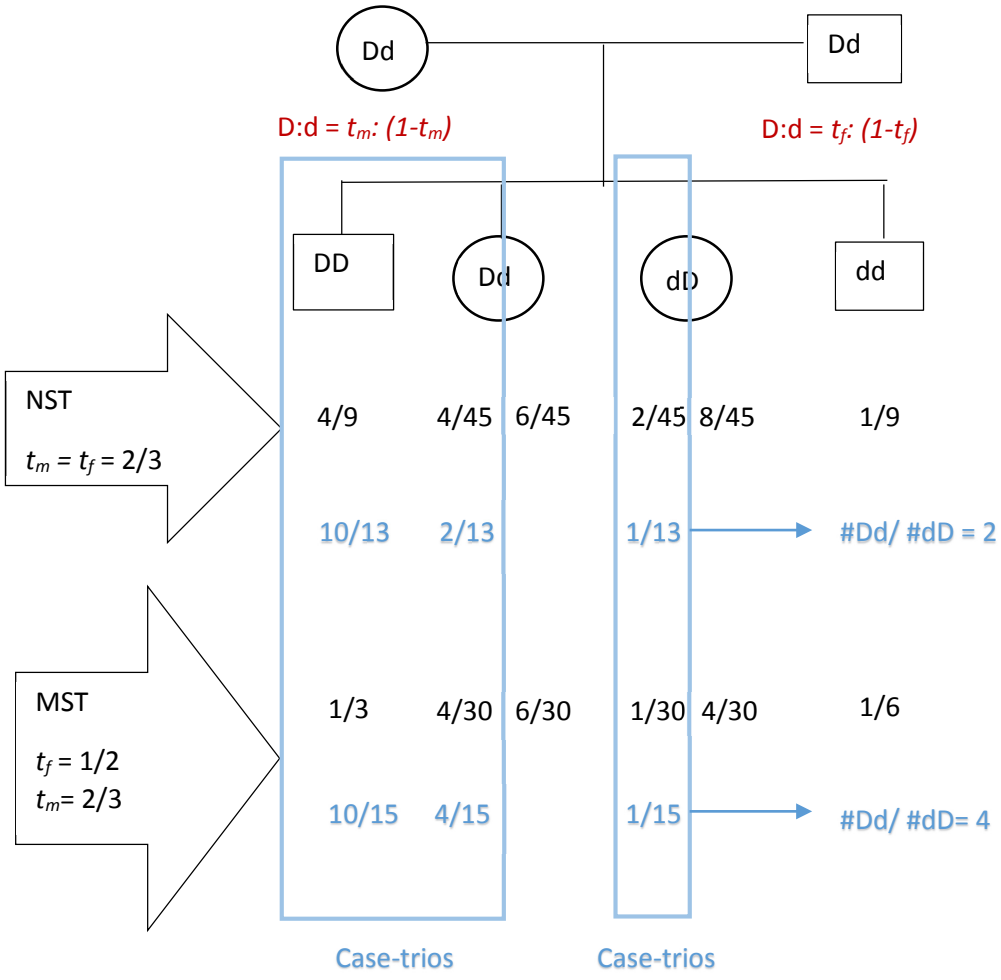
In this paper, we define the genotype using the additive model, counting the number of copies of the minor allele. We investigated the association between a bi-allelic disease susceptibility locus (DSL) and a disease, using the loglinear model with child and imprinting effect variables as proposed by Weinberg et al. [44], but using the parameterization as in the later work by Weinberg [45]. To adjust for ST, we added to the model a parental-specific offset parameter that depends on both the maternal and paternal minor allele transmission probabilities, noted as t_m and t_f, respectively. We assumed that these two sex-of-parent-specific variables can be computed from available control-trios datasets. In this model, the response variable is the number of trio (counts) for each of the 16 mother-father-child (MFC) genotype categories, as described in Table 5.1. Note that the mother-father-child (MFC) category 111 (triply heterozygous trios) was divided into 2 categories: one for the heterozygous child inheriting the disease allele from the mother (111[M]), and the other for inheritance from the father (111[F]). Assuming mating symmetry these 16 categories can be subdivided into 6 parental mating types as shown in Table 5.1.

Figure 5.1: Scenario with TRD, $f_2 = f_{1M}$ (maternal penetrance) $= f_{1F}$ (paternal penetrance) $= 1$, $f_0 = 0$ (dominant disease)



With NST, the proportion of Dd and dD individuals in the whole population are the same (2/9). When only case-trios are sampled (blue box), the proportion of Dd and dD are still the same, but they represent a greater proportion in the case-trios sample at 1/4. However, when there is MST and t_m: $(1-t_m)$ = 2:1, the ratio of Dd to dD individuals becomes 2:1 in both the whole population (1/3 to 1/6) and the case-trios sample (2/5 to 1/5), because maternal disease allele D is twice likely to be over-transmitted than paternal ones.

Figure 5.2 Scenario with TRD, $f_2 = 1$, f_{1M} (maternal penetrance) $= 0.4$, f_{1F} (paternal penetrance) $= 0.2$, T (imprinting factor) $= f_{1M}/f_{1F} = 2$, $f0 = 0$



Assuming NST with imprinting, the proportion of diseased Dd and dD individuals in the whole population are no longer 2/9, but are 4/45 and 2/45, respectively, because of the different penetrance values for Dd ($f_{1M} = 0.4$) and dD ($f_{1F} = 0.2$) genotypes. The proportion of Dd and dD individuals that are not diseased are 6/45 and 8/45, respectively. As a result, the proportion of Dd and dD individuals in the case-trios sample (blue box) are 2/13 and 1/13, respectively. Therefore, the ratio of Dd to dD individuals is at 2:1, because imprinting factor of maternal vs paternal expression is 2. When there is both MST ($t_f = 1/2$ and $t_m = 2/3$) and imprinting ($f_{1M}/f_{1F} = 2$), the diseased Dd and dD individuals in the whole population are now 4/30 and 1/30, respectively. The proportion of Dd and dD individuals that are not diseased are 6/30 and 4/30, respectively. In the case-trios sample (blue box), proportion of Dd and dD individuals are 4/15 and 1/15, respectively. The ratio of Dd to dD individuals is now 4:1. This is the combined result of imprinting and MST because maternal disease allele is twice likely to be over-transmitted and induces twice the gene expression level compared to paternal ones.

5.4.1 Parameterization schemes

We now briefly address the two parameterization schemes suggested by Weinberg et al. [44] and Weinberg [45]. The original parameterization scheme uses 4 parameters in the model: two for child effect with genotype 1 (R_1) and 2 (R_2), and two for imprinting effect of mother (I_M) and father (I_F) [44]. The second parameterization scheme [45] uses only 3 parameters: relative risk (RR) of genotype 1 child with inherited disease allele from father (R_1), RR of genotype 2 child with both the maternal and paternal imprinting effect (R_2), and risk ratio of maternal vs paternal imprinting effect (T). The latter was suggested to replace the first approach, and it is important to note that the interpretation of the R_1 and R_2 parameters differ between the two approaches.

Using the first approach [44], the parameters described above can be incorporated into the penetrance equations as:

$$f_2 = f_0 R_2 I_M I_F$$

$$f_{1M} = f_0 R_1 I_M$$

$$f_{1F} = f_0 R_1 I_F$$

where f_2 is the penetrance for child of genotype 2, f_{1M} is the penetrance for child of genotype 1 with disease allele inherited from the mother, f_{1F} is the penetrance for child of genotype 1 with disease allele inherited from the father, and f_0 is the penetrance of genotype 0 child. Since f_0 does not depend on any of the 4 parameters, we have only 3 equations but 4 parameters to estimate. Therefore, one of the 4 parameters is unidentifiable.

By using the second parameterization approach [45], the parameters described above can be incorporated into the penetrance equations as:

$$f_2 = f_0 R_2 \tag{5.1}$$

$$f_{1M} = f_0 R_1 T \tag{5.2}$$

$$f_{1F} = f_0 R_1 \quad (5.3)$$

We then have 3 equations and 3 parameters, which makes each parameter identifiable. Ainsworth et al. [155] stated that parameterization of the imprinting parameter in the second approach [45] is biologically unintuitive because the imprinting factor is only present in and can only be estimated by child genotype 1 category (C = 1); therefore is not seen in child genotype 2 category, whereas in fact, biologically, imprinting effect is present in child genotype 2 as well. While this is true, we are more interested to know how much more (or less) likely it is for the child to have the disease when the minor allele is inherited from the mother compared to the father. Therefore, for the purpose of our study, the second parameterization is intuitive in the interpretation of the measures of our interest, and is also more parsimonious with each parameter identifiable. In the following, models and results are presented under the second parameterization approach.

5.4.2 Loglinear model from Weinberg et al. (1998) with only child and imprinting variables

The 16 MFC categories loglinear model with child and imprinting effects using the second parameterization scheme can be written as:

$$log\ E[n_{MFC}|D] = \gamma_6 + \sum_{j=1}^{5} \gamma_j I_{[S=j]} + \beta_1 I[C=1] + \beta_2 I[C=2] + \zeta_M I_{[C=1,maternal]} \quad (5.4)$$

where M, F, and C represent the mother, father and child genotypes, respectively; n_{MFC} is the number of trios with genotypes MFC, and D is the disease status of the child. The $\gamma_6 + \gamma_j$ terms are the regression coefficients for the first 5 parental mating types in Table 5.1; γ_6 is the intercept corresponding to the 6th mating type MF=00. The indicator variable $I_{[C=1,maternal]}$ is 1 for a heterozygous child inheriting disease allele from the mother.

The β_1 and β_2 parameters are the regression coefficients for child genotypes 1 and 2. We denote $R_1 = R_{1F} = exp(\beta_1)$, which corresponds to the RR of child with 1 copy of disease allele inherited from the father, and zero copy from the mother; $R_{1M} = exp(\beta_1 + \zeta_M)$ which corresponds to the RR of child with 1 copy of disease allele inherited from the mother and zero copy from the father; $R_2 = exp(\beta_2)$ as the RR of child inheriting 1 copy of disease allele from both parents. R_{1F}, R_{1M} and R_2 are the relative risks with respect to baseline genotype 0.

$T = exp(\zeta_M)$ is the relative risk of maternal vs paternal expression at a neighbouring disease gene for genotype 1 child. When $T = exp(\zeta_M) = 1$, there is no parent-of-origin effect, whilst when $exp(\zeta_M) > 1$, a child with 1 maternally inherited disease allele has higher risk than a child with 1 paternally inherited disease allele. On the other hand, when $exp(\zeta_M) < 1$, a child with 1 maternally inherited disease allele has lower risk than child with 1 paternally inherited disease allele. This model denoted as model 1, is valid only when there is Mendelian inheritance. The complete derivation of this model is shown in Chapter 5.7.1.

Table 5.1: Relative Risk and imprinting parameterization

Stratum	MFC Genotype	Stratum frequency (μ_{MF}) under HWE	Probability of transmission (τ_{MFC})		Weinberg et al. (1998) parameterization		Weinberg (1999) parameterization	
			TRD	Mendelian	RR	Imprinting	RR	Imprinting
1	222	p^4	*1*	1	R_2	$I_M I_F$	R_2	1
2	212	$2p^3(1-p)$	t_f	1/2	R_2	$I_M I_F$	R_2	1
	211	$2p^3(1-p)$	$1-t_f$	1/2	R_1	I_M	R_1	T
	122	$2p^3(1-p)$	t_m	1/2	R_2	$I_M I_F$	R_2	1
	121	$2p^3(1-p)$	$1-t_m$	1/2	R_1	I_F	R_1	1
3	201	$p^2(1-p)^2$	*1*	1	R_1	I_M	R_1	T
	021	$p^2(1-p)^2$	*1*	1	R_1	I_F	R_1	1
4	112	$4p^2(1-p)^2$	$t_m t_f$	1/4	R_2	$I_M I_F$	R_2	1
	111[M]	$4p^2(1-p)^2$	$t_m(1-t_f)$	1/4	R_1	I_M	R_1	T
	111[F]	$4p^2(1-p)^2$	$t_f(1-t_m)$	1/4	R_1	I_F	R_1	1
	110	$4p^2(1-p)^2$	$(1-t_m)(1-t_f)$	1/4	1	1	1	1
5	101	$2p(1-p)^3$	t_m	1/2	R_1	I_M	R_1	T
	100	$2p(1-p)^3$	$1-t_m$	1/2	1	1	1	1
	011	$2p(1-p)^3$	t_f	1/2	R_1	I_F	R_1	1
	010	$2p(1-p)^3$	$1-t_f$	1/2	1	1	1	1
6	000	$(1-p)^4$	*1*	1	1	1	1	1

5.4.3 Loglinear model with child and imprinting variables and ST offset

Extending Model 1 to account for ST, we obtain the following model:

$$log\{E\,[n_{MFC}|D]\} = \alpha_6 + \sum_{j=1}^{5} \alpha_j\, I_{[S=j]} + log\,\tau_{MFC} + \beta_1 I_{[C=1]} + \beta_2 I_{[C=2]} + \zeta_M I_{[C=1,maternal]} \quad (5.5)$$

where the notation and regression parameters are defined in the same way as model 1, except for α_j which includes the mating type frequency but not the transmission probability *P[C|MF]*. The ST offset $\tau_{MFC} = P[C|MF]$ captures the sex-of-parent-specific transmission probability and is defined as in Table 5.1. The derivation of this loglinear model is also shown in Chapter 5.7.1, and is denoted as model 2.

5.4.4 Simulation set up

In our previous investigation on association studies with case-trios, we observed the inflation in Type 1 error, RR estimates and RR p-values when NST was not accounted for (Chapter 4). Inflation of RR was shown to be exponential in scale. There was also power loss when NST is in the opposite direction of the association signal (Chapter 4). Here we wanted to assess the extent of the RR and Type 1 error inflation with different combinations of maternal and paternal transmission probabilities of the minor allele.

We assumed that the genetic marker is the DSL under investigation with minor allele frequency (MAF) 0.1 and no recombination. We simulated a population of 100,000 trios with symmetric parental mating. We then sampled 500 case-trios from the simulated population to measure the association signal with the genetic marker. The maternal (t_m) and paternal (t_f) transmission probabilities of minor allele are known a priori in model 2, assuming that they could be obtained for example, from control-trios samples in existing databases. Penetrance for genotype 1 individuals inheriting disease allele from the father is f_1 (equivalent to f_{1F} noted previously), f_2 for genotype 2 individuals, and f_0 for genotype 0 individuals. Finally, the simulation parameter which indicates the ratio of maternal vs paternal expression at a neighbouring disease gene for genotype 1 individuals is denoted as *g*.

5.4.4.1 Scenarios of association and ST

We investigated three association setups: 1) the genetic marker is not associated with the disease and there is no imprinting ($f_0 = f_1 = f_2 = 0.1$ and $g = 1$), 2) the genetic marker is associated with the disease in the opposite direction of the TRD and there is no imprinting, 3) the genetic marker is associated with the disease in the opposite direction of the TRD and there is imprinting.

For each of the three setups, we tested a variety of TRD scenarios described in Table 5.2. Using these 7 scenarios, we then applied models 1 and 2 to the dataset, and observed the changes between models 1 and 2 with respect to the estimation of R_1 (RR for genotype 1 individuals inheriting disease allele from the father), R_2 (RR for genotype 2 individuals) and T (RR for maternal vs paternal inheritance of disease allele in genotype 1 individuals).

Table 5.2: Different scenarios of TRD for each of the three association setups

		Transmission probability	
scenario	Type of TRD	t_m	t_f
1	NST	0.3	0.3
2	NST	0.5	0.5
3	NST	0.7	0.7
4	PST	0.5	0.3
5	PST	0.5	0.7
6	MST	0.3	0.5
7	MST	0.7	0.5

5.4.4.2 Assessing inflation or attenuation of regression parameters, inflation of Type 1 error, sensitivity and specificity of models 1 and 2

We intend to measure the impact of t_m and t_f on the regression parameters R_1, R_2 and T when there is NST, or MST or PST on a continuum of t_m and t_f values. Using the 1st setup where there is no association or imprinting, we measured the inflation ratio for the 3 regression parameters contrasting model 1 with model 2 with $t_m = t_f$ ranging from 0.1 to 0.9 (NST). Then, we fixed t_f at 0.5 and tested t_m from 0.1 to 0.9 (MST). Similarly, we fixed t_m at 0.5 and set t_f ranging from 0.1 to 0.9 (PST). We also assess the Type 1 error, using the 1st setup for the 3 types of TRD at sample size 100, 300 and 500. Finally, to measure the sensitivity and specificity of models 1 and 2, we used the 2nd and 3rd setups, and plotted the receiver operating characteristic (ROC) curves.

5.5 Results

5.5.1 Impact of ST adjustment on R_1 and R_2

We observed the results for the 7 scenarios shown in Table 5.3 for the 1st simulation setup, and results for the 2nd and 3rd simulation setups are the same. The results showed that whenever there is an over-transmission, correctly adjusting for TRD reduces the estimates of R_1 and R_2, and this is reversed when there is under-transmission. In Figure 5.3A, NST led to the greatest inflation among all 3 types of TRD because both parents are over- or under-transmitting. Also, MST led to smaller inflation in R_1 than PST because R_1 primarily measures the RR genotype 1 with child inheriting disease allele from the father. The change in R_1 with respect to t_m was due to the change in the baseline risk of genotype 0 individuals. MST and PST nearly coincided with each other as seen in Figure 5.3B, since t_m and t_f were interchangeable when both parents transmitted a disease allele (see Table 5.1).

Table 5.3: Change in R_1, R_2 and T after correction with ST offset for the 7 different TRD scenarios using model 2, when there was no true association between marker and disease, nor there was imprinting effect on the marker (1st setup).

	Transmission probability		After correction for ST		
scenario	t_m	t_f	R_1	R_2	T
1	0.3	0.3	/	/	-
2	0.5	0.5	-	-	-
3	0.7	0.7	\	\	-
4	0.5	0.3	/	/	\
5	0.5	0.7	\	\	/
6	0.3	0.5	/	/	/
7	0.7	0.5	\	\	\

Notation:

/ = increased;

\ = decreased;

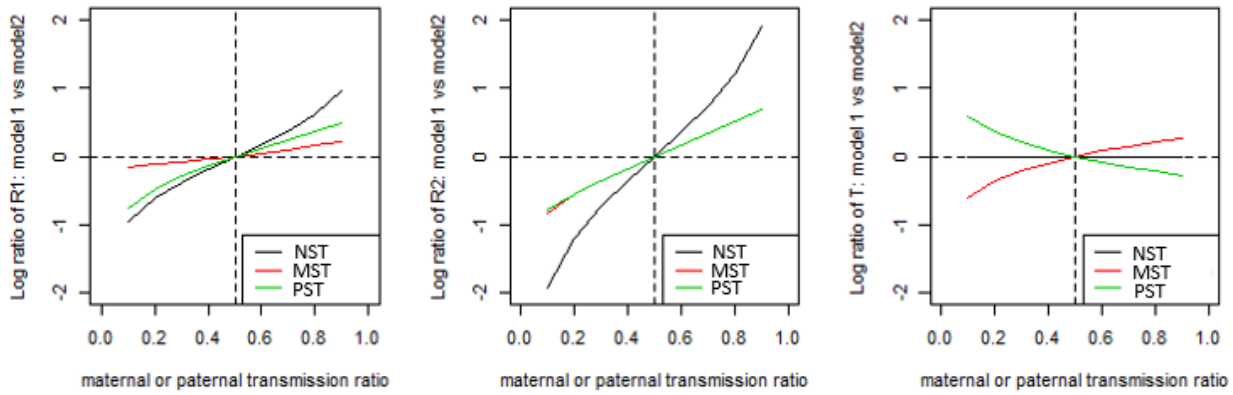
- = unchanged.

5.5.2 Impact of ST adjustment on imprinting parameter T

In scenarios 1 to 3 described in Table 5.2 we observed that the imprinting parameter T remained unchanged when $t_m = t_f$ (NST) after adjustment because maternal effect and paternal effect cancelled out each other in the ratio. In Figure 5.3C, we see that NST has no impact on T. When $t_f < 0.5$ (scenario 4), the paternally inherited disease allele appeared to be associated with lower risk, this led to an apparent smaller paternal expression. Since paternal effect is in the denominator of regression parameter T, T decreased when model is correctly adjusted by t_f. On the other hand, when $t_f > 0.5$ (scenario 5), T increased when model is correctly adjusted by t_f. Similar relationship is also shown in Figure 3C where smaller t_f leads to a larger T, and larger t_f a smaller T. Maternal expression is measured in the numerator of T, hence the trend is reversed in scenarios 6 and 7, and in Figure 5.3C for MST.

Figure 5.3: Inflation and attenuation of R_1, R_2 and T

A R_1 B R_2 C T



Note:

NST: $t_m = t_f$ from 0.1 to 0.9;

MST: t_m = 0.1 to 0.9, t_f = 0.5;

PST: t_f = 0.1 to 0.9, t_m = 0.5.

5.5.3 Inflation of Type 1 error

We plotted the theoretical and empirical Type 1 error separately for each of the 3 types of TRD (NST, MST and PST). The theoretical (Figure 5.4) and empirical Type 1 error (Figure 5.5) matched with each other well. We see that Type 1 error was inflated more and more severely as t_m and t_f became more skewed in model 1. MST (Figure 5.5B) and PST (Figure 5.5C) plots for model 1 are similar and had a more gradual climb in Type 1 error compared to NST (Figure 5.5A) because the combined effect is greater than either maternal or paternal TRD alone. For model 2, this inflation in Type 1 error is removed.

Figure 5.4: Theoretical Type 1 error ($f_0 = f_1 = f_2 = g = 1$)

A NST B MST C PST

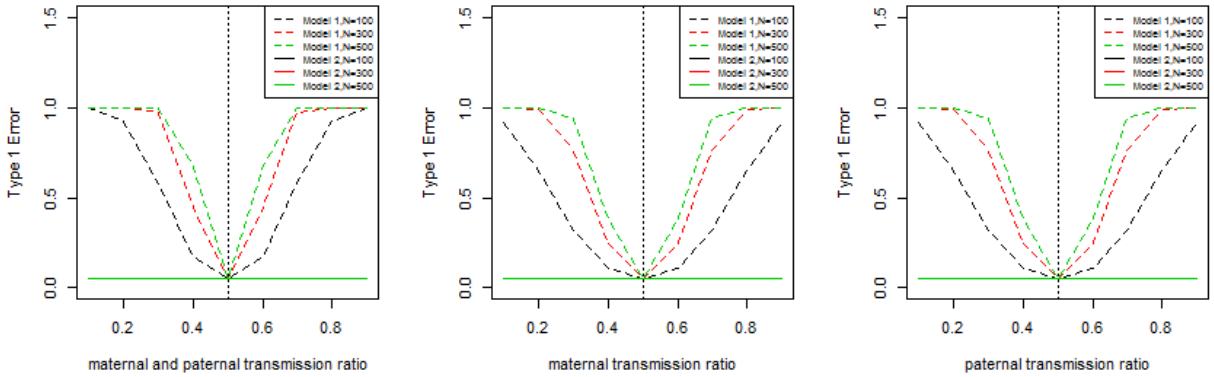
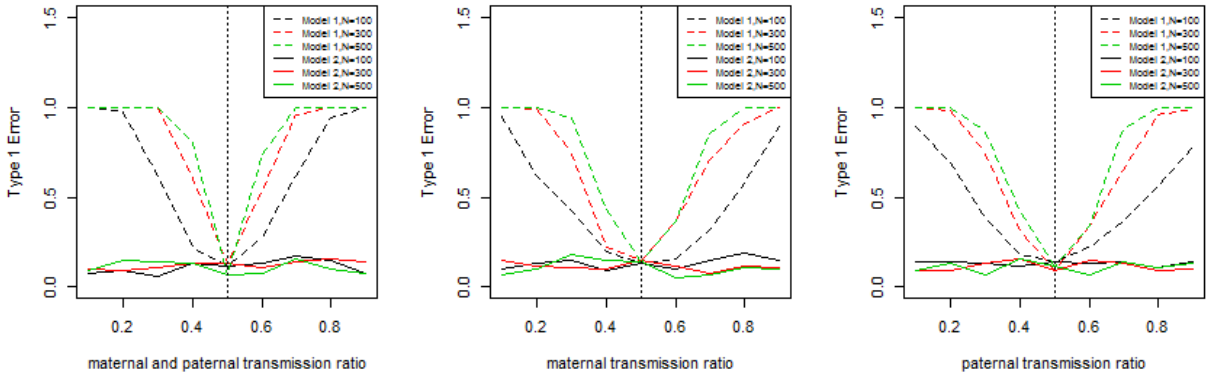


Figure 5.5: Empirical Type 1 error ($f_0 = f_1 = f_2 = g = 1$)

A NST B MST C PST



5.5.4 Sensitivity and Specificity of models 1 and 2

The ROC curves illustrating the sensitivity and specificity of models 1 and 2, under the scenario of a weak association between disease and DSL, is shown in Figure 5.6.

In the example shown in Figure 5.6A, two dataset were simulated: (1) no association with NST ($t_m = t_f = 0.4$) and (2) weak association ($f_0 = 0.11, f_1 = 0.13, f_2 = 0.15$) with NST ($t_m = t_f = 0.4$) but no imprinting ($g=1$). We see that for model 1, true positives in dataset 2 are attenuated by the NST, leading to poor sensitivity. At the same time, false positives in dataset 1 are inflated because of NST, leading to a poor specificity and an AUC of 0.31. On the other hand, adjusting NST for model 2 led to an AUC of 0.65.

For the example in Figure 5.6B, two other datasets are simulated: (1) no association with PST ($t_m = 0.5, t_f = 0.3$) and (2) weak association ($f_0 = 0.11, f_1 = 0.13, f_2 = 0.15$) and imprinting ($g = 0.6$) with PST ($t_m = 0.5, t_f = 0.3$). We see that true positives in dataset 2 are attenuated by PST, and false positives in dataset 1 are inflated, leading to poor sensitivity and specificity, and an AUC of 0.33. However, model 2 yielded an AUC of 0.69 since the bias due to PST is adjusted.

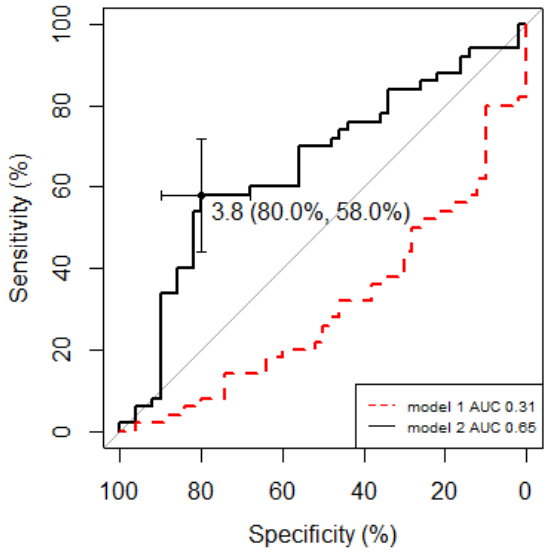
We simulated similar scenarios with a stronger association signal ($f_0 = 0.1, f_1 = 0.2, f_2 = 0.3$). The AUC for model 2 are close to 1, when there is either NST or PST. However, the AUC for model 1 remained around 0.3 for both NST and PST (results not shown here).

5.6 Discussion

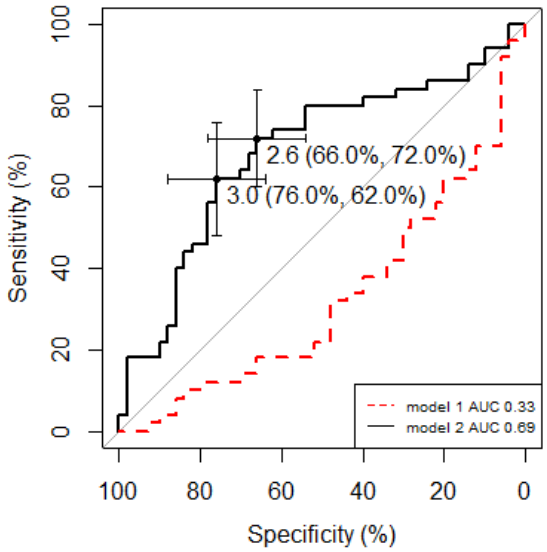
The inflation and attenuation of R_1 and R_2 as a result of change in t_m and t_f due to the presence of MST or PST are similar to what we observed for NST in previous work. Restoration of the true parameter estimates can be achieved using the sex-of-parent-specific transmission offset in a similar fashion. When MST and PST occur in the presence of imprinting effect, the measured parameter T could be masked. For example, if mother over-transmits or father under-transmits when there is paternal over-expression, the imprinting effect will not be observed as significant in model 1. Similarly, if mother under-transmits or father over-transmits when there is maternal over-expression, the measured imprinting parameter will be less significant either when MST/PST is not adjusted for.

Figure 5.6: ROC curve for weak association

A NST, with association and no imprinting



B PST, with association and maternal imprinting



(A) $t_m = t_f = 0.4, f_0 = 0.11, f_1 = 0.13, f_2 = 0.15, g = 1$

(B) $t_m = 0.5, t_f = 0.3, f_0 = 0.11, f_1 = 0.13, f_2 = 0.15, g = 0.6$

The reduction of the imprinting effect due to MST or PST is considerably more problematic than inflation because imprinting does not conventionally lie within the scope of genetic association studies. If preliminary results on imprinting are negative, it might be unlikely to pursue the investigation, while in fact, imprinting could be masked due to a ST. Therefore, when one wants to investigate the presence of imprinting effect, loglinear model 2 with ST offset adjustment should be considered as the first option for detecting true signals in association studies.

There are other study designs proposed to measure parent-of-origin (imprinting) effect. A popular design is to use case-mother duos, which are easier to recruit, instead of case-trios. For example, Ainsworth et al. [155] collapsed Weinberg et al. 15 [44] and 16 [45] mother-father-child (MFC) categories into 7 categories which are identified only by maternal and child genotype (MC). Even though it is easier to recruit mother-fetal pairs than case-parent trios, there is a difference of 8

parameters that can be estimated in the case-trios study design compared to mother-child duos, which allows more genetic or non-genetic factors to be considered in the model. Ainsworth et al. [155] also relies on the prior knowledge of minor allele frequency and mating type frequencies, which requires extra recruitments of unrelated controls to estimate MAF, and parents of controls and/or control-mother pairs to estimate mating type frequencies, to successfully fit a non-saturated model. Robustness against population stratification can also be affected. Most importantly, the paternal transmission of allele cannot be traced. Therefore, such models are not appropriate for the purpose of our study.

Genomic imprinting is an important epigenetic effect. More than 1% of all mammalian genes are believed to be imprinted. A database is available for imprinted genes [43] (http://igc.otago.ac.nz/) which provides a more comprehensive understanding of how genes behave under the influence of imprinting effect. Therefore, it is crucial to address the aspect of ST in order to correctly characterize the functions of genes, and their mechanisms of inheritance.

A limitation of our study is that the ST probabilities t_m and t_f used to adjust for MST and PST need to be computed separately from a control-trios population. We rely on the availability of such control-trios population recruited in consortia such as the HapMap project. The complete coverage of the human genome has now been made possible by the whole genome sequencing (WGS) technology. With this knowledge, we believe that majority of the TRD loci could be identified and assessed, once such control-trios data becomes available.

5.7 Appendix

5.7.1: Derivation of models 1 (without ST offset) and 2 (with ST offset)

5.7.1.1 Derivation of the general loglinear model

Let M, F, and C represent the mother, father and child genotypes respectively. The 16 MFC genotype categories are described in Table 5.1. Let n_{MFC} represents the number of trios with genotypes MFC, n the sample size, and D the disease status of the child, the probability of each MFC cell in Table 5.1 can be written as:

$$P[MFC|D] = E\left[\frac{n_{MFC}}{n}|D\right] = \frac{P[D|MFC]P[C|MF]P[MF]}{P[D]} \quad (5.6)$$

where

P[D|MFC] = Probability that the child is affected given a trio genotype MFC

P[C|MF] = Probability that the child genotype is C given parental genotypes MF

P[MF] = Probability of mating type MF for the parents

P[D] = disease prevalence

Since we assume that there is imprinting effect on the disease status of the child, *P[D|MFC]* can no longer be simplified to *P[D|C]*, as it depends on both parental genotypes. For C = 0, *P[D|MFC]* = f_0 . Equation 5.6 can be re-written as:

$$log\left\{E\left[\frac{n_{MFC}}{n}|D\right]\right\} = log\ P[D|MFC] + log\ P[C|MF] + log\ P[MF] - log\ P[D] \quad (5.7)$$

where $P[D|MFC] = f_0 R_c T$ and R_c and T are listed as the last 2 columns of Table 5.1.

Using the notations $P[C|MF] = \tau_{MFC}$, $P[MF] = \mu_{MF}$ (see Table 5.1), and $P[D] = d$, we obtain:

$$log\{E\ [n_{MFC}|D]\} = log(f_0 R_c T) + log\ \tau_{MFC} + log\ \mu_{MF} + log\ n - log\ d$$

$$= log\left(\frac{f_0 n}{d}\right) + log\ \tau_{MFC} + log\ \mu_{MF} + log(R_c) + log(T)$$

$$= log\left(\frac{f_0 n}{d}\right) + log\ \tau_{MFC} + log\ \mu_{MF} + \beta_c + \eta \quad (5.8)$$

where $\beta_c = log(R_c)$ and $\eta = log(T)$, of which the latter depends on the genotype combination of the trio. Model 1 described in the paper corresponds to the scenario where $t_m = t_f = 0.5$ (Mendelian transmission), which are substituted into τ_{MFC}. Model 2 corresponds to the scenario

where t_m and t_f can be different and are not restricted to 0.5 taking on values between 0 and 1, excluding 0 and 1.

5.7.1.2 Statistical equation for model 1

In order to fit the model described in equation 5.8, we use different grouping schemes for models 1 and 2. For model 1, the terms $log(\tau_{MFC})$ and $log\ (\mu_{MF})$ are grouped together as φ_{MFC}. Since t_m and t_f are assume to be 0.5 in this model, φ_{MFC} is the same within each mating type stratum (Table 5.1). We use S to be the indicator for each mating type stratum, then $\varphi_{MFC} = \varphi_S$, where S ranges from 1 to 6. Since imprinting parameter exists only for genotype child 1 categories (C = 1) when the disease allele is inherited from the mother, we can write:

$$\eta = \zeta_M I_{[C=1,maternal]}$$

To derive the statistical equation for model 1, equation 5.8 can be re-written as

$$log\{E\ [n_{MFC}|D]\} = log\left(\frac{f_0 n}{d}\right) + \sum_j \varphi_j I_{[S=j]} + \beta_1 I_{[C=1]} + \beta_2 I_{[C=2]} + \zeta_M I_{[C=1,maternal]}$$

We can then absorb the constant term $\frac{f_0 n}{d}$ into the summation of φ_j terms and have

$$log\{E\ [n_{MFC}|D]\} = \sum_j log\left[\left(\frac{f_0 n}{d}\right) exp(\varphi_j)\right] I_{[S=j]} + \beta_1 I_{[C=1]} + \beta_2 I_{[C=2]} + \zeta_M I_{[C=1,maternal]}$$

By noting γ_j as the first term of the above equation, model 1 can be written as:

$$log\{E\ [n_{MFC}|D]\} = \sum_j \gamma_j I_{[S=j]} + \beta_1 I_{[C=1]} + \beta_2 I_{[C=2]} + \zeta_M I_{[C=1,maternal]}$$

Since there are 6 strata of MF mating types, we fit the model with an intercept for stratum 6 mating type and obtained:

$$log\{E\ [n_{MFC}|D]\} = \gamma_6 + \sum_{j=1}^{5} \gamma_j I_{[S=j]} + \beta_1 I_{[C=1]} + \beta_2 I_{[C=2]} + \zeta_M I_{[C=1,maternal]} \qquad (5.9)$$

5.7.1.3 Statistical equation for model 2

For model 2, we separate the terms $log(\tau_{MFC})$ and $log\ (\mu_{MF})$, and replace $log(\tau_{MFC})$ by an offset given specific values of t_m and t_f (Table 5.1), and again estimate $log\ (\mu_{MF}) = log(\mu_j)$. Therefore, equation 5.8 can be re-written as:

$$log\{E\ [n_{MFC}|D]\} = log\left(\frac{f_0 n}{d}\right) + \sum_j log\ \mu_j\ I_{[S=j]} + log\ \tau_{MFC} + \beta_1 I_{[C=1]} + \beta_2 I_{[C=2]} + \zeta_M I_{[C=1,maternal]}$$

$$= \sum_j log\left(\frac{f_0 n}{d}\right) \mu_j\ I_{[S=j]} + log\ \tau_{MFC} + \beta_1 I_{[C=1]} + \beta_2 I_{[C=2]} + \zeta_M I_{[C=1,maternal]}$$

Replacing $log\left(\frac{f_0 n}{d}\right) \mu_j$ as α_j, model 2 can be written as:

$$log\{E\ [n_{MFC}|D]\} = \sum_j \alpha_j\ I_{[S=j]} + log\ \tau_{MFC} + \beta_1 I_{[C=1]} + \beta_2 I_{[C=2]} + \zeta_M I_{[C=1,maternal]}$$

We then fit the model with an intercept, and obtain:

$$log\{E\ [n_{MFC}|D]\} = \alpha_6 + \sum_{j=1}^{5} \alpha_j\ I_{[S=j]} + log\ \tau_{MFC} + \beta_1 I_{[C=1]} + \beta_2 I_{[C=2]} + \zeta_M I_{[C=1,maternal]} \quad (5.10)$$

The final statistical formula for model 1 is written in equation (5.9) and for model 2 in equation (5.10).

5.7.2: Non-Central Chi-square Likelihood for model 1 (without ST offset) and model 2 (with ST offset)

To perform the Likelihood Ratio Test (LRT) in assessing significance of association between the disease phenotype and DSL, we set up a null model for both models 1 and 2 with null hypothesis $H_0 : \beta_1 = \beta_2 = \zeta_M = 0$. The corresponding LRT test statistic, which is the difference in deviance between null and full model, has an asymptotic Chi-Square distribution with 3 degrees of freedom

accounting for the extra terms R_1, R_2 and T. Agresti [161] showed that when the alternative hypothesis is true, the resulting LRT is a chi-square statistic with a non-centrality parameter (NCP):

$$\lambda = 2n \sum_{MFC} \pi_{MFC}(M_a) log \left(\frac{\pi_{MFC}(M_a)}{\pi_{MFC}(M_0)}\right)$$

where $\pi_{MFC}(M_a)$ is the true probability of each cell with MFC combination, and $\pi_{MFC}(M_0)$ is the probability under the null hypothesis. We also denoted the degree of freedom as υ, which is 3 in our LRT because for the 3 variables R_1, R_2 and T present in the alternative model but not in the corresponding null model.

To calculate Type 1 error and power comparable to our theoretical values, we require the exact likelihood. Our likelihood for the alternative hypothesis shown in equation 5.6 can be written as:

$$\pi_{MFC}(M_a) = \frac{f_0 R_c T \tau_{MFC} \mu_{MF}}{d}$$

In the presence of TRD, even when the null hypothesis is true, the LRT still has a non-Central Chi-square distribution. The null model is different for models 1 and 2 because TRD is being adjusted in the offset of model 2 but not in model 1. Under the null hypothesis, $P[D|MFC] = P[D]$, and hence, $f_0 R_c T/d = 1$. The likelihoods for models 1 and 2 under null hypothesis are:

$$\pi_{MFC}(M_{01}) = \mu_{MF}\, \tau_{MFC}\, [0.5]$$

and

$$\pi_{MFC}(M_{02}) = \mu_{MF}\, \tau_{MFC}[t]$$

Under the alternative hypothesis, NCP for model 1 is:

$$\lambda_1 = 2n \sum_{MFC} \frac{f_0 R_c T \tau_{MFC} \mu_{MF}[t]}{d}\, log \left(\frac{f_0 R_c T \tau_{MFC}[t]}{\tau_{MFC}[0.5]\, d}\right) \quad (5.11)$$

and the NCP for model 2 is:

$$\lambda_2 = 2n \sum_{MFC} \frac{f_0 R_c T \tau_{MFC} \mu_{MF}[t]}{d} \; log\left(\frac{f_0 R_c T}{d}\right) \tag{5.12}$$

When t is not equal to 0.5, even though there is no association signal, the LRT is still a NCP chi-square statistic. The NCP for model 1 is 0 when both t = 0.5 (Mendelian transmission) and $\frac{f_0 R_c T}{d}$=1 (no association). Therefore, null hypothesis for model 1 requires both Mendelian transmission and no association between disease and DSL. However, since TRD has already been adjusted for in model 2, the NCP is 0 when $\frac{f_0 R_c T}{d}$=1 (no association).

Chapter 6

Summary and discussion

The role of TRD in the formation and maintenance of the human gene pool is considerably obscure. Human studies on TRD have mainly be prompted by successful findings in plant and animal studies. Multiple diseases associated with TRD loci have been found, but links between the mechanisms of TRD and the disease etiology have not been established, except perhaps for conditions related to embryo viability. The prevalence of TRD has not yet been determined by genetic studies and hence, the impact of TRD on either common or rare diseases is largely unknown. However, with the availability of next generation sequencing technology and large-scale recruiting effort such as the HapMap project on case- and control-trios, the possibility of mapping all TRD loci will be possible.

There are various forms of TRD, as described in Chapter 3, each dictated by specific biological mechanisms. We selected two simplest types of TRD, the NST and ST in order to demonstrate its effect on the results of family-based genetic association studies. Most of the existing statistical methodologies have a common assumption on Mendelian inheritance, and the models are not valid if this assumption is violated. The loglinear model is a convenient statistical tool for us to assess the effect of TRD on association results in case-parent study design. The readily available component in the likelihood function provides a natural way of extending the model to accommodate the effect of TRD, and to correct for it. Fitting the loglinear model with an offset does not require more computing time, and hence, can be applied to a large scale association study with the whole genome sequenced data and a large sample size. It offers a simple solution to the identification of an additional source of bias which could potentially confirm or refute study results from current literature.

Modeling ST in the loglinear model, however, poses a challenge. The imprinting effect could also co-exist with other epigenetic effect such as maternal effect, or maternal-fetal genotype interaction. Incorporating these factors into the loglinear model has not been investigated in this thesis, but is likely to be pursued in the continuing development of the method in the future. Currently, we have exclusively modeled imprinting and child effects. Although adding maternal and maternal-fetal

interaction effects into the loglinear model does not likely require more complex theoretical basis, the saturation of model parameters can constraint the generalization of the method. This might require changing the study design to acquire more information and degrees of freedom. Also, since these factors interplay with each other, identifying and dissecting the exact effect size in a single model could be complicated. Interpretation on the resulting parameter estimates might also require further biological evidence.

Future work in generalizing this method to a wider context and scale will be made possible with the availability of appropriate datasets and advancement in the knowledge of human genetics in general. The research carried out in this thesis provides evidence of the impact of TRD on genetic studies and a proof of concept that such effect can be adjusted to restore correct inference. Implication on existing findings in current literature will unfold as research progresses.

TRD is an under-explored phenomenon with features that can impact studies in three different genetic fields. The prospect of increasing awareness and understanding of TRD can produce major breakthroughs in these areas, such as re-assessing current research findings on DSL, identifying rare variants, and developing the link between TRD mechanisms and various disease etiologies. These could lead to more accurate and comprehensive knowledge about the relationships between our genome and a vast array of human diseases.

REFERENCES

1. Huang, L.O., A. Labbe, and C. Infante-Rivard, *Transmission ratio distortion: review of concept and implications for genetic association studies.* Hum Genet, 2013. **132**(3): p. 245-63.
2. Tsui, L.C., et al., *Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker.* Science., 1985. **230**(4729): p. 1054-7.
3. Gusella, J.F., et al., *A polymorphic DNA marker genetically linked to Huntington's disease.* Nature., 1983. **306**(5940): p. 234-8.
4. Walsh, T. and M.C. King, *Ten genes for inherited breast cancer.* Cancer Cell., 2007. **11**(2): p. 103-5.
5. McPherson, R., et al., *A common allele on chromosome 9 associated with coronary heart disease.* Science., 2007. **316**(5830): p. 1488-91.
6. Willer, C.J., et al., *Newly identified loci that influence lipid concentrations and risk of coronary artery disease.* Nat Genet., 2008. **40**(2): p. 161-9.
7. Hampe, J., et al., *A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1.* Nat Genet., 2007. **39**(2): p. 207-11.
8. Parkes, M., et al., *Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility.* Nat Genet., 2007. **39**(7): p. 830-2.
9. Rioux, J.D., et al., *Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis.* Nat Genet., 2007. **39**(5): p. 596-604.
10. Gudmundsson, J., et al., *Common sequence variants on 2p15 and Xp11.22 confer susceptibility to prostate cancer.* Nat Genet., 2008. **40**(3): p. 281-3.
11. Stacey, S.N., et al., *Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer.* Nat Genet., 2007. **39**(7): p. 865-9.
12. Wolpin, B.M., et al., *Genome-wide association study identifies multiple susceptibility loci for pancreatic cancer.* Nat Genet., 2014. **46**(9): p. 994-1000. doi: 10.1038/ng.3052.
13. Hakonarson, H., et al., *A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene.* Nature., 2007. **448**(7153): p. 591-4.
14. Todd, J.A., et al., *Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes.* Nat Genet., 2007. **39**(7): p. 857-64.
15. Saxena, R., et al., *Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels.* Science., 2007. **316**(5829): p. 1331-6.
16. Scott, L.J., et al., *A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants.* Science., 2007. **316**(5829): p. 1341-5.
17. Sladek, R., et al., *A genome-wide association study identifies novel risk loci for type 2 diabetes.* Nature., 2007. **445**(7130): p. 881-5.
18. Purcell, S.M., et al., *Common polygenic variation contributes to risk of schizophrenia and bipolar disorder.* Nature., 2009. **460**(7256): p. 748-52. doi: 10.1038/nature08185.
19. Shi, J., et al., *Common variants on chromosome 6p22.1 are associated with schizophrenia.* Nature., 2009. **460**(7256): p. 753-7. doi: 10.1038/nature08192.
20. Shi, Y., et al., *Common variants on 8p12 and 1q24.2 confer risk of schizophrenia.* Nat Genet., 2011. **43**(12): p. 1224-7.
21. Stefansson, H., et al., *Common variants conferring risk of schizophrenia.* Nature., 2009. **460**(7256): p. 744-7. doi: 10.1038/nature08186.
22. Cichon, S., et al., *Genome-wide association study identifies genetic variation in neurocan as a susceptibility factor for bipolar disorder.* Am J Hum Genet., 2011. **88**(3): p. 372-81.
23. Ferreira, M.A., et al., *Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder.* Nat Genet., 2008. **40**(9): p. 1056-8.
24. Eichler, E.E., et al., *Missing heritability and strategies for finding the underlying causes of complex disease.* Nat Rev Genet., 2010. **11**(6): p. 446-50.

25. Gibson, G., *Rare and common variants: twenty arguments.* Nat Rev Genet., 2011. **13**(2): p. 135-45.
26. Manolio, T.A., et al., *Finding the missing heritability of complex diseases.* Nature., 2009. **461**(7265): p. 747-53.
27. Zuk, O., et al., *Searching for missing heritability: designing rare variant association studies.* Proc Natl Acad Sci U S A., 2014. **111**(4): p. E455-64.
28. Boycott, K.M., et al., *Rare-disease genetics in the era of next-generation sequencing: discovery to translation.* Nat Rev Genet., 2013. **14**(10): p. 681-91. doi: 10.1038/nrg3555.
29. He, Z., et al., *Rare-variant extensions of the transmission disequilibrium test: application to autism exome sequence data.* Am J Hum Genet., 2014. **94**(1): p. 33-46.
30. Li, B. and S.M. Leal, *Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data.* Am J Hum Genet., 2008. **83**(3): p. 311-21.
31. Bush, W.S. and J.H. Moore, *Chapter 11: Genome-Wide Association Studies.* PLoS Computational Biology, 2012. **8**(12): p. e1002822.
32. Bosker, F.J., et al., *Poor replication of candidate genes for major depressive disorder using genome-wide association data.* Mol Psychiatry, 2011. **16**(5): p. 516-532.
33. Ioannidis, J.P., et al., *Replication validity of genetic association studies.* Nat Genet., 2001. **29**(3): p. 306-9.
34. Price, A.L., et al., *Principal components analysis corrects for stratification in genome-wide association studies.* Nat Genet., 2006. **38**(8): p. 904-9.
35. Devlin, B. and K. Roeder, *Genomic control for association studies.* Biometrics., 1999. **55**(4): p. 997-1004.
36. Spielman, R.S., R.E. McGinnis, and W.J. Ewens, *Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM).* American Journal of Human Genetics, 1993. **52**(3): p. 506.
37. Horvath, S., X. Xu, and N.M. Laird, *The family based association test method: strategies for studying general genotype--phenotype associations.* Eur J Hum Genet., 2001. **9**(4): p. 301-6.
38. Laird, N.M. and C. Lange, *Family-based designs in the age of large-scale gene-association studies.* Nat Rev Genet, 2006. **7**(5): p. 385-394.
39. Martin, E.R., et al., *A test for linkage and association in general pedigrees: the pedigree disequilibrium test.* Am J Hum Genet., 2000. **67**(1): p. 146-54.
40. Guilmatre, A. and A.J. Sharp, *Parent of origin effects.* Clin Genet., 2012. **81**(3): p. 201-9.
41. Lawson, H.A., J.M. Cheverud, and J.B. Wolf, *Genomic imprinting and parent-of-origin effects on complex traits.* Nat Rev Genet., 2013. **14**(9): p. 609-17. doi: 10.1038/nrg3543.
42. Lokody, I., *Gene expression: Consequences of parent-of-origin effects.* Nat Rev Genet, 2014. **15**(3): p. 145-145.
43. Morison, I.M., C.J. Paton, and S.D. Cleverley, *The imprinted gene and parent-of-origin effect database.* Nucleic Acids Res., 2001. **29**(1): p. 275-6.
44. Weinberg, C.R., 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, 1998. **62**(4): p. 969-78.
45. Weinberg, C.R., *Methods for Detection of Parent-of-Origin Effects in Genetic Studies of Case-Parents Triads.* American Journal of Human Genetics, 1999. **65**(1): p. 229-235.
46. Cordell, H.J., B.J. Barratt, and D.G. Clayton, *Case/pseudocontrol analysis in genetic association studies: A unified framework for detection of genotype and haplotype associations, gene-gene and gene-environment interactions, and parent-of-origin effects.* Genetic Epidemiology, 2004. **26**(3): p. 167-85.
47. Cordell, H.J. and D.G. Clayton, *A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes.* Am J Hum Genet., 2002. **70**(1): p. 124-41.

48. Self, S.G., et al., *On estimating HLA/disease association with application to a study of aplastic anemia.* Biometrics., 1991. **47**(1): p. 53-61.
49. Schaid, D.J., *General score tests for associations of genetic markers with disease using cases and their parents.* Genet Epidemiol, 1996. **13**(5): p. 423-49.
50. Gjessing, H.K. and R.T. Lie, *Case-parent triads: estimating single- and double-dose effects of fetal and maternal disease gene haplotypes.* Annals of human genetics, 2006. **70**(Pt 3): p. 382-96.
51. Kistner, E.O., C. Infante-Rivard, and C.R. Weinberg, *A method for using incomplete triads to test maternally mediated genetic effects and parent-of-origin effects in relation to a quantitative trait.* American journal of epidemiology, 2006. **163**(3): p. 255-61.
52. Kistner, E.O., M. Shi, and C.R. Weinberg, *Using cases and parents to study multiplicative gene-by-environment interaction.* American journal of epidemiology, 2009. **170**(3): p. 393-400.
53. Shi, M., David M. Umbach, and Clarice R. Weinberg, *Identification of Risk-Related Haplotypes with the Use of Multiple SNPs from Nuclear Families.* American Journal of Human Genetics, 2007. **81**(1): p. 53-66.
54. Sinsheimer, J.S., C.G. Palmer, and J.A. Woodward, *Detecting genotype combinations that increase risk for disease: maternal-fetal genotype incompatibility test.* Genetic Epidemiology, 2003. **24**(1): p. 1-13.
55. Kistner, E.O. and C.R. Weinberg, *Method for using complete and incomplete trios to identify genes related to a quantitative trait.* Genet Epidemiol., 2004. **27**(1): p. 33-42.
56. Kistner, E.O. and C.R. Weinberg, *A method for identifying genes related to a quantitative trait, incorporating multiple siblings and missing parents.* Genet Epidemiol., 2005. **29**(2): p. 155-65.
57. Wheeler, E. and H.J. Cordell, *Quantitative trait association in parent offspring trios: Extension of case/pseudocontrol method and comparison of prospective and retrospective approaches.* Genet Epidemiol., 2007. **31**(8): p. 813-33.
58. Clayton, D.G., *Program define pseudocc in Stata.* https://www-gene.cimr.cam.ac.uk/staff/clayton/software/stata/genassoc/pseudocc.ado, 2003.
59. Zhou, J.Y., Y.Q. Hu, and W.K. Fung, *A simple method for detection of imprinting effects based on case-parents trios.* Heredity (Edinb), 2007. **98**(2): p. 85-91.
60. Hu, Y.Q., J.Y. Zhou, and W.K. Fung, *An extension of the transmission disequilibrium test incorporating imprinting.* Genetics, 2007. **175**(3): p. 1489-504.
61. Hu, Y.Q., et al., *The transmission disequilibrium test and imprinting effects test based on case-parent pairs.* Genetic Epidemiology, 2007. **31**(4): p. 273-87.
62. Xia, F., J.Y. Zhou, and W.K. Fung, *A powerful approach for association analysis incorporating imprinting effects.* Bioinformatics, 2011. **27**(18): p. 2571-7.
63. Xia, F., J.Y. Zhou, and W.K. Fung, *Powerful tests for association on quantitative trait loci incorporating imprinting effects.* J Hum Genet., 2013. **58**(6): p. 384-90.
64. Zhou, J.Y., et al., *Detection of parent-of-origin effects based on complete and incomplete nuclear families with multiple affected children.* Hum Hered, 2009. **67**(1): p. 1-12.
65. Becker, T., M.P. Baur, and M. Knapp, *Detection of parent-of-origin effects in nuclear families using haplotype analysis.* Hum Hered, 2006. **62**(2): p. 64-76.
66. Zhou, J.Y., et al., *Detection of parent-of-origin effects in complete and incomplete nuclear families with multiple affected children using multiple tightly linked markers.* Hum Hered, 2009. **67**(2): p. 116-27.
67. Zhou, J.Y., et al., *Detection of parent-of-origin effects using general pedigree data.* Genetic Epidemiology, 2010. **34**(2): p. 151-8.
68. Zhou, J.Y., et al., *A powerful parent-of-origin effects test for qualitative traits incorporating control children in nuclear families.* J Hum Genet, 2012. **57**(8): p. 500-7.
69. He, F., et al., *Detection of parent-of-origin effects for quantitative traits in complete and incomplete nuclear families with multiple children.* American journal of epidemiology, 2011. **174**(2): p. 226-33.

70. Pardo-Manuel de Villena, F. and C. Sapienza, *Nonrandom segregation during meiosis: the unfairness of females.* Mammalian Genome, 2001. **12**(5): p. 331-339.
71. Naumova, A.K., et al., *Parental origin-dependent, male offspring-specific transmission-ratio distortion at loci on the human X chromosome.* American Journal of Human Genetics, 1998. **62**(6): p. 1493-9.
72. Paterson, A., L. Sun, and X.Q. Liu, *Transmission ratio distortion in families from the Framingham Heart Study.* BMC genetics, 2003. **4**(Suppl 1): p. S48.
73. Paterson, A., et al. *Transmission-ratio distortion in the Framingham Heart Study*. 2009. BioMed Central Ltd.
74. Santos, P.S., et al., *Assessment of transmission distortion on chromosome 6p in healthy individuals using tagSNPs.* European Journal of Human Genetics, 2009. **17**(9): p. 1182-9.
75. Zollner, S., et al., *Evidence for extensive transmission distortion in the human genome.* American Journal of Human Genetics, 2004. **74**(1): p. 62-72.
76. Infante-Rivard, C. and C.R. Weinberg, *Parent-of-origin transmission of thrombophilic alleles to intrauterine growth-restricted newborns and transmission-ratio distortion in unaffected newborns.* American journal of epidemiology, 2005. **162**(9): p. 891-7.
77. Meyer, W.K., et al., *Evaluating the evidence for transmission distortion in human pedigrees.* Genetics, 2012. **191**(1): p. 215-32.
78. Greenwood, C.M. and K. Morgan, *The impact of transmission-ratio distortion on allele sharing in affected sibling pairs.* American Journal of Human Genetics, 2000. **66**(6): p. 2001-4.
79. Hastings, I.M., *Germline selection: population genetic aspects of the sexual/asexual life cycle.* Genetics, 1991. **129**(4): p. 1167-76.
80. Yang, L., et al., *Parental effect of DNA (Cytosine-5) methyltransferase 1 on grandparental-origin-dependent transmission ratio distortion in mouse crosses and human families.* Genetics, 2008. **178**(1): p. 35-45.
81. Dean, N.L., et al., *Transmission ratio distortion in the myotonic dystrophy locus in human preimplantation embryos.* European Journal of Human Genetics, 2006. **14**(3): p. 299-306.
82. Riess, O., et al., *Transmission distortion of the mutant alleles in spinocerebellar ataxia.* Human genetics, 1997. **99**(2): p. 282-4.
83. Chevin, L.M. and F. Hospital, *The hitchhiking effect of an autosomal meiotic drive gene.* Genetics, 2006. **173**(3): p. 1829-32.
84. Huang , L.O., A. Labbe, and C. Infante-Rivard, *Impact of Transmission Ratio Distortion on the interpretation of genetic association studies and evolution of population parameters.* 6th Annual Genetic Epidemiology and Statistical Genetic Meeting, Abstract, 2011.
85. Crow, J.F., *The ultraselfish gene.* Genetics, 1988. **118**(3): p. 389.
86. Hurst, G.D. and J.H. Werren, *The role of selfish genetic elements in eukaryotic evolution.* Nat Rev Genet, 2001. **2**(8): p. 597-606.
87. The International HapMap Consortium, *A haplotype map of the human genome.* Nature, 2005. **437**(7063): p. 1299-320.
88. Hanchard, N., et al., *An investigation of transmission ratio distortion in the central region of the human MHC.* Genes and immunity, 2005. **7**(1): p. 51-58.
89. Naumova, A., et al., *Transmission ratio distortion of X chromosomes among male offspring of females with skewed X inactivation.* Developmental Genetics, 1995. **17**(3): p. 198-205.
90. Naumova, A.K., C.M. Greenwood, and K. Morgan, *Imprinting and deviation from Mendelian transmission ratios.* Genome, 2001. **44**(3): p. 311-20.
91. Bauer, H., et al., *The Nucleoside Diphosphate Kinase Gene Nme3 Acts as Quantitative Trait Locus Promoting Non-Mendelian Inheritance.* PLoS Genet, 2012. **8**(3).
92. Bauer, H., et al., *The t-complex-encoded guanine nucleotide exchange factor Fgd2 reveals that two opposing signaling pathways promote transmission ratio distortion in the mouse.* Genes Dev, 2007. **21**(2): p. 143-7.

93. Casellas, J., et al., *Genome scans for transmission ratio distortion regions in mice.* Genetics, 2012. **191**(1): p. 247-59.
94. Haston, C.K., D.G. Humes, and M. Lafleur, *X chromosome transmission ratio distortion in Cftr +/- intercross-derived mice.* BMC Genet, 2007. **8**: p. 23.
95. LeMaire-Adkins, R. and P.A. Hunt, *Nonrandom segregation of the mouse univalent X chromosome: evidence of spindle-mediated meiotic drive.* Genetics, 2000. **156**(2): p. 775.
96. Martin-DeLeon, P.A., et al., *Spam1-associated transmission ratio distortion in mice: elucidating the mechanism.* Reprod Biol Endocrinol, 2005. **3**: p. 32.
97. Schulz, R., et al., *Nondisjunction and transmission ratio distortion ofChromosome 2 in a (2.8) Robertsonian translocation mouse strain.* Mammalian Genome, 2006. **17**(3): p. 239-47.
98. Taveau, M., et al., *Bidirectional transcriptional activity of the Pgk1 promoter and transmission ratio distortion in Capn3-deficient mice.* Genomics, 2004. **84**(3): p. 592-5.
99. Underkoffler, L.A., et al., *Transmission Ratio Distortion in Offspring of Mouse Heterozygous Carriers of a (7.18) Robertsonian Translocation.* Genetics, 2005. **169**(2): p. 843-8.
100. Veron, N., et al., *Retention of gene products in syncytial spermatids promotes non-Mendelian inheritance as revealed by the t complex responder.* Genes Dev, 2009. **23**(23): p. 2705-10.
101. Wu, G., et al., *Maternal Transmission Ratio Distortion at the Mouse Om Locus Results From Meiotic Drive at the Second Meiotic Division.* Genetics, 2005. **170**(1): p. 327-34.
102. Eversley, C.D., et al., *Genetic mapping and developmental timing of transmission ratio distortion in a mouse interspecific backcross.* BMC Genet, 2010. **11**: p. 98.
103. Novitski, E., *Non-random disjunction in Drosophila.* Genetics, 1951. **36**(3): p. 267.
104. Sturtevant, A., *Preferential segregation in triplo-IV females of Drosophila melanogaster.* Genetics, 1936. **21**(4): p. 444.
105. Zimmering, S., *A genetic study of segregation in a translocation heterozygote in Drosophila.* Genetics, 1955. **40**(6): p. 809.
106. Aparicio, J.M., et al., *Evidence of subtle departures from Mendelian segregation in a wild lesser kestrel (Falco naumanni) population.* Heredity, 2010. **105**(2): p. 213-9.
107. Shemer, R., et al., *Structure of the imprinted mouse Snrpn gene and establishment of its parental-specific methylation pattern.* Proceedings of the National Academy of Sciences, 1997. **94**(19): p. 10267-10272.
108. Croteau, S., et al., *Inheritance patterns of maternal alleles in imprinted regions of the mouse genome at different stages of development.* Mammalian genome : official journal of the International Mammalian Genome Society, 2002. **13**(1): p. 24-9.
109. Sazhenova, E.A. and I.N. Lebedev, *[Epimutations of the KCNQ1OT1 imprinting center of chromosome 11 in early human embryo lethality].* Genetika, 2008. **44**(12): p. 1609-16.
110. Lange, K., *Mathematical and Statistical Methods for Genetic Analysis.* New York: Springer-Verlag, 1997.
111. De Rango, F., et al., *A novel sampling design to explore gene-longevity associations: the ECHA study.* European Journal of Human Genetics, 2007. **16**(2): p. 236-242.
112. Eaves, I.A., et al., *Transmission ratio distortion at the INS-IGF2 VNTR.* Nature genetics, 1999. **22**(4): p. 324.
113. Honeywell, C., et al., *Apparent transmission distortion of a pericentric chromosome one inversion in a large multi-generation pedigree.* Am J Med Genet A, 2012. **158A**(6): p. 1262-8.
114. Blyth, K., et al., *Runx2 in normal tissues and cancer cells: A developing story.* Blood Cells Mol Dis, 2010. **45**(2): p. 117-23.
115. Paterson, A.D. and A. Petronis, *Transmission ratio distortion in females on chromosome 10p11 p15.* American journal of medical genetics, 1999. **88**(6): p. 657-661.
116. Bettencourt, C., et al., *Segregation distortion of wild-type alleles at the Machado-Joseph disease locus: a study in normal families from the Azores islands (Portugal).* J Hum Genet, 2008. **53**(4): p. 333-9.

117. Shoubridge, C., et al., *Is there a Mendelian transmission ratio distortion of the c.429_452dup(24bp) polyalanine tract ARX mutation?* European Journal of Human Genetics, 2012.
118. Mitchell, A.A., D.J. Cutler, and A. Chakravarti, *Undetected genotyping errors cause apparent overtransmission of common alleles in the transmission/disequilibrium test.* Am J Hum Genet., 2003. **72**(3): p. 598-610.
119. Paterson, A.D., et al., *Transmission-ratio distortion in the Framingham Heart Study.* BMC Proc., 2009. **3**(Suppl 7): p. S51.
120. Becker, T., et al., *Transmission ratio distortion and maternal effects confound the analysis of modulators of cystic fibrosis disease severity on 19q13.* European Journal of Human Genetics, 2007. **15**(7): p. 774-8.
121. Imboden, M., et al., *Female predominance and transmission distortion in the long-QT syndrome.* New England Journal of Medicine, 2006. **355**(26): p. 2744-51.
122. Magee, A.C. and A.E. Hughes, *Segregation distortion in myotonic dystrophy.* J Med Genet, 1998. **35**(12): p. 1045-6.
123. Botta, A., et al., *Transmission ratio distortion in the spinal muscular atrophy locus: data from 314 prenatal tests.* Neurology, 2005. **65**(10): p. 1631-5.
124. Liu, L.Y., et al., *Transmission distortion in Crohn's disease risk gene ATG16L1 leads to sex difference in disease association.* Inflamm Bowel Dis, 2012. **18**(2): p. 312-22.
125. Friedrichs, F., et al., *Evidence of transmission ratio distortion of DLG5 R30Q variant in general and implication of an association with Crohn disease in men.* Human genetics, 2006. **119**(3): p. 305-11.
126. Henckaerts, L., et al., *Transmission ratio distortion of DLG5 R30Q: evidence for prenatal selection?* Inflamm Bowel Dis, 2010. **16**(6): p. 910-1.
127. Klopocki, E., et al., *Duplications of BHLHA9 are associated with ectrodactyly and tibia hemimelia inherited in non-Mendelian fashion.* J Med Genet, 2012. **49**(2): p. 119-25.
128. Deng, L., et al., *Constructing an initial map of transmission distortion based on high density HapMap SNPs across the human autosomes.* J Genet Genomics, 2009. **36**(12): p. 703-9.
129. Evans, D., et al., *A note on the power to detect transmission distortion in parent-child trios via the transmission disequilibrium test.* Behavior genetics, 2006. **36**(6): p. 947-950.
130. Haig, D. and A. Grafen, *Genetic scrambling as a defence against meiotic drive*.* Journal of theoretical Biology, 1991. **153**(4): p. 531-558.
131. Polanski, A., *Dynamic balance of segregation distortion and selection maintains normal allele sizes at the myotonic dystrophy locus* 1.* Mathematical biosciences, 1998. **147**(1): p. 93-112.
132. Westendorp, R., et al., *Optimizing human fertility and survival.* Nat Med, 2001. **7**(8): p. 873.
133. Maher, B., *Personal genomes: The case of the missing heritability.* Nature, 2008. **456**(7218): p. 18-21.
134. Bodmer, W. and C. Bonilla, *Common and rare variants in multifactorial susceptibility to common diseases.* Nature genetics, 2008. **40**(6): p. 695-701.
135. Gorlov, I.P., et al., *Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms.* The American Journal of Human Genetics, 2008. **82**(1): p. 100-112.
136. Kryukov, G.V., L.A. Pennacchio, and S.R. Sunyaev, *Most rare missense alleles are deleterious in humans: implications for complex disease and association studies.* The American Journal of Human Genetics, 2007. **80**(4): p. 727-739.
137. Cirulli, E.T. and D.B. Goldstein, *Uncovering the roles of rare variants in common disease through whole-genome sequencing.* Nat Rev Genet, 2010. **11**(6): p. 415-425.
138. Li, B. and S.M. Leal, *Discovery of Rare Variants via Sequencing: Implications for the Design of Complex Trait Association Studies.* PLoS Genet, 2009. **5**(5): p. e1000481.
139. Vogl, C. and S. Xu, *Multipoint mapping of viability and segregation distorting loci using molecular markers.* Genetics, 2000. **155**(3): p. 1439-47.

140. Casellas, J., et al., *A flexible bayesian model for testing for transmission ratio distortion.* Genetics, 2014. **198**(4): p. 1357-67.
141. Deng, H.W. and W.M. Chen, *The power of the transmission disequilibrium test (TDT) with both case-parent and control-parent trios.* Genetical research, 2001. **78**(3): p. 289-302.
142. Labbe, A., L.O. Huang, and C. Infante-Rivard, *Transmission Ratio Distortion: A Neglected Phenomenon with Many Consequences in Genetic Analysis and Population Genetics*, in *Epigenetics and Complex Traits*, A.K. Naumova and C.M.T. Greenwood, Editors. 2013, Springer: New York Heidelberg Dordrecht London. p. 265-285.
143. Lazzeroni, L.C. and K. Lange, *A conditional inference framework for extending the transmission/disequilibrium test.* Hum Hered, 1998. **48**(2): p. 67-81.
144. Rabinowitz, D. and N. Laird, *A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information.* Hum Hered, 2000. **50**(4): p. 211-23.
145. Infante-Rivard, C., et al., *Absence of association of thrombophilia polymorphisms with intrauterine growth restriction.* New England Journal of Medicine, 2002. **347**(1): p. 19-25.
146. Kvasnicka, J., et al., *[Prevalence of thrombophilic mutations of FV Leiden, prothrombin G20210A and PAl-1 4G/5G and their combinations in a group of 1450 healthy middle-aged individuals in the Prague and Central Bohemian regions (results of FRET real-time PCR assay)].* Cas Lek Cesk, 2012. **151**(2): p. 76-82.
147. Sapru, A., et al., *4G/5G polymorphism of plasminogen activator inhibitor-1 gene is associated with mortality in intensive care unit patients with severe pneumonia.* Anesthesiology, 2009. **110**(5): p. 1086-91.
148. Alfirevic, Z., et al., *Frequency of factor II G20210A, factor V Leiden, MTHFR C677T and PAI-1 5G/4G polymorphism in patients with venous thromboembolism: Croatian case control study.* Biochemia Medica, 2010. **20**(2): p. 229-35.
149. Ariens, R.A., et al., *Role of factor XIII in fibrin clot formation and effects of genetic polymorphisms.* Blood, 2002. **100**(3): p. 743-54.
150. Kawamura, Y., et al., *Gadolinium-phthalein complexone as a contrast agent for hepatobiliary MR imaging.* Journal of computer assisted tomography, 1989. **13**(1): p. 67-70.
151. Ulvik, A., et al., *Simultaneous determination of methylenetetrahydrofolate reductase C677T and factor V G1691A genotypes by mutagenically separated PCR and multiple-injection capillary electrophoresis.* Clin Chem, 1998. **44**(2): p. 264-9.
152. Infante-Rivard, C., et al., *Thrombophilic polymorphisms and intrauterine growth restriction.* Epidemiology., 2005. **16**(3): p. 281-7.
153. Mirea, L., et al., *Strategies for genetic association analyses combining unrelated case-control individuals and family trios.* American journal of epidemiology, 2012. **176**(1): p. 70-9.
154. Yang, J. and S. Lin, *Robust partial likelihood approach for detecting imprinting and maternal effects using case-control families.* 2013: p. 249-268.
155. Ainsworth, H.F., et al., *Investigation of maternal effects, maternal-fetal interactions and parent-of-origin effects (imprinting), using mothers and their offspring.* Genetic Epidemiology, 2011. **35**(1): p. 19-45.
156. Chen, J., H. Zheng, and M.L. Wilson, *Likelihood ratio tests for maternal and fetal genetic effects on obstetric complications.* Genet Epidemiol., 2009. **33**(6): p. 526-38. doi: 10.1002/gepi.20405.
157. Shi, M., et al., *Making the most of case-mother/control-mother studies.* Am J Epidemiol., 2008. **168**(5): p. 541-7. doi: 10.1093/aje/kwn149.
158. Vermeulen, S.H., et al., *A hybrid design: case-parent triads supplemented by control-mother dyads.* Genet Epidemiol., 2009. **33**(2): p. 136-44.
159. van Den Oord, E.J. and J.K. Vermunt, *Testing for linkage disequilibrium, maternal effects, and imprinting with (In)complete case-parent triads, by use of the computer program LEM.* American Journal of Human Genetics, 2000. **66**(1): p. 335-8.

160. Infante-Rivard, C., et al., *Thrombophilic polymorphisms and intrauterine growth restriction.* Epidemiology, 2005. **16**(3): p. 281-7.
161. Agresti, A., *Building and Applying Logistic Regression Models*, in *Cateigorical Data Analysis 2nd Ed*. 2002, John Wiley & Sons, Inc. p. 243-244.
162. de Villena, F. and C. Sapienza, *Transmission ratio distortion in offspring of heterozygous female carriers of Robertsonian translocations.* Human genetics, 2001. **108**(1): p. 31-36.
163. Zheng, Y., et al., *Spam1 (PH-20) mutations and sperm dysfunction in mice with the Rb (6.16) or Rb (6.15) translocation.* Mammalian Genome, 2001. **12**(11): p. 822-829.
164. Diplas, A.I., et al., *Differential expression of imprinted genes in normal and IUGR human placentas.* Epigenetics., 2009. **4**(4): p. 235-40.
165. Hitchins, M.P. and G.E. Moore, *Genomic imprinting in fetal growth and development.* Expert Rev Mol Med., 2002. **4**(11): p. 1-19.