MAKING ROBUST USE OF PARENTAL GENOTYPE DATA FOR FINDING EFFECTS OF VARIANTS ON THE X CHROMOSOME

Alison Sara Wise

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Public Health in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2015

Approved by:

Rebecca Fry

Amy H. Herring

Danyu Lin

Min Shi

Wei Sun

Clarice R. Weinberg

Fei Zou

ABSTRACT

Alison Sara Wise: Making Robust Use of Parental Genotype Data for Finding Effects of Variants of the X Chromosome
(Under the direction of Clarice R. Weinberg)

The X chromosome is generally understudied in association studies, in part because the analyst has limited methodological options. We are developing statistical methods for causal association for single nucleotide polymorphisms (SNP markers) on the X. The focus of our work is on case-parent triad association studies. Most current family-based methods extend the transmission disequilibrium test (TDT) to the X chromosome. We propose a new method to study association in case-parent triads: the parent-informed likelihood ratio test for the X chromosome (PIX-LRT). Our method provides estimation of relative risks and takes advantage of parental genotype information and the sex of the affected offspring. Under a *parental allelic exchangeability* assumption for the X, if for a given locus case-parent triads are complete, the parents of affected offspring provide an independent replication sample for the estimates based on transmission distortion to the affected offspring. For each offspring sex we can combine the parent-level and the offspring-level information to form a likelihood ratio test statistic; we then combine the two to form a single composite test statistic, which we show offers better power than existing methods.

Maternal SNP effects can influence the development and later health of the offspring through prenatal effects, regardless of which alleles are transmitted by the mother to her offspring. Previously, using triads alone, no method had been developed without an

assumption of Hardy-Weinberg Equilibrium (HWE) to test maternal effects on the X chromosome. For the second project we extended PIX-LRT to discover maternal X-chromosome SNP effects.

Our third project concerns the identification and estimation of effects of X haplotypes. For case-parent triads, the X-chromosome haplotype phases can be inferred. With phase information, as is available when triad genotypes are nonmissing, the problem can be managed via an extension of the PIX-LRT from a two-allele problem to a k-allele problem, where the "alleles" are now the existing haplotypes at the locus under study. The extended approach relies on a permutation-based p-value based on the most significant individual haplotype effect. Our methods are applied to a dataset consisting of over 2000 triads in which the affected offspring have an oral cleft.

ACKNOWLEDGMENTS

I would like to thank my dissertation advisor, Dr. Clarice Weinberg for her mentoring, guidance and encouragement throughout this process. Our meetings, talks and walks around the lake have been key to both my professional and personal growth. I wish to thank Dr. Min Shi, who has been instrumental throughout the entire dissertation process, for her support and advice. I would like to thank Dr. Amy Herring for her counsel and support throughout my entire graduate school experience. I would also like to thank Drs. Rebecca Fry, Danyu Lin, Wei Sun and Fei Zou for being on my thesis committee.

I would also like to thank the staff, faculty and students at UNC and the NIEHS. Melissa Hobgood, Veronica Stallings, Rita Ross and Parham Shaw, thank you for always being available to chat and troubleshoot problems. To Decal, my UNC friends, both in the cave and out, and my NIEHS lunch and walk buddies, you have helped keep me sane throughout this process.

Lastly, I could not have completed this work without the love and support of my family. To my parents, Ken and Jackie, and my sister Jenny, thank you for your encouragement in this and in all things I do. When I struggled during this process you were there for me, only a phone call away to offer encouragement or listen patiently. My greatest thanks goes to my husband John Tumbleston, whom I met during this crazy process and, with all the ups and downs, stuck with me.

TABLE OF CONTENTS

LIST OF TABLES ix

LIST OF FIGURES x

LIST OF ABBREVIATIONS xi

CHAPTER 1: INTRODUCTION AND LITERATURE REVIEW 1

1.1 INTRODUCTION 1

1.2 LITERATURE REVIEW 3

1.2.1 Genetic Background 3

1.2.2 X-Chromosome Inactivation 6

1.2.3 Family Based X-Chromosome Extensions 7

1.2.4 X-Chromosome Maternal Effects 12

1.2.5 X-Chromosome Haplotype Effects 15

1.2.6 Oral Cleft Data 19

1.3 PROPOSED RESEARCH 20

CHAPTER 2: PIX-LRT: A PARENT-INFORMED TEST FOR SNPs ON THE X CHROMOSOME USING CASE-PARENT TRIADS 23

2.1 INTRODUCTION 23

2.2 SUBJECTS AND METHODS 26

2.2.1 Case-Parent Design and Assumptions 26

2.2.2 Modification of the X-LRT to Achieve Robustness 27

2.2.3 PIX-LRT Statistic 31

2.2.4 Type I Error Rate and Power Calculations 35

2.2.5 Oral Cleft Data 36

2.3 RESULTS 38

2.3.1 Noncentrality Parameters 38

2.3.2 Oral Cleft 43

2.4 DISCUSSION 47

CHAPTER 3: PIX-LRT EXTENSIONS FOR MATERNAL EFFECTS OF GENETIC VARIANTS ON THE X CHROMOSOME 53

3.1 INTRODUCTION 53

3.2 SUBJECTS AND METHODS 56

3.2.1 Case-Parent Design and Assumptions 56

3.2.2 PIX-LRT Extension to Maternal Effects 57

3.2.3 Type I Error and Power Calculations 60

3.2.4 Oral Cleft Data 62

3.3 RESULTS 63

3.3.1 Noncentrality Parameters 63

3.3.2 Oral Cleft 67

3.4 DISCUSSION 68

CHAPTER 4: FAMILY BASED X-CHROMOSOME HAPLOTYPE ANALYSIS USING PARENT INFORMATION 70

4.1 INTRODUCTION 70

4.2 SUBJECTS AND METHODS 72

4.2.1 Case-Parent Design and Assumptions 72

4.2.2 PIX-LRT Extension to Haplotype Analysis 73

4.2.3 Type I Error and Power Calculations 76

4.2.4 Oral Cleft Data 79

4.3 RESULTS 81

4.3.1 Simulation Output 81

4.3.2 Oral Cleft 85

4.4 DISCUSSION 88

CHAPTER 5: CONCLUSION 91

APPENDIX A: TEST FOR THE PARENTAL ALLELIC EXCHANGEABILITY ASSUMPTION 93

APPENDIX B: CLOSED FORM SOLUTIONS FOR THE SSX-LRT 95

B.1 TRIADS WITH AFFECTED SONS 95

B.2 TRIADS WITH AFFECTED DAUGHTERS 96

APPENDIX C: CLOSED FORM SOLUTIONS FOR THE PARENT-ONLY ANALYSIS 100

C.1 TRIADS WITH AFFECTED SONS 100

C.2 TRIADS WITH AFFECTED DAUGHTERS 102

APPENDIX D: CLOSED FORM SOLUTIONS FOR THE PIX-LRT 106

D.1 TRIADS WITH AFFECTED SONS 106

D.2 TRIADS WITH AFFECTED DAUGHTERS 107

APPENDIX E: D ACKNOWLEDGEMENT 110

REFERENCES 111

LIST OF TABLES

Table 2.1: Probabilities of mating pairs conditional on mating sum with and without parental allelic exchangeability (exch) 27

Table 2.2: For affected sons and daughters, case-parent genotype probabilities using transmission information 29

Table 2.3: Relative risks and mating type probabilities associated with parental sum given affected offspring 33

Table 2.4: For affected sons and daughters, case-parents genotype probabilities using parental sum information 34

Table 2.5: Case-parent families by cleft type, gender and ancestry 37

Table 2.6: Noncentrality parameter and corresponding Type I error rates for X-LRT 39

Table 2.7: PIX-LRT analysis results of SNP rs5981162, located in the intergenic region between *ENFB1* and *PJA1* at basepair 68318753 45

Table 2.8: Top 5 CL/P SNPs from our PIX-LRT analysis and from Patel et al. 46

Table 3.1: For affected sons and daughters, case-parents triad frequencies under an assumption of parental allelic exchangeability 60

Table 3.2: Noncentrality parameters and corresponding Type I error rates in parentheses for PIX-LRT and HAPLIN 64

Table 4.1: Haplotype frequencies for the different scenarios used in the simulations 77

Table 4.2: Complete case-parent families by cleft type, gender and ancestry 79

Table 4.3: Simulated Type I error rates for X-haplotype methods 81

Table 4.4: Most significant haplotypes associated with oral cleft based on PIX-LRT 86

Table 4.5: Cross table of SNPs s6627483 and rs5970137 87

LIST OF FIGURES

Figure 2.1: Power estimates as a function of minor allele frequency of X-LRT and PIX-LRT 40

Figure 2.2: Noncentrality parameter estimates as a function of relative risk 41

Figure 2.3: Noncentrality parameter estimates as a function of missing parental genotypes using the Expectation-Maximization (EM) algorithm..... 42

Figure 2.4: Q-Q plot of $-\log10(p)$ as calculated from the test of parental allelic exchangeability 43

Figure 2.5: Individual single nucleotide polymorphism significance of the cleft example..... 44

Figure 2.6: Assessment of concordance through comparison of the parent-only Z scores and transmission (SSX) Z scores..... 48

Figure 3.1: Noncentrality parameters as a function of maternal relative risk..... 65

Figure 3.2: Noncentrality parameter estimates as a function of fraction of families missing parental genotype using the Expectation-Maximization algorithm..... 66

Figure 3.3: Individual single nucleotide polymorphism significance of maternal genotype for the cleft example..... 67

Figure 4.1: Power estimates as a function of risk haplotype frequency 83

Figure 4.2: Fraction of times PIX-LRT nominates the risk haplotype amongst significant simulations 84

Figure 4.3: Individual haplotype significance of the cleft examples..... 85

LIST OF ABBREVIATIONS

APL	Association in the Presence of Linkage Test
EM	Expectation-maximization
FBAT	Family-Based Association Tests
GWAS	Genome-wide Association Study
HWE	Hardy-Weinberg Equilibrium
IBD	Identity by Descent
LRT	Likelihood Ratio Test
NCP	Noncentrality Parameter
MAF	Minor Allele Frequency
PIX-LRT	Parent-informed X Chromosome Likelihood Ratio Test
SNP	Single Nucleotide Polymorphism
TDT	Transmission Disequilibrium Test
PDT	Pedigree Disequilibrium Test
XCI	X-chromosome Inactivation
XS-TDT	X-linked Sibling Transmission/Disequilibrium Test
X-APL	X Chromosome Association in the Presence of Linkage
X-PDT	X Chromosome Pedigree Disequilibrium Test
XTDT	X Chromosome Transmission/Disequilibrium Test
XRC-TDT	Reconstruction-combined TDT for X-Chromosome Markers
XMPDT	X Chromosome Monte Carlo Pedigree Disequilibrium Test

CHAPTER 1: INTRODUCTION AND LITERATURE REVIEW

1.1 Introduction

The X chromosome is unique in that males have only one, maternally-derived copy, while females are diploid. Regions on the X chromosome have been identified in association with several diseases, including Duchenne muscular dystrophy (Abbadi, Philippe et al. 1994), Parkinson's disease (Nemeth, Nolte et al. 1999, Scott, Nance et al. 2001, Pankratz, Nichols et al. 2003) and autism (Shao, Wolpert et al. 2002, Vincent, Melmer et al. 2005, Piton, Gauthier et al. 2011). However, the X lags behind its autosomal counterparts in association and linkage findings. In a review of every genome-wide association (GWAS) paper published from January 2010 to December 2011 and included in the NHGRI GWAS Catalog, Wise (Wise, Gyi et al. 2013) found only 33% of the reported studies had included the X chromosome. This inattention is in part due to the need to use statistical methods specific for X-linked markers.

Two popular study designs used in genetic association studies are the family-based and case-control study designs. In a family-based study, investigators collect genotype information from related individuals. A common family-based dataset, particularly in young-onset diseases, consists of genotype information from a case and his/her mother and father (a case-parent triad). In a case-control study, investigators collect genotypes from unrelated cases and controls.

Family-based studies provide some benefits over traditional case-control studies.

Family-based tests of linkage and association have the advantage of being robust to *population structure*, also known as *stratification* and *admixture*. Bias due to population stratification occurs when subpopulations exist with different minor allele frequencies and different baseline risks of disease. When studying a genetic marker in case-control studies, if population structure is present, an association between a disease and a marker can be spuriously detected in the absence of any causal association between the disease and the marker. As an example, consider a population with two subpopulations in which the first subpopulation has a higher prevalence of disease and a higher prevalence of the variant allele than the second population. Because cases are more likely to come from the first population and have the variant allele, it will appear that the variant allele is associated with the disease. However, if we instead have family data, then methods can use the non-transmitted alleles from parents to offspring as genetically-matched controls to the transmitted alleles, so that bias due to population is totally avoided.

Another benefit to family-based studies over traditional case-control studies is the ability to study maternal effects. The maternal genotype can influence the intra-uterine environment both directly and through its role in modulating the metabolism and effects of toxic exposures. Therefore, in genetic studies it may be of interest to investigate maternal effects, especially effects on early-onset disease, such as the birth defect oral cleft. In case-control studies, maternal effects are not directly study-able, and they confound results because the case's genome and the mother's are causally correlated. A researcher cannot distinguish if an observed association is due to the maternal genotype, the fetal genotype, or some combination of the two.

In this document, we will discuss statistical methods for assessing association in the

possible presence of linkage for markers on the X chromosome. We will introduce methods that are applicable to case-parent triads. Under an assumption of "parental allelic exchangeability", we will show that there is information in the parents of affected offspring that has not been used by previous methods. To detect the effect of fetal SNPs, we combine this parental information with transmission information in a new powerful method, the Parent-Informed likelihood ratio test for the X chromosome (PIX-LRT). In a second project and again under the assumption of "parental allelic exchangeability", we demonstrate how one can test for maternal SNP effects on the X chromosome in case-parent triad data. Previously, using triads alone, no method has been developed without an assumption of Hardy-Weinberg equilibrium (HWE) to test maternal effects on the X chromosome. Lastly, we use the information in the parents to improve upon methods that test for X haplotypes. Our methods will be applied to a dataset consisting of over 2000 triads in which the affected offspring have an oral cleft.

1.2 Literature Review

This chapter presents a review of the literature pertaining to analyzing the X chromosome in family data. Some background is given on methods for analyzing autosomal SNP markers in family-based studies, as they are basis for many of the X-chromosome extensions. We include a review of X-chromosome inactivation, a phenomenon that can influence how markers on the X chromosome are thought about and statistically modeled in an analysis. Additionally, we give background on the family-based study of oral cleft to which our methods will be applied.

1.2.1 Genetic Background

Each X chromosome, as with other chromosomes, consists of double-stranded DNA.

At each base pair position along each strand of the DNA, resides one of four nucleic acids: Adenine (A), Cytosine (C), Guanine (G) and Thymine (T). These base pairs code for genetic information. A genetic marker is a single nucleotide or DNA sequence at a particular location. The different forms of the markers are referred to as alleles.

A single nucleotide polymorphism is a form of genetic marker in which the single nucleotide at a base pair varies across individuals within the population. For instance, there may be a SNP where 20% of chromosomes in a population have Adenine at a given base pair, while the other 80% have Guanine. Adenine would then be considered the minor allele (or variant allele) in this population, with a minor allele frequency of 0.20. A particular ordered sequence of alleles that are located near each other on a chromosome and are inherited together as a linked string is referred to as a *haplotype*.

If at a marker there are two possible variants, as mentioned in the SNP scenario above, the marker is di-allelic. For a di-allelic marker, denote the minor and major allele *a* and *A*, respectively. For autosomal (in the nuclear DNA but not on either sex chromosome) di-allelic markers, each individual carries two copies on their two chromosomes, one from their mother and one from their father, and individuals can be either homozygous (carrying two of the same alleles, *AA* or *aa*) or heterozygous (carrying the two different alleles, *Aa*). For the X chromosome, females are either homozygous or heterozygous; however, males carry only one allele (*A* or *a*) and are therefore *hemizygous*.

One assumption in most family-based methods, such as the transmission disequilibrium test (the TDT, to be described), is that there is Mendelian (random) transmission of one of the two copies of the allele. That is, the two alleles a parent carries (or the mother carries in the case of the X chromosome) are equally likely to be transmitted to

the offspring. For the X chromosome, this means that if the mother is *Aa* and the father is *A*, their daughter is equally likely to be *Aa* or *AA*.

Most family-based methods do not assume that a marker under study is in HWE, which would likely be violated in the presence of population admixture (incomplete mixing). HWE states that if p is the prevalence of allele *A* then the probabilities of genotypes *aa*, *aA* and *AA* are $(1-p)^2$, $2p(1-p)$ and p^2, respectively. Allelic frequencies in such a population will remain the same from generation to generation if there is random mating and no disturbing influences. By contrast, if there are distinct genetic subpopulations and members of a subpopulation do not mate at random but mate selectively favoring their own subpopulation then that marker will not be in HWE.

Within a pair of chromosomes, during meiosis markers on one chromosome may be split up and recombined with markers on the other chromosome in a process called *crossover*; this is a *recombination* event. Markers that are proximally close on a chromosome are less likely to be split up (and more likely to be inherited together) than markers that are farther apart, which are less likely to be inherited together. Markers that are correlated because historically they have been inherited together more than at random are said to be in *linkage*. Most genome-wide association studies (GWAS) are based on markers that do not include any actual causative SNPs but may include nearby SNPs that are in linkage disequilibrium with an actual susceptibility SNP and consequently associated with the disease outcome. Many family-based studies test for association between a variant marker and a disease in the presence of possible linkage between the variant and the disease marker. Nevertheless linkage without association is hard to construct and more typical scenarios in nature would involve both linkage and association.

1.2.2 X-Chromosome Inactivation

To equalize the effective gene dosage in XX females compared to XY males, a process called X-chromosome inactivation (XCI) takes place early in female embryonic development. In each somatic cell one X chromosome becomes transcriptionally inactive at random, forming what is called a Barr body (Barr and Bertram 1949, Lyon 2002). For the most part, the same X remains inactive through all future cell divisions of that lineage. Lyon (Lyon 2002) pointed out that when considering diseases influenced by X-linked genes, heterozygous females are mosaics, with two types of cells, having gene expression governed by only one or the other X chromosome. Some structures, such as intestinal crypts and thyroid nodules arise from a single cell (i.e. they are monoclonal), so that an individual crypt or nodule has all its cells with the same X active. Conversely, muscle fibers arise from multiple cells (i.e. they are polyclonal), so both kinds of cells are present in a single fiber. Because the fraction expressing a particular X can vary randomly across fibers, in heterozygotes with Duchenne muscular dystrophy (an X-linked disease) one sees gradations of effect among muscle fibers (Lyon 2002).

Because X inactivation occurs early in embryologic development, when there are few cells available to undergo that binomial selection, some females may have skewed XCI, with more of one X expressed than the other X. This can make a heterozygous carrier of an X disease gene have gene expression (hence disease phenotype) that is similar to a homozygote. This phenomena has been studied with monozygotic twins in X-linked diseases such as Duchenne muscular dystrophy (Abbadi, Philippe et al. 1994) and red-green color vision deficiency (Jorgensen, Philip et al. 1992). Additionally, there are regions of the X that "escape" inactivation and are expressed on both the active and the (mostly) inactive X. Carrel

found that in a study of 94 genes spanning the X chromosome in 40 human samples, only about 65 percent of genes were subject to inactivation in all heterozygous samples, and 20 percent were inactivated in some, but not all samples. Carrel also found that the majority of X-linked genes with Y homology (including pseudoautosomal genes) escape XCI, but that these regions did not explain all of the escaped genes (Carrel and Willard 2005).

If exactly half of each of the two kinds of X are expressed, then the relative risk of the variant allele in males may be similar to the relative risk of homozygotes with two variant alleles in females. However, if a marker escapes X-inactivation and both of the two kinds of X are expressed, then the relative risk of the variant alleles in males may be similar to the relative risk of one variant allele in females. Because the specific inactivation profile is often unknown for a marker under study on the X chromosome, it may unwise for an analyst to impose a parametric relationship between the effects an allele has in males and the effects it has in females. Additionally, methods appropriate for the X chromosome may not be appropriate for markers on the X chromosome with Y homology, as males are able to have a double dose if an analog locus exists on the Y.

1.2.3 Family Based X-Chromosome Extensions

Most family-based methods available for X-chromosome analysis are extensions of autosomal methods. In this section we introduce these autosomal methods and their extensions. The TDT was introduced by Spielman, McGinnis and Ewen to detect autosomal SNPs associated with disease in case-parent family triads (Spielman, McGinnis et al. 1993). The TDT validly tests for association between a marker and a disease in the presence of possible linkage (that is, tests a null that there is either no linkage or no association with the disease) or can be considered a valid test of both linkage and association (that is, tests a null

that there is neither linkage nor association with the disease). The method is robust to population structure (Spielman and Ewens 1996). The TDT works by measuring the apparent transmission distortion between the two alleles from heterozygous parents to affected offspring. For a di-allelic locus with possible alleles *A* and *a*, let *b* be the number of times the *a* allele was transmitted (*A* was not transmitted) from a heterozygous parent and *c* be the number of times the *a* was not transmitted (*A* was transmitted). Then the TDT test statistic is:

$$\chi^2_{TDT} = \frac{(b-c)^2}{b+c}$$

This is "McNemar's test" based on transmission-discordant pairs and under a null hypothesis of no linkage and no within-family association this statistic has a central chi-squared distribution with one degree of freedom.

The TDT requires genotypic data from both parents; this is a particular drawback for diseases that occur later in life, when parental data may not be obtainable. Therefore, Spielman and Ewens developed the sibling TDT (S-TDT), which is appropriate when genotype data is available for affected and unaffected siblings. The S-TDT compares the observed number of a variant allele in affected siblings with the expected number under no linkage/association, conditional on the distribution of the allele in the sibships. The S-TDT and TDT can be combined into a single test statistic, "combined TDT" or C-TDT, when some families have genotyped parents and others have genotyped siblings but without parents (Spielman and Ewens 1998). Knapp introduced the reconstruction-combination TDT (RC-TDT) (Knapp 1999), which uses genotyped offspring to reconstruct missing parental genotypes and corrects for biases resulting from the reconstruction.

The APL method (association in the presence of linkage) was developed (Martin, Bass et al. 2003) for nuclear families that may have multiple affected offspring. APL estimates missing parent pairs by conditioning on offspring genotypes and an identity by descent (IBD) statistic (IBD refers to a segment of DNA that has the same ancestral origin in two individuals). To account for the difficulty in calculating the variance when different family structures are present, APL uses a bootstrap method. To handle general pedigrees, where nuclear families may be genetically related, and therefore cannot be considered independent, the pedigree disequilibrium test (PDT) was developed (Martin, Monks et al. 2000). The PDT compares the number of variant alleles transmitted within a pedigree to the number not transmitted.

FBAT (Family-Based Tests of Association) (Laird, Horvath et al. 2000) is a broad class of family based association tests that can use either dichotomous or measured phenotypes, nuclear families or sibships, and allows for additive, dominant or recessive models. FBAT defines a general class of test statistics as a function of offspring genotypes and phenotypes. The distribution of the test statistic is calculated by treating offspring genotypes as random and conditioning on the minimal sufficient statistics (Rabinowitz and Laird 2000). When parental genotypes are available, the traits and parental genotypes constitute the minimal sufficient statistic.

The TDT, S-TDT and RC-TDT were extended to X-linked markers with the XTDT, XS-TDT and XRC-TDT (Horvath, Laird et al. 2000). These extensions account for the fact that only the mother can be heterozygous, as fathers have only one X chromosome. Additionally, for sons, genotype data from fathers is not useful (uninformative), as the father does not transmit an X to his son. For the XS-TDT, to account for the different number of

alleles in males and females and the difference in disease prevalence between the sexes, sib pairs are partitioned by sex.

The APL method was extended to X-linked markers (X-APL) for nuclear families where more than one offspring may be affected (Chung, Morris et al. 2007). When parents are missing, to estimate parental genotypes, in addition to offspring genotypes and IBD, the sexes of the affected offspring are also conditioned on. X-APL also suggests running a sex stratified analysis to account for different SNP effects in males and females. PDT X-chromosome extensions are the XPDT and the Monte Carlo PDT (XMCPDT) (Ding, Lin et al. 2006). Like the XS-TDT, the XPTD excludes sib pairs of different sexes, losing information. To account for missing parental genotypes XMCPDT uses allele frequency estimates to impute missing parental genotypes. A version of FBAT (Laird, Horvath et al. 2000) can also be used for X chromosome analysis and generalizes the XTDT. One drawback to these methods is that while they provide tests for linkage/association, they do not provide relative risk estimates for the genotypes.

For autosomal markers, Weinberg et al. proposed a likelihood-based log-linear multinomial modeling approach for nuclear families (Weinberg, Wilcox et al. 1998). The approach does not require Hardy-Weinberg equilibrium and offers robustness against bias due to population stratification. This Poisson regression model also allows for the estimation of disease relative risks under a co-dominant (two-degrees of freedom), a dominant, a recessive or a log additive effect (one-degree of freedom). For a di-allelic marker, let *M*, *F*, and *C* be the number of variant alleles carried by the mother, father, and child, respectively. Then $M, F,$ and $C \in \{0,1,2\}$. Under an assumption of mating symmetry (e.g. $p(M = 0, F = 1) = p(M = 1, F = 0)$), there are six potential mating types (unordered pairs of parental

genotypes). The log of the expected triad counts (E[$n_{M,F,C}$]), based on a multinomial, can be modeled as

$$\ln\{E[n_{M,F,C}]\} = \mu_i + \beta_1 I_{C=1} + \beta_2 I_{C=2} + \ln(2)\, I_{M=F=C=1}$$

Where μ_i serves to stratify families by mating type and ln(2) is an offset term needed because the frequency of *C=1* is doubled when *M=F=1*. I_K is a dummy variable which equals 1 when expression *K* is true. Exponentiating the maximum likelihood estimates of the betas ($e^{\widehat{\beta}}$) provides estimates of the relative risks. One appeal of this method is that the analysis can be run in standard statistical software with Poisson regression. Likelihood ratio tests (LRT) can be carried out to test the null of no linkage/association ($\beta_1 = \beta_2 = 0$). The expectation-maximization (EM) algorithm can be used to handle missing SNP genotypes (or individuals) (Weinberg 1999, Rampersaud, Morris et al. 2007). Another benefit of the log-linear model approach is that is allows for other types of effect estimates to be incorporated, such as maternal effects (discussed below).

Zhang and colleagues recently developed the X-LRT, a log-linear likelihood ratio test of linkage/association for X-linked markers (Zhang, Martin et al. 2008). As in the model originally proposed (Weinberg, Wilcox et al. 1998), the X-LRT conditions on parental mating type (the ordered pair of parental genotypes) of which there are six. However, X-LRT allows the two sexes to have different genotype risks. Zhang et al. identify four relative risks, which are assumed to be the same across mating type category. Compared to males with the minor allele, the four relative risks are: R_{AY} for males with the major allele, $R_{aa}$for females with two minor alleles, $R_{Aa}$for females with one minor allele and $R_{AA}$for females with two major alleles.

The X-LRT can test as a global null H_0: $R_{AA} = R_{Aa} = R_{aa}, R_{AY} = 1$. However, it can

also test for sex-specific effects, a female-specific null (H_0: $R_{AA} = R_{Aa} = R_{aa}$) and a male-specific null (H_0: $R_{AY} = 1$). The LRT statistics based on these three tests are asymptotic chi-squared with three, two and one degrees of freedom, respectively. While Zhang developed a package for the X-LRT, as with the log-linear model for autosomal markers, for complete triads this method can be run with widely available software.

The X-LRT allows for genotyped unaffected siblings and the EM algorithm can be used to enable the likelihood to be maximized despite some missing parental genotypes. This method performs well compared to the X-chromosome transmission-based methods mentioned above, and allows for male and female offspring to have separate relative risks. We will show that despite stratifying on parental mating types, this method is subject to bias unless the data are further stratified by the sex of the offspring.

The PIX-LRT method we develop in chapter 2 will provide robustness and also improve statistical power by making full use of parental genotype data. We do this by imposing a parental genetic exchangeability assumption for the X. We assume that in the source population, conditional on the three X chromosomes carried by the two parents, the one that happens to be carried by the father is a random choice among the 3.

1.2.4 X-Chromosome Maternal Effects

Maternal effects on the autosome have been identified in a number of childhood diseases, including childhood medulloblastoma (Lupo, Nousome et al. 2012), clubfoot (Weymouth, Blanton et al. 2011) and oral cleft (Jugessur, Shi et al. 2010). To date there are no methods available to test for maternal effects in case-parent triads on the X-chromosome without assuming HWE. In this section I will review robust methods that have been developed for family data for the autosomal chromosomes that do not assume HWE, and the

existing methods for the X-chromosome, which do assume HWE.

In family-based studies, researchers are able to study potential maternal effects associated with a maternal genotype marker for families where the offspring has developed a disease condition. Mitchell (Mitchell 1997) noted that an advantage of the TDT over case-control studies is that the TDT can potentially be used to differentiate between maternal and genotypic effects. She suggested applying the TDT in two stages to a triad family with maternal grandparents also genotyped. The TDT applied to the case-parent triad would measure the fetal genotypic effects. Then the TDT applied to the mother and her parents (now treating the mother as the case) would measure the maternally-mediated genotypic effect. Because the transmission of alleles from grandparents to parents and thence to offspring are independent events, the two TDT statistics are independent. One major drawback to this approach is that the grandparents must be genotyped. While conceptually appealing, this multi-generational family design may be very hard to implement, particularly for late-onset diseases.

For autosomal markers, the log-linear approach developed by Weinberg et al. (above) can also assess the effects of disease markers that act through maternal effects while adjusting for fetal effects (Wilcox, Weinberg et al. 1998). The test for maternal effects is not as robust as that for fetal effects because one must assume genetic mating symmetry for the parents, which in effect permits the paternal genotype to serve as control for the maternal genotype. Using the notation above, and if α_1 is the ln relative risk for a maternal effect associated with the mother carrying a single copy of the variant allele, and α_2 is the ln relative risk for a maternal effect associated with two copies of the variant allele (relative to no copies), then:

$$\ln E\left[n_{M,F,C}\right] = \mu_i + \beta_1 I_{C=1} + \beta_2 I_{C=2} + \alpha_1 I_{M=1} + \alpha_2 I_{M=2} + \ln(2)\, I_{M=F=C=1}$$

Maximum likelihood estimates can be calculated for $\alpha_1$and α_2, and exponentiating the estimates provide the estimated relative risks. The EM algorithm can be used to handle missing autosomal SNP genotypes (or individuals) (Weinberg 1999, Rampersaud, Morris et al. 2007). Unlike the method proposed by Mitchell, this method only requires case-parent triads and can account for missing data through use of the EM.

Sinsheimer, Palmer and Woodward introduced the maternal-fetal genotype test (MFG) to evaluate maternal-fetal incompatibility (Sinsheimer, Palmer et al. 2003). With certain diseases it may be of interest to study if the maternal genotype and the fetal genotype interact. For example, if the maternal-fetal genotype combination adversely affects the developing fetus (as can be the case with an Rh-negative mother pregnant with an Rh-positive fetus). Sinsheimer *et al*. extended the log-linear model with maternal and fetal effect parameters, mentioned above, to include parameters for incompatibility, and found an apparent interactive effect for schizophrenia.

HAPLIN (Gjessing and Lie 2006) is a likelihood-based method for analyzing maternal and fetal haplotypes in case-parent triads. HAPLIN is not robust to population stratification because it assumes HWE and does not condition on parental mating type. "Single-dose" effects (effects of one copy of the haplotype) of maternal haplotypes are measured in the model (as described in section 1.2.5). HAPLIN was expanded to measure effects of X haplotypes, if inherited by the offspring (Jugessur, Skare et al. 2012) (as described in section 1.2.5). This extension allows HAPLIN be used in analyzing maternal effects on the X chromosome (Myking, Boyd et al. 2013).

1.2.5 X-Chromosome Haplotype Effects

In autosomes, when studying haplotypes only the unphased genotypes (the genotype made up of the sum of the two haplotypes) are measured. For certain unphased genotypes it can be possible to reconstruct the haplotypes (for example if a person is homozygous at each SNP in the set then the two haplotypes are identical). However, this is typically not the case, so methods have been developed to handle phase ambiguity. Likelihood methods are commonly used for haplotype analysis. For example, Lin et al. provide a likelihood method (available in software HAPSTAT) to study haplotype-disease association in non-family based studies (cross-sectional, longitudinal, case-control and cohort) (Lin and Zeng 2005, Lin and Zeng 2006). For family-based studies, many methods have been developed to handle phase ambiguity. We introduce some of these methods here, and highlight those methods that have inspired X-chromosome extensions.

Dudbridge developed a likelihood-based association analysis for nuclear families and unrelated controls which is implemented in the software UNPHASED (Dudbridge 2008). UNPHASED can be run on either binary or continuous traits. Here we focus on case-parent triad families with binary traits. For complete data (with phase known) the likelihood has two factors; (1) the probability of the case genotype conditional on parental genotypes and offspring trait and (2) probability of the parental genotypes conditional on having a child with the trait. For known phases, these probabilities can be solved directly. However, for phases unknown (or other missing data), UNPHASED identifies all possible completions (sets of phased haplotypes) that are compatible with the observed data. If F, M and C denote the possible phased haplotype pairs and F', M', C' are either unphased or phased genotypes for fathers, mothers and cases, respectively, the possible completions are $\{f,m,c : Pr(F=f,$

M=m, C=c | observed *F', M', C')>0*}. The conditional probabilities are then calculated for the possible completions and included in the pseudo-complete-data likelihood. Individual haplotype effects, or an omnibus test of no haplotype effect can be tested. Dudbridge mentions possible coding schemes to reduce the number of parameters, such as assuming individual haplotypes are in HWE.

Additional likelihood approaches include TRANSMIT and PCPH. TRANSMIT was introduced by Clayton and measures transmission/disequilibrium of haplotypes to cases by conditioning cases on their parents (Clayton 1999). Clayton uses the likelihood to develop a score test of no haplotype effect. For incomplete data (phases unknown), the score vector is averaged over all possible parental haplotype configurations consistent with the observed data. PCPH, the projection conditional on parental haplotypes method is another likelihood method that was introduced by Allen and Satten (Allen and Satten 2007), which uses an estimating equation approach.

To avoid the hypothetically very large number of parameters that need estimating, many methods assume HWE for the haplotypes under study. In family-based designs, for haplotype analysis, the number of potential parental mating pairs becomes increasingly large with the number of haplotypes. The APL method described in section1.2.3 was extended to haplotypes in (Chung, Hauser et al. 2006). For haplotypes, APL assumes that there is no recombination between the markers within a family. Additionally, the EM is used to deal with phase ambiguity and account for missing genotypes. APL assumes HWE for haplotypes (though not for individual SNPs). For each of the n haplotypes, a test statistic T_h is calculated (a measure of the transmission of a haplotype compared to the expected transmission). A global test statistic G (a function of the T_h) is then calculated to measure an overall haplotype

effect. No relative risk estimates are calculated.

Sinsheimer and colleagues introduced the gamete competition model as a likelihood extension of the TDT that uses full pedigree data and can be applied to multiple alleles (Sinsheimer, Blangero et al. 2000). The model gives an estimate of the strength of transmission distortion to affected offspring for each for allele, which allows the alleles to be ranked. The method was extended to analyze haplotype data (Sinsheimer, McKenzie et al. 2001). To account for missing phase information, HWE is assumed and a quasi-Newton optimization algorithm is used for maximum likelihood estimation.

HAPLIN is a likelihood-based method that allows estimation of haplotype relative risks (Gjessing and Lie 2006). HAPLIN applies to case-parent triad data, but can also incorporate independent controls. HAPLIN assumes that haplotypes are in HWE. To restrict the number of parameters needed for all haplotype interactions, HAPLIN assumes a multiplicative model (the relative risk associated with any two haplotypes is the product of their individual risks). However, for each haplotype, there is a parameter to distinguish a double dose effect (the effect if a person has two copies of the same haplotype). The EM algorithm is used to maximize the likelihood when there are missing haplotypes.

Methods also exist for the autosome that do not require phase estimation. An example is the TRIad Multi-Marker method (TRIMM), introduced by Shi and colleagues (Shi, Umbach et al. 2007). For each case, *i*, the *complement* with the non-transmitted haplotypes is created as $M_i+F_i-C_i$. That complement vector corresponds to the set of genotypes that case should have been just as likely to inherit (under the null) for that set of loci from the same parents. A difference vector, with length equal to the number of SNPs in the set, is then constructed as the case minus the complement, which is $2C_i - M_i - F_i$. A vector of Z statistics is

calculated for each linked set of SNPs for all the families, based for each locus only on nonmissing and nonzero differences. The sign of each Z can be taken as nominating the allele that is showing evidence for an association with risk. An overall test statistic is the maximum Z score, max_Z^2. A permutation-based p-value for max_Z^2 is calculated by randomly assigning, within each family, case status to either the true case or their complement. To take advantage of the correlation that results from linkage between the haplotype SNPs, a Hotelling's T^2 statistic can also be calculated. Using permutations (again based on randomly assigned case/complement status), TRIMM combines the max_Z^2 and Hotelling's T^2 into a combined statistic.

For X-chromosome analysis, UNPHASED, X-APL, X-LRT and HAPLIN are the only haplotype methods that can specifically be applied to the X-chromosome. For X-chromosome analysis, UNPHASED (Dudbridge 2008), to avoid the issue of how to model a genetic effect if male and female affected offspring are combined in an analysis, conducts a separate analysis on each sex. As for haplotypes, Chung (Chung, Morris et al. 2007) also suggests running X-APL on affected males and females separately. X-LRT, as mentioned in section 1.2.3 also has a haplotype extension. However, this extension is limited to two-marker haplotypes. X-LRT tests all haplotypes simultaneously and assumes parental mating is random with respect to a haplotype, no recombination, and that haplotype penetrance is multiplicative for females. X-LRT for haplotypes also runs a separate analysis on males and females. For the X-chromosome, HAPLIN allows a range of X-chromosome models to be estimated depending on assumptions made about the allele effects in males compared to females (Jugessur, Skare et al. 2012). For example, under an X-inactivation assumption, one model constrains the relative risk in males with one copy of a haplotype to be the same as

that for females with two copies of the haplotype. In all but UNPHASED, the authors noted that on the X-chromosome, the phase of males is known, as they have only one X. They also noted that if both parental genotypes are available, the phase of a daughter's haplotypes can also be identified, because one of her haplotypes must be the same as her father's. None of these methods made use of the separate parental data based on how the haplotypes distribute across the parents of affected offspring.

1.2.6 Oral Cleft Data

Oral cleft is a common birth defect. Multiple genetic and environmental risk factors are thought to underlie oral cleft (Dixon, Marazita et al. 2011). For example, maternal smoking is a recognized risk factor (Wyszynski, Duffy et al. 1997). The fact that the recurrence risk for children born after a sibling with cleft is more than 30 suggests that genetics plays an important role in susceptibility to this birth defect (Sivertsen, Wilcox et al. 2008). The clefting phenotype is divided into two etiologically separate categories: cleft palate only (denoted CPO) and cleft lip with or without cleft palate (denoted CL/P). This phenotype split is based on genetic and embryological findings (Murray 2002). Oral cleft can occur with other abnormalities or as an isolated abnormality ("non-syndromic"). Within our research, we focus on non-syndromic oral cleft, however it is of interest to note that there are X-linked syndromes that feature cleft palate or cleft lip with or without palate. For example, mutations on the X-chromosome in the gene *EFNB1* are responsible for the majority of cases of craniofrontonasal syndrome (CFNS) (Twigg, Kan et al. 2004, Wieland, Jakubiczka et al. 2004), whose features can include cleft lip and palate, and mutations in the X chromosome gene *TBX22* specifically cause cleft palate and ankyloglossia (tongue-tie) (Marcano, Doudney et al. 2004).

We obtained access to data from the International Consortium to Identify Genes and Interactions Controlling Oral Clefts (see Appendix E for the dbGaP acknowledgement). The data were downloaded from the database of Genotypes and Phenotypes (dbGaP) (Mailman, Feolo et al. 2007) (Accession number: phs000094.v1.p1 (Beaty, Murray et al. 2010)). The data consists of 7089 study subjects, the majority of whom are members of a case-parent triad of either Caucasian or Asian ancestry. Cases had either non-syndromic cleft palate only, cleft lip only, or cleft lip with cleft palate. Subjects were genotyped for 592,532 SNPs (including SNPs on the X-chromosome), although not all of these SNPs passed quality inspection. Patel et al. (Patel, Beaty et al. 2013) analyzed 14,486 SNPs individually on the X-chromosome with FBAT using complete triads. Haplotype analysis was run on regions of interest using UNPHASED. Their findings suggested four X-linked markers in the *DMD* gene were associated with CL/P.

We are using these data both to simulate case-parent triads in a genetically realistic way and to reanalyze the X chromosome data for SNPs related to clefting. We begin by looking for effects of fetal inherited variants, then consider maternal variants, and finally consider haplotypes in the fetus.

1.3 Proposed Research

In Chapter 2, we develop a method for case-parent triads that incorporates parental information into X-chromosome analysis. We present a new method, the sex-stratified X-chromosome likelihood ratio test (SSX-LRT), which is similar to the X-LRT but provides robustness by allowing distinct mating type parameters for male versus female affected offspring. For studies using case-parent triads, we show that additional improvement in detecting markers associated with a trait is possible by exploiting genotype information in the

parents not used in previous methods. We exploit the fact that mothers and fathers of affected offspring are differentially enriched for susceptibility markers in a way that depends on the sex of the affected offspring. We demonstrate that an assumption of parental allelic "exchangeability" enables the added information to be captured in a way that resists bias due to population stratification. Consequently, regardless of what alleles parents transmit to their affected offspring, additional information can be robustly gleaned from the parental X genotypes to supplement the transmission-based SSX-LRT, creating the more powerful "parent-informed X-chromosome likelihood ratio test" (PIX-LRT).

The method has the additional advantage that a kind of replication is provided that is internal to the study. This replication is achieved by comparing the results from the transmission-based analysis to results from the parent-based analysis. If triad data are complete, the two sets of findings are statistically independent; hence concordance between those results strengthens confidence in the inference.

We initially describe the SSX-LRT and PIX-LRT for single X-linked SNP markers with complete genotype data. An extension of the approach then enables inclusion of families with missing individuals or sporadically missing SNP genotype data. We assess Type I error rates for SSX-LRT, PIX-LRT and X-LRT and compare power for the SSX-LRT, PIX-LRT and XTDT by calculating chi-squared noncentrality parameters based on expected counts (Agresti 2012). As an example, after using the data to test our parental exchangeability assumption, we apply the PIX-LRT to the oral cleft dataset (section 1.2.5) to analyze SNP markers on the X-chromosome. We conclude with a discussion of the advantages and limitations of PIX-LRT, and our SNP findings related to oral cleft. This work was published in *Frontiers in Genetics* and titled "Learning about the X from our parents" (Wise, Shi et al.

2015).

In Chapter 3, we extend PIX-LRT to assess possible maternal effects while controlling for fetal effects. Because PIX-LRT conditions on the sum of a mating pair (the number of variant alleles they carry collectively), the risk in offspring of mothers with no copies can be compared to offspring of mothers with one copy, and the risk in offspring of mothers with two copies can be compared to that in offspring of mothers with one copy. Therefore, under an assumption of parental allelic exchangeability, we are able to test for both fetal and maternal effects in the same model. The EM algorithm enables maximization of the likelihood despite the inclusion of some triads with missing genotypes. We will test for maternal effects using the oral cleft dataset.

In Chapter 4, we will incorporate parental information into haplotype analysis on the X chromosome. The X is unique in that, if complete case-parents triad genotype data is present, all haplotype phases can be determined. Therefore, for complete data the k-haplotype problem can be considered an extension of the two-allele problem addressed in Chapter 2 to a k-allele problem. Our approach considers each haplotype in turn, testing it against the aggregate of all others, which produces k p-values, each based on PIX-LRT run on the dichotomized haplotypes. We evaluate that set of p-values by means of an efficient permutation procedure, which works as follows. For each family we reform the parents (imposing exchangeability) and generate at random an offspring genotype of the same sex. PIX-LRT could then be run on the dichotomized haplotypes reducing the problem to a two-allele problem.

CHAPTER 2: PIX-LRT: A PARENT-INFORMED TEST FOR SNPs ON THE X CHROMOSOME USING CASE-PARENT TRIADS

We present a new method to study association in case-parent triads: the parent-informed likelihood ratio test for the X chromosome (PIX-LRT). Our method provides estimation of relative risks and takes advantage of parental genotype information and the sex of the affected offspring to increase statistical power to detect an effect. We apply PIX-LRT to publically available data from an international consortium of genotyped families affected by the birth defect oral cleft and find a strong, internally-replicated signal for a SNP marker related to cleft lip with or without cleft palate. The following chapter was published in *Frontiers in Genetics* in the article titled "Learning about the X from our parents" (Wise, Shi et al. 2015).

2.1 Introduction

The X chromosome is unique in that males have only one, maternally-derived copy, while females are diploid. As a form of dosage compensation, a random X is inactivated in each cell early in female embryonic development (Lyon 2002). Regions on the X chromosome have been identified in association with several diseases, including Parkinson's disease (Nemeth, Nolte et al. 1999, Scott, Nance et al. 2001, Pankratz, Nichols et al. 2003) and autism (Shao, Wolpert et al. 2002, Vincent, Melmer et al. 2005, Piton, Gauthier et al. 2011). However, the X lags behind its autosomal counterparts in association and linkage findings, in part due to the need to use methods specific for X-linked markers (Wise, Gyi et

al. 2013).

Most family-based methods available for X chromosome analysis are extensions of autosomal methods. The original transmission/disequilibrium test (TDT) was proposed to detect autosomal SNPs associated with disease in case-parent triads (Spielman, McGinnis et al. 1993). For studies that also include unaffected siblings and may or may not include parental genotyping, we have the sibling TDT (S-TDT) (Spielman and Ewens 1998) and the reconstruction-combination TDT (RC-TDT) (Knapp 1999). These family-based methods were extended to X-linked markers with the XTDT, XS-TDT and XRC-TDT (Horvath, Laird et al. 2000). A number of extensions have been developed to accommodate larger families (Ding, Lin et al. 2006, Chung, Morris et al. 2007). A version of FBAT (Laird, Horvath et al. 2000) can also be used for the X chromosome and generalizes the XTDT. These methods provide p-values to test for association, but they do not enable estimation of disease-related marker relative risks. The method we will propose is for case-parent triads, but accommodates triads with a missing individual.

Likelihood-based log-linear multinomial modeling approaches for nuclear families can use the EM algorithm to handle missing autosomal SNP genotypes (or individuals), and provide both robustness against bias due to population stratification and the opportunity to estimate disease-related marker relative risks (Weinberg, Wilcox et al. 1998, Rampersaud, Morris et al. 2007). HAPLIN is a likelihood-based method that is able to estimate relative risks for single SNPs and haplotypes on the autosomes and X chromosome (Gjessing and Lie 2006, Jugessur, Skare et al. 2012). However, as we are interested in methods that do not assume Hardy-Weinberg equilibrium (HWE), which HAPLIN requires, we will not discuss the method further.

The X-LRT, a log-linear likelihood ratio test of association for X-linked markers that does not assume HWE, was recently developed (Zhang, Martin et al. 2008). This method performs well compared to transmission-based methods, and allows male and female offspring to have separate relative risks. The X-LRT conditions on parental mating type (the pair of parental genotypes), which we will show can cause bias because families with female and male affected offspring are forced to share the same mating type parameters. We present a new method, the sex-stratified X chromosome likelihood ratio test (SSX-LRT), which prevents that bias by allowing distinct mating type parameters for male versus female affected offspring.

We show that additional improvement is possible by exploiting genotype information in the parents not used in previous methods. Mothers and fathers of affected offspring are differentially enriched for susceptibility markers depending on the sex of the affected offspring. We demonstrate that an assumption of parental allelic "exchangeability" enables the added information to be captured in a way that resists bias due to population stratification. Consequently, regardless of what alleles parents transmit to their affected offspring, additional information can be robustly gleaned from the parental X genotypes to supplement the transmission-based SSX-LRT, creating the "parent-informed X chromosome likelihood ratio test" (PIX-LRT).

In the following sections, we initially describe the SSX-LRT and PIX-LRT for single X-linked SNP markers with complete genotype data. An extension of the approach then enables inclusion of families with missing genotype data. We assess Type I error rates for SSX-LRT, PIX-LRT and X-LRT and compare power for the SSX-LRT, PIX-LRT and XTDT by calculating chi-squared noncentrality parameters based on expected counts

(Agresti 2012). As an example, we apply the PIX-LRT to family data from an oral cleft dataset to analyze SNP markers on the X chromosome. We conclude with a discussion of the advantages and limitations of PIX-LRT, and our SNP findings for cleft lip.

2.2 Subjects and Methods

2.2.1 Case-Parent Design and Assumptions

Consider a sample of case-parent triads who have all been genotyped at a di-allelic X locus. Let *M*, *F*, and *C* denote the number of copies of the variant (minor) allele in the mother, father and affected offspring (proband), respectively. We exclude regions on the X that correspond to a homologous region on Y, including the pseudo-autosomal regions and the X-transposed region (PARs, XTR). Then, $M \in \{0,1,2\}, F \in \{0,1\}, C \in \{0,1\}$ for male offspring, and $C \in \{0,1,2\}$ for female offspring. Consider tests of the null hypothesis that there is no association or no linkage against the alternative of association in the presence of linkage. Assume there is Mendelian transmission at that locus in the source population. Further assume parental allelic exchangeability in the source population, as in (Shi, Umbach et al. 2008); that is, within a mating pair, the variant alleles are randomly located across the three X chromosomes (see Table 2.1). This assumption, which is met under nonassortative mating within subpopulations at that locus, can be tested within the source population using the following model for the parental genotype count:

$$\ln\big(E[N_{M,F}|M+F]\big) = \log(\mu_{M+F}) + \alpha_1 I_{(M=1,F=0)} + \alpha_2 I_{(M=1,F=1)} + \log(2)\, I_{(M=1)} \qquad (2.1)$$

Here *E* denotes expected value and $N_{M,F}$ is the random multinomial count variable denoting the number of triads where the mother and father carry *M* and *F* copies of the variant allele, respectively. μ_{M+F} are nuisance parameters that stratify families by conditioning on the sum of parental genotypes and $\log(2)$ is an offset term required because there are two ways for *M*

to equal 1 (the variant allele can be on either chromosome). The parameters α_1 and α_2 are the log of half the odds that the mother carries 1 copy of the variant when the parents together have 1 and 2 copies, respectively. See Table 1. We can calculate a likelihood ratio test statistic for $\alpha_1 = \alpha_2 = 0$, which under exchangeability is distributed as a central chi-squared with two degrees-of-freedom (see Appendix A for closed-form solutions). Note that parental allelic exchangeability is much less restrictive than assuming Hardy-Weinberg equilibrium (HWE) because it must hold only within unknown genetic subpopulations. Lastly, as is generally required for family studies we assume that variants are not determinants of fetal survival or parental ability to reproduce.

Table 2.1: Probabilities of mating pairs conditional on mating sum when parental allelic exchangeability is present (exch), and when it is not (no exch).

M+F	M	F	Pr(M,F|M+F, exch)	Pr(M,F|M+F, no exch)
0	0	0	1	1
1	1	0	2/3	$2\exp(\alpha_1)/(1+2\exp(\alpha_1))$
	0	1	1/3	$1/(1+2\exp(\alpha_1))$
2	2	0	1/3	$1/(1+2\exp(\alpha_2))$
	1	1	2/3	$2\exp(\alpha_2)/(1+2\exp(\alpha_2))$
3	2	1	1	1

2.2.2 Modification of the X-LRT to Achieve Robustness

The X-LRT (Zhang, Martin et al. 2008) provides a powerful likelihood ratio test for triad data with affected sons and daughters and also allows one to estimate disease-related marker relative risks. A multinomial likelihood is expressed in terms of offspring genotype relative risks. X-LRT conditions on the parental mating type by including mating type parameters, but forces those parameters to be the same for families with affected male and female offspring. That approach can consequently be biased (shown in results) when

subpopulations have different minor allele frequencies and disease risks in males versus females (noncarriers) differ among subpopulations. This bias can also occur when recruitment rates for families with male versus female affected offspring differ across subpopulations with different minor allele frequencies. To remove this bias we stratify by both the parental mating type and the sex of the affected offspring (see Table 2.2). Let *aff* denote the event that the offspring is affected and define the relative risks, within parental mating type, as follows:

$$e^{\beta_1} = R_{G1} = \Pr(\text{aff}|\text{girl}, C = 1)/\Pr(\text{aff}|\text{girl}, C = 0)$$

$$e^{\beta_2} = R_{G2} = R_{G1}\Pr(\text{aff}|\text{girl}, C = 2)/\Pr(\text{aff}|\text{girl}, C = 1)$$

$$e^{\beta_3} = R_B = \Pr(\text{aff}|\text{boy}, C = 1)/\Pr(\text{aff}|\text{boy}, C = 0)$$

The analysis follows a multinomial for the counts based on both triad genotypes and sex (*g* for girl and *b* for boy) of the affected offspring ($N_{M,F,C,sex}$), modeled in a log-linear form, multiplying the sex-specific expected counts for each parental genotype pair by the probabilities shown in Table 2.2, as follows::

$$\ln\left(E\left[N_{M,F,C,sex}\right]\right) = \log(\gamma_{M,F,sex}) + \beta_1 I_{(C=1,sex=g)} + \beta_2 I_{(C=2,sex=g)} + \beta_3 I_{(C=1,sex=b)} \qquad (2.2)$$

Here $\gamma_{M,F,sex}$ are 12 nuisance parameters that serve to confer robustness against population stratification by stratifying families by both mating type and sex of affected offspring. Exponentiating the β's produces the relative risk estimates (e.g. $e^{\hat{\beta}_1} = \hat{R}_{G1}$). Inclusion of three unconstrained relative risk parameters allows one to avoid imposing an arbitrary relationship between the relative risks in boys and girls. We therefore refer to this method as the sex-stratified X-LRT (SSX-LRT). The corresponding log-likelihood for each sex (using the lower case "*n*" to denote observed counts of the variable *N*) is proportional to:

$$\sum_{M,F,C} n_{M,F,C,\text{sex}} \log(\Pr(M,F,C|\text{aff},\text{sex})) \quad (2.3)$$

Expression 2.3 can be rewritten as:

$$\sum_{M,F,C} n_{M,F,C,sex} \log(\Pr(C|M,F,\text{aff},\text{sex}) * \Pr(M,F|\text{aff},\text{sex})) \quad (2.4)$$

Table 2.2: For affected sons and daughters, case-parent genotype probabilities using transmission information.

			Affected Son			Affected Daughter	
M	F	C	Pr(C|M,F,b)	$E(N_{M,F,b})$	C	Pr(C|M,F,g)	$E(N_{M,F,g})$
0	0	0	1	γ_{00b}	0	1	γ_{00g}
0	1	0	1	γ_{01b}	0	1	γ_{01g}
1	0	0	$1/(1+R_B)$	γ_{10b}	0	$1/(1+R_{G1})$	γ_{10g}
		1	$R_B/(1+R_B)$		1	$R_{G1}/(1+R_{G1})$	
1	1	0	$1/(1+R_B)$	γ_{11b}	1	$R_{G1}/(R_{G1}+R_{G2})$	γ_{11g}
		1	$R_B/(1+R_B)$		2	$R_{G2}/(R_{G1}+R_{G2})$	
2	0	1	1	γ_{20b}	1	1	γ_{20g}
2	1	1	1	γ_{21b}	2	1	γ_{21g}

With complete data, the nuisance parameters do not need to be explicitly estimated when calculating the maximum likelihood estimates for the relative risks and the likelihood ratio test statistic. Closed form solutions to the maximum likelihood equations for the relative risks and the corresponding likelihood ratio test statistic under specified genetic models of inheritance are given in Appendix B. Note that triads with affected sons do not need genotyped fathers for the SSX-LRT.

When some genotype information is missing, we use the Expectation-Maximization (EM) algorithm as described in (Weinberg 1999). For the EM, the mating type parameters are needed to calculate the maximum likelihood estimators of the relative risks and the likelihood ratio test statistic. If two subpopulations have different minor allele frequencies and also different degrees of missingness, the missingness can be informative and use of the

EM can induce bias in the estimate. To avoid this bias, if the subpopulations are identifiable (e.g. the analyst can stratify on ancestry) the EM can be run on a likelihood that allows different mating type parameters for each.

To form a test based on both families of affected sons and families of affected daughters, we recommend forming a combined test statistic. Let X_B and X_G be the one degree-of-freedom LRT chi-squared statistics based on families with affected sons and daughters, respectively, where X_G is based on the coding: $R_{G1}^2 = R_{G2}$. Under the null the sum of the two (independent) test statistics has a chi-squared distribution with two degrees-of-freedom.

However, rather than just computing the sum we note that the most plausible departures from the null would involve scenarios where the boys and girls experience the same direction of effect, that is, the variant either increases risk for both or decreases risk for both. (In fact, because the two test statistics are statistically independent, one could regard families with affected daughters as a replication sample for findings based on families with affected sons.) Accordingly, the following construction exploits that directional agreement to enhance power (see (Zaykin 2011)) for a combined test. Take the square root of each chi-squared statistic and attach to that square root the sign corresponding to the direction of the estimated effect, S_B and S_G ("+1" for relative risk >1 and "-1" for relative risk <1). Under the combined null hypothesis the results will be two independent standard Gaussian statistics. Let N_B and N_G be the number of triads with a heterozygous mother and an affected son or daughter, respectively. The weighted combined Z statistic is constructed as follows:

$$Z_C = \frac{S_B\sqrt{N_B X_B} + S_G\sqrt{N_G X_G}}{\sqrt{N_B + N_G}} \sim N(0,1) \qquad (2.5)$$

Z_C^2 follows a central chi-squared distribution with one degree of freedom under the null and a noncentral chi-squared under alternatives, where the noncentrality parameter is:

$$\left[E\left\{\frac{S_B\sqrt{N_B X_B}+S_G\sqrt{N_G X_G}}{\sqrt{N_B+N_G}}\right\}\right]^2$$

Intuitively, if the number of informative families with an affected son is markedly different from the number with an affected daughter, this weighting scheme will favor the larger test statistic and sample size.

We focus on a sex-stratified analysis, but one could alternatively impose a relationship between R_B and R_{G1}, R_{G2}. Such a model can be fitted by use of widely available software (e.g. glm in (R Development Core Team 2013)) to maximize the multinomial likelihood and to estimate parameters. For example, under a simple model based on X-inactivation, one could argue for a two degree-of-freedom test with:

$$H_0\colon R_{G1}=R_{G2}=R_B=1 \quad (\beta_1=\beta_2=\beta_3=0\,)$$

$$H_a\colon R_{G2}=R_B\neq 1 \text{ or } R_{G1}\neq 1 \quad (\beta_2=\beta_3\neq 0 \text{ or } \beta_1\neq 0)$$

If we additionally assume a log-additive model in girls, we would have a one degree-of-freedom test with:

$$H_a\colon R_{G1}^2=R_{G2}=R_B\neq 1 \quad (2\beta_1=\beta_2=\beta_3\neq 0)$$

The log-linear form of the model would be:

$$\ln\left(E\left[n_{M,F,C,sex}\right]\right)=\log(\gamma_{M,F,sex})+\beta_1 C[I_{(sex=g)}+2I_{(sex=b)}]$$

Similar analyses that either simplify the parameterizations or are aimed at testing X-inactivation relationships (where $R_{G2}=R_B$ is the null hypothesis to be tested) can also be carried out in the context of the PIX-LRT method to be described.

2.2.3 PIX-LRT Statistic

The likelihood we will maximize is based on two separate factors, one that models transmissions conditional jointly on both the sex of the affected offspring and the parental

genotypes (cf. the stratum parameters in Table 2.2), i.e. *M*, *F* (as described above for SSX-LRT), and another that models *M*, *F* conditional on *M*+*F* and the sex of the affected offspring (cf. Table 2.1). The second, parental-information component is statistically independent of the transmission-based component, allowing parental data to provide a kind of internal replication. That parental piece has not been explicitly exploited by other methods.

The transmission-based part of the information is very much like that captured by the SSX-LRT just described in Section 2.2.2, through maximizing expression (2.4) above. But PIX-LRT will augment that by capturing information from the parents, rather than conditioning away that information (cf. the 12 stratification parameters included in Table 2.2), by instead conditioning more coarsely on the total number of copies of the variant carried by the two parents.

We begin with some intuition to clarify why there is information in how a fixed number of variant alleles (*M*+*F*) is distributed across the two parents. Under the null hypothesis, for a SNP on the X chromosome, one would expect neither the mother's two chromosomes nor the father's single chromosome to be enriched for either allele. However, suppose the variant is linked to risk of the disease. Because the mother of an affected son was the source of his only X, the mothers of affected sons should be enriched for that variant as compared to the fathers. Because a father of an affected daughter transmitted his only X to his daughter, whereas the mother could transmit either one of her two X's to her daughter, the fathers of affected daughters should be enriched compared to the mothers. These resulting opposing patterns of enrichment within the parents can be exploited by conditioning on the sex of the affected offspring and the number of variant alleles the parents carry (*M*+*F*),

taking advantage of our parental exchangeability assumption.

Specifically, one can augment the earlier analysis by incorporating the following log likelihood to capture the parent-only information:

$$\sum_{M,F} n_{M,F,sex} \log(\Pr(M,F|M+F,\text{aff},\text{sex}) * \Pr(M+F|\text{aff},\text{sex})) \quad (2.6)$$

The probabilities used in expression 2.6 are given in Table 2.3. For complete data, closed form maximum likelihood estimates of the relative risk and a likelihood ratio test statistic could be obtained from this method using only parents (see Appendix C). The EM can be used when genotype data is missing.

Table 2.3: Relative risks and mating type probabilities associated with parental sum given affected offspring.

				Affected Sons	Affected Daughters
M+F	*M*	*F*	Null Prob	Pr(M,F| M+F)	Pr(M,F|M+F)
0	0	0	1	1	1
1	1	0	2/3	$(1+R_B)/(2+R_B)$	$(1+R_{G1})/(1+2R_{G1})$
	0	1	1/3	$1/(2+ R_B)$	$R_{G1}/(1+2R_{G1})$
2	2	0	1/3	$R_B/(1+2R_B)$	$R_{G1}/(2R_{G1}+ R_{G1})$
	1	1	2/3	$(1+ R_B)/(1+2R_B)$	$(R_{G1}+R_{G2})/(2R_{G1}+ R_{G2})$
3	2	1	1	1	1

The combined likelihood that now includes both the parental data and the transmission data can be written as a multinomial (see Table 2.4) and modeled in a log-linear form as follows:

$$\ln\left(E\left[N_{M,F,C,sex}|M+F\right]\right) = \log(\mu_{M+F,sex}) + \beta_1 I_{(C=1,sex=g)} + \beta_2 I_{(C=2,sex=g)} + \beta_3 I_{(C=1,sex=b)} \quad (2.7)$$

As before, inclusion of three unconstrained relative risk parameters allows one to avoid imposing an arbitrary relationship on the relative risks in boys and in girls. The corresponding likelihood for each sex is then proportional to:

$$\sum_{M,F,C} n_{M,F,C,sex} \log(\Pr(M,F,C|M+F,\text{aff},\text{sex}) * \Pr(M+F|\text{aff},\text{sex})) \qquad (2.8)$$

Table 2.4: For affected sons and daughters, case-parents genotype probabilities using parental sum information.

	Affected Sons					Affected Daughters		
M+F	M	F	C	Pr(M,F,C|M+F)	$E(N_{M+F})$	C	Pr(M,F,C|M+F)	$E(N_{M+F})$
0	0	0	0	1	μ_{0b}	0	1	μ_{0g}
1	0	1	0	$1/(2+R_B)$	μ_{1b}	1	$R_{G1}/(1+2R_{G1})$	μ_{1g}
	1	0	0	$1/(2+R_B)$		1	$R_{G1}/(1+2R_{G1})$	
	1	0	1	$R_B/(2+R_B)$		0	$1/(1+2R_{G1})$	
2	1	1	0	$1/(1+2R_B)$	μ_{2b}	2	$R_{G2}/(2R_{G1}+R_{G2})$	μ_{2g}
	1	1	1	$R_B/(1+2R_B)$		1	$R_{G1}/(2R_{G1}+R_{G2})$	
	2	0	1	$R_B/(1+2R_B)$		1	$R_{G1}/(2R_{G1}+R_{G2})$	
3	2	1	1	1	μ_{3b}	2	1	μ_{3g}

For complete data, closed-form solutions to the maximum likelihood equations for the relative risks and the corresponding likelihood ratio test statistic are given in Appendix D. The number of informative families is greater for PIX-LRT than SSX-LRT; families where M=0, F=1 and M=2, F=0 are informative for PIX-LRT but not for SSX-LRT. The partial information can be used for all triads where at least one member has genotype data. A combined score can be calculated for PIX-LRT as was described for SSX-LRT. However N_B and N_G are now the number of informative families, that is, families for which M+F cannot be inferred to be 0 or 3 with an affected son or daughter. This method is available as an R

package at

http://www.niehs.nih.gov/research/resources/software/biostatistics/pixlrt/index.cfm.

2.2.4 Type I Error Rate and Power Calculations

The Type I error rate and the power are assessed by calculating the non-centrality parameter (NCP) for the distribution of a chi-squared likelihood ratio test statistic. Under the null hypothesis, the LRT statistic follows a central chi-squared distribution, which has an NCP of 0. The NCP is calculated by treating expected triad counts under the specified population structure as data used to fit the relevant models (O'Brien 1986, Agresti 2012). Values of noncentrality parameters can be translated to power values using the noncentral chi-squared distribution with the appropriate degrees of freedom.

To assess performance when there is admixture present in the population, we calculated the NCP for PIX-LRT, SSX-LRT, XTDT and X-LRT. Consider two scenarios, each with two subpopulations of equal size, with no effect of the variant allele in either sex. In the first scenario, one subpopulation has a minor allele frequency of 0.3, a disease risk of 0.02 in males, and 0.02 in females. The second subpopulation has a minor allele frequency of 0.2, a risk of 0.01 in males, and of 0.02 in females. A second scenario is similar except in the first subpopulation the disease risk is 0.03 in females and the second subpopulation has a disease risk of 0.02 in males. For computational convenience we assume HWE within each subpopulation. The expected counts were calculated for 1000 families with affected offspring. Non-centrality parameters were estimated for tests of (1) H_0: no effect in males or females; (2) H_{0m}: no effect in males; (3) H_{0f}: no effect in females.

We compare PIX-LRT to X-LRT for a scenario where both are valid. We consider a setting in which there are 1000 triads, $R_B$is 1.5, and $R_{G1}^2 = R_{G2} = 2$. In the non-carriers, the

disease risk in boys is twice that in girls. We calculate power (based on noncentrality parameters) as a function of minor allele frequencies. We choose an alpha level of $5x10^{-6}$ as this approximates the alpha 0.05 Bonferroni-corrected value needed for the X chromosome. We modify X-LRT to account for a log-additive dose effect in girls. Therefore, the X-LRT test for a fetal effect involves two degrees-of-freedom.

We also compare power of the PIX-LRT to the SSX-LRT and the XTDT, under a homogeneous population, for computational simplicity. To calculate the NCP for the XTDT we use the method proposed by Deng (Deng and Chen 2001). We do not include the other X chromosome extensions, as only complete triads are considered with no additional siblings or extended pedigrees.

For our power analysis we consider settings in which sex-specific tests are of interest to highlight similarities and differences between the two sexes. We consider the following 500-triad scenarios: affected male offspring and a minor allele frequency of either 0.3 or 0.1; affected female offspring, $R_{G1}^2 = R_{G2}$, and a minor allele frequency of either 0.3 or 0.1. We plot the noncentrality parameters as a function of the relative risk (R_{G1} for girls), and include the corresponding power for a one-degree-of-freedom LRT at alpha level $5x10^{-6}$.

To study the EM algorithm in PIX-LRT and SSX-LRT, we use the same scenarios and set R_B to 2, and R_{G1} to 2. We plot the noncentrality parameter (and the power at alpha level $5x10^{-6}$) as a function of the proportion of missing fathers, missing mothers, or a combination. For the combination scenarios, only one parent is missing, and twice as many fathers as mothers are assumed missing.

2.2.5 Oral Cleft Data

We applied PIX-LRT with the EM to the X chromosome data from the International

Consortium to Identify Genes and Interactions Controlling Oral Clefts. The data were downloaded from dbGaP (Mailman, Feolo et al. 2007) (Accession number: phs000094.v1.p1 (Beaty, Murray et al. 2010)). The data were previously analyzed by Patel et al. (Patel, Beaty et al. 2013) using FBAT (Laird, Horvath et al. 2000). Patel et al. (Patel, Beaty et al. 2013) used only complete triads and included all ethnicities in their joint analysis, whereas we included partial triads but only Asian (including Pacific Islanders) and Caucasian ethnicities. We analyzed 13283 SNPs on the X chromosome that had a minor allele frequency in the parents greater than 0.02, and had a unique mapping from the Illumina Human610-Quad v1.0 Build 36 to Build 37. For a family-wise alpha of 0.05 with a Bonferroni correction, the cutoff for the p-value is 3.74×10^{-6}.

We included all triads for which we have genotype data from the case, and the parents are not of differing ethnicity. 13% of the Asian triads and 21% of the Caucasian triads were incomplete. If multiple affected siblings were present, we randomly chose one sibling (27 siblings removed). We analyzed 1105 European families and 1286 Asian families. The clefting phenotype is divided into two categories: one is cleft palate only (denoted CPO) and the other is cleft lip with or without cleft palate (denoted CL/P). This phenotype split is based on genetic and embryological findings suggesting they are distinct (Murray 2002). The gender and cleft subtype breakdown is shown in Table 2.5. Note that CL/P predominantly affects boys while CPO is slightly more common in girls.

Table 2.5: Case-parent families by cleft type, gender and ancestry

	European		Asian		Total	
	Male	Female	Male	Female	Male	Female
Cleft Type						
CL/P	539	296	675	353	1214	649
CPO	132	138	103	155	235	293
Total by gender	671	434	778	508	1449	942
Total	1105		1286		2391	

CL/P is cleft lip with or without palate, CPO is cleft palate only

We first test to see if any SNPs violate parental allelic exchangeability with Equation 2.1, using all pairs of parents. We use a QQ plot of $-\log_{10}$(p-value) to look for violations in exchangeability. This allows an overall assessment of exchangeability, but we recognize that SNPs that are truly associated with oral cleft may tend to violate exchangeability.

We run PIX-LRT with the EM on Asian and Caucasian families together and separately, allowing for different mating type parameters for each ethnic category. When the analysis is on the individual populations, only SNPs with a MAF greater than 0.02 in each population are studied (11368 in Asians, 13156 in Caucasian). We test markers separately for cleft palate only (CPO) and cleft lip with or without palate (CL/P). The combined test statistic (1df) is used to combine information from families with affected sons and daughters. For female triads, we applied a log-additive risk model (1df). Plots of $-\log_{10}$(p-value) against the marker position along the X chromosome (as determined by Build 37) can identify regions of interest.

For CL/P, we compared our top five SNPs using PIX-LRT with EM to the top five identified in Patel et al. (Patel, Beaty et al. 2013). For these SNPs, we apply SSX-LRT and the parent-only analysis (Equation 2.5) to complete triads stratifying on sex of the affected offspring to better understand similarities and differences between our two results. SSX-LRT and the parent-only analysis are independent when complete triads are used, which enables estimates of relative risks to be compared in terms of agreement for parental versus offspring-based findings, and affected-boy families versus affected-girl families.

2.3 Results

2.3.1 Noncentrality Parameters

Under a null scenario where the relative risks are 1, the NCPs calculated for PIX-LRT and SSX-LRT (Equations 2.4 and 2.7) and XTDT are all zero, which ensures the nominal

Type I error rate. Table 2.6 displays the NCP and Type I errors calculated for the X-LRT for an admixed population. The NCPs are all greater than 0, implying inflated Type I error rates.

Table 2.6: Noncentrality parameter and corresponding Type I error rates for X-LRT. For 1000 triads for a null variant in an admixed population as calculated by X-LRT. Type I error rates for $\alpha = 0.05$ are shown in parenthesis. Scenario 1: First subpopulation has a MAF of 0.3 and a disease risk of 0.02 in males and females. Second subpopulation has a MAF of 0.2 and a disease risk of 0.01 in males and 0.02 in females. Scenario 2: First subpopulation has a MAF of 0.3 and a disease risk of 0.02 in males and 0.03 females. Second population has a MAF of 0.2 and a disease risk of 0.03 in males and 0.02 in females. H_0: no disease-locus effect in male or female, H_{0m}: no effect in males and test done in boy-affected families, H_{0f}: no effect in females and test done in girl-affected families.

	Scenario 1	Scenario 2
	X-LRT	X-LRT
H_0: $R_{G1} = R_{G2} = R_B = 1$	0.64 (0.09)	0.98 (0.11)
H_{0m}: $R_B = 1$	0.10 (0.06)	0.13 (0.07)
H_{0f}: $R_{G1} = R_{G2} = 1$	0.38 (0.08)	0.62 (0.10)

Figure 2.1 shows a plot of the power for 1000 triads at a Type I error rate of 5×10^{-6} with a range of minor allele frequencies. The one degree-of-freedom combined PIX-LRT analysis outperforms the two degree-of-freedom X-LRT analysis. Figure 2.2 shows plots of the estimated NCP and the corresponding power for 500 triads at a Type I error rate of 5×10^{-6} with varying disease relative risks. For complete triads, PIX-LRT has higher NCPs (and corresponding power) than both the SSX-LRT and the XTDT. The SSX-LRT and XTDT perform similarly (see Discussion). For instance, for 500 triads with affected sons and a SNP with a minor allele frequency of 0.3 and relative risk of 2, the PIX-LRT has an estimated NCP of 37.22 (power = 0.94), the SSX-LRT has an estimated NCP of 27.45 (power = 0.75) and the X-TDT has an estimated NCP of 26.92 (power = 0.73). For this scenario, the expected number of informative triads used in PIX-LRT is 347.31 compared to 242.31 in SSX-LRT and XTDT. These estimates decrease if the minor allele frequency is 0.1.

The NCP plots for 500 triads with affected daughters are similar to those of affected sons (Figure 2.2, right compared to left). Under a log-additive model for girls, if the disease

relative risk in sons equals the disease relative risk in heterozygous daughters ($R_B = R_{G1} = \sqrt{R_{G2}}$) then the estimated XTDT NCPs will be the same between the two sexes, and the estimated SSX-LRT NCPs will be the same (results not shown). The estimated PIX-LRT NCPs are close, but not identical. For R_{G1}=2 and a MAF of 0.3, the estimated NCP is 36.57 (power = 0.93). If instead, the disease relative risk in sons equals the disease relative risk in daughters with two copies of the variant allele, ($R_{B1} = R_{G2} = R_{G1}^2$), then for R_{G1}=$\sqrt{2}$, the estimated PIX-LRT NCP is 8.86 (power = 0.06). Under this scenario, triads with affected sons offer greater power than those with affected daughters.

Figure 2.1: Power estimates as a function of minor allele frequency of X-LRT and PIX-LRT. Each analysis is based on 1000 triads with affected sons and daughters. $R_B = 1.5$, $R_{G1}^2 = R_{G2} = 2$ and among non-carriers boys are twice as likely to have the disease as girls. Solid line represents PIX-LRT. Dashed line represents XLRT. PIX-LRT uses a 1 degree-of-freedom combined test, while X-LRT uses a 2 degree-of-freedom test.

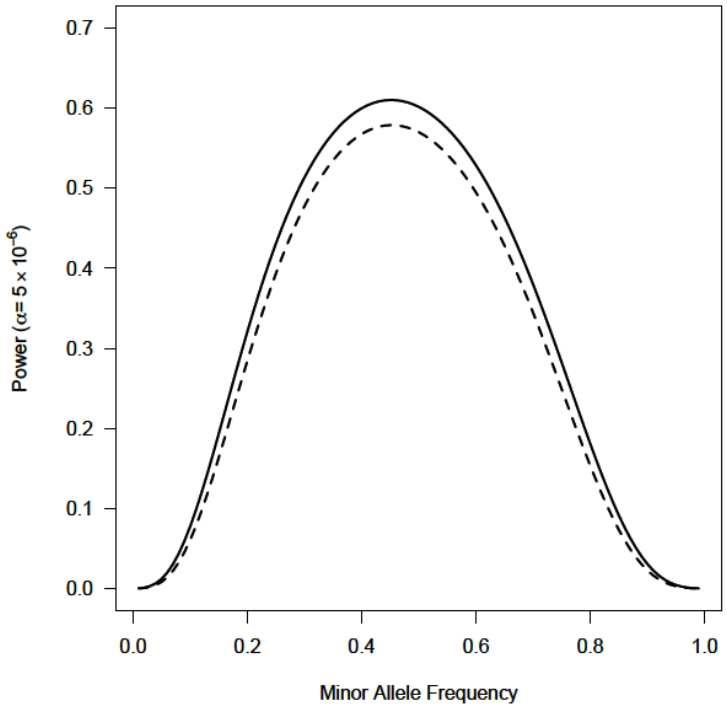
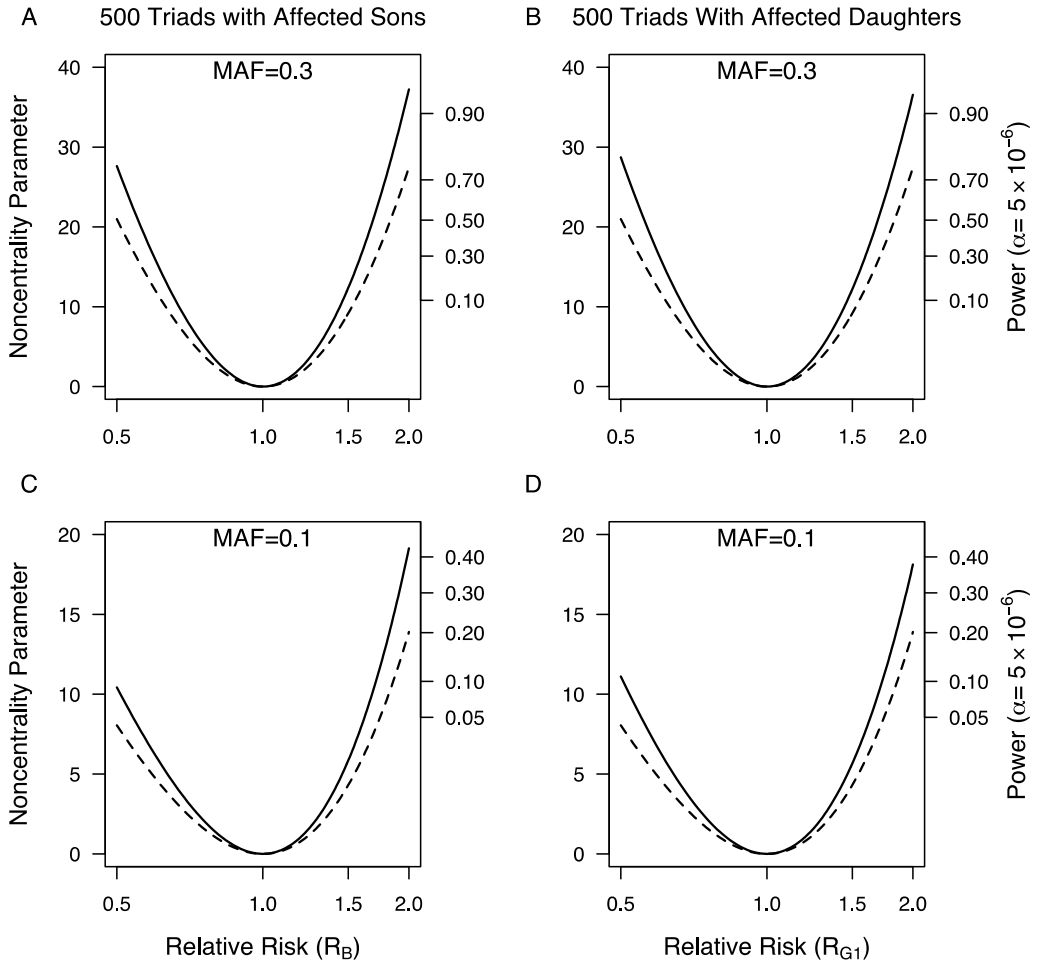


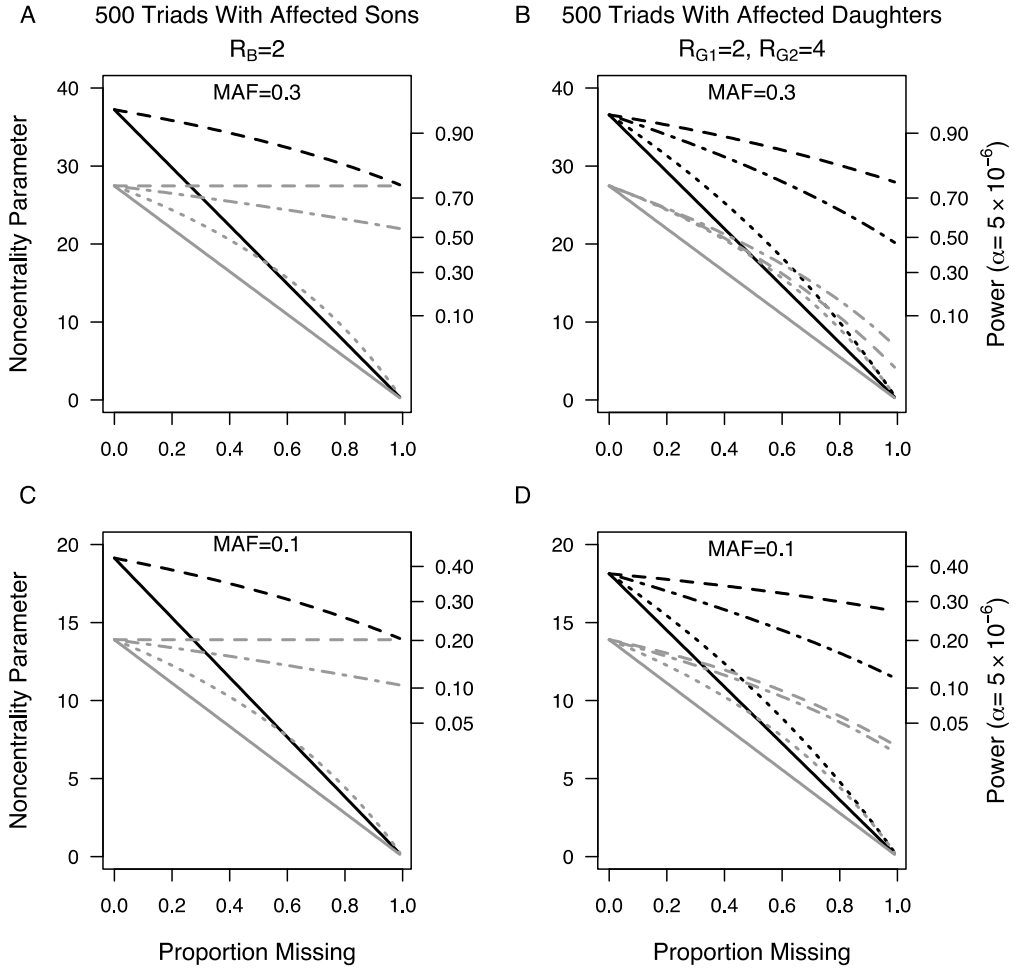
Figure 2.2: Noncentrality parameter estimates as a function of relative risk. (A, C) 500 triads with affected sons and (B, D) 500 triads with affected daughters. Minor allele frequencies of 0.3 (A, B) and 0.1 (C, D) are used. Solid lines represents PIX-LRT results. Dashed lines represents SSX-LRT and XTDT (plotted results were indistinguishable). PIX-LRT and SSX-LRT assume the relative risk in affected daughters is log additive in the number of copies of the minor allele.



Plots of the effect of missing genotype data on the estimated NCP and the corresponding power at a Type I error rate of $5x10^{-6}$ are shown in Figure 2.3. Regardless of minor allele frequency, for triads with sons, PIX-LRT with the EM algorithm works equally well when some mothers are missing as when some fathers are missing (proof not shown). The SSX-LRT does not lose any power when fathers of sons are missing, as the fathers are non-informative. When mothers of sons are missing, we see the greatest power loss. In triads

with daughters, regardless of minor allele frequency, more power can be recaptured from the EM when fathers are missing compared to mothers. This trend is seen in both PIX-LRT and SSX-LRT.

Figure 2.3: Noncentrality parameter estimates as a function of missing parental genotypes using the Expectation-Maximization (EM) algorithm. PIX-LRT and SSX-LRT were run on (A, C) 500 triads with affected sons and $R_B = 2$ and (B,D) 500 triads with affected daughters $R_{G1} = 2$ and $R_{G2} = 4$. Minor allele frequencies of 0.3 (A, B) and 0.1 (C, D) were used. Black lines represent PIX-LRT results. Gray lines represent SSX-LRT results. Solid lines represent results based on excluding incomplete triads (SSX-LRT can use triads with missing fathers). Dashed lines represent results based on triads with the fathers missing (for PIX-LRT this single line represents either parent missing). Dashed/dotted lines represent triads with either mother or father missing, with twice as many fathers missing than mothers. Dotted lines represent results based on triads with mothers missing.



2.3.2 Oral Cleft

The QQ plot to assess parental exchangeability in the SNPs is shown in Figure 2.4. Four SNPs (rs17269319, rs3747355, rs5906541, and rs12558269) are not shown because their p-values (as calculated from Equation 2.1) are extreme outliers, less than $1x10^{-16}$. No father was found to carry any of these SNPs, despite some missing fathers having evidently transmitted the allele to their daughter. We consequently had reason to doubt the quality of the genotyping for those SNPs and omitted them from further analysis. (Patel et al. (Patel, Beaty et al. 2013) also noted that rs17269319 and rs12558269 had poor intensity plots.) The remaining points fell nicely on the QQ plot, except for 5 SNPs (rs2710404, rs5921330, rs1573667, rs7060927, and rs2266806) that raised concern about the parental exchangeability assumption. If these SNPs had appeared as top SNPs in the PIX-LRT analysis, those findings would need a closer look.

Figure 2.4: QQ plot of $-\log_{10}(p)$ as calculated from the test of parental allelic exchangeability. 95% confidence intervals are shown. Four SNPs (rs17269319, rs3747355, rs5906541, and rs12558269) are not shown because of extremely low p-values. No fathers carried the minor alleles for these four SNPs and the quality of genotyping consequently appears to be inadequate.

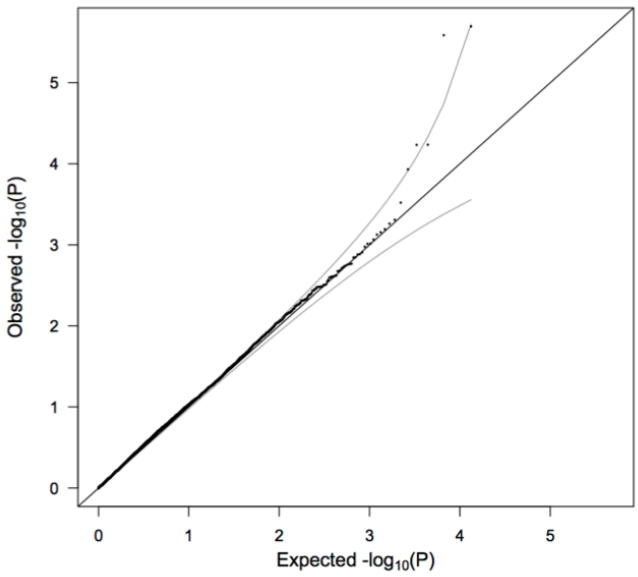
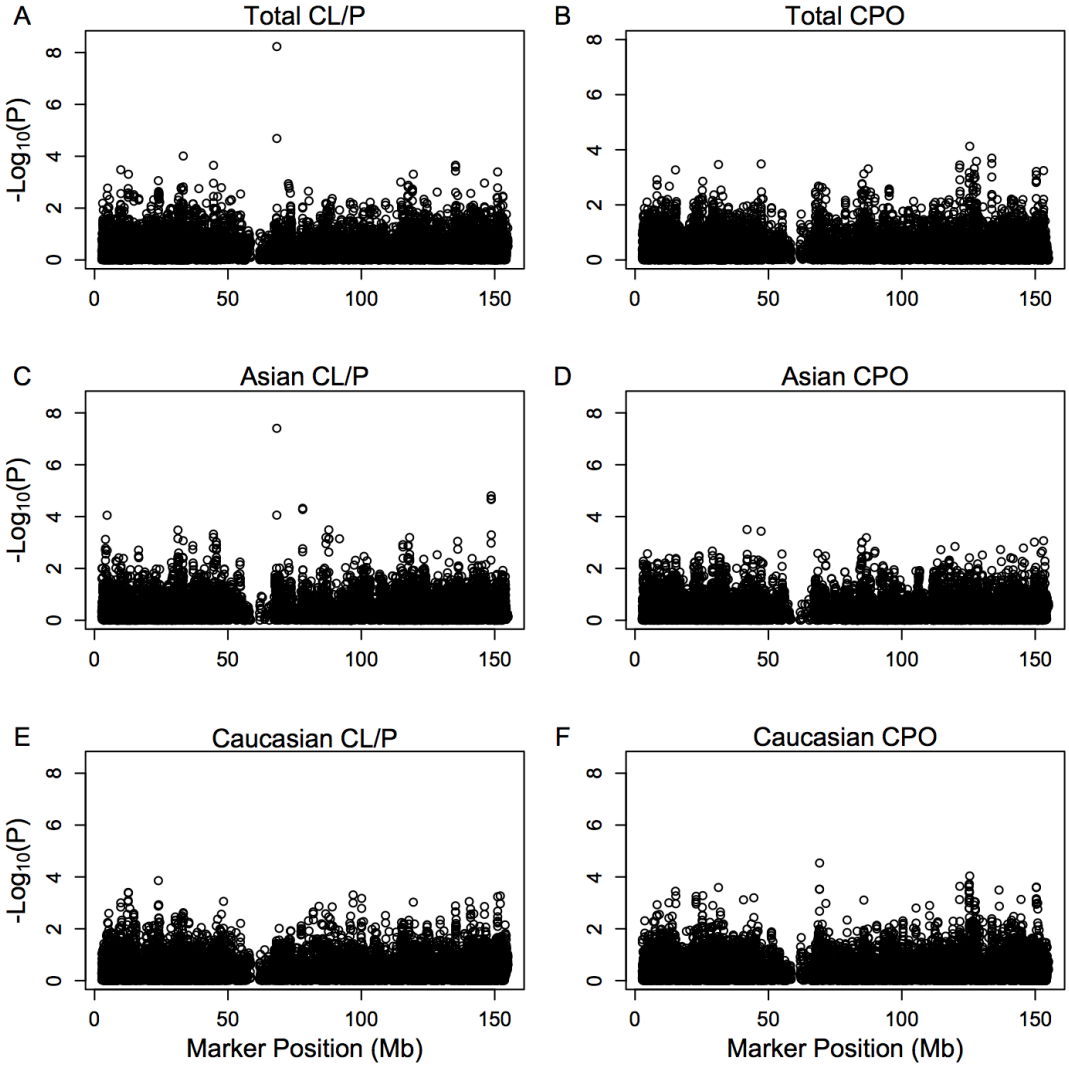


Figure 2.5 shows results of the PIX-LRT with EM analysis of the SNPs along the X chromosome for CL/P and CPO in Caucasians and Asians separately and combined. The CPO analysis did not produce results suggestive of a marker related to CPO and no SNPs had p-values below the Bonferroni-corrected 3.76×10^{-6}.

Figure 2.5: Individual single nucleotide polymorphism significance of the cleft example. The p-values (shown as $-\log_{10}(p)$) are calculated from PIX-LRT with the EM using dbGaP data from families with oral cleft. A log-additive model is assumed for the risk in affected daughters and a combined score is used to combine the sex-specific statistics. Models were run on cleft lip with or without cleft palate families amongst (A) Asians and Caucasians, (C) Asians only, (E) Caucasians only, as well as cleft palate only families amongst (B) Asian and Caucasians, (D) Asians only, (F) Caucasians only.



In CL/P analyses, we identified one SNP with a strong signal, rs5981162, the minor allele being associated with a decreased risk of cleft lip with or without palate (uncorrected p-value = 5.88×10^{-09}). The PIX-LRT estimated disease relative risk within the combined Asian and Caucasian populations was 0.48 for male offspring carrying the variant allele, and 0.56 for female offspring carrying one copy of the variant allele (0.32 for two copies), showing good concordance. Similar relative risks are estimated in separate analyses of the Asian and Caucasian populations (see Table 2.7). The evidence for an effect is particularly strong in the Asian population, which has a higher variant allele frequency, and hence more informative families than the Caucasian population. The effect estimates based on parents of girls and parents of boys were also in good agreement with the offspring-based estimates (see Table 2.8). By contrast the PIX-LRT analysis of rs5981162 with CPO shows no effect (see Table 2.7), suggesting phenotypic specificity. Additionally, the test for parental allelic exchangeability produced a p-value of 0.18 for rs5981162, suggesting no violation in the assumption.

Table 2.7: PIX-LRT analysis results of SNP rs5981162, located in the intergenic region between *ENFB1* and *PJA1* at basepair 68318753. PIX-LRT with the EM was run on Asian and Caucasians separately and together. A log-additive model was used for triads with affected daughters and the combined score was calculated with the results from the sex-stratified analysis.

Cleft	Population	MAF[B]	Inf. boy fams[A]	Inf. girl fams[A]	P-value	R_B	R_{G1}
CL/P	All	0.076	415	146	5.88×10^{-09}	0.48	0.56
	Asian	0.126	284	121	3.94×10^{-08}	0.49	0.54
	Caucasian	0.016	131	25	4.24×10^{-02}	0.38	0.72
CPO	All	0.076	78	65	0.544	0.85	0.91
	Asian	0.126	43	51	0.469	0.87	0.83
	Caucasian	0.016	35	14	0.895	0.77	2.35

[A] The number of informative triads at the marker.
[B]The minor allele frequency calculated from the parents in the population, not stratified by cleft type.

Table 2.8: Top 5 CL/P SNPs from our PIX-LRT analysis and from Patel et al. The top 5 from Patel et al. after excluding SNPs that raised genotyping concern. Parent informed X-LRT (PIX-LRT) with the EM, Sex stratified X-LRT (SSX-LRT) on complete data, and a parent only analysis on complete data are used to calculate relative risk for CL/P in the combined Asian and Caucasian families. Combined Z statistics are used as opposed to LRT statistics to show direction of effect and are calculated from the sex-stratified analysis, as mentioned in the methods section.

Our Top 5	Patel Top 5	Marker	Position	Gene	MAF[B]	Method	Comb. Z stat	R_B	R_{G1}
1	1	rs5981162	68318753	EFNB1, PJA1[A]	0.076	PIX EM	-5.82	0.48	0.56
						SSX	-4.79	0.48	0.58
						Parent only	-3.09	0.45	0.51
2	-	rs5980788	68315938	EFNB1, PJA1[A]	0.039	PIX EM	-4.26	0.49	0.54
						SSX	-4.01	0.45	0.53
						Parent only	-1.38	0.84	0.45
3	2	rs5928207	33244129	DMD	0.357	PIX EM	-3.90	0.73	0.87
						SSX	-4.72	0.67	0.74
						Parent only	-0.02	0.90	1.22
4	-	rs5930900	135296409	MAP7D3	0.370	PIX EM	3.69	1.22	1.30
						SSX	2.24	1.13	1.28
						Parent only	1.71	1.40	1
5	-	rs5905410	44584855	FUNDC1, DUSP21[A]	0.356	PIX EM	-3.69	0.81	0.72
						SSX	-3.39	0.81	0.67
						Parent only	-1.99	0.77	0.72
-	3	rs5928208	33253904	DMD	0.390	PIX EM	-3.09	0.76	0.92
						SSX	-4.60	0.65	0.75
						Parent only	1.34	1.07	1.57
-	4	rs6631759	33239353	DMD	0.289	PIX EM	-3.17	0.75	0.90
						SSX	-4.45	0.66	0.73
						Parent only	0.55	0.99	1.33
-	5	rs5971698	33245234	DMD	0.366	PIX EM	-3.08	0.76	0.93
						SSX	-4.33	0.65	0.83
						Parent only	1.02	1.09	1.32

[A] This marker lies in the intergenic region between the genes shown.
[B]The minor allele frequency calculated from the parents in the population, not stratified by cleft type.

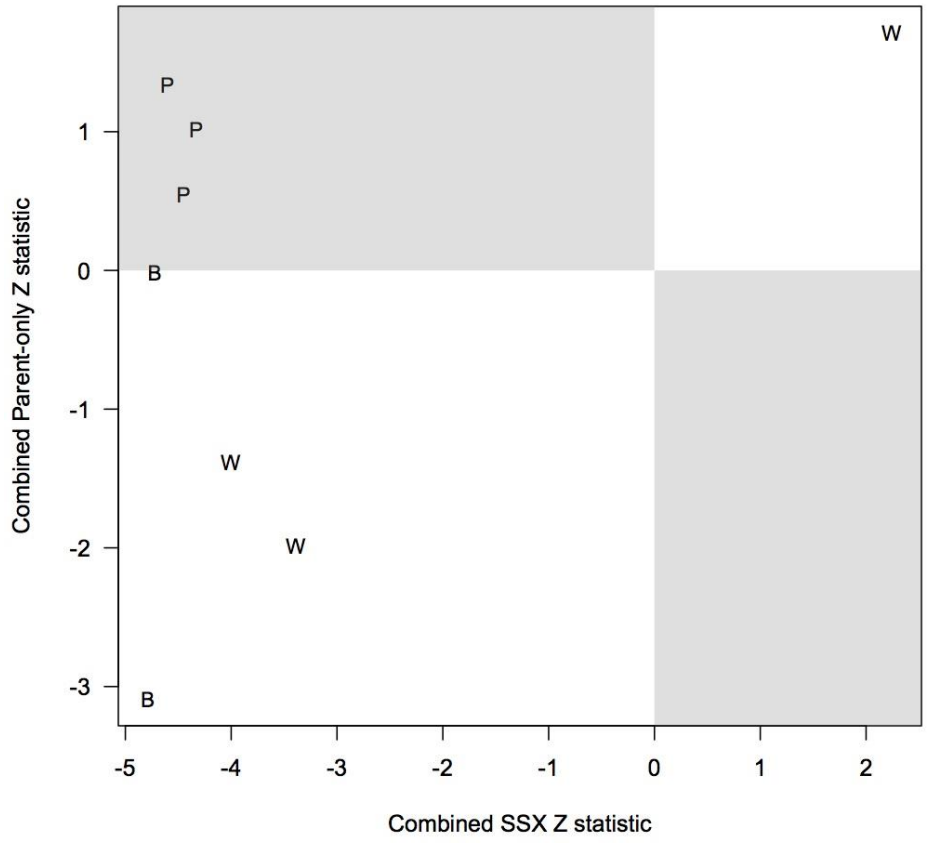
Table 2.8 compares the top 5 SNPs (based on the combined p-value) for CL/P within our analysis using PIX-LRT with EM and the top 5 SNPs based on the Patel et al. [28] analysis. The top 5 SNPs from our analysis all showed no violation in the test of parental allelic exchangeability (the smallest p-value was 0.18). The SNPs rs17269319, rs5906541

and rs12558269 were excluded for quality control reasons, as discussed above (see Table 2.8 and Discussion). Two SNPs were in the top 5 under both analyses: rs5928207 and rs5981162 (our top hit). All triads for the 8 SNPs were analyzed with PIX-LRT with EM, and separate analyses using SSX-LRT and the parent only method, carried out to assess agreement, were based only on complete triads, to guarantee statistical independence. The combined Z-score (Equation 2.5) is shown in the table. Figure 2.6 plots the offspring-based SSX versus the parent-only Z-score. SNPs that ranked high in the Patel analysis also have large SSX Z-scores. Evidence for a true hit is strengthened if the signs of the two independent statistics are in agreement, i.e. the points should ideally fall in the southwest or northeast quadrants of the figure. PIX-LRT identifies SNPs that have high offspring-based SSX and parent-only Z-scores in the same directions (i.e. concordance), the white-background quadrants in the plot. The PIX-LRT top hit, rs5981162, is in the southwest quadrant, showing strong evidence of a protective effect in both the offspring-based SSX and the parent-based analysis.

2.4 Discussion

We have introduced new methods to analyze SNPs on the X chromosome: the SSX-LRT and the PIX-LRT. The SSX-LRT allows for stratification by sex of the affected offspring, which is based on the X-LRT but confers robustness against population stratification. The PIX-LRT then builds on the SSX-LRT by incorporating additional information in the parental genotypes that previous methods have not exploited. This information allows PIX-LRT to gain substantial power in identifying SNPs on the X chromosome associated with disease risk.

Figure 2.6: Assessment of concordance through comparison of the parent-only Z scores and transmission (SSX) Z scores. The figure shows the top five single nucleotide polymorphism (SNP) hits from the PIX-LRT analysis and Patel el al. in Asian and Caucasian families with cleft lip with or without palate. We excluded SNPs in the Patel et al. analysis that raised quality control concerns. The parent-only and SSX analyses assume the relative risk in affected daughters is log additive in the number of copies of the variant allele. A combined score is used to combine the sex-specific statistics. "B" represents SNPs in the top 5 under both analyses, "P" represents SNPs that were in the top 5 for Patel et al. but not for us, "W" represents SNPs that were in the top 5 for our analysis but not Patel et al.



For situations in which both PIX-LRT and X-LRT are appropriate, under an assumed log-additive model for girls, the combined PIX-LRT outperforms X-LRT. The combined PIX-LRT enables a one degree-of-freedom test to be run. No assumption about the relationship between the male and female relative risks is made. Under this scenario, the X-LRT loses some power because it is a two degree-of-freedom test. For the X-LRT to be a one

degree-of-freedom test, a relationship between the boy and girl relative risk must be asserted, and such a model may be mis-specified. It should be noted, however, that if the directions of the relative risks in boys and girls are opposite, then PIX-LRT loses power, while X-LRT does not.

As we showed in the results section, the parent-only portion of the PIX-LRT can also be used independently of the offspring-based transmission portion as a form of replication. This assessment of replication can only use complete triads however if the offspring-based and parent-based tests are to remain independent. The SNP that we identified as strongly protective based on PIX-LRT showed replication both across ethnic groups, across boys versus girls and across parent-based results versus offspring-based results, strengthening evidence for effect.

In general, with use of the EM algorithm more power is recaptured with missing fathers than with missing mothers (cf. Figure 2.3). This difference is driven by the daughter cases. For daughters, we can infer the father's genotype as long as the mother and daughter are not both heterozygous. However, with only the father's and the daughter's genotypes, we cannot know the mother's genotype. For boys, in a transmission-based test (e.g. SSX-LRT), only the mothers are informative, so missing fathers do not affect the power of the test. However, in PIX-LRT, fathers are informative, and so when fathers of sons are missing, power is lost. For sons, when either parent is missing, the genotype of the complete triad cannot be known. For families with one parent and an affected son, the parents turn out to be equally informative (proof not shown).

While we demonstrated use of the EM algorithm for triads with a missing parent, there are circumstances where the genotype for the affected offspring might be missing. For

example, in studying a defect such as anencephaly, following prenatal diagnosis a medically-indicated abortion might have been conducted. For families where only the parental genotype data is available, if the sex is known, the parent-only portion of the PIX-LRT can still be used in analyses of potential effects of variants on the X chromosome.

When, as in the oral cleft data used, families have missing parents, the EM enables use of their information. However, use of the EM can induce bias if a population has multiple subpopulations with both the minor allele frequencies and the extent of missingness varying across subpopulations. This bias is not specific to our method, and can be avoided via stratification if the subpopulations are identifiable.

For X-chromosome-wide association studies using case-parent triads, the power to detect an effect is influenced by the sex of the affected offspring. If the disease relative risk for a heterozygous female is less than that for a male carrier, as may be the case due to X-inactivation, the estimated power derived from the PIX-LRT, SSX-LRT and X-TDT would typically be less for triads with daughters than for those with sons (Figure 2.2). Furthermore, for both SSX-LRT and PIX-LRT, missing mothers are at least as costly as missing fathers in their effects on power (Figure 2.3).

Some limitations deserve mention. The PIX-LRT estimates can be biased if an allele violates the parental exchangeability assumption, in which case the SSX-LRT may be a more appropriate method. In analyzing the oral cleft data we excluded the small fraction of differing-ethnicity parents, but including them did not noticeably affect the exchangeability QQ plot (data not shown). Transmission-based tests may also be biased if the violation is due to genotyping error or because the SNP is associated with fetal survival. If a SNP affects risk through a maternal effect (Wilcox, Weinberg et al. 1998), the parental contribution to the

PIX-LRT results may be biased. Current research is extending the PIX-LRT to accommodate maternal effects.

Furthermore, the PIX-LRT and other X-chromosome methods are not suitable for the pseudo-autosomal regions (PAR) and the X-chromosome-transposed region (XTR). These regions have homologous regions on the Y chromosome, so that a male can have two copies of a SNP.

The NCP estimates obtained from SSX-LRT and the XTDT are similar because these two tests are, respectively, the likelihood and score test for the same model. Schaid et al. (Schaid and Sommer 1994) showed that the TDT is the score test for a logistic regression allele dosage model (log additive). One can similarly show this for the XTDT.

We applied PIX-LRT to an international consortium of genotyped families affected by the birth defect oral cleft. In a previous analysis of the data, some of the most significant SNPs identified by Patel et al. [28] were not as significant when analyzed with PIX-LRT. An example is SNP rs5928208, which showed weaker results with PIX-LRT because the effect seen from the transmission analysis was not evident in the parent-only analysis. The top two SNPs in Patel, rs5906541 and rs17269319, and also rs3747355 and rs12558269, violated the mating exchangeability assumption. A harder look at the family genotypes was revealing in that their apparent absence in the fathers and the sons who were genotyped (as opposed to their inferred presence in fathers who were missing) raised concerns over the quality of genotyping for those SNPs.

With PIX-LRT, we identified rs5981162 as having a strong and protective effect on cleft lip with or with palate. This SNP was ranked fairly high in the previous analysis of the data by Patel et al. [28], but PIX-LRT estimated sex-specific relative risks and found

estimation concordance and a stronger p-value signal. The rs5981162 SNP is located between genes *EFNB1* and *PJA1*, and is downstream of these two genes. *EFNB1* is known to play a role in facial development: mutations on *EFNB1* are responsible for the majority of cases of craniofrontonasal syndrome (CFNS) (Twigg, Kan et al. 2004, Wieland, Jakubiczka et al. 2004), whose features can include cleft lip and palate. The SNP rs5981162 may potentially be located in a regulatory region of the *EFNB1* gene and functional studies could be illuminating.

CHAPTER 3: PIX-LRT EXTENSIONS FOR MATERNAL EFFECTS OF GENETIC VARIANTS ON THE X CHROMOSOME

In this chapter we extend the method developed in Chapter 2, PIX-LRT, to enable identification of effects of variants on the maternal X chromosome that can influence the later health of the offspring by modifying the prenatal environment. By taking advantage of an assumption of allelic exchangeability, the proposed method is able to distinguish such maternal effects from effects due to fetal inherited variants, and can provide estimates of relative risks. We apply PIX-LRT to publically available data from an international consortium of genotyped families affected by the birth defect oral cleft to test for potential maternal effects.

3.1 Introduction

The maternal genotype is of obvious relevance for pregnancy complications like preterm birth and pre-eclampsia. It can also influence the intra-uterine environment both directly and through its role in modulating the metabolism and effects of feto-toxic exposures. It is consequently of interest to investigate effects of variants carried by the mother (regardless of their transmission to the affected offspring), especially in relation to effects on early-onset disease, such as on mental illnesses, childhood cancers, and birth defects. Using family-based studies, maternal effects of variants on the autosome have been shown to be associated with a number of childhood diseases, including childhood medulloblastoma (Lupo, Nousome et al. 2012), clubfoot (Weymouth, Blanton et al. 2011) and oral cleft (Jugessur, Shi et al. 2010).

Mitchell (Mitchell 1997) first noted that an advantage of the family-based study design over the case-control study design, was that it could potentially be used to differentiate between maternal and fetal effects. In case-control studies that do not genotype the mothers, such distinctions cannot be made and maternal effects cannot be directly studied. Such effects can confound results, because the genome of the affected offspring and that of the mother are causally correlated. By contrast, with a family-based design, researchers can probe potential maternal effects. Mitchell suggested applying the TDT to a 3-generation study in which cases, parents, and maternal grandparents are genotyped. If a maternal effect is involved in the etiology, a causative variant allele will have been preferentially transmitted from heterozygous maternal grandparents to the affected child's mother. While conceptually appealing, this multi-generational family design may be logistically very hard to implement, because the maternal grandparents may be hard to locate or unwilling to be studied.

For case-parent triads and autosomal markers, Weinberg, Wilcox et al. noted that if one can assume mating symmetry in parental genotypes, grandparents are not needed. One can extend the log-linear model to detect the effects that act through maternal mechanisms, while adjusting for fetal genotype effects (Weinberg, Wilcox et al. 1998, Wilcox, Weinberg et al. 1998). For the log-linear model as applied to the autosome, the assumption of genetic mating symmetry for the parents in effect permits the paternal genotype to serve as control for the maternal genotype. This model enables maternal-effect relative risks to be estimated under an assumption that the effect is recessive, dominant, log-additive or fully unconstrained, i.e. co-dominant. The EM (expectation-maximization) algorithm can be used to handle missing autosomal SNP genotypes (or individuals) (Weinberg 1999, Rampersaud,

Morris et al. 2007). Unlike the method proposed by Mitchell, this method only requires case-parent triads and can account for missing data through use of the EM. This model allows for population structure, as HWE is not assumed, though mating must be non-assortative within genetic subpopulations with respect to the variants under study.

Although those methods have been widely applied to studies of the autosome, there is a dearth of methods to study maternal X chromosome genetic effects. The X chromosome is unique in that females are diploid while males have only one, maternally-derived copy. There is random inactivation of an X in each cell early in female embryonic development as a form of dosage compensation (Lyon 2002). To date the only method available to test for maternal effects in case-parent triads is HAPLIN (Gjessing and Lie 2006), which requires the assumption of Hardy-Weinberg Equilibrium (HWE) and consequently is not robust against population stratification. HAPLIN is a likelihood-based method for analyzing maternal and fetal haplotypes in case-parent triads. Single-dose effects (effects of one copy of the haplotype) and double-dose effects of maternal haplotypes (or single SNPs) can be estimated as relative risks using the model (as described in section 1.2.5). A version of HAPLIN that has been developed for the X chromosome allows the user to analyze maternal effects of X chromosome variants (Myking, Boyd et al. 2013).

We developed PIX-LRT (the parent-informed X chromosome likelihood ratio test)(Chapter 2) as a method to measure fetal SNP effects of X chromosome variants using information from both the transmission of a variant X allele from parents to affected offspring, and information related to the distribution across parents' genotypes specific to the sex of the affected offspring. An assumption of "parental allelic exchangeability" enables the added parental information to be captured in a way that resists bias due to genetic population

stratification. Here we show that the same assumption also allows an extension of PIX-LRT to distinguish maternal from fetal effects and enables estimation of relative risks for maternally-mediated effects.

In the following sections, we initially describe the PIX-LRT extension for testing maternal effects of single X-linked SNP markers when case-parent genotype data are complete. We show that without the assumption of parental allelic exchangeability the maternal effects could not be statistically identified. The EM algorithm can be used to maximize the likelihood when some families have missing SNP genotype data. We assess Type I error rates and power for testing maternal effects with PIX-LRT by calculating chi-squared noncentrality parameters based on expected counts (Agresti 2012). We consider scenarios in which population structure is present, and in which there may also be varying degrees of a concomitant direct effect of the inherited allele on the offspring. As an example application, we apply the PIX-LRT to family data from a large family-based oral cleft dataset to analyze maternal effects of SNP markers on the X chromosome. We conclude with a discussion of the advantages and limitations of using PIX-LRT to study maternal effects, and also discuss our SNP findings.

3.2 Subjects and Methods

3.2.1 Case-Parent Design and Assumptions

We consider a sample of genotyped case-parent triads, where all sampled offspring have been diagnosed with the condition of interest. For a di-allelic locus, let M, F, and C denote the number of copies of the variant (minor) allele in the mother, father and affected offspring (proband), respectively. We exclude the pseudo-autosomal regions and the X-transposed region (PARs, XTR), as these regions on the X correspond to a homologous

region on the Y. Then $M \in \{0,1,2\}, F \in \{0,1\}, C \in \{0,1\}$ for male offspring, and $C \in \{0,1,2\}$ for female offspring.

We make similar assumption as in Chapter 2. We assume there is Mendelian transmission at the locus in the source population and that, although the condition under study may reduce the likelihood of survival to birth, neither the maternal nor the fetal genotype influences that survival likelihood, conditional on the occurrence of the condition. Further assume parental allelic *exchangeability* in the source population. This assumption was assumed for the autosome in Min et. al (Shi, Umbach et al. 2008) and states that, within a mating pair, the total copies the father and mother carry of the variant allele are randomly located across their three X chromosomes. Note that parental allelic exchangeability is much less restrictive than HWE because it must hold only within mating pairs and allows allele frequencies to differ across genetic subpopulations. We will also initially assume that a maternal effect has the same multiplicative effect on risk for male and female offspring.

3.2.2 PIX-LRT Extension to Maternal Effects

The method we will describe extends PIX-LRT to enable study of maternal effects. Briefly, for case-parent triads, PIX-LRT takes advantage of information in parents to improve the power to detect an effect of an X variant in the offspring inherited genotype (see Chapter 2). Under the null hypothesis that a SNP on the X chromosome is unrelated to risk, one would expect neither the mothers nor the fathers to be enriched for either allele. However, if there is a fetal genetic effect at that locus, we showed that selection based on affected offspring induces a distortion in the distribution of the marker in the parents. The direction of this distortion depends on the sex of the affected offspring, and can be exploited under parental allelic exchangeability. We showed that if we condition on the total number of

alleles carried by the parents, *M+F*, the information from asymmetries in the parental X genotypes can be combined with transmission information from parents to affected offspring via a log-linear model.

If there is a maternal effect, then even if there is no effect of the fetal genotype that effect will produce a distortion of the allelic distribution across the parents. For example, consider a variant allele that if carried by the mother, increases the fetus' risk of disease. Then the mothers of affected offspring will be enriched for this variant compared to their mates. Unlike the fetal effect, this distortion will be in the same direction in parents of affected sons and affected daughters. If an assumption of parental allelic exchangeability is made, and we condition on (*M+F*), we can measure this distortion. However, if instead we condition on the mating type, (*M, F*), we cannot identify a maternal effect, but only the fetal genotype effect, for which only heterozygous mothers are informative. Because a maternal effect and a fetal effect will both cause distortion, when testing for a maternal effect, (unless it is known that there is no fetal effect) it can be important to allow for a fetal genotype effect in the model.

Let "aff" denote the event that the offspring (or pregnancy) is affected and define the relative risks for fetal and maternal genotypes as follows:

$$e^{\beta_1} = R_{G1} = \Pr(\text{aff}|\text{girl}, C = 1, M, F) / \Pr(\text{aff}|\text{girl}, C = 0, M, F)$$

$$e^{\beta_2} = R_{G2} = \mathrm{R_{G1}}\Pr(\text{aff}|\text{girl}, C = 2, M, F) / \Pr(\text{aff}|\text{girl}, C = 1, M, F)$$

$$e^{\beta_3} = R_B = \Pr(\text{aff}|\text{boy}, C = 1, M, F) / \Pr(\text{aff}|\text{boy}, C = 0, M, F)$$

$$e^{\alpha_1} = R_{M1} = \Pr(\text{aff}|M = 1, C, F) / \Pr(\text{aff}|M = 0, C, F)$$

$$e^{\alpha_2} = R_{M2} = \Pr(\text{aff}|M = 2, C, F) / \Pr(\text{aff}|M = 0, C, F)$$

A likelihood that includes both fetal and maternal effects can be written as a multinomial and modeled in a log-linear form as follows:

$$\ln\left(E\left[N_{M,F,C,sex}|M+F\right]\right)$$
$$= \log\left(\mu_{M+F,sex}\right) + \beta_1 I_{(C=1,sex=g)} + \beta_2 I_{(C=2,sex=g)} + \beta_3 I_{(C=1,sex=b)} +$$
$$\alpha_1 I_{(M=1)} + \alpha_2\, I_{(M=2)} \qquad (3.1)$$

Here, $\mu_{(M+F,sex)}$ are nuisance parameters that serve to stratify families by conditioning on the sum of parental genotypes. (This model is equivalent to a polytomous logistic regression model that conditions on the total number of families with the given *M+F*.) The expected counts are shown in Table 3.1. In the above models $I_{(K)}$ is an indicator variable equal to 1 if *K* is true, and 0 otherwise. The corresponding log likelihood would be:

$$\ell \sim \sum_{M,F,C,sex} n_{M,F,C,sex} \log(\Pr(M,F,C|M+F,\text{aff},\text{sex}) * \Pr(M+F|\text{aff},\text{sex})) \qquad (3.2)$$

If no assumption is made about the relative sizes of the maternal relative risks for a mother with 1 versus 2 copies of a variant allele, then a test for co-dominant maternal effects will be:

$$H_0: R_{M1} = R_{M2} = 1 \quad (\alpha_1 = \alpha_2)$$
$$H_a: R_{M1} \neq 1 \text{ or } R_{M2} \neq 1 \quad (\alpha_1 \neq 0 \text{ or } \alpha_2 \neq 0)$$

Under this fully general genomic model the likelihood ratio test statistic given by -2($\ell(H_0) - \ell(H_a)$) is distributed chi-square with two degrees of freedom, where $\ell(H_0)$ is the maximized likelihood under the null and $\ell(H_a)$ is the maximized likelihood under the co-dominant alternative. This statistic and the relative risk estimates can be calculated with standard generalized linear model (GLM) software. The maternal effect can also be tested under a log-additive, dominant or recessive model.

Table 3.1: For affected sons and daughters, case-parents triad frequencies under an assumption of parental allelic exchangeability.

			Affected Sons		Affected Daughters	
M+F	M	F	C	Triad Frequency	C	Triad Frequency
0	0	0	0	μ_{0b}	0	μ_{0g}
1	0	1	0	μ_{1b}	1	μ_{1g}
	1	0	0	$\mu_{1b}R_{M1}$	1	$\mu_{1g}R_{M1}$
	1	0	1	$\mu_{1b}R_BR_{M1}$	0	$\mu_{1g}R_{G1}R_{M1}$
2	1	1	0	$\mu_{2b}R_{M1}$	2	$\mu_{2g}R_{G2}R_{M1}$
	1	1	1	$\mu_{2b}R_BR_{M1}$	1	$\mu_{2g}R_{G1}R_{M1}$
	2	0	1	$\mu_{2b}R_BR_{M2}$	1	$\mu_{2g}R_{G1}R_{M2}$
3	2	1	1	$\mu_{3b}R_BR_{M2}$	2	$\mu_{3g}R_{G2}R_{M2}$

3.2.3 Type I Error and Power Calculations

As in Chapter 2, we calculate the power and the Type I error rate by calculating the noncentrality parameter (NCP) for the distribution of a chi-squared likelihood ratio test statistic. Under the null hypothesis of no maternal effect, the LRT statistic follows a central chi-squared distribution, which has an NCP of 0. The NCP is calculated by treating expected triad counts under the specified population structure as data used to fit the relevant models (O'Brien 1986, Agresti 2012).Values of noncentrality parameters can be translated to power values using the noncentral chi-squared distribution with the appropriate degrees of freedom and multiplying any given NCP by the ratio of the sample size contemplated to the sample size used in the calculation.

Under the null hypothesis of no maternal effect, we will show that the LRT statistic follows a central chi-squared distribution, which has an NCP of 0, even when genetic subpopulations are present in the population. As would be expected, because it relies on HWE, HAPLIN is biased under such a situation. Consider two scenarios where

subpopulations are present within a population: (1) there are no effects of the fetal genotype ($R_{G1} = R_{G2} = R_B = 1$) nor are there maternal effects ($R_{M1} = R_{M2} = 1$); or (2) there are fetal ($R_{G1}^2 = R_{G2} = 2, R_B = 1.5$) effects but no maternal effects. In both scenarios there are 1000 families with affected offspring, one subpopulation has a minor allele frequency of 0.1, a disease risk of 0.02 in males, and 0.02 in females. The second subpopulation (of the same size as the first) has a minor allele frequency of 0.2, a risk of 0.01 in males, and of 0.02 in females.

In addition to calculating NCP (and the corresponding Type 1 error rate), we will also perform 10000 simulations to estimate power at a 0.05 alpha level under a range of alternative scenarios. We use a co-dominant model for maternal effects, so that a two degree-of-freedom LRT chi-squared test statistic is calculated. The current version of the HAPLIN software in R (R Development Core Team 2013) allows for maternal effects to be tested on the X chromosome, but imposes some limitations to the flexibility of the model. For instance, in its current implementation, male and female triads can either be run in separate models, or a relationship between the relative risks can be assumed. We wanted to be able to include males and females in the same model without imposing a relationship between their relative risks (necessitating three parameters). We therefore recoded the HAPLIN procedure in R with the glm procedure to allow for this less restrictive scenario. To verify our coding, we compared results from our code with those from HAPLIN under the constraints HAPLIN currently assumes. Results were in agreement.

We also compare the power of PIX-LRT to HAPLIN, under a homogeneous population. For comparability we assume HWE so that HAPLIN is unbiased, which should confer an advantage to HAPLIN. For the power simulations we consider 1000 triads

consisting of a mix of affected sons and daughters. We test for maternal effects in the presence of log-additive fetal effects ($R_{G1}^2 = R_{G2} = 2, R_B = 1.5$) but (given that analysts lack omniscience) in the analysis we allow three parameters for the fetal effects. In our scenarios, for non-carriers of both sexes, the risk of disease is the same. We assume a log-additive maternal effect ($R_{M1}^2 = R_{M2}$). We plot the noncentrality parameters as a function of the relative risk (R_{M1}) for two scenarios; one where the variant allele frequency is 0.05 and another where the allele frequency is 0.2. In the plot we include the corresponding power for a one-degree-of-freedom LRT at alpha level $5x10^{-6}$. We use an alpha level of $5x10^{-6}$ because this approximates the alpha 0.05 Bonferroni-corrected value that would be needed for the X chromosome based on the Illumina Human610-Quad v1.0 Build 36.

To include families where some individuals are missing, we use the EM algorithm with PIX-LRT. We calculate and plot the power for an analysis that simply discards the incomplete families, to assess the loss of power, and we quantify the recapture of power that can be achieved by using the EM. Specifically, we plot the noncentrality parameter (and the power at alpha level $5x10^{-6}$) as a function of the proportion of missing fathers, missing mothers, or a combination. For the combination scenarios, only one parent is missing, and twice as many fathers as mothers are assumed missing. As above, we will assume fetal effects are present ($R_{G1}^2 = R_{G2} = 2, R_B = 1.5$) and analyze the data using 3 risk parameters for the fetal effects. Furthermore, we set the minor allele frequency to 0.2, the relative risk associated with the mother carrying one copy of the variant allele to 1.7, and that for two variants to $2.89 = 1.7^2$ (thereby imposing a log-additive effect).

3.2.4 Oral Cleft Data

Under a log-additive model (1 df) for maternal effects and a co-dominant model to

allow for possible fetal genetic effects, we use PIX-LRT with the EM to test for maternal effects in the oral cleft dataset. Details of this dataset are described in Chapter 2. Asian and Caucasian triads will be analyzed separately and together. When analyzed together, we allow for different mating type parameters for each ethnic category to ensure validity. Additionally, we test markers separately for cleft palate only (CPO) and cleft lip with or without palate (CL/P) as distinct phenotypes. Because we include more variables in the model than we did in Chapter 2 and sparseness of data may become a problem here, we increase the minor allele cutoff to 0.05. As a result, there are 10571 SNPs that pass this screen amongst Asians, 12417 SNPs amongst Caucasians, and 12365 SNPs amongst the combined populations. The appropriate alpha for a Bonferroni-corrected family-wise error rate of 0.05 are 4.73×10^{-6}, 4.03×10^{-6} and 4.04×10^{-6}, respectively. As before, we exclude SNPs that raise quality control concerns. Plots of $-\log_{10}$(p-value) against the marker position along the X chromosome (as determined by Build 37) are constructed as an X-based Manhattan plot to display the results.

3.3 Results

3.3.1 Noncentrality Parameters

Table 3.2 displays noncentrality and power results. Under a null scenario where the maternal genotype relative risks are 1, the NCPs calculated for PIX-LRT are zero, which ensures the nominal Type I error rate. Within an admixed population, whether fetal effects are absent (scenario 1) or present (scenario 2), a test for maternal effects using PIX-LRT produces NCPs of 0. Based on 10000 simulations, for a nominal alpha of 0.05, the Type I error rates were 0.0512 and 0.0518 for scenarios 1 and 2, respectively, which are statistically compatible with a true rate of 0.05. By contrast, HAPLIN produces NCPs greater than 0, implying inflated Type I error rates. There is greater inflation in the presence of fetal effects; based on 10000 simulations, for a nominal alpha of 0.05, the Type I error rates were 0.088

and 0.099 for scenarios 1 and 2, respectively.

Table 3.2: Noncentrality parameters and corresponding Type I error rates in parentheses for PIX-LRT and HAPLIN. Simulated Type I error rates for 1000 triads for a null maternal variant in an admixed population[A] as calculated by PIX-LRT and HAPLIN[B].

	PIX-LRT		HAPLIN[B]	
Scenario	NCP	Simulation	NCP	Simulation
1. $R_{G1} = R_{G2} = R_B = 1$	0 (0.05)	0.048	0.504 (0.090)	0.088
2. $R_{G1}^2 = R_{G2} = 2$, $R_B = 1.5$	0 (0.05)	0.049	0.624 (0.100)	0.099

A 2 degree-of-freedom test is performed for maternal effects. NCP corresponding Type I error rate for $\alpha = 0.05$ are show in parenthesis. For simulations, Type I error rate for $\alpha = 0.05$ calculated from 10000 runs. Scenario 1: no fetal effects present. Scenario 2: fetal effects present.
[A]First subpopulation has a MAF of 0.1, a disease risk of 0.02 in males and females. Second subpopulation has a MAF of 0.2 and a disease risk of 0.01 in males and 0.02 in females.
[B]HAPLIN recoded in R to allow 3 parameters for fetal effects.

Figure 3.1 shows plots of the estimated NCP and the corresponding power for studies with1000 triads and applying a Type I error rate of $5x10^{-6}$ with varying maternal effect relative risks. For detecting maternal effects, the trade-off between assumptions and power is evident; PIX-LRT (which is more generally valid) has NCPs (and corresponding power) that are less than that of HAPLIN. For instance, for 1000 triads with affected offspring and a SNP with a minor allele frequency of 0.2, if the relative risk associated with the mother carrying one copy of the variant allele is 1.7, then PIX-LRT has an estimated NCP of 27.97 (power = 0.77) and the HAPLIN an estimated NCP of 32.11 (power = 0.86). These estimates are both smaller for smaller minor allele frequencies.

Figure 3.2 shows the effect of missing genotype data on the estimated NCP and the corresponding power at a Type I error rate of $5x10^{-6}$. The EM recaptures more power when mothers are missing compared to fathers, suggesting that fathers are slightly more informative than mothers for identifying maternally-mediated effects. For example, when 20% of mothers are missing, the NCP when the EM is used is 26.63 (power = 0.72), whereas when 20% of fathers are missing the NCP is 25.79 (power = 0.69). However, in both of these

cases, the EM allows for an increase in power; if only complete families had been analyzed in these situations the NCP would be 22.38 (power = 0.57).

Figure 3.1: Noncentrality parameters as a function of maternal relative risk. Shown for three relative risks, 0.2 (top), 0.1 (middle) and 0.05 (bottom). Each is calculated for 1000 triads with sons and daughters. We test for maternal effects in the presence of log-additive fetal effects ($R_{G1}^2 = R_{G2} = 2, R_B = 1.5$) but allow three parameters for the fetal effects. Non-carrier risk in males and females is the same. Solid line is PIX-LRT, dashed line is modified HAPLIN.

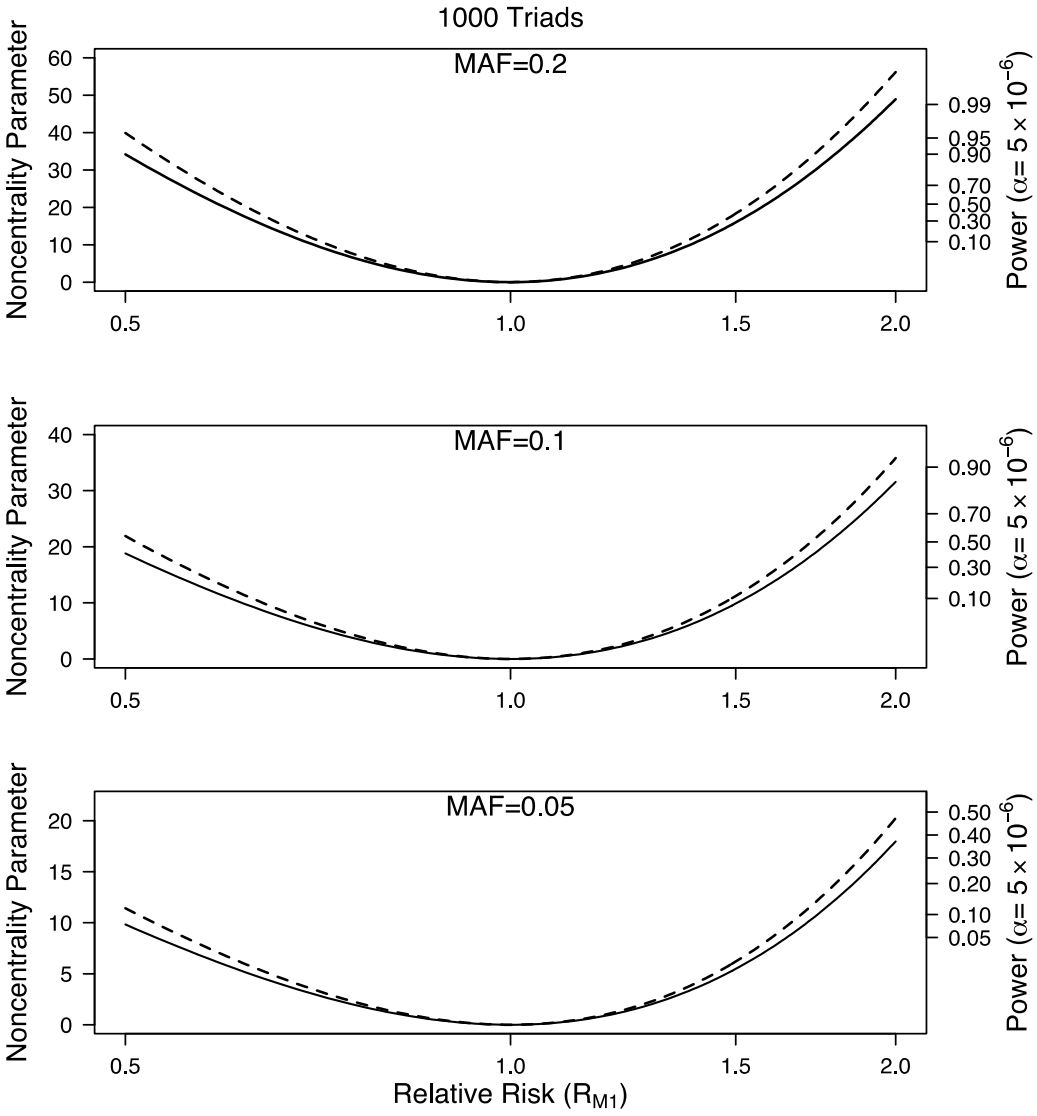
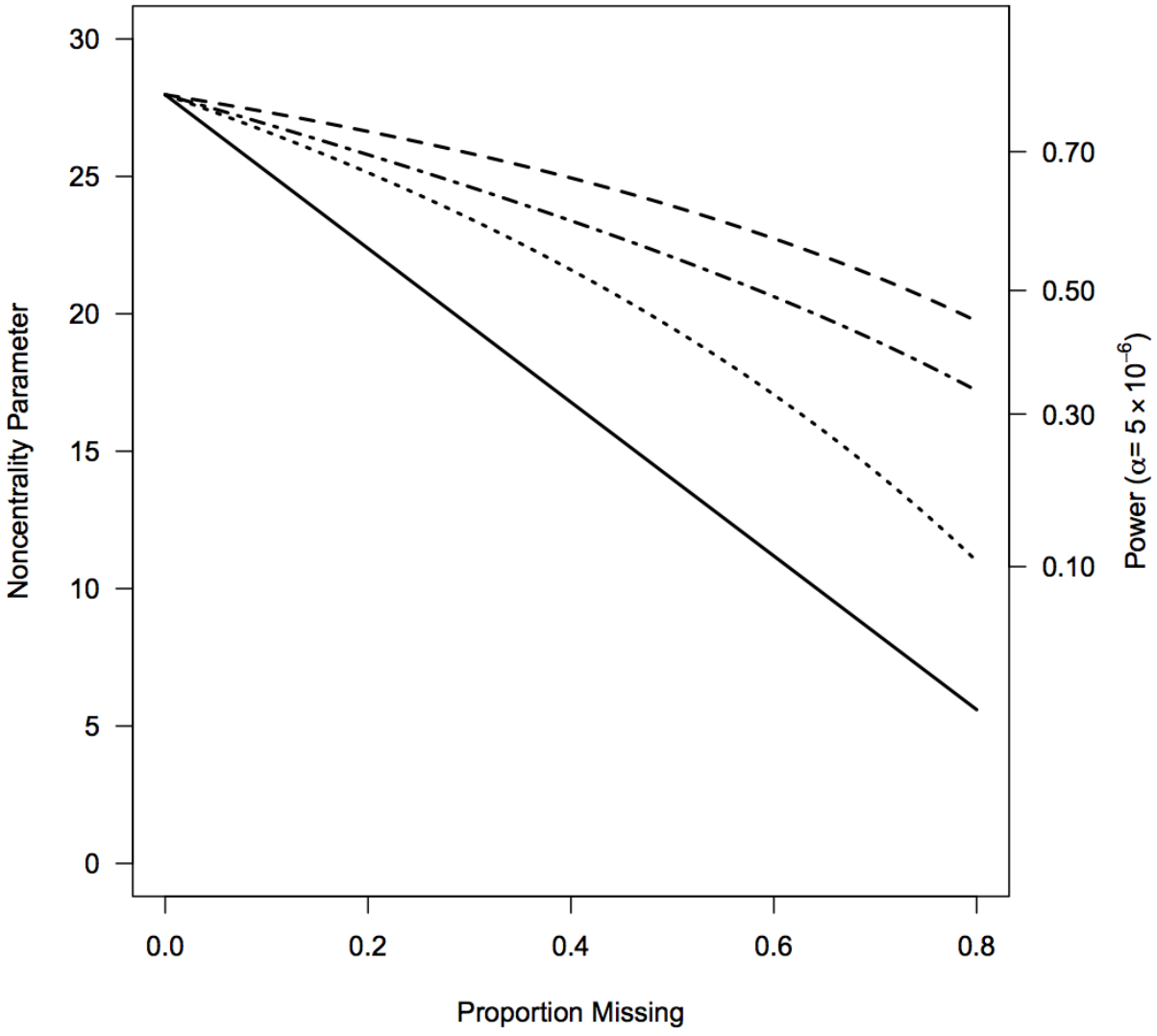


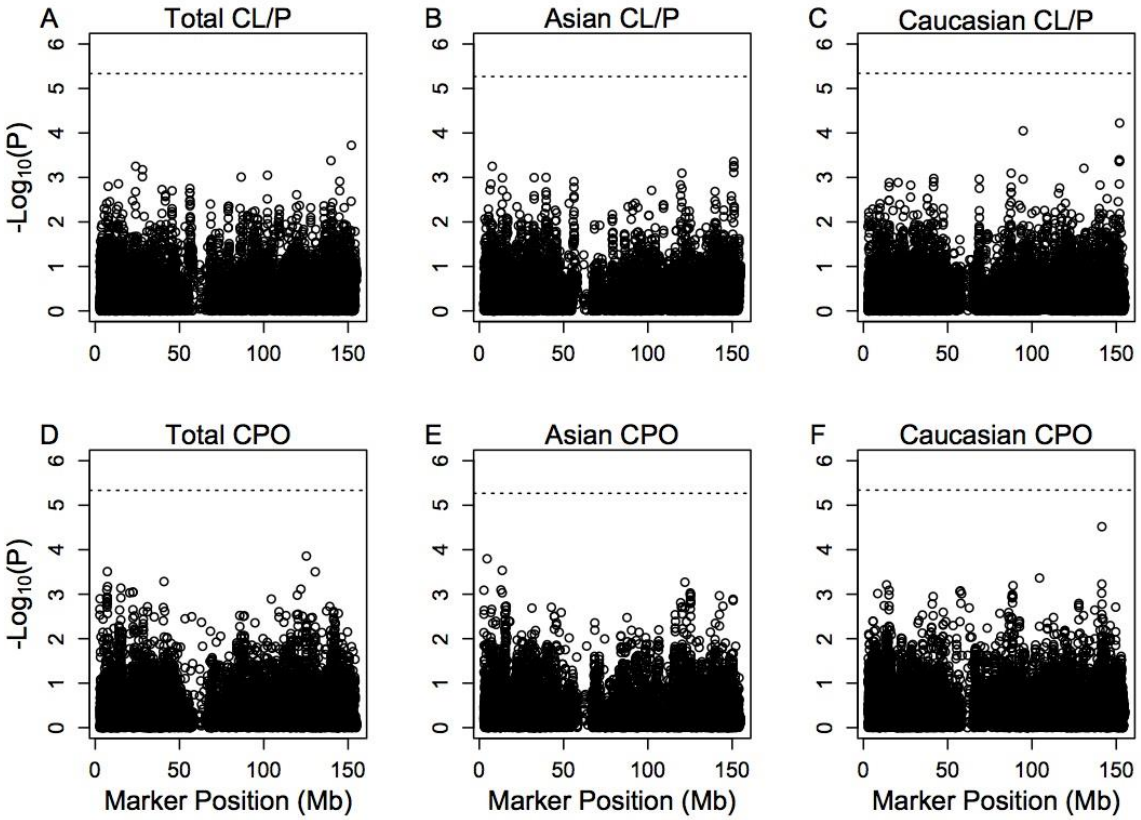
Figure 3.2: Noncentrality parameter estimates as a function of fraction of families missing parental genotype using the Expectation-Maximization algorithm. PIX-LRT was run on 1000 triads to detect maternal effects in the presence of fetal effects ($R_{G1}^2 = R_{G2} = 2, R_B = 1.5$). A log-additive model is used for maternal effects with $R_{M1} = 1.7$ and $R_{M2} = 1.7^2$. Non-carrier risk in males and females is the same. Minor allele frequency of 0.2 was used. Solid line represents results based on excluding incomplete triads. Dashed line represents results based on triads with the mother missing. Dashed/dotted line represents triads with either mother or father missing, with twice as many fathers missing as mothers. Dotted line represents results based on triads with only father missing.



3.3.2 Oral Cleft

Figure 3.3 shows results of the PIX-LRT with EM analysis of the oral cleft dataset for maternal genotype effects of SNPs on the X chromosome adjusting for possible fetal effects. Results are shown separately for CL/P and CPO in Caucasians and Asians separately and combined. No analysis produced results suggestive of an association between a marker and either CL/P or CPO. No SNPs had p-values below the Bonferroni cut-off.

Figure 3.3: Individual single nucleotide polymorphism significance of maternal genotype for the cleft example. The p-values (shown as $-\log_{10}(p)$) are calculated from PIX-LRT with the EM using dbGaP data from families with oral cleft. A log additive model is assumed for the risk in affected daughters and a composite score is used to combine the sex-specific statistics. Models were run on cleft lip with or without cleft palate families amongst (A) Asians and Caucasians, (C) Asians only, (E) Caucasians only, as well as cleft palate only families amongst (B) Asian and Caucasians, (D) Asians only, (F) Caucasians only. Dashed line is the Bonferroni correction for an alpha of 0.05.



3.4 Discussion

We have previously (Chapter 2) introduced a method to analyze SNPs on the X chromosome for a possible association with risk of disease when inherited by the fetus: the PIX-LRT. PIX-LRT incorporates both transmission and parental genotype information to create a powerful test to identify SNPs on the X chromosome associated with disease risk. We have here extended PIX-LRT to distinguish maternal from fetal effects and provide estimates of relative risks for maternally-mediated effects. An assumption of parental allelic exchangeability allows PIX-LRT to resist bias due to genetic population stratification. Furthermore, when, as in the oral cleft dataset, families have missing parents, PIX-LRT takes advantage of the EM algorithm so that information from these families is captured.

To our knowledge, PIX-LRT and HAPLIN are the only tools that can analyze maternal effects on the X chromosome in case-parent triads. When we compare the power to detect a maternal effect, HAPLIN performs better than PIX-LRT, however, the cost is loss of robustness against population stratification. HAPLIN relies on Hardy-Weinberg equilibrium, so, unlike PIX-LRT, is not robust. As shown in the results section, when sub-populations are present, PIX-LRT retains the nominal type 1 error rate, while HAPLIN is biased.

Some limitations deserve mention. The PIX-LRT estimates can be biased if an allele violates the parental exchangeability assumption. While this assumption can be tested (see Chapter 2), if a particular allele confers risk specifically when carried by the mother, this allele will also appear to violate the exchangeability assumption, so distinguishing a true maternal effect from such a violation can be tricky. The current PIX-LRT model to identify maternal effects will also be biased if there is a parent-of-origin effect, i.e. if the penetrance of a fetally-inherited disease variant depends on whether the variant came from the mother or

father. It should be noted that under conditional where the parental exchangeability assumption is violated, the HWE assumption would also be violated. Therefore, HAPLIN will also be biased in these instances.

We tested for maternal effects in genotyped families affected by the birth defect oral cleft. We found no evidence of a maternal genotype associated with CL/P nor CPO. These null results were consistent amongst Asians and Caucasians.

CHAPTER 4: FAMILY BASED X-CHROMOSOME HAPLOTYPE ANALYSIS USING PARENT INFORMATION

Thus far we have discussed and introduced methods for single marker analysis for SNPs on the X chromosome. In this chapter we generalize PIX-LRT to account for haplotypes. We use the term "haplotypes" loosely, to mean a set of polymorphic loci with linkage tight enough that the probability of recombination in one mitosis is effectively 0.0. In contrast with the autosome, for complete triads there is no phase ambiguity and haplotypes are identifiable on the X chromosome. We take advantage of this property to develop a method for identifying X haplotypes that are associated with risk. As before, we demonstrate our method by applying it to publically available data from an international consortium of genotyped families affected by the birth defect oral cleft.

4.1 Introduction

Haplotype analysis can be more powerful than SNP analysis in the presence of multiple susceptibility alleles in the autosome because simultaneous marker information is captured (Akey, Jin et al. 2001, Morris and Kaplan 2002). Sets of SNPs that are in LD within a gene-coding region can also have a joint effect on the structure of the protein product so such analyses can potentially be highly informative with respect to mechanisms of effect. When studying autosomal haplotypes, only the unphased genotypes (the sum of the two haplotypes) can typically be measured. For certain genotypes there will be phase ambiguity as it can be impossible to reconstruct the haplotype (for example if a person is heterozygous at more than one SNP). Therefore, for both family-based studies and non-family based

studies of the autosome, methods have been developed to account for phase ambiguity (Clayton 1999, Lin and Zeng 2005, Chung, Hauser et al. 2006, Lin and Zeng 2006, Allen and Satten 2007). For family-based methods, the TRIad Multi-Marker method (TRIMM), introduced by Shi and colleagues (Shi, Umbach et al. 2007) allows for haplotype analysis without phase estimation and TRIMMest (Shi, Umbach et al. 2009) extends the method to enable estimation of the relative risk for a candidate haplotype.

By contrast, the X chromosome is unique and wonderful in that, as we will show, if complete case-parents genotype data is present there is no phase ambiguity. There are four methods available to analyze haplotypes on the X chromosome in nuclear families: the X-LRT (Zhang, Martin et al. 2008), the X-APL (Chung, Morris et al. 2007), UNPHASED (Dudbridge 2008) and HAPLIN (Gjessing and Lie 2006, Jugessur, Skare et al. 2012). UNPHASED and HAPLIN were originally developed to analyze variants on the autosomes. HAPLIN, X-LRT and UNPHASED are all likelihood-based methods that provide estimates of the haplotypes relative risks relative to a reference haplotype. X-APL cannot provide haplotype relative risk estimates. X-APL and UNPHASED are designed for nuclear families with one or more affected siblings. HAPLIN is designed for case-parent triads but can also be used for case-control data and case-parent control-parent triads. X-LRT analyzes case-parent triads, and can use sibling data to account for missing genotypes. However, the X-LRT method is limited to two-marker haplotypes, and we will not consider it further here. It should be noted that, by assuming HWE, all four of these methods can account for missing genotype data. Currently, the method we present only handles complete triads, but offers robustness because HWE is not required.

PIX-LRT (the parent-informed X chromosome likelihood ratio test)(Chapter 2) is a

method to measure SNP effects of fetally-inherited X chromosome variants. PIX-LRT uses information from both the transmission of a variant X allele from parents to affected offspring, and information from the parental genotypes. An assumption of "parental allelic exchangeability" enables the added parental information to be captured in a way that resists bias due to genetic population stratification. Parental exchangeability is here generalized to apply to sets of SNPs in high LD as follows: we assume that the three haplotypes carried by each set of parents are randomly distributed, two to the mother and one to the father. Here we generalize PIX-LRT to allow association studies of haplotype effects for the X. The extended approach relies on a permutation-based p-value based on the most significant individual haplotype effect.

In the following sections, we describe the PIX-LRT extension for testing haplotype effects on the X chromosome in case-parent triads. We compare the performance of our method to HAPLIN, UNPHASED and X-APL using simulations to assess Type I error rates and power. As an illustrative example, we apply PIX-LRT to data from a family-based oral cleft dataset to analyze haplotypes on the X chromosome. We conclude with a discussion of the advantages and limitations of using PIX-LRT to study haplotypes.

4.2 Subjects and Methods

4.2.1 Case-Parent Design and Assumptions

We consider a sample of genotyped case-parent triads, where all sampled offspring have been diagnosed with the condition of interest. For fathers and sons, the haplotype is directly measured, as males only have one X chromosome. For mothers and daughters, the measurable genotype is the summed combination of the haplotypes from each of their two X chromosomes. We can identify the individual haplotypes in females if we assume there is no recombination within the variants considered. To see why, consider that each female

offspring has exactly inherited her father's complete X. If we subtract the father's genotype from the daughter's (summed) genotype, we can infer her maternally-inherited haplotype at any linked set of loci. Hence we can also identify the two haplotypes carried by the mother, by subtracting that inferred haplotype from the summed maternal genotype. For triads with male offspring (who only have one haplotype) we also know exactly the two haplotypes carried by the mother, again by subtraction.

The assumptions required are similar to those exploited in Chapters 2 and 3. We assume there is Mendelian transmission of the haplotype in the source population. We also assume parental allelic exchangeability in the source population, as described in Chapter 2. We also assume that the variants are not determinants of fetal survival or parental ability to reproduce. As before, we exclude the pseudo-autosomal regions and the X-transposed region (PARs, XTR), as these regions on the X can meiotically cross over with a homologous region on the Y.

4.2.2 PIX-LRT Extension to Haplotype Analysis

If we have identified a haplotype of interest prior to the analysis, and are interested in testing that specific haplotype against all other haplotypes, the analysis is straightforward. We can think of the nominated haplotype as the variant haplotype. Once the individual haplotypes are identified (as described above in section 4.2.1), the mothers and daughters have either 0,1 or 2 copies of the variant haplotype, while fathers and sons have either 0 or 1. This situation then becomes analogous to the di-allelic marker case described in Chapter 2 and we can proceed with the PIX-LRT analysis, except that the haplotype of interest now serves as our variant "allele".

The most common scenario will be an exploratory analysis. Suppose we are considering a particular set of SNPs and we have no prior basis for considering one set of them to be the candidate haplotype of interest. Then we would need to test for the global null that none of the haplotypes for that set of markers is associated with the disease. For example, for a set of 4 markers we would need to consider all 16 possible haplotypes in this analysis. Our method considers each in turn, takes the best, and then uses a permutation-based procedure to develop a valid test that accounts for the fact that we are doing 16 overlapping tests. The method follows three steps. In Step 1, we use subtraction to convert the genotypes to the corresponding haplotypes and pairs of haplotypes, as described in section 4.2.1 above.

In Step 2, we identify a test-statistic based on the observed data. For each haplotype, we run PIX-LRT on that haplotype compared to all other haplotypes grouped together. Once we have dichotomized the haplotypes, as mentioned above, it is straightforward to run the PIX-LRT of Chapter 2. For the sake of simplicity, we perform the PIX-LRT analysis with a log-additive effect for girls and a one-degree-of-freedom combined test for boy and girls. For haplotypes with N di-allelic SNPs there would be 2^N tests. However, we may want to apply a filter based on frequency, requiring, for example, that we only test the subset of haplotypes that have prevalence at least 1%. Suppose a total of G distinct haplotypes survive that filter. For those G haplotypes, there are G test statistics and corresponding p-values. We take the minimum p-value from those G analyses as our single test statistic. Denote this test statistic as T^*, which selection also serves to designate the haplotype with the statistically strongest association with the disease. From the PIX-LRT analysis, we can also estimate the relative risk associated with carrying this haplotype compared to the other haplotypes.

In Step 3, we permute the data to compute a permutation-based p-value, using the following procedure. For each permutation of the data, we refer back to our parental allelic exchangeability assumption, which assumes that under the null hypothesis alleles (or haplotypes) are randomly distributed across the parents in a mating pair. Therefore, for each family, we randomly assign the three parental haplotypes (denote them M_1, M_2 and F, respectively), to form a permuted mother and father (M_1', M_2' and F', respectively). Next, for triads with sons, we create a permuted son by randomly assigning to him either M_1' or M_2'. For triads with daughters, we create a permuted daughter by assigning to her F' and a random choice of either M_1' or M_2'. Once we have permuted each of the original triads in our dataset, we run Step 2 on our permuted dataset and calculate a permutation-based test statistic T_i which is the minimum p-value from among the haplotypes being tested (where i is the i^{th} permutation run). The T_i permutation-based test statistic values are independently and identically distributed from the null distribution, against which we can compare the observed-data test statistic, T^*. That is, the permutation-based p-value is $\Pr[T_i < T^*]$, estimated by the proportion of simulations that yield a smaller p-value than the observed-data result. For our simulations, which assess power at a single location, we run 1000 permutations.

In a GWAS of the X chromosome where we use a sliding window to find association hot spots, we are only interested in identifying and estimating the very lowest p-values. In that setting the following simplification will help to improve computational efficiency. For each given set of linked SNPs being tested, we run permutations until a total of 4 permutation-based test statistics are less than T^*. We can estimate our p-value to be 4 divided by the number of permutation runs required to reach 4 that fall below T^*. For example, if the

fourth event where $T_i < T^*$ occurs at run number $i = 2000$, then the estimated p-value is 4/2000 = 0.002. Since the number of "failures" to the fourth "success" follows a negative binomial distribution with parameter equal to the p-value, this ratio gives a maximum likelihood estimate of the p-value $\Pr[T_i < T^*]$. Under a global null the expected number of permutation runs required for each haplotype considered is 8, which will cut down on computation time dramatically. If a permutation p-value is estimated by that procedure to be less than 10^{-5}, we then run 5,000,000 permutations on this haplotype to get a more accurate p-value. For a Bonferroni-corrected p-value of 4×10^{-6}, based on 5,000,000 permutations, if the number of permuted test statistics less than the observed test statistics is 20 or less, the permutation p-value will be significant.

4.2.3 Type I Error and Power Calculations

We use simulations to compare X-APL, UNPHASED, HAPLIN and PIX-LRT under null scenarios with and without HWE. We are interested in testing the global null that no haplotype is associated with the disease, compared to an alternative where at least one haplotype is associated with the disease. HAPLIN and UNPHASED have a suite of options to choose from for selected analyses to be run. For both of these methods, we use their global test for comparison. HAPLIN, X-APL and UNPHASED have global tests where the degrees of freedom for the chi-squared test statistic equal the number of haplotypes minus one. The default for HAPLIN is to remove families that carry haplotypes that have a frequency less than 1%. For X-APL and UNPHASED, which also have this parameter option, for comparability we set the programs to remove families with these rare haplotypes as well. This option has the potential to increase the power to detect an effect, as the number of tests, and hence the degrees of freedom, decreases. HAPLIN also allows the user to specify the

relationship between the effect of a male carrying one copy of the variant and a female carrying one or two copies. We set the "comb.sex" option to "double", which sets the effect of males having one copy of the variant equal to that for females having two, adjustsing for X inactivation.

For the null simulation under HWE, four markers are simulated to generate 16 haplotypes. The haplotype frequencies we use for the null simulations are shown in Table 4.1, scenario 1. Each dataset contains 1000 families and we simulate 1000 datasets. Males and females have the same risk of disease. To simulate the scenario in which HWE is violated, we include two genetically different subpopulations. The first subpopulation has the haplotype frequencies as in scenario 1 of Table 4.1, while the second population has frequencies as in scenario 2, as shown. The risk of disease in the second population is three times that in the first population. Again, 1000 simulations are run on a population of 1000 families.

Table 4.1: Haplotype frequencies for the different scenarios used in the simulations. For the haplotype notation, "1" refers to the SNP in the haplotype carrying the minor allele. For example, "1100", is the haplotype where the first two SNPs carry the minor allele, and the second two SNPs do not. When a risk haplotype is involved, haplotype "1100", shown in bold, is the risk haplotype.

Haplotype	Frequencies		
	Scenario 1	Scenario 2	Scenario 3
0000	0.2401	0.4096	0.343
0001	0.1029	0.1024	0
0010	0.1029	0.1024	0.147
0011	0.0441	0.0256	0
0100	0.1029	0.1024	0.147
0101	0.0441	0.0256	0
0110	0.0441	0.0256	0.063
0111	0.0189	0.0064	0
1000	0.1029	0.1024	0.147
1001	0.0441	0.0256	0
1010	0.0441	0.0256	0.063
1011	0.0189	0.0064	0
1100	**0.0441**	0.0256	**0.063**
1101	0.0189	0.0064	0
1110	0.0189	0.0064	0.027
1111	0.0081	0.0016	0

We simulated 1000 data sets to estimate the power at a 0.05 alpha level under a range of alternative scenarios. We assume HWE and simulated 1000 families in each data set. We consider six risk scenarios (referred to as A-F). In all risk scenarios, we designate haplotype "1100" to be the risk haplotype (with several vectors of haplotype frequencies, as shown in Table 4.1). In scenarios A and B, the relative risk associated with disease for a boy carrying the risk haplotype compared to the other haplotypes (R_B) is 1.5. For girls, we assume a log-additive models, so that the relative risk associated with a disease for a girl carrying one copy of a risk haplotype compared to the nonrisk haplotypes (R_{G_1}) is square root of 1.5 and for a girl carrying two copies (R_{G_2}) is 1.5. In scenarios C and D, we consider a model where PIX-LRT, HAPLIN and UNPHASED are misspecified. In simulating the data, we set the true $R_B = R_{G_2} = 1.5$ and $R_{G_1} = 1$. The three methods assume a log-additive effect in girls, whereas the true model is recessive. In scenarios E and F, the "risk" haplotype has a protective effect. We set $R_{G_2} = R_{G_1}^2 = R_B = 1/1.5$.

In the scenarios A, C and E, all 16 haplotypes can occur in the population. We initially set our haplotype frequencies to be the same as the null situation under HWE above (scenario 1 in Table 4.1). In the scenarios B, D and F, only 8 haplotypes can occur in the population. To achieve this we give the 4th SNP a minor allele frequency of zero (scenario 3 in Table 4.1). We evaluate the power under a range of risk haplotype frequencies, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 and 0.9. After modifying the risk haplotype frequency, we rescale the remaining haplotypes so the sum of all frequencies equals 1. Lastly, for the simulations PIX-LRT identifies as significant at 0.05, we calculate how often the designated risk haplotype ("1100") is nominated as the risk haplotype, i.e. is the one with the minimum

individual p-value (T^*).

4.2.4 Oral Cleft Data

We apply the haplotype-based PIX-LRT to the X chromosome data from the International Consortium to Identify Genes and Interactions Controlling Oral Clefts, as was described in Chapter 2.2.5. A complete haplotype analysis has not previously been performed on this dataset; however, Patel et al. (Patel, Beaty et al. 2013) previously analyzed a selection of the X haplotypes using those data. They used UNPHASED (Dudbridge 2008) to analyze combinations of the 25 SNPs in the Duchenne muscular dystrophy (DMD) gene, because individual SNPs in that gene showed strong associations for the phenotype cleft lip with or without cleft palate (CL/P).

For our analysis, we here consider only complete triads of Asian (including Pacific Islanders) and Caucasian ethnicities. We analyze Asian and Caucasians family triads separately and combined. Additionally, we test haplotypes separately for cleft palate only (CPO) and cleft lip with or without palate (CL/P), based on evidence that those phenotypes have distinct genetic etiologies (Murray 2002). The gender and cleft subtype breakdown is shown in Table 4.2. Notice that the two phenotypes differ in the sex ratio of affected offspring.

Table 4.2: Complete case-parent families by cleft type, gender and ancestry.

	European		Asian		Total	
	Male	Female	Male	Female	Male	Female
Cleft Type						
CL/P	424	240	575	312	999	552
CPO	105	107	93	140	198	247
Total by gender	529	347	668	452	1197	799
Total	876		1120		1996	

CL/P is cleft lip with or without palate, CPO is cleft palate only

We use a sliding window approach to analyze haplotypes, by using in turn sets of 4 neighboring (in the panel available) SNPs on the X chromosome. We first filter down the panel by considering only those SNPs with a minor allele frequency in parents greater than 0.05, and also restricting to those with a unique mapping from the Illumina Human610-Quad v1.0 Build 36 to Build 37. We also exclude SNPs for which we had genotyping concerns (rs17269319, rs3747355, rs5906541, rs12558269) (see Chapter 2 for details). As a result, there are 10,571 SNPs that pass this screening among Asians, 12,417 SNPs amongst Caucasians, and 12,365 SNPs amongst the combined populations. For 4-SNP moving window haplotype analyses, the number of haplotype tests is then the total number of SNPs minus 3. The appropriate alpha for a Bonferroni-corrected family-wise error rate of 0.05 for Asians, Caucasians and the combined sample is consequently 4.73×10^{-6}, 4.03×10^{-6} and 4.04×10^{-6}, respectively.

As discussed above, we are only interested in identifying haplotypes with strong evidence for association, and consequently we do not need to perform a large number of permutations for each set of 4 SNPs. For each haplotype we run permutations until 4 of the permutation test statistics are less than the observed test statistic. If a permutation p-value is less than 10^{-5}, we then run additional permutations on this haplotype until the total is 5,000,000 to get a more accurate p-value. If the number of permuted test statistics less than the observed test statistics is 23 or fewer for the Asian population or 20 or fewer for the Caucasian of combined population, the permutation p-value will be significant at the Bonferroni-corrected p-value. To display results, we construct plots of $-\log_{10}$(p-value) against the marker position of the first SNP in the haplotype along the X chromosome (as determined by Build 37).

4.3 Results

4.3.1 Simulation Output

Under a null scenario of no association between the haplotype and disease, we used computer simulations to compare PIX-LRT, HAPLIN, X-APL and UNPHASED. Based on 2000 data sets with 1000 families each we ran simulations at a nominal alpha of 0.05 under a scenario where HWE was imposed, and then repeated that for a scenario where there was instead population stratification. The results are shown in Table 4.3. When HWE is imposed, the type I error rates of all four methods are close to the nominal level 0.05. When there is population stratification and HWE is violated, PIX-LRT, HAPLIN and X-APL again have type I errors close to the nominal levels. UNPHASED however appears to have inflated type I error, with an estimate of 0.062 (SE= 0.0064).

Table 4.3: Simulated Type I error rates for X-haplotype methods. 2000 datasets simulated with1000 triads for a null variant when Hardy-Weinberg equilibrium is assumed (HWE) and when it is violated (no HWE).

	PIX-LRT	HAPLIN	X-APL	UNPHASED
HWE	0.046	0.052	0.055	0.058
NO HWE	0.055	0.045	0.053	0.065

We next compared the power of the four methods under different risk scenarios, with results shown in Figure 4.1. In three risk scenarios (A,C and E), we consider four markers that produce 16 haplotypes in the simulated population. HWE was imposed for all. The frequency of the risk haplotype in the population ranges from 0.05 to 0.9. In all three of these scenarios, although the risk models differ, the relationship between the 4 methods is similar. PIX-LRT generally (except when frequency of the risk haplotype is very small or large) outperforms HAPLIN, X-APL and UNPHASED. Also, X-APL and UNPHASED perform similarly.

In scenarios B, D, and F, four markers produce only 8 haplotypes in the simulated population. PIX-LRT still outperforms X-APL and UNPHASED (Figure 4.1). However, we see that the difference in power between HAPLIN and PIX-LRT is very small. In this scenario, X-APL, UNPHASED and HAPLIN analyses are all based on a 7 degree-of-freedom chi-squared test, compared to the 15 degree-of-freedom test above. As a result, UNPHASED and HAPLIN perform much better in this scenario. We do not see much improvement in X-APL. PIX-LRT also has greater power when there are fewer haplotypes present.

Restricting attention to the subset of simulations where the PIX-LRT analysis was significant at an alpha of 0.05, we wanted to see how often the risk haplotype was correctly identified. PIX-LRT accurately identified the true risk haplotype with high accuracy, as shown in Figure 4.2, which shows the fraction of times the risk haplotype was identified as the most significant result. PIX-LRT generally has higher than 90% accuracy. Not surprisingly, the accuracy appears to decrease with the power.

4.3.2 Oral Cleft

The results of the PIX-LRT haplotype analyses along the X chromosome for oral cleft are shown in Figure 4.3. In the Caucasian subsample alone, no significant associations were detected between any haplotypes and either CL/P or CPO; no p-values fell below the Bonferroni-corrected cut-off of 4.03×10^{-6}. We also did not find significant results between the total population and CL/P and the Asian population and CPO.

However, we did identify three haplotypes with significant association with oral cleft. Details of these three haplotypes are shown in Table 4.4. The three haplotypes all had p-values beneath the 0.05 Bonferroni-adjusted significance level within each race by oral cleft

analysis. However, if we were to adjust for all six analyses (three race breakdowns and two forms of cleft), these results would not be significant.

Figure 4.1: Power estimates as a function of risk haplotype frequency. The level of significance is set at alpha = 0.05. Each analysis is based on 1000 datasets consisting of 1000 triads with affected sons and daughters and a designated risk haplotype "1100". (A,B) $R_{G1}^2 = R_{G2} = R_B = 1.5$. (C,D) $R_{G1}^2 = 1, R_{G2} = R_B = 1.5$ (E,F) $R_{G1}^2 = R_{G2} = R_B = 1/1.5$. For scenarios A,C,E, all 16 haplotypes from a 4 SNP window exist in the population. For scenarios B,D,F, 8 haplotypes from a 4 SNP window exist in the population. Solid line represents PIX-LRT. Dashed line represents HAPLIN. Dotted line represents X-APL. Dot/dash line represent UNPHASED.

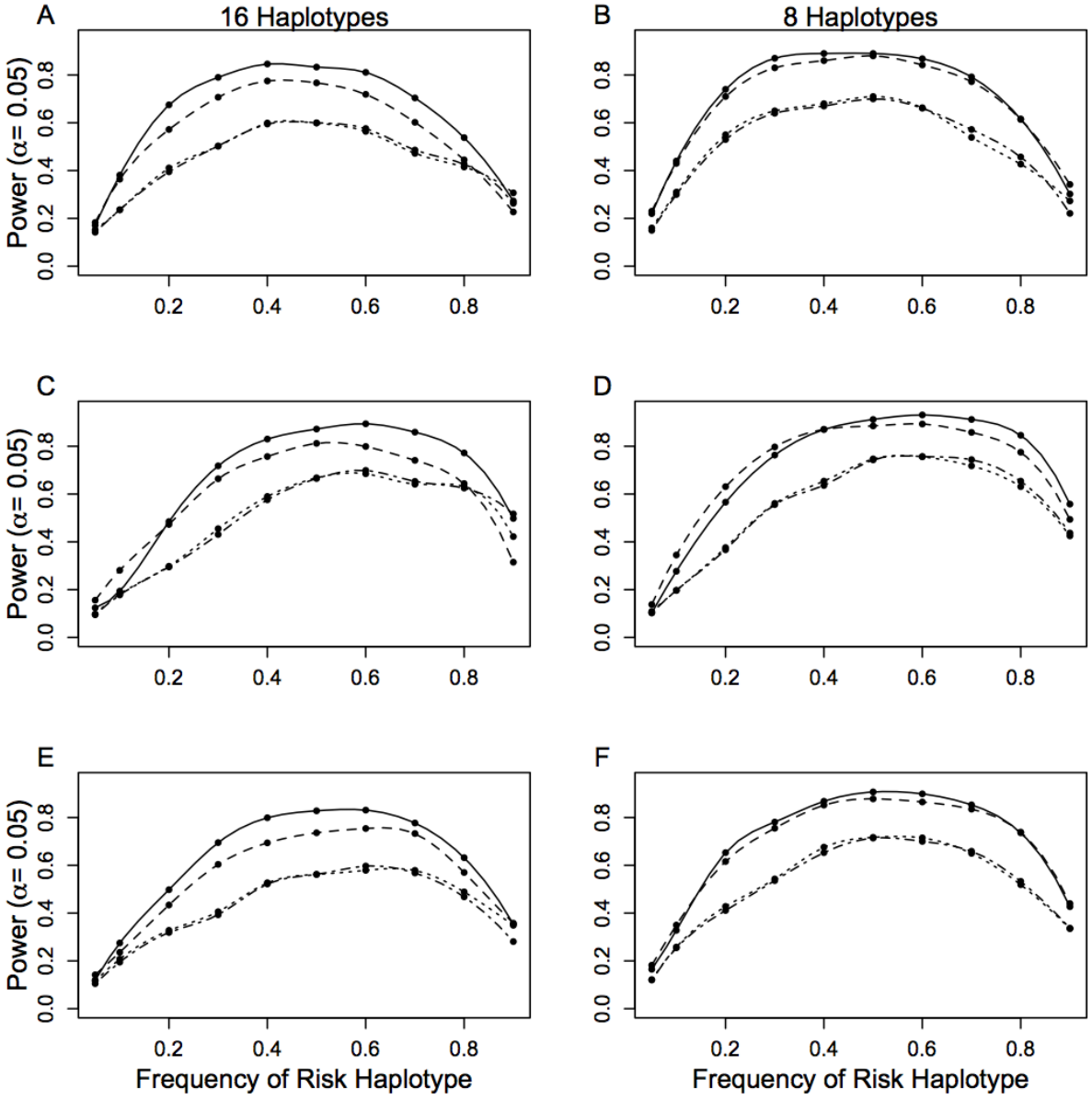


Figure 4.2: Fraction of times PIX-LRT nominates the risk haplotype amongst significant simulations. Each analysis is based on 1000 datasets consisting of 1000 triads with affected sons and daughters and a designated risk haplotype "1100". (A,B) $R^2_{G1} = R_{G2} = R_B = 1.5$. (C,D) $R^2_{G1} = 1, R_{G2} = R_B = 1.5$ (E,F) $R^2_{G1} = R_{G2} = R_B = 1/1.5$. For scenarios A,C,E, all 16 haplotypes from a 4 SNP window exist in the population. For scenarios B,D,F, 8 haplotypes from a 4 SNP window exist in the population.

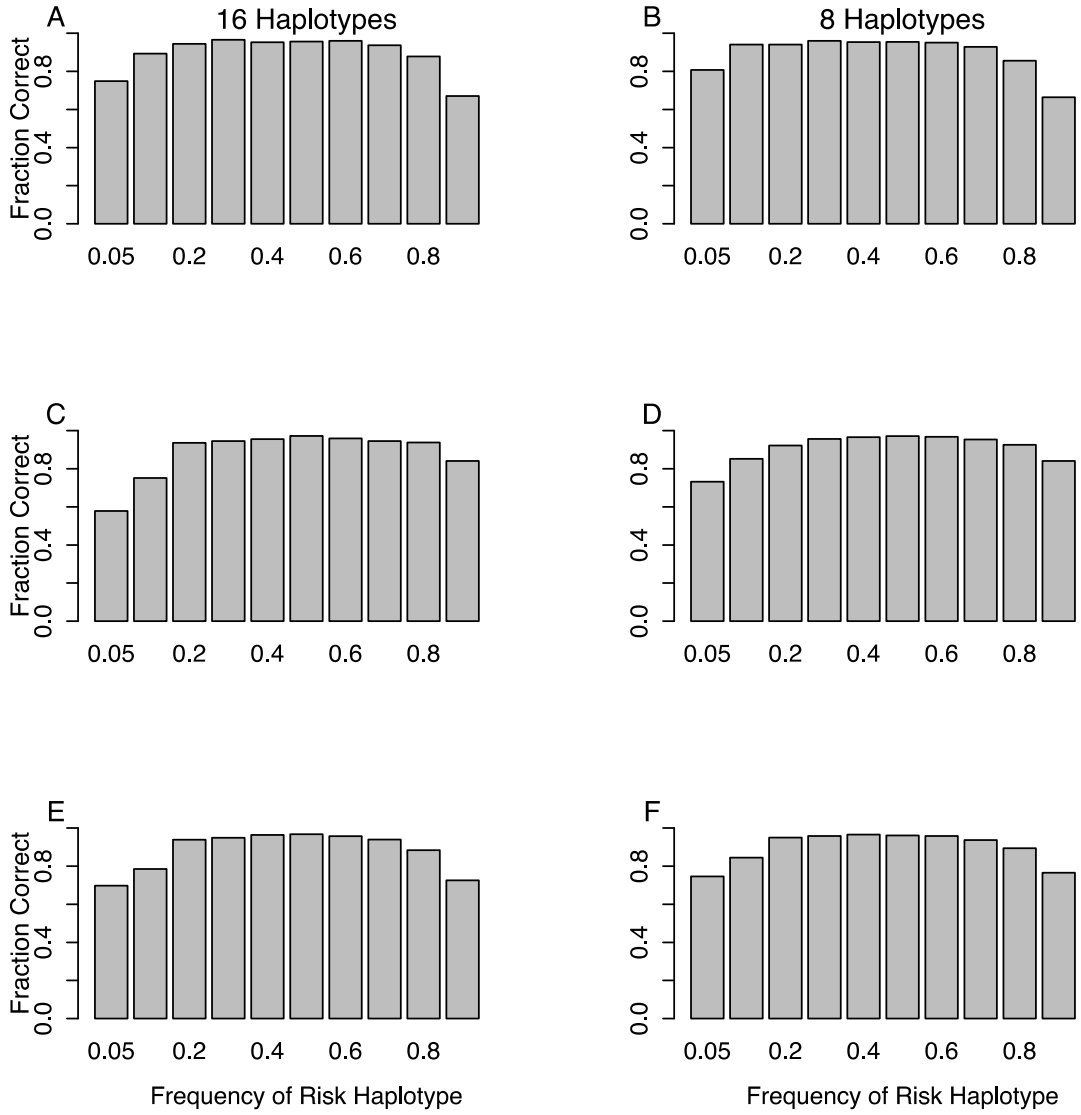


Figure 4.3: Individual haplotype significance of the cleft examples. The p-values (shown as $-\log_{10}(p)$) are calculated from PIX-LRT applied to haplotypes consisted of 4 SNPs using dbGaP data from families with oral cleft. Models were run on cleft lip with or without cleft palate families amongst (A) Asians and Caucasians, (C) Asians only, (E) Caucasians only, as well as cleft palate only families amongst (B) Asian and Caucasians, (D) Asians only, (F) Caucasians only. The dashed horizontal lines are the Bonferroni-corrected p-values at an alpha of 0.05, where the adjustment is specific to the panel of findings.

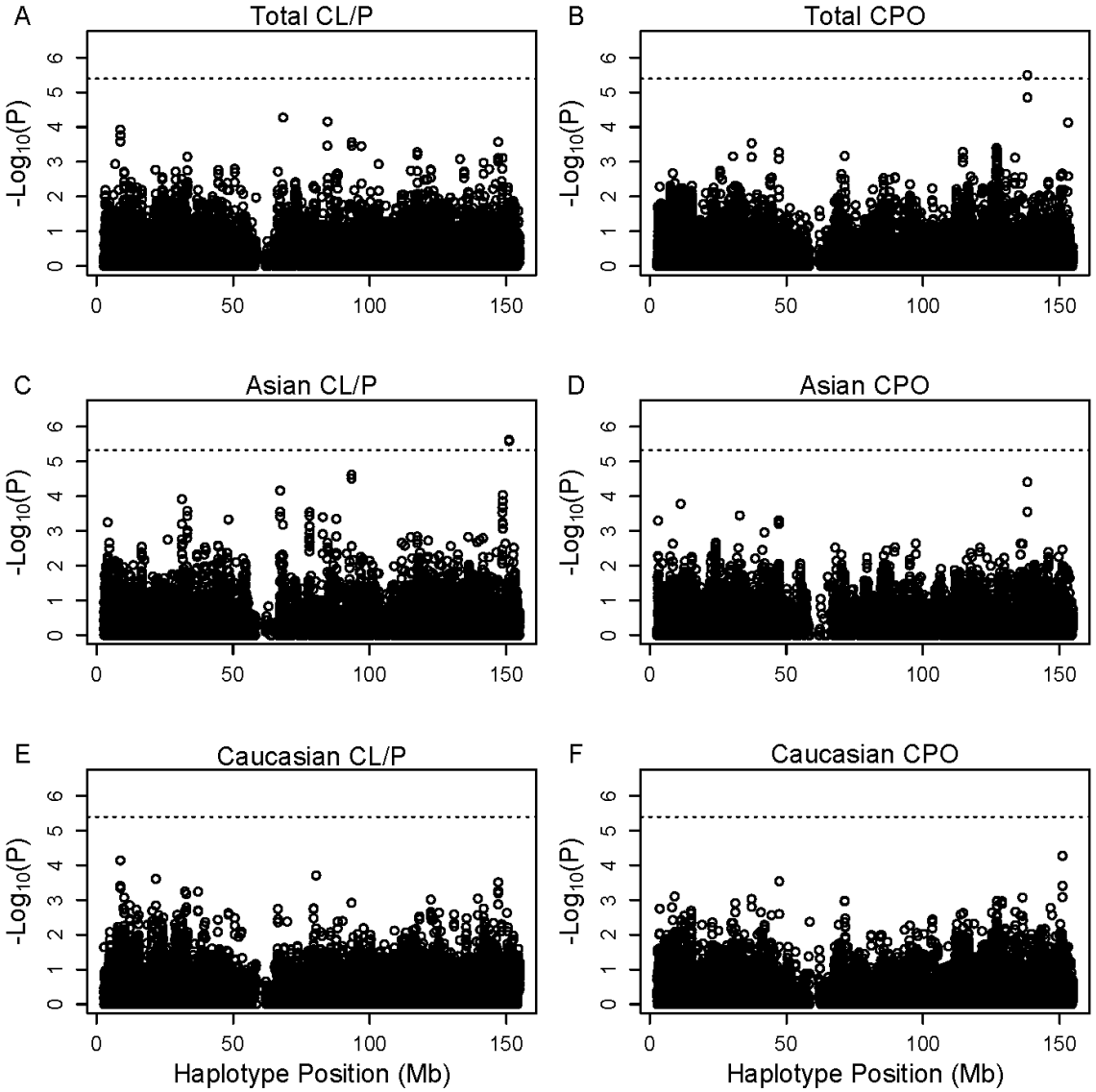


Table 4.4: Most significant haplotypes associated with oral cleft based on PIX-LRT. Haplotypes shown here had an initial permutation based p-value less than 10^{-5}. Subsequently, a p-value based on 5 million permutations was calculated (P-Values). The 4 SNPs in the haplotype and the frequency of the risk haplotype (Freq) in the parents and the location[A] are shown. The populations (Pop) the association was seen in, and whether the association was with cleft lip with or without palate (CL/P) or cleft palate only (CPO) are displayed.

SNP1	SNP2	SNP3	SNP4	P-Value	Location	Freq	Pop	Cleft
rs17002006	rs5976286	rs5931572	rs2886973	$3.2x10^{-6}$	138232190-138236688[B]	0.020	Total	CPO
rs12843815	rs6627483	rs5970136	rs5970137	$2.4x10^{-6}$	151022400-151040060[C]	0.026	Asian	CL/P
rs6627483	rs5970136	rs5970137	rs964180	$2.6x10^{-6}$	151038722-151041999[C]	0.027	Asian	CL/P

[A]Location is based on the position of the first and last SNP in the haplotype based Illumina Human610-Quad v1.0 Build 37.
[B]These 4 SNPs are located in the *FGF13* gene.
[C]These 4 SNPs are located between genes *CNGA2* and *MAGEA4*.

The first of these haplotypes consists of the SNPs rs17002006, rs5976286, rs5931572, and rs2886973. If we denote "1" to represent the minor allele of a SNP, then the haplotype "1101" is associated with CPO in the combined Asian and Caucasian populations (p-value = $3.2x10^{-6}$) (see Figure 4.3B and Table 4.4). This haplotype shows evidence of a strong protective effect in both males and females (relative risks are 0.01 and 0 for girls and boys respectively. The relative risk of 0 in boy is because no affected sons carried that haplotype, although fathers and mothers did). The haplotype spans 4498 base pairs on the Fibroblast Growth Factor 13 gene (*FGF13*). None of the 4 SNPs in this haplotype displays a marginal SNP association with CPO (all 4 p-values are greater than 0.05). The haplotype also shows a strong, though not significant, protective association in the Asian population alone (p-value $=3.9x10^{-5}$). Additionally, the overlapping haplotype consisting of SNPs rs5976286, rs5931572, rs2886973, and rs2213408 has a strong protective association in the combined population, with a p-value of $1.4x10^{-5}$.

Table 4.5: Cross table of SNPs s6627483 and rs5970137. Number of fathers, mothers, sons and daughters who carry each combination of SNPs s6627483 and rs5970137 in complete families. NA refers to missing genotype.0,1,2 refer to the number of minor allele copies carried.

	rs6627483	rs5970137			
		0	**1**	**2**	**NA**
Fathers	**0**	1639	1	-	0
	1	267	0	-	0
	NA	0	89	-	0
Sons	**0**	983	0	-	0
	1	156	0	-	0
	NA	1	57	-	7
Mothers	**0**	1339	143	0	1
	1	443	0	0	0
	2	36	24	0	0
	NA	0	0	10	0
Daughters	**0**	525	56	0	0
	1	184	0	0	0
	2	13	16	0	0
	NA	1	0	4	0

In the Asian population we identified two overlapping 4-SNP haplotypes with a strong positive association with cleft lip with or without palate (see Figure 4.3C and Table 4.4). These haplotypes consisted of the five SNPs rs12843815, rs6627483, rs5970136, rs5970137, and rs964180 (with rs6627483, rs5970136 and rs5970137 in both). These SNPs are in the intergenic regions of melanoma antigen family A, 4 (*MAGEA4*) and cyclic nucleotide gated channel alpha 2 (*CNGA2*). Upon inspecting these haplotypes we noted that in complete families (when no genotypes were missing for any of the 4 SNPs), there were no males who carried the minor allele of rs5970137. However, when we looked at complete families for the single SNPs, males did carry the minor allele. A cross-table of the genotypes of rs5970137 and rs6627483 revealed that in all but one instance where a male carried the minor allele of rs5970137, the rs6627483 genotype was missing. In 147 males who had a missing rs6627483 genotype, but non-missing rs5970137 genotype, all but one carried the minor allele at rs5970137 (Table 4.5). The two loci are 1338 bases apart and it is unclear why

the genotype at one would affect the genotype assayability at the other. This pattern of missingness raises quality control concerns for haplotypes containing rs6627483 and rs5970137 as well as doubts over the significance of these two significant findings.

4.4 Discussion

We have previously (Chapter 2) introduced PIX-LRT, a method to analyze SNPs on the X chromosome for a possible association with risk of disease when inherited by the fetus. In this chapter we have generalized PIX-LRT from a single-SNP analysis method to a haplotype method. The haplotype analysis constructs a permutation-based p-value based on the most significant individual haplotype effect.

As shown in the results section, we used simulations to compare PIX-LRT to other family-based haplotype analysis methods for the X chromosome: HAPLIN, X-APL and UNPHASED. Under a null scenario in which no haplotype is associated with the disease, when the population is in the Hardy Weinberg equilibrium, all four methods displayed appropriate Type I error. When population stratification was present, PIX-LRT, X-APL and HAPLIN displayed appropriate Type I error under the null. UNPHASED had slightly inflated Type I error.

When we compared the power to detect a haplotype effect on the X chromosome for the four methods, PIX-LRT performed strongly. Using simulations, we found that when 16 haplotypes (consisting of 4 SNPs) were present in the population, and one haplotype conferred risk, PIX-LRT outperformed the other methods. HAPLIN had the second highest power, followed by X-APL and lastly, UNPHASED. When only 8 haplotypes were present, PIX-LRT and HAPLIN performed similarly, while X-APL and UNPHASED had lower power. HAPLIN, UNPHASED and X-APL analyses are based on chi-squared statistics with

the degrees of freedom equal to the number of haplotypes minus 1. Therefore, it is not surprising to see that the power advantage of PIX-LRT is greater when the number of haplotypes is also greater.

There are some areas in which our ongoing research will improve upon the current haplotype-based PIX-LRT. If a haplotype affects risk through a maternal effect, PIX-LRT results will be biased. In Chapter 3 we showed how for a single-SNP analysis PIX-LRT could be extended to accommodate maternal effects. We plan to similarly extend the haplotype PIX-LRT to identify and measure maternal effects. Furthermore, although in this chapter we have shown examples involving one single risk haplotype, our ongoing research is focused on improving the ability of PIX-LRT to detect an effect when there are multiple risk haplotypes. Lastly, the haplotype-based PIX-LRT only uses complete families. In the single SNP analysis we used the EM algorithm to accommodate missing families, we could similarly incorporate the EM algorithm into the haplotype analysis.

We applied PIX-LRT to the X chromosome of a dataset from an international consortium of genotyped families affected by the birth defect oral cleft. We looked at all haplotypes containing 4 sequential SNPs in the available platform. We identified a novel haplotype on the *FGF13* gene that has a significant association with cleft palate only in the combined Asian and Caucasian population. The haplotype showed a strongly protective effect in the population.

Additionally, in the Asian population we found 2 overlapping haplotypes with a significant association with cleft lip with or without palate. However, upon further inspection of these haplotypes and their SNPs, we noticed an odd relationship between two of the SNPs, whenever males had the minor allele for one SNP, the genotype was missing the second

SNP. This finding raises quality control concerns for the dataset over this region of the chromosome, and highlights the importance of a careful QC critique of apparently significant findings.

CHAPTER 5: CONCLUSION

In this dissertation, we have introduced a new method for studying the effects of genetic variants on the X chromosome in case-parent triads. The X chromosome is often overlooked in genome wide association studies and we aim to introduce a powerful and straightforward method to analyze it for causative variants. Most X-chromosome methods for family data use transmission-based information to test for association between a variant and disease. Our new method, the "parent-informed X-chromosome likelihood ratio test" (PIX-LRT), takes advantage of information in the parental genotypes in addition to transmission-based information. This parental information has not previously been exploited by other methods. Our method is able to use this information robustly under an assumption of parental allelic exchangeability. This assumption is weaker than an assumption of Hardy-Weinberg equilibrium, as it allows for population stratification.

In Chapter 2 we introduced the "sex-stratified X-chromosome likelihood ratio test" (SSX-LRT), an analysis tool that can be used when the assumption of parental allelic exchangeability is violated and outlined the details of PIX-LRT for a di-allelic single nucleotide polymorphism (SNP). We demonstrated the increased power PIX-LRT has over the transmission-based methods, how relative risks are calculated and how the EM algorithm can be used to include triads with missing genotype data.

In Chapter 3 we demonstrated how PIX-LRT can be extended to analyze maternal genetic effects. By taking advantage of an assumption of allelic exchangeability, the method

is able to distinguish such maternal effects from effects due to fetal inherited variants, and can provide estimates of relative risks.

In Chapter 4 we generalized the PIX-LRT method to study haplotype effects. We used a permutation-based method to calculate the p-value for haplotypes. For scenarios with a single risk haplotype, PIX-LRT often outperforms other X chromosome haplotype methods. We commented on future research in this area, which includes maternal haplotype effects and adjusting the method to account for scenarios with multiple risk haplotypes.

Lastly, in each Chapter 2-4, we showed how the method could be applied to a dataset through use of a publically available oral cleft dataset. In Chapter 2, we found a SNP located between genes *EFNB1* and *PJA1* with a minor allele that showed a strong protective effect for the cleft lip with or without cleft palate phenotype. In Chapter 3, we found that within the dataset there was no evidence of a maternal genetic effect. In Chapter 4, we identified a haplotype on the *FGF13* gene with evidence of a protective effect for the cleft palate only phenotype. In addition, we demonstrated the importance of checking the underlying assumptions of the models and the quality of the data used before interpreting analytical results. We were able to identify SNPs and haplotypes that had been reported as risk-related but that raised QC concerns.

APPENDIX A: TEST FOR THE PARENTAL ALLELIC EXCHANGEABILITY ASSUMPTION

For the following, as in the main paper, we define *M* and *F* as the number of variant alleles carried by the mother and father. To test for parental allelic exchangeability define:

$$Q_1 = \Pr(M = 1|\mathrm{M} + \mathrm{F} = 1)$$
$$Q_2 = \Pr(M = 1|\mathrm{M} + \mathrm{F} = 2)$$

If we refer to Table 1 in the main paper, then $Q_1 = \frac{2\exp(\alpha_1)}{1+2\exp(\alpha_1)}$ and $Q_2 = \frac{2\exp(\alpha_2)}{1+2\exp(\alpha_2)}$. Under the null of parental allelic exchangeability, we expect:

$$\Pr(M = 1|\mathrm{M} + \mathrm{F} = 1) = 2\Pr(M = 0|\mathrm{M} + \mathrm{F} = 1)$$
$$\Pr(M = 1|\mathrm{M} + \mathrm{F} = 2) = 2\Pr(M = 2|\mathrm{M} + \mathrm{F} = 2)$$

Therefore, we are interested in the following test:

$$H_0: Q_1 = Q_2 = 2/3$$
$$H_A: Q_1 \neq 2/3 \text{ or } Q_2 \neq 2/3$$

Define:

- n_1 = the number of parents where *M+F=1*
- x_{01} = the number of parents where *M+F=1* and *M* = 0
- x_{10} = the number of parents where *M+F=1* and *M* = 1
- n_2 = the number of parents where *M+F=2*
- x_{20} = the number of parents where *M+F=2* and *M*=2
- x_{11} = the number of parents where *M+F=2* and *M*=1

Then we have the following binomial model:

$$p(x_{10}, x_{11}|n_1, n_2, Q_1, Q_2) = \binom{n_1}{x_{10}}(Q_1)^{x_{10}}(1-Q_1)^{n_1-x_{10}}\binom{n_2}{x_{11}}(Q_2)^{x_{11}}(1-Q_2)^{n_2-x_{11}}$$
$$= \binom{n_1}{x_{10}}(Q_1)^{x_{10}}(1-Q_1)^{x_{01}}\binom{n_2}{x_{11}}(Q_2)^{x_{11}}(1-Q_2)^{x_{20}}$$

We differentiate the log likelihood:

$$\ell \sim x_{10}\log(Q_1) + x_{01}\log(1-Q_1) + x_{11}\log(Q_2) + x_{20}\log(1-Q_2)$$

The maximum likelihood estimates are $Q_1$and$Q_2$are:

$$\hat{Q}_1 = \frac{x_{10}}{x_{10}+x_{01}} = \frac{x_{10}}{n_1}$$

$$\hat{Q}_2 = \frac{x_{11}}{x_{11}+x_{20}} = \frac{x_{11}}{n_2}$$

The likelihood ratio test statistic (LRTS) is then as follows:

$$
\begin{aligned}
LRTS &= -2\left(\ell\left(Q_1 = Q_2 = \frac{2}{3}\right) - \ell\left(Q_1 = \hat{Q}_1, Q_2 = \hat{Q}_2\right)\right) \\
&= -2\left(x_{10}\log\left(\frac{2}{3}\right) + x_{01}\log\left(\frac{1}{3}\right) + x_{11}\log\left(\frac{2}{3}\right) + x_{20}\log\left(\frac{1}{3}\right) - x_{10}\log\left(\hat{Q}_1\right)\right. \\
&\qquad \left. - x_{01}\log\left(1-\hat{Q}_1\right) - x_{11}\log\left(\hat{Q}_2\right) - x_{20}\log\left(1-\hat{Q}_2\right)\right) \\
&= -2\left(x_{10}\log\left(\frac{2}{3\hat{Q}_1}\right) + x_{01}\log\left(\frac{1}{3\left(1-\hat{Q}_1\right)}\right) + x_{11}\log\left(\frac{2}{3\hat{Q}_2}\right)\right. \\
&\qquad \left. + x_{20}\log\left(\frac{1}{3\left(1-\hat{Q}_2\right)}\right)\right) \\
&= -2\left(x_{10}\log\left(\frac{2n_1}{3x_{10}}\right) + x_{01}\log\left(\frac{n_1}{3x_{01}}\right) + x_{11}\log\left(\frac{2n_2}{3x_{11}}\right) + x_{20}\log\left(\frac{n_2}{3x_{20}}\right)\right)
\end{aligned}
$$

The $LRTS$ is distributed chi-squared with 2 degree of freedom under parental allelic exchangeability.

APPENDIX B: CLOSED FORM SOLUTIONS FOR THE SSX-LRT

As in the main paper, we define *M*, *F*, and *C* as the number of variant alleles carried by the mother, father and child and we define the relative risk of being affected (aff), conditional on mating type (*M*, *F*) to control for population stratification, as:

$$R_B = \Pr(\text{aff}|\text{boy}, C = 1) \,/\Pr(\text{aff}|\text{boy}, C = 0)$$

$$R_{G1} = \Pr(\text{aff}|\text{girl}, C = 1) \,/\Pr(\text{aff}|\text{girl}, C = 0)$$

$$R_{G2} = R_{G1} * \Pr(\text{aff}|\text{girl}, C = 2) \,/\Pr(\text{aff}|\text{girl}, C = 1)$$

B.1 Triads with affected sons

We are interested in the following hypothesis test:

$$H_0: R_B = 1$$
$$H_A: R_B \neq 1$$

Define:

- n = the number of triads with a heterozygous mother (M=1)
- x = the number of triads with a heterozygous mother and an affected son with the variant allele (M=1, C=1)

Referring to Table 2 in the main paper, we have the following binomial model:

$$p(x|n, R_B) = \binom{n}{x}\left(\frac{R_B}{1+R_B}\right)^x\left(\frac{1}{1+R_B}\right)^{n-x} \quad \text{(B. 1)}$$

To find the maximum likelihood estimate of R_B ($\hat{R}_B$), we differentiate the log likelihood function (ignoring the constant terms) that corresponds to equation B.1, set it to 0 and solve for $\hat{R}_B$:

$$\ell \sim x\log(R_B) - n\log(1+R_B)$$

$$\frac{d\ell}{dR_B} = \frac{x}{R_B} - \frac{n}{1+R_B}$$

$$0 = \frac{x}{\hat{R}_B} - \frac{n}{1+\hat{R}_B}$$

$$= \left(1+\hat{R}_B\right)x - \hat{R}_B n$$

$$\hat{R}_B = \frac{x}{n-x}$$

The likelihood ratio test statistic (*LRTS*) is then as follows:

$$\begin{aligned} LRTS &= -2\left(\ell(R_B = 1) - \ell\left(R_B = \hat{R}_B\right)\right) \\ &= -2\left(x\log\left(\frac{1}{2}\right) + (n-x)\log\left(\frac{1}{2}\right) - x\log\left(\frac{\hat{R}_B}{1+\hat{R}_B}\right) - (n-x)\log\left(\frac{1}{1+\hat{R}_B}\right)\right) \\ &= -2\left(n\log\left(\frac{1}{2}\right) - x\log\left(\hat{R}_B\right) + n\log\left(1+\hat{R}_B\right)\right) \\ &= -2\left(n\log\left(\frac{n}{2(n-x)}\right) - x\log\left(\frac{x}{n-x}\right)\right) \end{aligned}$$

The $LRTS$ is distributed chi-squared with 1 degree of freedom under the null.

B.2 Triads with affected daughters

We are interested in the following hypothesis:

$$\begin{aligned} &H_0: R_{G1} = R_{G2} = 1 \\ &H_A: R_{G1} \neq 1 \text{ or } R_{G2} \neq 1 \end{aligned}$$

Define:

- n_1 = the number of triads with M=1 and F=0
- x_1 = the number of triads with M=1, F=0 and C = 1
- n_2 = the number of triads with M=1 and F=1
- x_2 = the number of triads with M=1, F=1 and C = 2

Referring to Table 2 in the main paper, we have the following binomial model:

$$\begin{aligned} &p(x_1, x_2 | n_1, n_2, R_{G1}, R_{G2}) \\ &= \binom{n_1}{x_1}\left(\frac{R_{G1}}{1+R_{G1}}\right)^{x_1}\left(\frac{1}{1+R_{G1}}\right)^{n_1-x_1}\binom{n_2}{x_2}\left(\frac{R_{G2}}{R_{G1}+R_{G2}}\right)^{x_2}\left(\frac{R_{G1}}{R_{G1}+R_{G2}}\right)^{n_2-x_2} \end{aligned} \quad \text{(B. 2)}$$

To find the maximum likelihood estimate of $R_{G1}$and R_{G2} ($\hat{R}_{G1}$ and $\hat{R}_{G2}$), we differentiate the log likelihood function that corresponds to equation B.2, set it to 0 and solve for $\hat{R}_{G1}$ and $\hat{R}_{G2}$:

$$\ell \sim (x_1 + n_2 - x_2)\log(R_{G1}) - n_1\log(1+R_{G1}) + (x_2)\log(R_{G2}) - n_2\log(R_{G1}+R_{G2})$$

To solve for $\hat{R}_{G2}$:

$$\frac{d\ell}{dR_{G2}} = \frac{x_2}{R_{G2}} - \frac{n_2}{R_{G1}+R_{G2}}$$

$$0 = \frac{x_2}{R_{G2}} - \frac{n_2}{R_{G1} + R_{G2}}$$

$$= (R_{G1} + R_{G2})(x_2) - R_{G2}(n_2)$$

$$\hat{R}_{G2} = \frac{R_{G1}(x_2)}{n_2 - x_2}$$

To solve for $\hat{R}_{G1}$:

$$\frac{d\ell}{dR_{G1}} = \frac{x_1 + n_2 - x_2}{R_{G1}} - \frac{n_1}{1 + R_{G1}} - \frac{n_2}{R_1 + \hat{R}_{G2}}$$

$$= \frac{x_1 + n_2 - x_2}{R_{G1}} - \frac{n_1}{1 + R_{G1}} - \frac{n_2}{R_{G1}\left(1 + \frac{x_2}{n_2 - x_2}\right)}$$

$$= \frac{x_1 + n_2 - x_2}{R_{G1}} - \frac{n_1}{1 + R_{G1}} - \frac{n_2 - x_2}{R_{G1}}$$

$$= \frac{x_1}{R_{G1}} - \frac{n_1}{1 + R_{G1}}$$

$$0 = \left(1 + \hat{R}_{G1}\right)x_1 - \hat{R}_{G1}n_1$$

$$\hat{R}_{G1} = \frac{x_1}{n_1 - x_1}$$

So, given $\hat{R}_{G1}$:

$$\hat{R}_{G2} = \frac{x_1 x_2}{(n_1 - x_1)(n_2 - x_2)}$$

The likelihood ratio test statistic (to be compared to a 2 DF chi-squared) is then as follows:

$$LRTS = -2\left(\ell(R_{G1} = R_{G2} = 1) - \ell\left(R_{G1} = \hat{R}_{G1}, R_{G2} = \hat{R}_{G2}\right)\right)$$

$$= -2\left(x_1 \log\left(\frac{1}{2}\right) + (n_1 - x_1)\log\left(\frac{1}{2}\right) + x_2 \log\left(\frac{1}{2}\right) + (n_2 - x_2)\log\left(\frac{1}{2}\right)\right.$$

$$\left. - x_1 \log\left(\frac{\hat{R}_{G1}}{1 + \hat{R}_{G1}}\right) - (n_1 - x_1)\log\left(\frac{1}{1 + \hat{R}_{G1}}\right) - x_2 \log\left(\frac{\hat{R}_{G2}}{\hat{R}_{G1} + \hat{R}_{G2}}\right)\right.$$

$$\left. - (n_2 - x_2)\log\left(\frac{\hat{R}_{G1}}{\hat{R}_{G1} + \hat{R}_{G2}}\right)\right)$$

$$= -2\left(n_1 \log\left(\frac{1 + \hat{R}_{G1}}{2}\right) + n_2 \log\left(\frac{\hat{R}_{G1} + \hat{R}_{G2}}{2\hat{R}_{G1}}\right) - x_1 \log\left(\hat{R}_{G1}\right) - x_2 \log\left(\frac{\hat{R}_{G2}}{\hat{R}_{G1}}\right)\right)$$

$$= -2\left(n_1 \log\left(\frac{n_1}{2(n_1 - x_1)}\right) + n_2 \log\left(\frac{n_2}{2(n_2 - x_2)}\right) - x_1 \log\left(\frac{x_1}{n_1 - x_1}\right) - x_2 \log\left(\frac{x_2}{n_2 - x_2}\right)\right)$$

We can also consider a **log-additive** model such that:

$$H_0: R_{G1} = R_{G2} = 1$$
$$H_A: R_{G1}^2 = R_{G2} \neq 1$$

Then we have the following binomial model:

$$\begin{aligned} & p(x_1, x_2 | n_1, n_2, R_1, R_2) \\ &= \binom{n_1}{x_1}\left(\frac{R_{G1}}{1 + R_{G1}}\right)^{x_1}\left(\frac{1}{1 + R_{G1}}\right)^{n_1 - x_1}\binom{n_2}{x_2}\left(\frac{R_{G2}}{R_{G1} + R_{G2}}\right)^{x_2}\left(\frac{R_{G1}}{R_{G1} + R_{G2}}\right)^{n_2 - x_2} \\ &= \binom{n_1}{x_1}\left(\frac{R_{G1}}{1 + R_{G1}}\right)^{x_1}\left(\frac{1}{1 + R_{G1}}\right)^{n_1 - x_1}\binom{n_2}{x_2}\left(\frac{R_{G1}^2}{R_{G1} + R_{G1}^2}\right)^{x_2}\left(\frac{R_{G1}}{R_{G1} + R_{G1}^2}\right)^{n_2 - x_2} \\ &= \binom{n_1}{x_1}\left(\frac{R_{G1}}{1 + R_{G1}}\right)^{x_1}\left(\frac{1}{1 + R_{G1}}\right)^{n_1 - x_1}\binom{n_2}{x_2}\left(\frac{R_{G1}}{1 + R_{G1}}\right)^{x_2}\left(\frac{1}{1 + R_{G1}}\right)^{n_2 - x_2} \\ &= \binom{n_1}{x_1}\binom{n_2}{x_2}\left(\frac{R_{G1}}{1 + R_{G1}}\right)^{x_1 + x_2}\left(\frac{1}{1 + R_{G1}}\right)^{n_1 + n_2 - x_1 - x_2} \end{aligned} \quad \text{(B. 3)}$$

To find the maximum likelihood estimate of $R_{G1}$and R_{G2} ($\hat{R}_{G1}$ and $\hat{R}_{G2}$), we differentiate the log likelihood function that corresponds to equation B.3, set it to 0 and solve for $\hat{R}_{G1}$ and $\hat{R}_{G2}$:

$$\ell \sim (x_1 + x_2)\log(R_{G1}) - (n_1 + n_2)\log(1 + R_{G1})$$

$$\frac{d\ell}{dR_{G1}} = \frac{x_1 + x_2}{R_{G1}} - \frac{n_1 + n_2}{1 + R_{G1}}$$

$$0 = \frac{x_1 + x_2}{\hat{R}_{G1}} - \frac{n_1 + n_2}{1 + \hat{R}_{G1}}$$

$$\hat{R}_{G1} = \frac{x_1 + x_2}{n_1 + n_2 - x_1 - x_2}$$
$$\hat{R}_{G2} = \hat{R}_{G1}^2$$

The likelihood ratio test statistic (to be compared to a 1 DF chi-squared) is then as follows:

$$LRT = -2\left(\ell(R_1 = R_2 = 1) - \ell\left(R_1 = \hat{R}_{G1}, R_2 = \hat{R}_{G1}^2\right)\right)$$

$$= -2\left((x_1 + \mathrm{x}_2)\log\left(\frac{1}{2}\right) + (n_1 + n_2 - x_1 - x_2)\log\left(\frac{1}{2}\right) \right.$$
$$\left. - (x_1 + \mathrm{x}_2)\log\left(\frac{\hat{R}_{G1}}{1+\hat{R}_{G1}}\right) - (n_1 + n_2 - x_1 - x_2)\log\left(\frac{1}{1+\hat{R}_{G1}}\right)\right)$$
$$= -2\left((n_1 + \mathrm{n}_2)\log\left(\frac{1+\hat{R}_{G1}}{2}\right) - (x_1 + x2)\log\left(\hat{R}_{G1}\right)\right)$$
$$= -2\left((n_1 + \mathrm{n}_2)\log\left(\frac{n_1 + n_2}{2(n_1 + n_2 - x_1 - x_2)}\right) - (x_1 + x_2)\log\left(\frac{x_1 + x_2}{n_1 + n_2 - x_1 - x_2}\right)\right)$$

Similarly, we can calculate the maximum likelihood estimates and likelihood ratio test statistic for a **dominant** model where:

$$H_0: R_{G1} = R_{G2} = 1$$
$$H_A: R_{G1} = R_{G2} \neq 1$$

The MLEs of $R_{G1}$and $R_{G2}$are:

$$\hat{R}_{G2} = \hat{R}_{G1} = \frac{x_1}{n_1 - x_1}$$

The likelihood ratio test statistic (to be compared to a 1 DF chi-squared) is:

$$LRT = -2\left(n_1\log\left(\frac{n_1}{2(n_1 - x_1)}\right) - x_1\log\left(\frac{x_1}{n_1 - x_1}\right)\right)$$

We calculate the maximum likelihood estimates and likelihood ratio test statistic for a **recessive** model where:

$$H_0: R_{G1} = R_{G2} = 1$$
$$H_A: R_{G1} = 1, R_{G2} \neq 1$$

The MLE of R_{G2} is:

$$\hat{R}_{G2} = \frac{x_2}{n_2 - x_2}$$

The likelihood ratio test statistic (to be compared to a 1 DF chi-squared) is:

$$LRT = -2\left(n_2\log\left(\frac{n_2}{2(n_2 - x_2)}\right) - x_1\log\left(\frac{x_2}{n_2 - x_2}\right)\right)$$

Note that families in which the father carries the variant allele are not information under a dominant model. And only families in which the father carries the variant allele are informative under a recessive model.

APPENDIX C: CLOSED FORM SOLUTIONS FOR THE PARENT-ONLY ANALYSIS

As in the main paper, we define *M*, *F*, and *C* as the number of variant alleles carried by the mother, father and child and we define the relative risk of being affected (aff), conditional on mating type (*M*, *F*) to control for population stratification, as:

$$R_B = \Pr(\text{aff}|\text{boy}, C = 1) \,/\Pr(\text{aff}|\text{boy}, C = 0)$$

$$R_{G1} = \Pr(\text{aff}|\text{girl}, C = 1) \,/\Pr(\text{aff}|\text{girl}, C = 0)$$

$$R_{G2} = R_{G1} * \Pr(\text{aff}|\text{girl}, C = 2) \,/\Pr(\text{aff}|\text{girl}, C = 1)$$

C.1 Triads with affected sons

We are interested in the following hypothesis test:

$$H_0: R_B = 1$$
$$H_A: R_B \neq 1$$

Define:

- n_1 = the number of triads with $M+F=1$
- x_1 = the number of triads with $M+F=1$ and $M=1$ and $F=0$
- n_2 = the number of triads with $M+F=2$
- x_2 = the number of triads with $M+F=2$ and $M=2$ and $F=0$

More generally, we have the following model (see Table 3 in paper):

$$p(x_1, x_2|n_1, n_2, s) = \binom{n_1}{x_1}\left(\frac{1+R_B}{2+R_B}\right)^{x_1}\left(\frac{1}{2+R_B}\right)^{n_1-x_1}\binom{n_2}{x_2}\left(\frac{R_B}{1+2R_B}\right)^{x_2}\left(\frac{1+R_B}{1+2R_B}\right)^{n_2-x_2} \quad (C.1)$$

To find the maximum likelihood estimate of R_B ($\hat{R}_B$), we differentiate the log likelihood function (ignoring the constant terms) that corresponds to C.1, set it to 0 and solve for $\hat{R}_B$:

$$\ell \sim x_1 \log(1+R_B) - n_1 \log(2+R_B) + x_2 \log(R_B) + (n_2 - x_2)\log(1+R_B) - n_2 \log(1+2R_B)$$

$$\frac{d\ell}{ds} = \frac{x_1 + n_2 - x_2}{1+R_B} - \frac{n_1}{2+R_B} + \frac{x_2}{R_B} - \frac{2n_2}{1+2R_B}$$

$$0 = \frac{x_1 + n_2 - x_2}{1+\hat{R}_B} - \frac{n_1}{2+\hat{R}_B} + \frac{x_2}{\hat{R}_B} - \frac{2n_2}{1+2\hat{R}_B}$$

$$= \left(2\hat{R}_B + 5\hat{R}_B^2 + 2\hat{R}_B^3\right)(x_1 + n_2 - x_2) - \left(\hat{R}_B + 3\hat{R}_B^2 + 2\hat{R}_B^3\right)(n_1)$$
$$- \left(2 + 7\hat{R}_B + 7\hat{R}_B^2 + 2\hat{R}_B^3\right)(x_2) - \left(2\hat{R}_B + 3\hat{R}_B^2 + \hat{R}_B^3\right)(2n_2)$$

$$= a_0 + a_1\hat{R}_B + a_2\hat{R}_B^2 + \hat{R}_B^3$$

where:

$$a_0 = \frac{x_2}{x_1 - n_1}$$

$$a_1 = \frac{2x_1 + 5x_2 - n_1 - 2n_2}{2x_1 - 2n_1}$$

$$a_2 = \frac{5x_1 + 2x_2 - 3n_1 - n_2}{2x_1 - 2n_1}$$

We can use Cardano's folmula to solve for the cubic. This solution was published in the 1500's by Gerolamo Cardano in *Ars Magna* (Cardano 1545, Cardano and Witmer 1993)

We are interested in the positive root for $\hat{R}_B$:

$$Q = \frac{3a_1 - {a_2}^2}{9}$$

$$R = \frac{9a_2a_1 - 27a_0 - 2{a_2}^3}{54}$$

$$D = Q^3 + R^2$$

$$S = \sqrt[3]{R + \sqrt{D}}$$

$$T = \sqrt[3]{R - \sqrt{D}}$$

The three roots are:

$$\left(-\frac{1}{3}\right)a_2 + (S + T)$$

$$\left(-\frac{1}{3}\right)a_2 - \frac{1}{2}(S + T) + \frac{1}{2}i\sqrt{3}(S - T)$$

$$\left(-\frac{1}{3}\right)a_2 - \frac{1}{2}(S + T) - \frac{1}{2}i\sqrt{3}(S - T)$$

Let $\hat{R}_B$ be the positive, real root (if D>0, the first root). An approach using trigonometry was later developed as well and can be used to avoid the imaginary number (Nickalls 2006).

The likelihood ratio test statistic (to be compared to a 1 DF chi-squared) is as follows:

$$
\begin{aligned}
LRTS &= -2\left(\ell(R_B = 1) - \ell\left(R_B = \hat{R}_B\right)\right) \\
&= -2\left(x_1 \log(2) - n_1 \log(3) + x_2 \log\left(\frac{1}{2}\right) - n_2 \log\left(\frac{3}{2}\right) - x_1 \log\left(1 + \hat{R}_B\right)\right. \\
&\qquad \left. + n_1 \log\left(2 + \hat{R}_B\right) - x_2 \log\left(\frac{\hat{R}_B}{1 + \hat{R}_B}\right) + n_2 \log\left(\frac{1 + 2\hat{R}_B}{1 + \hat{R}_B}\right)\right) \\
&= -2\left(x_1 \log\left(\frac{2}{1 + \hat{R}_B}\right) + n_1 \log\left(\frac{2 + \hat{R}_B}{3}\right) - x_2 \log\left(\frac{2\hat{R}_B}{1 + \hat{R}_B}\right)\right. \\
&\qquad \left. + n_2 \log\left(\frac{2\left(1 + 2\hat{R}_B\right)}{3\left(1 + \hat{R}_B\right)}\right)\right)
\end{aligned}
$$

C.2 Triads with affected daughters

We are interested in the following hypothesis:

$$
\begin{aligned}
H_0&: R_{G1} = R_{G2} = 1 \\
H_A&: R_{G1} \neq 1 \text{ or } R_{G2} \neq 1
\end{aligned}
$$

Define:

- n_1 = the number of triads where $M+F=1$
- x_1 = the number of triads where $M+F=1$ and $M = 1$ and $F = 0$
- n_2 = the number of triads where $M+F=2$
- x_2 = the number of triads where $M+F=2$ and $M = 2$ and $F = 0$

More generally, we have the following model (see Table 3 in paper):

$$
\begin{aligned}
&p(x_1, x_2 | n_1, n_2, R_{G1}, R_{G2}) \\
&= \binom{n_1}{x_1}\left(\frac{1 + R_{G1}}{1 + 2R_{G1}}\right)^{x_1}\left(\frac{R_{G1}}{1 + 2R_{G1}}\right)^{n_1 - x_1}\binom{n_2}{x_2}\left(\frac{R_{G1}}{2R_{G1} + R_{G2}}\right)^{x_2}\left(\frac{R_{G1} + R_{G2}}{2R_{G1} + R_{G2}}\right)^{n_2 - x_2} \qquad (\text{C.2})
\end{aligned}
$$

To find the maximum likelihood estimate of $R_{G1}$and R_{G2} ($\hat{R}_{G1}$ and $\hat{R}_{G2}$), we differentiate the log likelihood function that corresponds to equation C.2, set it to 0 and solve for $\hat{R}_{G1}$ and $\hat{R}_{G2}$:

$$
\begin{aligned}
\ell \sim & x_1 \log(1 + R_{G1}) + (x_2 + n_1 - x_1)\log(R_{G1}) - n_1 \log(1 + 2R_{G1}) \\
& + (n_2 - x_2)\log(R_{G1} + R_{G2}) - n_2 \log(2R_{G1} + R_{G2})
\end{aligned}
$$

$$\frac{d\ell}{dR_{G2}} = \frac{n_2 - x_2}{R_{G1} + R_{G2}} - \frac{n_2}{2R_{G1} + R_{G2}}$$

$$0 = \frac{n_2 - x_2}{R_{G1} + \hat{R}_{G2}} - \frac{n_2}{2R_{G1} + \hat{R}_{G2}}$$

$$= \left(2R_{G1} + \hat{R}_{G2}\right)(n_2 - x_2) - \left(R_{G1} + \hat{R}_{G2}\right)(n_2)$$

$$= \hat{R}_{G2}(x_2) + R_{G1}(2x_2 - n_2)$$

$$\hat{R}_{G2} = R_{G1}\frac{n_2 - 2x_2}{x_2}$$

$$\frac{d\ell}{dR_{G1}} = \frac{x_1}{1 + R_{G1}} + \frac{x_2 + n_1 - x_1}{R_{G1}} - \frac{2n_1}{1 + 2R_{G1}} + \frac{n_2 - x_2}{R_{G1} + \hat{R}_{G2}} - \frac{2n_2}{2R_{G1} + \hat{R}_{G2}}$$

$$= \frac{x_1}{1 + R_{G1}} + \frac{x_2 + n_1 - x_1}{R_{G1}} - \frac{2n_1}{1 + 2R_{G1}} + \frac{n_2 - x_2}{R_{G1}\left(1 + \frac{n_2 - 2x_2}{x_2}\right)} - \frac{2n_2}{R_{G1}\left(2 + \frac{n_2 - 2x_2}{x_2}\right)}$$

$$= \frac{x1}{1 + R_{G1}} + \frac{x_2 + n_1 - x_1}{R_{G1}} - \frac{2n_1}{1 + 2R_{G1}} + \frac{x_2}{R_{G1}} - \frac{2x_2}{R_{G1}}$$

$$= \frac{x_1}{1 + R_{G1}} + \frac{n_1 - x_1}{R_{G1}} - \frac{2n_1}{1 + 2R_{G1}}$$

$$0 = (R_{G1} + 2R_{G1}^2)x_1 + (1 + 3R_{G1} + 2R_{G1}^2)(n_1 - x_1) - (R_{G1} + R_{G1}^2)2n_1$$

$$= R_{G1}(n_1 - 2x_1) + n_1 - x_1$$

$$\hat{R}_{G1} = \frac{x_1 - n_1}{n_1 - 2x_1}$$

The likelihood ratio test statistic (to be compared to a 2 DF chi-squared) is:

$$LRTS = -2\left(l(R_{G1} = R_{G2} = 1) - l\left(R_{G1} = \hat{R}_{G1}, R_{G2} = \hat{R}_{G2}\right)\right)$$

$$= -2\left(x_1 \log(2) - n_1 \log(3) - x_2 \log(2) + n_2 \log\left(\frac{2}{3}\right) - x_1 \log\left(\frac{1 + \hat{R}_{G1}}{\hat{R}_{G1}}\right)\right.$$

$$\left. - n_1 \log\left(\frac{\hat{R}_{G1}}{1 + 2\hat{R}_{G1}}\right) - x_2 \log\left(\frac{\hat{R}_{G1}}{\hat{R}_{G1} + \hat{R}_{G2}}\right) - n_2 \log\left(\frac{\hat{R}_{G1} + \hat{R}_{G2}}{2\hat{R}_{G1} + \hat{R}_{G2}}\right)\right)$$

$$= -2\left(x_1 \log\left(\frac{2(n_1 - x_1)}{x_1}\right) - n_1 \log\left(\frac{3(n_1 - x_1)}{n_1}\right) - x_2 \log\left(\frac{2x_2}{n_2 - x_2}\right) + n_2 \log\left(\frac{2n_2}{3(n_2 - x_2)}\right)\right)$$

We can also consider a **log-additive** model such that:

$$H_0: R_{G1} = R_{G2} = 1$$
$$H_A: R_{G1}^2 = R_{G2} \neq 1$$

More generally, we have the following model:

$$p(x_1, x_2 | n_1, n_2, R_{G1}, R_{G2})$$
$$= \binom{n_1}{x_1}\left(\frac{1 + R_{G1}}{1 + 2R_{G1}}\right)^{x_1}\left(\frac{R_{G1}}{1 + 2R_{G1}}\right)^{n_1 - x_1}\binom{n_2}{x_2}\left(\frac{R_{G1}}{2R_{G1} + R_{G2}}\right)^{x_2}\left(\frac{R_{G1} + R_{G2}}{2R_{G1} + R_{G2}}\right)^{n_2 - x_2}$$

$$= \binom{n_1}{x_1}\left(\frac{1 + R_{G1}}{1 + 2R_{G1}}\right)^{x_1}\left(\frac{R_{G1}}{1 + 2R_{G1}}\right)^{n_1 - x_1} \times \binom{n_2}{x_2}\left(\frac{R_{G1}}{2R_{G1} + R_{G1}^2}\right)^{x_2}\left(\frac{R_{G1} + R_{G1}^2}{2R_{G1} + R_{G1}^2}\right)^{n_2 - x_2}$$

$$= \binom{n_1}{x_1}\left(\frac{1 + R_{G1}}{1 + 2R_{G1}}\right)^{x_1}\left(\frac{R_{G1}}{1 + 2R_{G1}}\right)^{n_1 - x_1}\binom{n_2}{x_2}\left(\frac{1}{2 + R_{G1}}\right)^{x_2}\left(\frac{1 + R_{G1}}{2 + R_{G1}}\right)^{n_2 - x_2} \quad \text{(C. 3)}$$

To find the maximum likelihood estimate of $R_{G1}$and R_{G2} ($\hat{R}_{G1}$ and $\hat{R}_{G2}$), we differentiate the log likelihood function that corresponds to equation C.3, set it to 0 and solve for $\hat{R}_{G1}$ and $\hat{R}_{G2}$:

$$\ell \sim x_1 \log(1 + R_{G1}) + (n_1 - x_1)\log(R_{G1}) - n_1 \log(1 + 2R_{G1}) + (n_2 - x_2)\log(1 + R_{G1}) - n_2 \log(2 + R_{G1})$$

$$\frac{d\ell}{dR_{G1}} = \frac{x_1 + n_2 - x_2}{1 + R_{G1}} - \frac{n_2}{2 + R_{G1}} + \frac{n_1 - x_1}{R_{G1}} - \frac{2n_1}{1 + 2R_{G1}}$$

$$0 = \frac{x_1 + n_2 - x_2}{1 + \hat{R}_{G1}} - \frac{n_2}{2 + \hat{R}_{G1}} + \frac{n_1 - x_1}{\hat{R}_{G1}} - \frac{2n_1}{1 + 2\hat{R}_{G1}}$$

$$= \left(2\hat{R}_{G1} + 5\hat{R}_{G1}^2 + 2\hat{R}_{G1}^3\right)(x_1 + n_2 - x_2) - \left(\hat{R}_{G1} + 3\hat{R}_{G1}^2 + 2\hat{R}_{G1}^3\right)(n_2) - \left(2 + 7\hat{R}_{G1} + 7\hat{R}_{G1}^2 + 2\hat{R}_{G1}^3\right)(n_1 - x_1) - \left(2\hat{R}_{G1} + 3\hat{R}_{G1}^2 + \hat{R}_{G1}^3\right)(2n_1)$$

$$= a_0 + a_1\hat{R}_{G1} + a_2\hat{R}_{G1}^2 + \hat{R}_{G1}^3$$

where:

$$a_0 = \frac{x_1 - n_1}{x_2}$$

$$a_1 = \frac{2x_2 + 5x_1 - n_2 - 3n_1}{2x_2}$$

$$a_2 = \frac{5x_2 + 2x_1 - 2n_2 - n_1}{2x_2}$$

See Appendix C.1 for how to solve for $\hat{R}_{G1}$, then $\hat{R}_{G2} = \hat{R}_{G1}^2$.

The likelihood ratio test statistic (to be compared to a 1 DF chi-squared) is:

$$\begin{aligned} LRTS &= -2\left(\ell(R_{G1} = R_{G2} = 1) - \ell\left(R_{G1} = \hat{R}_{G1}, R_{G2} = \hat{R}_{G1}^2\right)\right) \\ &= -2\left(x_1 \log(2) - n_1 \log(3) + x_2 \log\left(\frac{1}{2}\right) - n_2 \log\left(\frac{3}{2}\right) - x_1 \log\left(\frac{1 + \hat{R}_{G1}}{\hat{R}_{G1}}\right)\right. \\ &\qquad \left. + n_1 \log\left(\frac{1 + 2\hat{R}_{G1}}{\hat{R}_{G1}}\right) - x_2 \log\left(\frac{1}{1 + \hat{R}_{G1}}\right) + n_2 \log\left(\frac{2 + \hat{R}_{G1}}{1 + \hat{R}_{G1}}\right)\right) \\ &= -2\left(x_1 \log\left(\frac{2\hat{R}_{G1}}{1 + \hat{R}_{G1}}\right) + n_1 \log\left(\frac{1 + 2\hat{R}_{G1}}{3\hat{R}_{G1}}\right) + x_2 \log\left(\frac{1 + \hat{R}_{G1}}{2}\right)\right. \\ &\qquad \left. + n_2 \log\left(\frac{2\left(2 + \hat{R}_{G1}\right)}{3\left(1 + \hat{R}_{G1}\right)}\right)\right) \end{aligned}$$

Similarly for triads with affected daughters, we can calculate the maximum likelihood estimates and likelihood ratio test statistic for a **dominant** or **recessive** model (results not shown).

APPENDIX D: CLOSED FORM SOLUTIONS FOR THE PIX-LRT

In this section we define a likelihood that involves both the transmission-based information and the parental information. As in the main paper, we define *M*, *F*, and *C* as the number of variant alleles carried by the mother, father and child and we define the relative risk of being affected (aff), conditional on mating type (*M*,*F*) to control for population stratification, as:

$$R_B = \Pr(\text{aff}|\text{boy}, C = 1)\ /\Pr(\text{aff}|\text{boy}, C = 0)$$

$$R_{G1} = \Pr(\text{aff}|\text{girl}, C = 1)\ /\Pr(\text{aff}|\text{girl}, C = 0)$$

$$R_{G2} = R_{G1} * \Pr(\text{aff}|\text{girl}, C = 2)\ /\Pr(\text{aff}|\text{girl}, C = 1)$$

D.1 Triads with affected sons

We are interested in the following hypothesis test:

$$H_0: R_B = 1$$
$$H_A: R_B \neq 1$$

Define:

- n_1 = the number of triads where $M+F=1$
- x_1 = the number of triads where $M+F=1$ and $C=1$
- n_2 = the number of triads where $M+F=2$
- x_2 = the number of triads where $M+F=2$ and $C=1$

We have the following model:

$$p(x_1, x_2|n_1, n_2, R_B) = \binom{n_1}{x_1}\left(\frac{R_B}{2+R_B}\right)^{x_1}\left(\frac{2}{2+R_B}\right)^{n_1-x_1}\binom{n_2}{x_2}\left(\frac{2R_B}{1+2R_B}\right)^{x_2}\left(\frac{1}{1+2R_B}\right)^{n_2-x_2}$$

The likelihood and ML estimate are as follows:

$$\ell \sim x_1 \log(R_B) - n_1 \log(2+R_B) + x_2 \log(R_B) - n_2 \log(1+2R_B)$$

$$\frac{d\ell}{dR_B} = \frac{x_1+x_2}{R_B} - \frac{n_1}{2+R_B} - \frac{2n_2}{1+2R_B}$$

$$0 = \left(2+5\hat{R}_B+2\hat{R}_B^2\right)(x_1+x_2) - \left(\hat{R}_B+2\hat{R}_B^2\right)(n_1) - \left(2\hat{R}_B+\hat{R}_B^2\right)(2n_2)$$

$$= \hat{R}_B^2\left(2(x_1+x_2-n_1-n_2)\right) + \hat{R}_B(5x_1+5x_2-n_1-4n_2) + 2(x_1+x_2)$$

If:

$a = 2(x_1 + x_2 - n_1 - n_2)$
$b = 5x_1 + 5x_2 - n_1 - 4n_2$
$c = 2(x_1 + x_2)$

Then $\hat{R}_B = \frac{(-b-\sqrt{b^2-4ac})}{2a}$ and the LRT statistic is:

The likelihood ratio test statistic (to be compared to a 1 DF chi-squared) is:

$$LRTS = -2\left(\ell(R_B = 1) - \ell\left(R_B = \hat{R}_B\right)\right)$$
$$= -2\left(x_1 \log(1) - n_1 \log(3) + x_2\log(1) - n_2\log(3) - x_1\log\left(\hat{R}_B\right) + n_1\log\left(2 + \hat{R}_B\right)\right.$$
$$\left. + x_2\log\left(\hat{R}_B\right) + n_2\log\left(1 + 2\hat{R}_B\right)\right)$$
$$= -2\left(x_1 \log\left(\frac{1}{\hat{R}_B}\right) + n_1 \log\left(\frac{2 + \hat{R}_B}{3}\right) + x_2\log\left(\frac{1}{\hat{R}_B}\right) + n_2\log\left(\frac{1 + \hat{R}_B}{3}\right)\right)$$

D.2 Triads with affected daughters

Define:

- n_1 = the number of triads where $M+F = 1$
- x_1 = the number of triads where $M+F = 1$ and $C = 1$
- n_2 = the number of triads where $M+F = 2$
- x_2 = the number of triads where $M+F = 2$ and C = 2

We are interested in testing the following hypothesis:

$$H_0: R_{G1} = R_{G2} = 1$$
$$H_A: R_{G1} \neq 1, R_{G2} \neq 1$$

We have the following model:

$$p(x_1, x_2 | n_1, n_2, R_{G1}, R_{G2})$$
$$= \binom{n_1}{x_1}\left(\frac{2R_{G1}}{1 + 2R_{G1}}\right)^{x_1}\left(\frac{1}{1 + 2RG_1}\right)^{n_1 - x_1}\binom{n_2}{x_2}\left(\frac{R_{G2}}{2R_{G1} + R_{G2}}\right)^{x_2}\left(\frac{2R_{G1}}{2R_{G1} + R_{G2}}\right)^{n_2 - x_2}$$

The likelihood and ML estimates for R_1 and R_2 are as follows:

$$\ell \sim (x_1 + n_2 - x_2)\log(R_{G1}) - n_1 \log(1 + 2R_{G1}) + x_2 \log(R_{G2}) - n_2 \log(2R_{G1} + R_{G2})$$

$$\frac{d\ell}{dR_{G2}} = \frac{x_2}{R_{G2}} - \frac{n_2}{2R_{G1} + R_{G2}}$$

$$0 = x_2\left(2R_{G1} + \hat{R}_{G2}\right) - n_2\left(\hat{R}_{G2}\right)$$

$$\hat{R}_{G2} = \frac{2R_{G1}x_2}{n_2 - x_2}$$

$$\frac{d\ell}{dR_{G1}} = \frac{x_1 + n_2 - x_2}{R_{G1}} - \frac{2n_1}{1 + 2R_{G1}} - \frac{2n_2}{2R_{G1} + R_{G2}}$$

$$\begin{aligned} 0 &= \frac{x_1 + n_2 - x_2}{\hat{R}_{G1}} - \frac{2n_1}{1 + 2\hat{R}_{G1}} - \frac{2n_2}{2\hat{R}_{G1} + \hat{R}_{G2}} \\ &= \frac{x_1 + n_2 - x_2}{\hat{R}_{G1}} - \frac{2n_1}{1 + 2\hat{R}_{G1}} - \frac{2n_2}{2\hat{R}_{G1}\left(\frac{n_2}{n_2 - x_2}\right)} \\ &= \left(1 + 2\hat{R}_{G1}\right)x_1 - 2\hat{R}_{G1}n_1 \end{aligned}$$

$$\hat{R}_{G1} = \frac{x_1}{2(n_1 - x_1)}$$

The likelihood ratio test statistic (to be compared to a 1 DF chi-squared) is:

$$\begin{aligned} LRTS = & -2\left(\ell(R_{G1} = R_{G2} = 1) - \ell\left(R_{G1} = \hat{R}_{G1}, R_{G2} = \hat{R}_{G2}\right)\right) \\ &= -2\Bigg(x_1\log(1) - n_1\log(3) + x_2\log(1) - n_2\log(3) - x_1\log\left(\hat{R}_{G1}\right) \\ &\qquad + n_1\log\left(1 + 2\hat{R}_{G1}\right) - x_2\log\left(\frac{\hat{R}_{G2}}{\hat{R}_{G1}}\right) + n_2\log\left(\frac{2\hat{R}_{G1} + \hat{R}_{G2}}{\hat{R}_{G1}}\right)\Bigg) \\ &= -2\left(x_1\log\left(\frac{2(n_1 - x_1)}{x_1}\right) + n_1\log\left(\frac{n_1}{3(n_1 - x_1)}\right) + x_2\log\left(\frac{n_2 - x_2}{2x_2}\right)\right. \\ &\qquad \left. + n_2\log\left(\frac{2n_2}{3(n_2 - x_2)}\right)\right) \end{aligned}$$

We can also consider a **log-additive** model such that:

$$H_0: R_{G1} = R_{G2} = 1$$
$$H_A: R_{G1}^2 = R_{G2} \neq 1$$

Our model is as follows:

$$\begin{aligned} & p(x_1, x_2 | n_1, n_2, R_1, R_2) \\ &= \binom{n_1}{x_1}\left(\frac{2R_{G1}}{1 + 2R_{G1}}\right)^{x_1}\left(\frac{1}{1 + 2R_{G1}}\right)^{n_1 - x_1}\binom{n_2}{x_2}\left(\frac{R_{G2}}{2R_{G1} + R_{G2}}\right)^{x_2}\left(\frac{2R_{G1}}{2R_{G1} + R_{G2}}\right)^{n_2 - x_2} \\ &= \binom{n_1}{x_1}\left(\frac{2R_{G1}}{1 + 2R_{G1}}\right)^{x_1}\left(\frac{1}{1 + 2R_{G1}}\right)^{n_1 - x_1}\binom{n_2}{x_2}\left(\frac{R_1^2}{2R_1 + R_1^2}\right)^{x_2}\left(\frac{2R_1}{2R_1 + R_1^2}\right)^{n_2 - x_2} \\ &= \binom{n_1}{x_1}\left(\frac{2R_1}{1 + 2R_1}\right)^{x_1}\left(\frac{1}{1 + 2R_1}\right)^{n_1 - x_1}\binom{n_2}{x_2}\left(\frac{R_{G1}}{2 + R_{G1}}\right)^{x_2}\left(\frac{2}{2 + R_{G1}}\right)^{n_2 - x_2} \end{aligned}$$

The likelihood and ML estimates for R_1 and R_2 are as follows:

$$\ell \sim (x_1)\log(R_{G1}) - n_1\log(1 + 2R_{G1}) + x_2\log(R_{G1}) - n_2\log(2 + R_{G1})$$

$$\frac{d\ell}{dR_{G1}} = \frac{x_1 + x_2}{R_{G1}} - \frac{2n_1}{1 + 2R_{G1}} + \frac{n_2}{2 + R_{G1}}$$

$$0 = \frac{x_1 + x_2}{\hat{R}_{G1}} - \frac{2n_1}{1 + 2\hat{R}_{G1}} + \frac{n_2}{2 + \hat{R}_{G1}}$$

$$= \left(2 + 5\hat{R}_{G1} + 2\hat{R}_{G1}^2\right)(x_1 + x_2) - \left(2\hat{R}_{G1} + \hat{R}_{G1}^2\right)(2n_1) - \left(\hat{R}_{G1} + 2\hat{R}_{G1}^2\right)(n_2)$$

$$= a\hat{R}_{G1}^2 + b\hat{R}_{G1} + c$$

Where:
$a = 2(x_1 + x_2 - n_1 - n_2)$
$b = 5x_2 + 5x_1 - 4n_1 - n_2$
$c = 2(x_1 + x_2)$

So $\hat{R}_{G1} = \frac{-b - \sqrt{b^2 - 4ac}}{2a}$ and $\hat{R}_{G2} = \hat{R}_{G1}^2$. The LRT statistic (to be compared to a 1 DF chi-squared) is:

$$LRTS = -2\left(\ell(R_{G1} = R_{G2} = 1) - \ell\left(R_{G1} = \hat{R}_{G1}, R_{G2} = \hat{R}_{G1}^2\right)\right)$$

$$= -2\left(x_1\log(1) - n_1\log(3) + x_2\log(1) - n_2\log(3) - x_1\log\left(\hat{R}_{G1}\right) + n_1\log\left(1 + 2\hat{R}_{G1}\right) - x_2\log\left(\hat{R}_{G1}\right) + n_2\log\left(2 + \hat{R}_{G1}\right)\right)$$

$$= -2\left(x_1\log\left(\frac{1}{\hat{R}_{G1}}\right) + n_1\log\left(\frac{1 + 2\hat{R}_{G1}}{3}\right) + x_2\log\left(\frac{1}{\hat{R}_{G1}}\right) + n_2\log\left(\frac{2 + \hat{R}_{G1}}{3}\right)\right)$$

Similarly for triads with affected daughters, we can calculate the maximum likelihood estimates and likelihood ratio test statistic for a **dominant** or **recessive** model (results not shown).

APPENDIX E: D ACKNOWLEDGEMENT

The details of the collection and methods for samples used in this oral cleft study are described by Beaty *et al.* (Beaty, Murray et al. 2010). The data sets used for the analyses described in this manuscript were obtained through dbGaP at www.ncbi.nlm.nih.gov/gap through accession number phs000094.v1.p1. Funding support for the study entitled 'International Consortium to Identify Genes and Interactions Controlling Oral Clefts' was provided by several previous grants from the National Institute of Dental and Craniofacial Research (NIDCR), including: R21-DE-013707, R01-DE-014581, R37-DE-08559, P50-DE-016215, R01-DE-09886, R01-DE-012472, R01-DE-014677, R01-DE-016148, R21-DE-016930; R01-DE-013939. Additional support was provided in part by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences, the Smile Train Foundation for recruitment in China and a Grant from the Korean government. The genome-wide association study, also known the Cleft Consortium, is part of the Gene Environment Association Studies (GENEVA) program of the trans-NIH Genes, Environment and Health Initiative [GEI] supported by U01-DE-018993. Genotyping services were provided by the Center for Inherited Disease Research (CIDR). CIDR is funded through a federal contract from the National Institutes of Health (NIH) to The Johns Hopkins University, contract number HHSN268200782096C. Assistance with genotype cleaning, as well as with general study coordination, was provided by the GENEVA Coordinating Center (U01-HG-004446) and by the National Center for Biotechnology Information (NCBI).

REFERENCES

Abbadi, N., C. Philippe, M. Chery, H. Gilgenkrantz, F. Tome, H. Collin, D. Theau, D. Recan, O. Broux, M. Fardeau and et al. (1994). "Additional case of female monozygotic twins discordant for the clinical manifestations of Duchenne muscular dystrophy due to opposite X-chromosome inactivation." Am J Med Genet **52**(2): 198-206.

Agresti, A. (2012). Categorical Data Analysis. New Jersey, Wiley.

Akey, J., L. Jin and M. Xiong (2001). "Haplotypes vs single marker linkage disequilibrium tests: what do we gain?" Eur J Hum Genet **9**(4): 291-300.

Allen, A. S. and G. A. Satten (2007). "Inference on haplotype/disease association using parent-affected-child data: the projection conditional on parental haplotypes method." Genet Epidemiol **31**(3): 211-223.

Barr, M. L. and E. G. Bertram (1949). "A morphological distinction between neurones of the male and female, and the behaviour of the nucleolar satellite during accelerated nucleoprotein synthesis." Nature **163**(4148): 676.

Beaty, T. H., J. C. Murray, M. L. Marazita, R. G. Munger, I. Ruczinski, J. B. Hetmanski, K. Y. Liang, T. Wu, T. Murray, M. D. Fallin, R. A. Redett, G. Raymond, H. Schwender, S. C. Jin, M. E. Cooper, M. Dunnwald, M. A. Mansilla, E. Leslie, S. Bullard, A. C. Lidral, L. M. Moreno, R. Menezes, A. R. Vieira, A. Petrin, A. J. Wilcox, R. T. Lie, E. W. Jabs, Y. H. Wu-Chou, P. K. Chen, H. Wang, X. Ye, S. Huang, V. Yeow, S. S. Chong, S. H. Jee, B. Shi, K. Christensen, M. Melbye, K. F. Doheny, E. W. Pugh, H. Ling, E. E. Castilla, A. E. Czeizel, L. Ma, L. L. Field, L. Brody, F. Pangilinan, J. L. Mills, A. M. Molloy, P. N. Kirke, J. M. Scott, M. Arcos-Burgos and A. F. Scott (2010). "A genome-wide association study of cleft lip with and without cleft palate identifies risk variants near MAFB and ABCA4." Nat Genet **42**(6): 525-529.

Cardano, G. (1545). Hieronymi Cardani, praestantissimi mathematici, philosophi, ac medici, artis magnae, siue, De regulis algebraicis lib. unus : qui & totius operis de arithmetica, quod opus perfectum inscripsit, est in ordine decimus. Norimbergae, Per Ioh. Petreium excusum.

Cardano, G. and T. R. Witmer (1993). Ars magna, or, The rules of algebra. New York, Dover.

Carrel, L. and H. F. Willard (2005). "X-inactivation profile reveals extensive variability in X-linked gene expression in females." Nature **434**(7031): 400-404.

Chung, R. H., E. R. Hauser and E. R. Martin (2006). "The APL test: extension to general nuclear families and haplotypes and examination of its robustness." Hum Hered **61**(4): 189-199.

Chung, R. H., R. W. Morris, L. Zhang, Y. J. Li and E. R. Martin (2007). "X-APL: an improved family-based test of association in the presence of linkage for the X chromosome." Am J Hum Genet **80**(1): 59-68.

Clayton, D. (1999). "A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission." Am J Hum Genet **65**(4): 1170-1177.

Deng, H. W. and W. M. Chen (2001). "The power of the transmission disequilibrium test (TDT) with both case-parent and control-parent trios." Genet Res **78**(3): 289-302.

Ding, J., S. Lin and Y. Liu (2006). "Monte Carlo pedigree disequilibrium test for markers on the X chromosome." Am J Hum Genet **79**(3): 567-573.

Dixon, M. J., M. L. Marazita, T. H. Beaty and J. C. Murray (2011). "Cleft lip and palate: understanding genetic and environmental influences." Nat Rev Genet **12**(3): 167-178.

Dudbridge, F. (2008). "Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data." Hum Hered **66**(2): 87-98.

Gjessing, H. K. and R. T. Lie (2006). "Case-parent triads: estimating single- and double-dose effects of fetal and maternal disease gene haplotypes." Ann Hum Genet **70**(Pt 3): 382-396.

Horvath, S., N. M. Laird and M. Knapp (2000). "The transmission/disequilibrium test and parental-genotype reconstruction for X-chromosomal markers." Am J Hum Genet **66**(3): 1161-1167.

Jorgensen, A. L., J. Philip, W. H. Raskind, M. Matsushita, B. Christensen, V. Dreyer and A. G. Motulsky (1992). "Different patterns of X inactivation in MZ twins discordant for red-green color-vision deficiency." Am J Hum Genet **51**(2): 291-298.

Jugessur, A., M. Shi, H. K. Gjessing, R. T. Lie, A. J. Wilcox, C. R. Weinberg, K. Christensen, A. L. Boyles, S. Daack-Hirsch, T. T. Nguyen, L. Christiansen, A. C. Lidral and J. C. Murray (2010). "Maternal genes and facial clefts in offspring: a comprehensive search for genetic associations in two population-based cleft studies from Scandinavia." PLoS One **5**(7): e11493.

Jugessur, A., O. Skare, R. T. Lie, A. J. Wilcox, K. Christensen, L. Christiansen, T. T. Nguyen, J. C. Murray and H. K. Gjessing (2012). "X-linked genes and risk of orofacial clefts: evidence from two population-based studies in Scandinavia." PLoS One **7**(6): e39240.

Knapp, M. (1999). "The transmission/disequilibrium test and parental-genotype reconstruction: the reconstruction-combined transmission/ disequilibrium test." Am J Hum Genet **64**(3): 861-870.

Laird, N. M., S. Horvath and X. Xu (2000). "Implementing a unified approach to family-based tests of association." Genet Epidemiol **19 Suppl 1**: S36-42.

Lin, D. Y. and D. Zeng (2005). "Maximum likelihood methods for haplotype sharing studies." Genetic Epidemiology **29**(3): 265-265.

Lin, D. Y. and D. Zeng (2006). "Likelihood-based inference on haplotype effects in genetic association studies." Journal of the American Statistical Association **101**(473): 89-104.

Lupo, P. J., D. Nousome, M. F. Okcu, M. Chintagumpala and M. E. Scheurer (2012). "Maternal variation in EPHX1, a xenobiotic metabolism gene, is associated with childhood medulloblastoma: an exploratory case-parent triad study." Pediatr Hematol Oncol **29**(8): 679-685.

Lyon, M. F. (2002). "X-chromosome inactivation and human genetic disease." Acta Paediatr Suppl **91**(439): 107-112.

Mailman, M. D., M. Feolo, Y. Jin, M. Kimura, K. Tryka, R. Bagoutdinov, L. Hao, A. Kiang, J. Paschall, L. Phan, N. Popova, S. Pretel, L. Ziyabari, M. Lee, Y. Shao, Z. Y. Wang, K. Sirotkin, M. Ward, M. Kholodov, K. Zbicz, J. Beck, M. Kimelman, S. Shevelev, D. Preuss, E. Yaschenko, A. Graeff, J. Ostell and S. T. Sherry (2007). "The NCBI dbGaP database of genotypes and phenotypes." Nat Genet **39**(10): 1181-1186.

Marcano, A. C., K. Doudney, C. Braybrook, R. Squires, M. A. Patton, M. M. Lees, A. Richieri-Costa, A. C. Lidral, J. C. Murray, G. E. Moore and P. Stanier (2004). "TBX22 mutations are a frequent cause of cleft palate." J Med Genet **41**(1): 68-74.

Martin, E. R., M. P. Bass, E. R. Hauser and N. L. Kaplan (2003). "Accounting for linkage in family-based tests of association with missing parental genotypes." Am J Hum Genet **73**(5): 1016-1026.

Martin, E. R., S. A. Monks, L. L. Warren and N. L. Kaplan (2000). "A test for linkage and association in general pedigrees: the pedigree disequilibrium test." Am J Hum Genet **67**(1): 146-154.

Mitchell, L. E. (1997). "Differentiating between fetal and maternal genotypic effects, using the transmission test for linkage disequilibrium." Am J Hum Genet **60**(4): 1006-1007.

Morris, R. W. and N. L. Kaplan (2002). "On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles." Genet Epidemiol **23**(3): 221-233.

Murray, J. C. (2002). "Gene/environment causes of cleft lip and/or palate." Clin Genet **61**(4): 248-256.

Myking, S., H. A. Boyd, R. Myhre, B. Feenstra, A. Jugessur, A. S. Devold Pay, I. H. Ostensen, N. H. Morken, T. Busch, K. K. Ryckman, F. Geller, P. Magnus, H. K. Gjessing, M. Melbye, B. Jacobsson and J. C. Murray (2013). "X-chromosomal maternal and fetal SNPs and the risk of spontaneous preterm delivery in a Danish/Norwegian genome-wide association study." PLoS One **8**(4): e61781.

Nemeth, A. H., D. Nolte, E. Dunne, S. Niemann, M. Kostrzewa, U. Peters, E. Fraser, E. Bochukova, R. Butler, J. Brown, R. D. Cox, E. R. Levy, H. H. Ropers, A. P. Monaco and U. Muller (1999). "Refined linkage disequilibrium and physical mapping of the gene locus for X-linked dystonia-parkinsonism (DYT3)." Genomics **60**(3): 320-329.

Nickalls, R. W. D. (2006). "Viète, Descartes and the cubic equation." Mathematical Gazette **90**(203-208).

O'Brien, R. G. (1986). Using the SAS system to perform power analyses for log-linear models. Proc. 11th Annual SAS Users Group Conference.

Pankratz, N., W. C. Nichols, S. K. Uniacke, C. Halter, J. Murrell, A. Rudolph, C. W. Shults, P. M. Conneally, T. Foroud and G. Parkinson Study (2003). "Genome-wide linkage analysis and evidence of gene-by-gene interactions in a sample of 362 multiplex Parkinson disease families." Hum Mol Genet **12**(20): 2599-2608.

Patel, P. J., T. H. Beaty, I. Ruczinski, J. C. Murray, M. L. Marazita, R. G. Munger, J. B. Hetmanski, T. Wu, T. Murray, M. Rose, R. J. Redett, S. C. Jin, R. T. Lie, Y. H. Wu-Chou, H. Wang, X. Ye, V. Yeow, S. Chong, S. H. Jee, B. Shi and A. F. Scott (2013). "X-linked markers in the Duchenne muscular dystrophy gene associated with oral clefts." Eur J Oral Sci **121**(2): 63-68.

Piton, A., J. Gauthier, F. F. Hamdan, R. G. Lafreniere, Y. Yang, E. Henrion, S. Laurent, A. Noreau, P. Thibodeau, L. Karemera, D. Spiegelman, F. Kuku, J. Duguay, L. Destroismaisons, P. Jolivet, M. Cote, K. Lachapelle, O. Diallo, A. Raymond, C. Marineau, N. Champagne, L. Xiong, C. Gaspar, J. B. Riviere, J. Tarabeux, P. Cossette, M. O. Krebs, J. L. Rapoport, A. Addington, L. E. Delisi, L. Mottron, R. Joober, E. Fombonne, P. Drapeau and G. A. Rouleau (2011). "Systematic resequencing of X-chromosome synaptic genes in autism spectrum disorder and schizophrenia." Mol Psychiatry **16**(8): 867-880.

R Development Core Team (2013). R: A language and enviroment for statistical computing. Vienna, Austria, R Foundation for Statistical Computing.

Rabinowitz, D. and N. Laird (2000). "A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information." Hum Hered **50**(4): 211-223.

Rampersaud, E., R. W. Morris, C. R. Weinberg, M. C. Speer and E. R. Martin (2007). "Power calculations for likelihood ratio tests for offspring genotype risks, maternal effects, and parent-of-origin (POO) effects in the presence of missing parental genotypes when unaffected siblings are available." Genet Epidemiol **31**(1): 18-30.

Schaid, D. J. and S. S. Sommer (1994). "Comparison of statistics for candidate-gene association studies using cases and parents." Am J Hum Genet **55**(2): 402-409.

Scott, W. K., M. A. Nance, R. L. Watts, J. P. Hubble, W. C. Koller, K. Lyons, R. Pahwa, M. B. Stern, A. Colcher, B. C. Hiner, J. Jankovic, W. G. Ondo, F. H. Allen, Jr., C. G. Goetz, G. W. Small, D. Masterman, F. Mastaglia, N. G. Laing, J. M. Stajich, B. Slotterbeck, M. W. Booze, R. C. Ribble, E. Rampersaud, S. G. West, R. A. Gibson, L. T. Middleton, A. D. Roses, J. L. Haines, B. L. Scott, J. M. Vance and M. A. Pericak-Vance (2001). "Complete genomic screen in Parkinson disease: evidence for multiple genes." JAMA **286**(18): 2239-2244.

Shao, Y., C. M. Wolpert, K. L. Raiford, M. M. Menold, S. L. Donnelly, S. A. Ravan, M. P. Bass, C. McClain, L. von Wendt, J. M. Vance, R. H. Abramson, H. H. Wright, A. Ashley-Koch, J. R. Gilbert, R. G. DeLong, M. L. Cuccaro and M. A. Pericak-Vance (2002).

"Genomic screen and follow-up analysis for autistic disorder." Am J Med Genet **114**(1): 99-105.

Shi, M., D. M. Umbach, S. H. Vermeulen and C. R. Weinberg (2008). "Making the most of case-mother/control-mother studies." Am J Epidemiol **168**(5): 541-547.

Shi, M., D. M. Umbach and C. R. Weinberg (2007). "Identification of risk-related haplotypes with the use of multiple SNPs from nuclear families." Am J Hum Genet **81**(1): 53-66.

Shi, M., D. M. Umbach and C. R. Weinberg (2009). "Using case-parent triads to estimate relative risks associated with a candidate haplotype." Ann Hum Genet **73**(Pt 3): 346-359.

Sinsheimer, J. S., J. Blangero and K. Lange (2000). "Gamete-competition models." Am J Hum Genet **66**(3): 1168-1172.

Sinsheimer, J. S., C. A. McKenzie, B. Keavney and K. Lange (2001). "SNPs and snails and puppy dogs' tails: analysis of SNP haplotype data using the gamete competition model." Ann Hum Genet **65**(Pt 5): 483-490.

Sinsheimer, J. S., C. G. Palmer and J. A. Woodward (2003). "Detecting genotype combinations that increase risk for disease: maternal-fetal genotype incompatibility test." Genet Epidemiol **24**(1): 1-13.

Sivertsen, A., A. J. Wilcox, R. Skjaerven, H. A. Vindenes, F. Abyholm, E. Harville and R. T. Lie (2008). "Familial risk of oral clefts by morphological type and severity: population based cohort study of first degree relatives." BMJ **336**(7641): 432-434.

Spielman, R. S. and W. J. Ewens (1996). "The TDT and other family-based tests for linkage disequilibrium and association." Am J Hum Genet **59**(5): 983-989.

Spielman, R. S. and W. J. Ewens (1998). "A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test." Am J Hum Genet **62**(2): 450-458.

Spielman, R. S., R. E. McGinnis and W. J. Ewens (1993). "Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM)." Am J Hum Genet **52**(3): 506-516.

Twigg, S. R., R. Kan, C. Babbs, E. G. Bochukova, S. P. Robertson, S. A. Wall, G. M. Morriss-Kay and A. O. Wilkie (2004). "Mutations of ephrin-B1 (EFNB1), a marker of tissue boundary formation, cause craniofrontonasal syndrome." Proc Natl Acad Sci U S A **101**(23): 8652-8657.

Vincent, J. B., G. Melmer, P. F. Bolton, S. Hodgkinson, D. Holmes, D. Curtis and H. M. Gurling (2005). "Genetic linkage analysis of the X chromosome in autism, with emphasis on the fragile X region." Psychiatr Genet **15**(2): 83-90.

Weinberg, C. R. (1999). "Allowing for missing parents in genetic studies of case-parent triads." Am J Hum Genet **64**(4): 1186-1193.

Weinberg, C. R., A. J. Wilcox and R. T. Lie (1998). "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." Am J Hum Genet **62**(4): 969-978.

Weymouth, K. S., S. H. Blanton, M. J. Bamshad, A. E. Beck, C. Alvarez, S. Richards, C. A. Gurnett, M. B. Dobbs, D. Barnes, L. E. Mitchell and J. T. Hecht (2011). "Variants in genes that encode muscle contractile proteins influence risk for isolated clubfoot." Am J Med Genet A **155A**(9): 2170-2179.

Wieland, I., S. Jakubiczka, P. Muschke, M. Cohen, H. Thiele, K. L. Gerlach, R. H. Adams and P. Wieacker (2004). "Mutations of the ephrin-B1 gene cause craniofrontonasal syndrome." Am J Hum Genet **74**(6): 1209-1215.

Wilcox, A. J., C. R. Weinberg and R. T. Lie (1998). "Distinguishing the effects of maternal and offspring genes through studies of "case-parent triads"." Am J Epidemiol **148**(9): 893-901.

Wise, A. L., L. Gyi and T. A. Manolio (2013). "eXclusion: toward integrating the X chromosome in genome-wide association analyses." Am J Hum Genet **92**(5): 643-647.

Wise, A. S., M. Shi and C. R. Weinberg (2015). "Learning about the X from our parents." Front Genet **6**: 15.

Wyszynski, D. F., D. L. Duffy and T. H. Beaty (1997). "Maternal cigarette smoking and oral clefts: a meta-analysis." Cleft Palate Craniofac J **34**(3): 206-210.

Zaykin, D. V. (2011). "Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis." J Evol Biol **24**(8): 1836-1841.

Zhang, L., E. R. Martin, R. H. Chung, Y. J. Li and R. W. Morris (2008). "X-LRT: a likelihood approach to estimate genetic risks and test association with X-linked markers using a case-parents design." Genet Epidemiol **32**(4): 370-380.