Spring 2017

Family-based association studies of autism in boys via facial-feature clusters

Luke Andrew Settles

Department:

FAMILY-BASED ASSOCIATION STUDIES OF AUTISM IN BOYS VIA FACIAL-FEATURE CLUSTERS

by

LUKE ANDREW SETTLES

A THESIS

Presented to the Graduate Faculty of the

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

in

APPLIED MATHEMATICS WITH A STATISTICS EMPHASIS

2017

Approved by

Dr. Gayla Olbricht, Advisor
Dr. Robert Paige
Dr. Donald Wunsch

ABSTRACT

Autism spectrum disorder (ASD) refers to a set of developmental disorders with varied attributes. Due to its substantial heterogeneity in terms of behavioral and clinical phenotypes, it is challenging to discern the genetic biomarkers behind ASD, even though the disease is known to be genetic in nature. This serves as a motivation to detect relationships between single nucleotide polymorphisms (SNPs) and a causal autism disease susceptibility locus (DSL) within more homogeneous subgroups. Recently, clinically meaningful subclassifications of ASD have been discovered utilizing facial features of prepubescent boys. Therefore, through the employment of data from 44 prepubertal Caucasian boys with ASD belonging to one of the three facial-feature clusters and their immediate family, we attempt to identify possible genetic markers corresponding to the varying phenotypes of ASD. We utilize tools from family-based association studies for their ability to detect both linkage and association while being most powerful for rare diseases. The transmission disequilibrium test (TDT) and the family-based association test (FBAT) are implemented for the combined ASD and cluster-membership phenotypes; these tests use affected offspring and all offspring, respectively. We also carry out a screening method involving conditional power estimation and a rank-weighting step addressing the multiple testing problem. In each of the analyses conducted, there is not sufficient evidence to conclude that any of the 2828 SNPs included in the study are linked and associated with a DSL corresponding to the phenotype being tested. In order to increase the low statistical power due to small sample sizes, we recommend to recruit additional boys with ASD, determine the facial-feature cluster to which they belong, and genotype the boy and both his parents. There is no need to genotype any unaffected offspring, because their contributions to the test statistic are minor.

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Gayla Olbricht, for her assistance and guidance in the classroom, in research, and professionally. I also wish to extend gratitude to Dr. Tayo Obafemi-Ajayi and Dr. Donald Wunsch for detecting facial-feature clusters of boys with ASD and their curiosity concerning potential underlying genetic factors. I appreciate the assistance of Dr. Obafemi-Ajayi and Cynthia Germeroth in determining and obtaining the relevant genetic data. Additionally, I want to acknowledge Dr. Robert Paige for serving on my committee and sharing his expertise in and out of the classroom. Randy Haffer was invaluable in his assistance with computational methodology and resources. I also appreciate the prompt responses and assistance with the PBAT software from Dr. Matthew McQueen. Additionally, I am grateful for the financial support of the Missouri S&T Chancellor's Fellowship. Finally, I am appreciative of the support my family and friends have provided throughout my studies.

TABLE OF CONTENTS

Page

ABSTRACT iii

ACKNOWLEDGMENTS iv

LIST OF ILLUSTRATIONS viii

LIST OF TABLES ix

SECTION

1. MOTIVATION 1

2. BACKGROUND 4

2.1. BIOLOGICAL BACKGROUND 4

2.1.1. Introduction to Genetic Association Studies 6

2.1.2. Linkage versus Association 8

2.1.3. Modes of Inheritance and Genotype Coding 12

2.2. DATA 14

2.2.1. Genetic Data for Association Studies 14

2.2.2. Quality Control of SNP Data 17

2.2.3. Autism Spectrum Disorder (ASD) Data and Quality Control 19

2.3. STATISTICAL BACKGROUND 22

2.3.1. Chi-Square Test of Independence 22

2.3.2. McNemar's Test 25

2.3.3. The Multiple-Testing Problem 28

3. HYPOTHESIS TESING IN FAMILY-BASED DESIGNS 33

3.1. HYPOTHESES FOR FAMILY-BASED DESIGNS 33

3.2. THE TRANSMISSION DISEQUILIBRIUM TEST (TDT) 35

3.2.1. Theory of the TDT 35

3.2.2. Comparing the Use of the Exact and Approximate Distributions of the TDT 40

3.2.3. Interpreting the Results of the TDT for the ASD Data 46

3.3. THE FAMILY-BASED ASSOCIATION TEST (FBAT) 46

3.3.1. FBAT Theory 46

3.3.2. Obtaining the TDT from the FBAT 53

3.3.3. Extensions of the FBAT 56

3.4. CONDITIONAL POWER FOR FAMILY-BASED ASSOCIATION TESTS 59

4. FAMILY-BASED ASSOCIATION STUDIES FOR FACIAL-FEATURE CLUSTERS AND ASD 62

4.1. THE CLUSTER MEMBERSHIP PHENOTYPES 62

4.2. WITHIN-CLUSTER ANALYSIS VIA THE EXACT TDT 64

4.3. INCORPORATION OF UNAFFECTED OFFSPRING WITH THE C2MP . 66

4.3.1. Motivation for Focusing on the C2MP 66

4.3.2. Phenotypic Assumptions 67

4.3.3. Test-Statistic Considerations 68

4.3.4. Results from Incorporating Unaffected Offspring 69

5. SCREENING FOR FAMILY-BASED DESIGNS 71

5.1. CONDITIONAL POWER ESTIMATION WHERE ALL PROBANDS ARE AFFECTED 72

5.1.1. Setting-Specific Test Statistic 72

5.1.2. Estimating the Conditional Marker Density for the Conditional Power 73

5.2. WEIGHTING THE SIGNIFICANCE LEVEL 83

5.3. SCREENING FOR THE WITHIN-CLUSTER ANALYSES VIA THE EXACT TDT 87

5.4. SCREENING FOR THE C2MP FOCUS VIA THE FBAT 91

5.5. EXAMINING THE CONDITIONAL POWER ESTIMATES 91

6. DISCUSSION AND CONCLUSIONS 97

6.1. SUMMARY 97

6.2. FUTURE WORK 98

BIBLIOGRAPHY 102

VITA 106

LIST OF ILLUSTRATIONS

Figure		Page
1.1	Example Facial Features	3
2.1	Single Nucleotide Polymorphism (SNP) and Alleles	5
3.1	Histogram of the Number of Heterozygous Parents.	41
3.2	Exact vs. Approximate *p*-Values.	42
3.3	Change in *p*-Value vs. the Number of Heterozygous Parents.	43
3.4	Histograms of Approximate and Exact *p*-Values for the TDT.	45
5.1	Histograms of Conditional Power Estimates.	94
5.2	Conditional Power Estimate vs. Number of Informative Families (NIF).	96

LIST OF TABLES

Table		Page
2.1	Haplotype Frequencies under Linkage Disequilibrium.	11
2.2	Example PED File Layout.	16
2.3	Example MAP File Layout.	17
2.4	Cluster Membership Count Summary	19
2.5	Summary of PLINK Commands for Quality Control.	21
2.6	Summary of Quality Control Results.	22
2.7	A General Contingency Table.	23
2.8	The 2×2 Contingency Table.	26
2.9	Outcomes of Testing m Hypotheses.	30
3.1	The General Contingency Table for the TDT.	37
3.2	Example Set of Families for the TDT.	38
3.3	Classifying Parental Contributions for the Contingency Table for the TDT.	38
3.4	The Contingency Table for the TDT Example.	39
3.5	SNPs Closest to Significance via Exact TDT without Considering Facial Clusters.	47
3.6	Coding the Genotype based on the Mode of Inheritance.	49
3.7	Contributions to the FBAT Test Statistic with Additive Mode of Inheritance.	54
4.1	Summary of Phenotypes.	63
4.2	SNPs Closest to Significance via the Exact TDT within Facial Clusters.	65
4.3	SNPs Closest to Significance via the FBAT with Unaffected Offspring for the ASDP$*$C2MP.	70
5.1	Genetic Probabilities for Parental Mating Types Under the Additive Mode of Inheritance.	79

5.2 Probabilities of Parental Mating Types Under Random Mating and HWE. 80

5.3 Partition Details for Exponential Weighting with $k_1 = 5$, $r = 2$, and $\alpha = 0.05$. .. 88

5.4 Screening Results for Within-Cluster Analyses via the Exact TDT. 89

5.5 Screening Results for the C2MP Focus..... 92

1. MOTIVATION

Autism spectrum disorder (ASD) refers to a set of developmental disorders with heterogeneous attributes, and it may be somewhat impairing to severely disabling. Common characteristics of those with ASD include: repetitive behaviors, social problems such as struggling to interact and communicate with others, and limited activities or interests. About one in 68 children are diagnosed with ASD with diagnosis being more likely among boys than girls. An additional risk factor is having parents with advanced age ("a mother who was 35 or older, and/or a father who was 40 or older when the baby was born") [1]. In fact, ASD is fundamentally a genetic disorder, because of a multitude of evidence of genetic factors and almost none of environmental factors [2, p. 379]. We are interested in identifying genetic biomarkers leading to this complex disease.

Humans share around 99.5% of their genetic information, so only a small fraction of our genetic material accounts for the tremendous variability between individuals [3]. One must search these differences, which commonly occur in the form of single nucleotide polymorphisms (SNPs), to identify genetic variants to which a disease may be attributed. The idea is that if individuals afflicted by a particular disease all possess a specific genetic sequence and those without the trait do not, then there is a relation between that genetic information and the disease. Statistically, the goal is to demonstrate an association between a genetic marker and a disease trait. This association may result from a genotyped SNP that is directly associated with the trait or one that is indirectly associated with the trait via association with an unknown disease susceptibility locus (DSL). Studies that investigate these connections are referred to as genetic association studies [4].

Due to the fact that there is substantial heterogeneity in terms of behavioral and clinical phenotypes of those with ASD, it has been challenging to discern the genetic markers behind and etiology of the disease [5, p. 2]. This serves as a motivation to detect linkage and association between SNPs and an autism DSL within more homogeneous subgroups as opposed to ASD as a whole. Utilization of the facial morphology to identify genetic markers is natural for ASD. The shape of one's face is associated with brain growth and development, and there are "anatomic abonormalit[ies] in the autistic brain" [6, 7]. Some suggest that the spectrum of ASD phenotypes correspond to changes in embryonic developmental patterns. These changes in brain development could therefore manifest itself in a person's facial structure. Thus, the same genes controlling the neurological characteristic of ASD may also be involved in facial morphology [6]. For instance, it has been hypothesized that "common autism-causing genes affect early brain development and simultaneously the facial phenotypes" [5, p. 2]. Furthermore, there has been progress in the identifying clinically meaningful subgroups of ASD using facial features, where the subgroups possess statistically significant differences in behavioral and clinical characteristics.

Aldridge et al. were the first to use facial morphology to determine clinical subgroups of autism. They utilized Caucasian males aged eight to twelve years old with ASD and employed a Euclidean distance matrix and principal components analysis to determine two groupings based on the subjects' facial features. One of these facially-determined groups exhibited a more severe form of ASD, e.g. increased language regression, lower IQs, and higher autism severity scores [5]. More recently, Obafemi-Ajayi et al. considered an overlapping sample of Caucasian boys; 52 from the Aldridge et al. study plus ten new boys were studied [5, 8]. Via geodesic facial distances and cluster analysis, they discovered three subgroupings; see Figure 1.1 for an example of the distances utilized. Importantly, there was a subgroup common to both investigations "characterized by lower IQs and

Vineland Adaptive behavior scores, severe autism symptoms measured by gold standard autism diagnostic measures (ADI-R and ADOS), and more than twice likelihood of early language regression" [8, p. 15]. Additionally, the authors postulated that yet another reason for a possible (epi)genetic underpinning for this subgroup is that relatively more severe measurements of subclincial autistic traits in the mothers were present [8, p. 10].

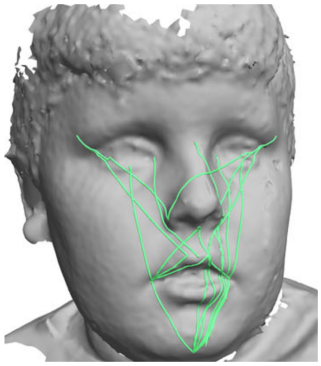


Figure 1.1. Example Facial Features. Obafemi-Ajayi et al. considered three-dimensional facial surface distances to determine facial-feature clusters [8].

Hence, through the established facially-delineated clusters, we hope to determine possible genetic markers corresponding to their varied ASD phenotypes. In order to accomplish this objective, we plan to use tools from genetic association studies, in particular those from family-based association analyses which will utilize the information from the nuclear families, on genetic data from some of the subjects in the studies discussed above.

2. BACKGROUND

In this chapter, we will discuss the biological, data, and statistical background relevant to the investigation of ASD and facial-feature cluster membership via genetic association studies.

2.1. BIOLOGICAL BACKGROUND

In this section, we introduce key biological concepts that are needed to study genetic association. First, a location on a chromosome that can have two or more possible variants is referred to as a polymorphic genetic locus, or more simply a polymorphism. Alleles are those disparate variants at the locus. A gene or specific locus on a gene that possesses variants associated with a disease is named a disease susceptibility locus (DSL); this is what researchers are interested in finding [9]. Typically, the relationship is assumed to be causal [9, p. 20].

It is necessary to distinguish genetic data between individuals; loci that enable this are called genetic markers. A single nucleotide polymorphism (SNP) is the most fundamental type of genetic marker and involves a variation in only a single base pair. The difference between an allele and a SNP merits additional discussion. An allele refers to a particular variant of a gene segment, whereas a SNP refers to the change in a single base pair. Therefore, an allele is determined by the SNP(s) in that section of the gene [9]. Refer to Figure 2.1. Usually, the allele is referred to by its characteristic base pair. From here on out, the term marker and SNP will be used interchangeably, because that is the only type of genetic marker we consider.

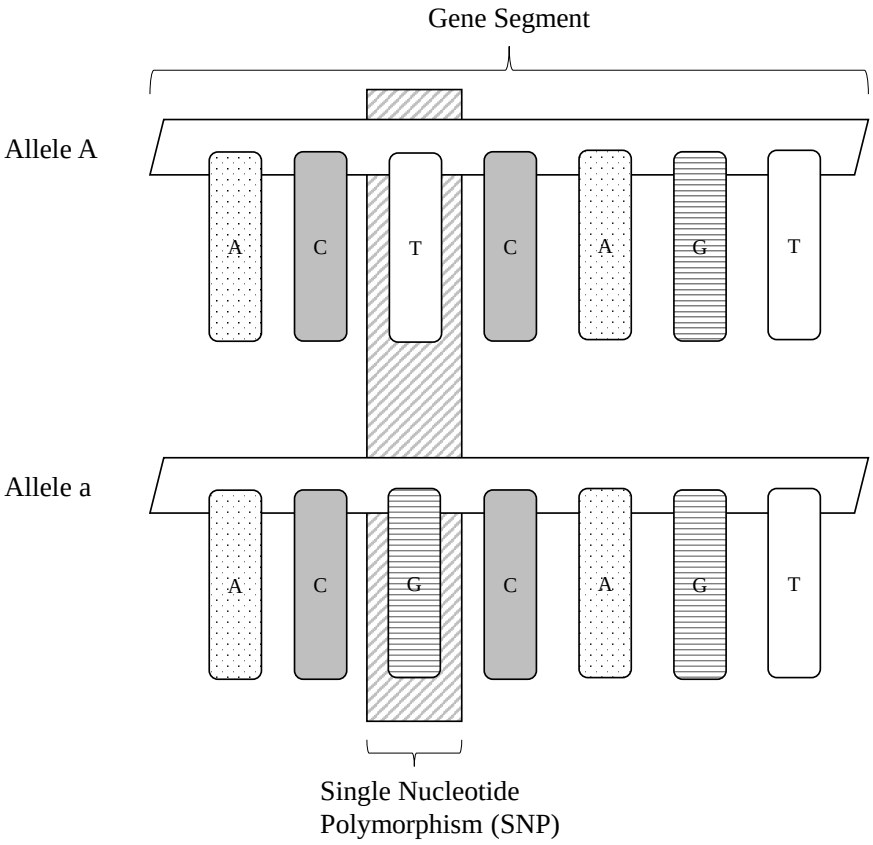


Figure 2.1. Single Nucleotide Polymorphism (SNP) and Alleles. A single nucleotide polymorphism (SNP) at the third base pair on a pair autosomal chromosomes is shown. It is a SNP, because there is variation in the nucleotide present at that location. The two variants of the gene segment are different alleles, denoted with A and a.

Various methods exist to identify genetic markers. In particular, this paper will focus on genotyping SNPs, which is addressed in Section 2.2. Note that there is a difference between genotyping and sequencing. Genotyping solely involves identifying the genetic variants that a particular person has; it only examines SNPs that are spread out over a gene or genome. Contrastingly, sequencing is concerned with ascertaining the exact base pairs in a region or over the entire genome. SNPs can be pinpointed from sequencing as well, but information on all base pairs, including those in regions that are common to all individuals, is also obtained [10].

One genetic principle, the Hardy-Weinberg equilibrium (HWE), is utilized regularly in statistical applications, because it can considerably simplify the theory and methods needed. It states that the allele frequency defines the distribution of genotypes, and the distribution does not change from generation to generation. There are numerous assumptions required for HWE to hold, but the principle approximates the genotype's distribution well even if some do not hold exactly. Formally, HWE exists when the population satisfies

$$P(\text{AA genotype}) = p^2$$
$$P(\text{Aa genotype}) = 2p(1-p)$$
$$P(\text{aa genotype}) = (1-p)^2$$

where p is the frequency of allele A [9, p. 36]. We will take advantage of the theoretical simplification that accompanies this principle in Chapter 5.

2.1.1. Introduction to Genetic Association Studies. The two overarching types of association studies are dictated by the subjects utilized. These types are often referred to as designs. The first, and more prevalent, design uses unrelated individuals and is referred to as a population or case-control design; some subjects have the disease of interest (cases) while

others do not (controls). Alternatively, instead of using unrelated individuals, family-based designs make use of families of diseased subjects. Population designs are more popular, because they typically require fewer resources to recruit and genotype the subjects. Despite this popularity disparity, family-based designs possess numerous advantages and are the focus of this work [11].

In regards to the biology, family-based designs are not affected by population substructure [11]. One type of population substructure, population admixture, involves individuals who "have a mixture of different genetic ancestries," which is quite realistic. This results in the likelihood of an individual possessing a particular allele to depend on their ancestral mixture when allele frequencies differ in the ancestral populations [9, p. 34]. When this fact is unaccounted for, association tests may result in a higher than expected number of false positives at a given significance level [9, p. 126]. In family-based studies, information about the genotypes of the parents and/or siblings, and the genetic background will not differ within the families. Therefore, they are robust to population substructure.

Also, in family-based designs, any declaration of significance indicates that both association and linkage are present, which avoids many misleading associations [11]. Linkage and association will be discussed in Section 2.1.2. Furthermore, these studies afford resolutions to the multiple-testing problem via some screening procedures [11]. The price paid for these advantages and robustness manifests itself in the reduced power of the test, in particular for common diseases; the case-control study has higher power in this setting. Nonetheless, the tests for family-based designs involving trios, the affected offspring (proband) and both parents, have power closest to that of population designs for common diseases among two other subtypes of family-based designs. In fact, trio designs are more powerful than case-control studies for rare diseases. Another disadvantage is that it is typically more difficult to recruit participates for family-based designs. For instance, for late-onset diseases like

Alzheimer's disease, the parents' genotypes may be unretrievable [9]. However, this is not the case for childhood diseases or developmental disorders, like ASD, where the parents are typically directly involved in the child's health care. Nonetheless, in general, it is more challenging to obtain large sample sizes with family-based designs [9].

Apart from the test statistics needed to establish association, genotyping quality control and the multiple-testing problem are two other statistical concerns. There exist assorted filtering methods to ensure the genotype data is of desirable quality; if these are not completed, the bias and power of the statistical test may be affected. The multiple-testing problem refers to the fact that conducting hypothesis tests of association for thousands of genetic markers (SNPs) will result in a higher experiment-wise Type I error rate than stated if additional controls are not implemented. Both of these concerns will be examined in more detail later.

2.1.2. Linkage versus Association. In order to find the location of disease genes in the genome, techniques of gene mapping are employed. The goal of gene mapping is to identify disease susceptibility loci (DSL) that are near genetic markers, which are simply sequences of DNA whose physical chromosomal location is known, chosen for the analysis. There two types of analyses that are of utmost importance for these types of studies: linkage analysis and association mapping. Linkage analysis is based on the study of relatives where "the genetic signal is detectable with markers even relatively far away from the disease gene." Its advantage is that it is more likely to require only a small number of genetic marker loci to cover the entire area under consideration; however, this large distance between markers and a DSL results in an imprecise estimate of the location of the DSL. On the other hand, association mapping is based on unrelated subjects and "genetic signal can be observed with marker data only in very close proximity to the disease gene." Even though this fact will necessitate the use of a larger quantity of genetic markers within

an equivalent genomic region of interest, this approach will yield a "more precise estimate of the genetic location of the DSL" [9].

Recombination events are the reason for the disparity in the detection range between linkage and association analyses [9]. Recombination is defined as any "production of new allele combinations," and is due to one of two processes: independent assortment or crossing-over [12, p. 121]. When alleles reside on different chromosomes, recombinants, the product of recombination, are due to independent assortment under Mendel's Second Law (the principle of independent assortment). Alleles present on the same chromosome, especially when they are located near each other, will tend to be inherited together which will not result in recombinants. The second process is crossing-over, which is where homologous chromosome pairs exchange parts during meiosis. When this occurs, the alleles will not be inherited together which allows recombination to take place [12].

The recombination fraction, or recombination frequency, (θ) ranges from 0 to 0.5 and represents the probability the recombination occurs between the two loci under consideration. One way to see the upper bound for this value is through Mather's Law. If P_0 is the probability of no crossovers occurring between two loci, then according to Mather's Law

$$\theta = \frac{1 - P_0}{2}.$$

Note that the farther two loci are from each other, the more likely a crossover may occur between them, which results in θ near 0.5. We may expect this upper bound to be attained when the two loci are on different chromosomes, because they would be governed by the principle of independent assortment [9].

Linkage between two loci is present when the recombination frequency is less than one-half, $\theta < 0.5$. This corresponds to a violation of Mendel's Second Law and implies that

recombination is due to crossing-over. Therefore, hypothesis testing for linkage involves a null hypothesis of no linkage and an alternative of the presence of linkage, i.e.

$$H_0 : \theta = \frac{1}{2} \text{ (recombination fraction is 0.5; no linkage)}$$
$$H_A : \theta < \frac{1}{2} \text{ (recombination fraction is less than 0.5; linkage present).}$$

Detecting linkage can be somewhat easy; e.g. it is possible to do so within a single family. However, the detection does not readily translate into precise knowledge concerning the DSL location [9]. Essentially, linkage tells us that the SNP and DSL are close enough that recombination occurs due to crossing-over and not independent assortment.

Before discussing association mapping in more detail, we introduce the fundamental underlying concept of linkage (dis)equilibrium. Consider the alleles A, a and B, b at two markers, and denote the relative frequency of the alleles at each marker to be P_A, P_a, P_B, and P_b. A haplotype is a "set of alleles at different loci that are present together on the same chromosome" and tend to be inherited together [11, p. 389]. In this setup, there are four haplotype combinations, each with relative frequencies P_{AB}, P_{Ab}, P_{aB}, and P_{ab}. For example, P_{aB} can be viewed as the probability that a haplotype randomly taken from the population at the loci will have alleles a and B. Under linkage equilibrium, an absence of association between alleles at the two loci, the relative frequencies behave like probabilities of independent events, e.g. $P_{aB} = P_a P_B$. In words, the haplotype relative frequency is simply the product of the individual allele frequencies at the loci. When the frequencies are displayed in a 2×2 table this coincides with independence as tested with a chi-square statistic from Section 2.3.1 [9].

When linkage equilibrium does not hold, we have linkage disequilibrium (LD). In this case, for example, $P_{aB} \neq P_a P_B$, and the amount by which these differ is of interest. The

LD coefficient does just that and is defined using the uppercase alleles,

$$\delta = P_{AB} - P_A P_B. \tag{2.1}$$

The 2×2 table of relative frequencies can seen in Table 2.1 When $\delta = 0$, the table represents the population allele frequencies under linkage equilibrium. One detail that cannot be overlooked is that P_{AB} must be nonnegative, which leads to defining bounds on δ. Another feature to recognize is the fact that the LD coefficient, δ, depends highly on the allele relative frequencies. Hence, there are issues concluding independence from $\delta = 0$, because extremely small values for the allele relative frequencies can result in values of δ that are nearly zero even when linkage disequilibrium is present. Due to this undesirable quality, alternative methods of quantifying LD exist [9].

Table 2.1. Haplotype frequencies under linkage disequilibrium (LD).

		B Locus		
	A Locus	B	b	Row Total
	A	$p_{AB} = p_A p_B + \delta$	$p_{Ab} = p_A p_b - \delta$	p_A
	a	$p_{aB} = p_a p_B - \delta$	$p_{ab} = p_a p_b + \delta$	p_a
Column Total		p_B	p_b	

With this information we can discuss association mapping. Despite the seemingly contradictory naming, linkage disequilibrium is at the heart of association mapping, not linkage analysis. When LD is present between a DSL and the marker, a relationship between the disease and the marker is anticipated. On the other hand, when LD is absent, we have independence of the loci, and therefore, we expect no association between the disease and the marker [9].

One would like the detection of association to imply that the SNP under consideration is actually associated with the disease. However, this is only the case when association is due to both LD and linkage. Note that one cannot use population association of a marker and DSL to infer linkage, because association may occur even when linkage is not present. In fact, association could be due to one of four different characteristics: population stratification, admixture, "confounding by extraneous variables," or linkage and LD ("lack of recombination between the two loci") [9, 11, 13]. Therefore, there are three ways association could occur in the absence of linkage; in each case, the discovered association would not actually imply association between the SNP and disease via the DSL. In other words, association without linkage is misleading.

2.1.3. Modes of Inheritance and Genotype Coding. Association studies aim to determine the presence of a relationship between a genetic sequence and a disease/trait. Nonetheless, the relationship itself will not be deterministic. Therefore, specification of a genetic model, which describes the probabilistic relationship between a genotype and phenotype, is necessary. Typically, the trait under consideration is dichotomous and can be coded as

$$Y = \begin{cases} 1 & \text{if the individual has the disease (affected)} \\ 0 & \text{if the individual does not have the disease (unaffected).} \end{cases}$$

The genotypes are intrinsically categorical in nature and are prescribed by the allele combination at the particular locus; we will use G to represent the genotype. Note that with Mendel's notation, alleles "A" and "a" denote the dominant and recessive alleles, respectively. However, we are yet to designate a relationship between alleles, and in practice this is unknown for most loci included in association studies. With this in mind, we will introduce a different representation; "A" and "a" will refer to the less frequent (minor) allele

and more frequent (normal) allele, respectively. Recall that the unknown DSL is assumed to have "a direct effect on the phenotype through some biological mechanism" [9, p. 20]. In other words, one assumes that the DSL has a causal relationship with the phenotype under consideration. In order to distinguish between the causal DSL and other loci, we will denote the alleles of the DSL with "D" and "d."

The locus of interest has some probabilistic effect on the phenotype Y which is conditional on the genotype G. The penetrance function describes this effect; for discrete phenotypes, it is a set of conditional probabilities that model the phenotypic distribution. For example, suppose the phenotype is not affected by a locus. In this case, the penetrance probabilities are identical; symbolically, $P(Y|G = DD) = P(Y|G = Dd) = P(Y|G = dd)$. In general, if the phenotype is dichotomous as above, then for each genotype G, $P(Y = 1|G) + P(Y = 0|G) = 1$ [9].

The dependency mechanism between the quantity of disease alleles and the distribution of Y is described by the mode of inheritance. There are four frequently employed modes of inheritance [9]:

1. Codominant: $P(Y = 1|G = DD) \neq P(Y = 1|G = Dd) \neq P(Y = 1|G = dd)$; there is no assumed relationship between these probabilities beyond inequality.
2. Dominant: $P(Y = 1|G = DD) = P(Y = 1|G = Dd)$; the presence of single allele, D, is enough to affect the disease.
3. Recessive: Two copies of the disease allele, D, are needed to affect disease risk.
4. Additive: The increase in disease risk is the same for each additional disease allele, D, present; this can be measured on two scales.
   (a) Linear scale: $P(Y = 1|G = Dd) = 0.5[P(Y = 1|G = DD) + P(Y = 1|G = dd)]$
   (b) Log (multiplicative) scale:
   $$P(Y = 1|G = Dd) = \sqrt{P(Y = 1|G = DD)P(Y = 1|G = dd)}$$

For each of these modes of inheritance, there is a distinct test for association; see Sections 7.1-7.2 of [9] for more details. Since the true genetic model is typically unknown, selecting which mode of inheritance to use can be troublesome. Tests for the codominant and additive model are the most popular among practitioners, because no matter the true underlying model, they are correlated and perform nearly as well as that for the appropriate model. However, it is necessary to use the recessive or codominant test if the recessive model is of interest [9]. From the perspective of statistical power, the additive model is decent no matter the true genetic mode of inheritance [14, p. 4]. With this as a motivation, we will utilized the additive model in this work.

Despite puissant advances in the field of genetic association studies over the past decades, there are still numerous challenges that remain to be addressed. One major issue is that the entire genetic variation in the human genome cannot solely be described by SNPs whose allele frequencies are larger than 5%. Rare variants are needed to attain the complete picture of heritability, but the common SNPs utilized in association studies are inadequate placeholders for these. The inability to incorporate all but basic covariates in the models for a particular phenotype is another limitation of the current methods; there are many complex factors that cannot be included. More work must be done to account for these deficiencies along with other issues not mentioned here [9].

2.2. DATA

First, we discuss the collection and format of the genetic data for association studies. Then the procedures for quality control are motivated. Finally, we introduce the ASD genetic data to used for the analyses in this work.

2.2.1. Genetic Data for Association Studies. The data needed for genetic association studies involves genotyped SNPs. The specific type of association study being

performed dictates the selection method and quantity of SNPs to be used; this an essential step to setup the study for success. For instance, if certain genes are known to have functions that could be related to the biology of the disease being considered, then the markers (SNPs) would be selected throughout that particular region. Another possible context is a follow up to a previous study that suggested genetic linkage between certain SNPs and the DSL; in that case, the previously identified SNPs are utilized. The final common circumstance is a genome-wide association study (GWAS). This involves genotyping and testing hundreds of thousands of SNPs across the entire genome for association with the disease of interest. For GWAS, one must select SNPs such that they are "sufficiently correlated, i.e. in strong linkage disequilibrium, with SNPs that will not be genotyped in the same region." The HapMap project is dedicated to cataloging linkage disequilibrium of the human genome in order to aid the selection of appropriate SNPs; this has played a prominent role in reducing the cost and time of genotyping SNPs [9].

Once the particular SNPs have been selected, the genotyping is carried out using SNP-chips. These chips are capable of genotyping hundreds of thousands of SNPs across the genome at the same time [9]. First, a DNA sample is extracted from the subject and divided into many unique strands. Each strand will contain a SNP of interest, and the SNP-chip array will contain probes, which are possible nucleotides of the SNP (C/G and A/T), that can match to the sample strands [15]. The extracted DNA segment is considerably amplified so that the automated genotyping process can identify each SNP from the intensity of its two alleles. From there, the observed data points are divided into three clusters, one for each possible allele combination (e.g. AA, Aa, aa). Those that are homozygous should have an intensity reading near zero for the allele they do not possess [9].

The genotyped SNP data is stored in a PED(igree) file where each row corresponds to a different individual; see Table 2.2 for an example of this format. The first six columns

contain the following ordered identifying information: family/pedigree ID (pID), individual ID (iID), father ID (fID), mother ID (mID), sex, and affection/phenotype status (aff). Note that an additional phenotype file is needed when working with disease traits that are not dichotomous, i.e. when one is interested in something other than affection status. When the affection status is of interest, the phenotype is coded 2, 1, and 0 for affected, unaffected, and unknown, respectively. Sex is coded 0, 1, and 2 for unknown, male, and female, respectively. If an individual has parents included in the study, then the cells in the father ID and mother ID columns for that row will contain the identifier corresponding to those individuals within the family ID [14].

Table 2.2. Example PED File Layout. This portion of a PED file shows the data format for two families and two bi-allelic markers (SNPs) called rs1 and rs2; their two alleles are denoted with A1 and A2. The genotype coding here uses the nucleotide characters.

pID	iID	fID	mID	sex	aff	rs1_A1	rs1_A2	rs2_A1	rs2_A2
11067	1	2	3	1	2	A	A	A	G
11067	2	0	0	1	1	A	A	A	G
11067	3	0	0	2	1	A	C	A	A
11465	1	2	3	1	2	A	C	A	A
11465	2	0	0	1	1	A	C	A	A
11465	3	0	0	2	1	C	C	A	A

All columns beyond the sixth one correspond to alleles of certain markers. Some software, such as PLINK, only allow these markers to be bi-allelic [16]. In this case, two adjacent columns will represent the two alleles of a given marker. The coding of the genotypes may be the nucleotide character or a numeric value, i.e A, C, G, T or 1, 2, 3, 4. Missing genotypes are coded as 0, and it should not be the case that only one allele is missing for a given individual [14, 16].

Table 2.3. Example MAP File Layout. This portion of a MAP file shows the marker name, the chromosome in which it is located, and its genetic position. These three columns must also be present MAP file, but others may be included.

Chromsome	Marker	Position
4	rs10003143	114272159
4	rs10007543	114196280
4	rs10013743	114498871
4	rs10015472	114180410

A MAP file may be used in addition to the PED file; refer to Table 2.3 for an example. The purpose of the MAP file is to relate the markers to a specific location on the genome; these are especially helpful when considering SNPs across the entire genome. Each row corresponds to a single SNP and the number of columns vary based on the software being utilized. At a minimum, columns for the marker name, the chromosome on which it is found, and the genetic/base-pair position (in bp units) are included. It is common, but not required, for the files to also contain the genetic position (in morgans). One must ensure that the number and names of markers in the PED file match those contained in the MAP file [14, 16]

2.2.2. Quality Control of SNP Data. As with any process, every step is prone to error. For instance, the DNA quality control and handling methods can affect the quality of the data. Additionally, some errors only pose problems in certain types of studies. For example, random genotyping error for each genotype occurs when the errors are independent of the genotype and phenotype of the subject, but the effect of this error has disparate consequences depending on the study design. The significance level used in testing for association in population-based studies is not affected by this random error. On

the other hand, it will elevate the significance level, leading to a greater number of false-positives, in family-based studies. Therefore, these errors must be handled appropriately. Furthermore, GWAS's contain more errors than smaller scale studies due to the sheer number of SNPs utilized, even if the error rate is small. Including SNPs and/or subjects with many genotyping errors may introduce systematic bias or reduce statistical power. For these and other reasons, it is vital to perform quality filtering of the data before analysis [9].

Various genotype quality control filters can be employed for this type of genetic data. Some are design-specific and others are ubiquitous. The first filter involves removing any subjects for which 2% or more of the SNP genotypes are missing; this would be indicative of a systematic genotyping error for that individual. Due to the fact that genotype clusters are determined by allele intensities, errors become more common when the minor allele, A, frequency is small. Therefore, the scarce homozygous genotype , AA, is troublesome to identify. This leads to the second filter, which states that any SNP with a minor allele frequency smaller than 5% should be excluded from the analysis. The third filter eliminates SNPs that violate the Hardy-Weinberg equilibrium with p-values less than 10^{-5} for the appropriate statistical test; this is only necessary for GWAS [9]. Also, there may be an issue with a particular SNP if many individuals are missing genotype data for that marker. If this missing rate is too high, e.g. larger than 3%, the SNP should be excluded from the analysis [17].

The next filter is specific to family-based designs. It may be the case, that an impossible Mendelian transmission appears in the data (e.g. parents Aa and AA have a child with aa). These types of errors are tallied for each SNP and family. A cutoff value is determined, which can depend on the sample and differ between studies; a conventional threshold is five errors. The SNPs and families that exceed the selected value are removed from the analysis [9, p. 180]. However, even if a SNP or individual does not exceed the

Mendelian error threshold, the offspring genotype is impossible given the parental genotype information. Therefore, the genotypes where Mendelian errors have occurred should be set to missing if the individual or SNP is still included in the analysis.

After filtering has been completed, the overall genotyping quality can be evaluated through various methods. One common method depends on the fact that all but a few SNPs will have no genetic association. Thus, the distribution of p-values should essentially follow that expected under the null hypothesis.

2.2.3. Autism Spectrum Disorder (ASD) Data and Quality Control. The data utilized in this thesis was obtained from the Thompson Center for Autism and Neurodevelopmental Disorders at the University of Missouri in Columbia, MO. It includes SNP genomic data for 44 male Caucasian boys, aged 8 to 12 years old, with ASD, both of their unaffected parents, and any siblings, which happened to all be unaffected. There are a total of 172 individuals for which both alleles of bi-allelic markers were genotyped. The 44 families being considered were grouped into three different clinical meaningful subgroups based on their facial features; they are the subgroup of the 52 considered by Obafemi et al. that had clean genetic data [8]. From here on out, these subgroups will be referred to as clusters. See Table 2.4 for how the 172 individuals are divided among the three clusters.

Table 2.4. Cluster Membership Count Summary. For each cluster, the counts for the number of families whose probands belong to the cluster, the number of siblings without ASD in those families, and the total number of offspring among all such families are displayed. Note that the number of families is equal to the number of probands because of the ascertainment condition.

Cluster	Families	Unaffected Siblings	Total Offspring
1	10	9	19
2	13	14	27
3	21	17	38
Overall	44	40	84

Based on past association studies, the Simons Foundation Autism Research Initiative (SFARI), a research program focused on all aspects of ASD, has identified various genes that seem to be associated with ASD. The two classifications with the most evidence of association are referred to as high confidence (category 1) and strong candidate (category 2). SFARI "considered a rigorous statistical comparison between cases and controls, yielding genome-wide statistical significance, with independent replication to be the strongest possible evidence for a gene. These criteria were relaxed slightly for category 2" [18]. This analysis will consider markers from genes with these classifications, so the focus is much narrower than a genome-wide scale. We aim to establish both association and linkage between markers, if any, and ASD. The markers cover 16 high-confidence genes and 35 strong-candidate genes. The unfiltered data includes 3335 markers where 544 belong to high-confidence genes and 2791 belong to strong-candidate genes as determined by the SFARI as of October 21, 2016 [18].

Quality control is implemented using the software PLINK, which is a "free, open-source whole genome association analysis toolset, designed to perform a range of basic, large-scale analyses in a computationally efficient manner" [16]. It has options available to specify inclusion thresholds for the data. The thresholds and utilized values are shown in Table 2.5; the command to specify the corresponding PLINK option for each threshold is mentioned. Motivation for each filter was provided in Section 2.2.2. We included the Hardy-Weinberg equilibrium threshold in this analysis because of an assumption in a screening method discussed in Chapter 5.

The following order for considered the filters, which follows that of PLINK, was utilized: `--mind` N, `--hwe` N, `--geno` N,`--maf` N, and `--me` N M. Note that PLINK formulates Mendelian-error threshold in terms of percentages (`--me` N M) where it excludes families with more than N% of the SNPs possessing Mendelian errors and excludes

SNPs with more than $M\%$ of trios possessing Mendelian errors [16]. Therefore, those percentages were specified after performing the anterior filters so that the count did not exceed five for a given family or marker. This was also confirmed by outputting a summary of the Mendelian errors.

Table 2.5. Summary of PLINK Commands for Quality Control.

Threshold	PLINK Command	Utilized Value
Missingness per individual	`--mind` N	0.02
Missingness per marker	`--geno` N	0.03
Minor allele frequency	`--maf` N	0.05
Hardy-Weinberg equilibrium	`--hwe` N	0.00001
Mendel error rate	`--me` N M	0.001767 0.113637

The filtering results for the autism data set under consideration are shown in Table 2.6. No marker or family exceeded five Mendelian errors, but the genotypes involved in the erroneous transmissions, both for the parents and offspring, were set to missing. Therefore, we reconsidered the missingness rates in the context of the original number of SNPs and individuals. However, still no individuals needed to be removed. On the other hand, 19 markers were eliminated. Of those, 17 exceeded the threshold before setting Mendelian errors to missing, while the other two markers did so afterwards. As for low MAF, 491 markers were excluded before handling Mendelian errors, and no additional ones violated the filtering criterion after the fact. Note that three of the eliminated SNPs violated the thresholds for both missingness and minor allele frequency. Hence, we can use all 172 individuals to test 2828 markers.

Table 2.6. Summary of Quality Control Results on the 3335 Original Markers for 172 Individuals.

Threshold	SNPs Removed	Individuals Removed
Missingness per individual	N/A	0
Missingness per marker	19	N/A
Minor allele frequency	491	N/A
Hardy-Weinberg equilibrium	0	N/A
Mendel error rate	0	0

2.3. STATISTICAL BACKGROUND

Now, we introduce some statistical concepts that provide the foundation for family-based association studies. First, contingency tables and the chi-square test of independence are explained. Then, we consider a specific contingency table setting that necessitates McNemar's test. Lastly, we discuss the multiple testing problem and its relevance to genetic studies.

2.3.1. Chi-Square Test of Independence. Categorical variables appear frequently in biostatistics, and in many cases, these can be viewed as criteria for classification. Other names for a categorical variable are nominal and polychotomous. When only two categories are possible, it is known as a binary or dichotomous variable. With a simple random sample, one can evaluate the dependence of the criteria on each other, especially when there are only two variables. Thus, it is possible to see if proportions of one variable differ for various levels of the remaining variable [19]. Put more succinctly, one can test if the classification criteria are independent [20]. In a statistical framework, the null hypothesis would be that the two nominal variables are independent. We will discuss a common test for this hypothesis: the Pearson's chi-square test of independence.

In order to do so, we first introduce contingency tables, which are common tabular displays for classification of this type. These tables have r rows and c columns where r and c are the number of levels of the first and second polychotomous variables; note that the order here is arbitrary. Marginal totals and a grand total are also included. A general contingency table is shown in Table 2.7.

Table 2.7. A General Contingency Table.

Classification Criterion 2	Classification Criterion 1					
	1	2	3	$\cdots$	c	Total
1	n_{11}	n_{12}	n_{13}	$\cdots$	n_{1c}	$n_{1\cdot}$
2	n_{21}	n_{22}	n_{23}	$\cdots$	n_{2c}	$n_{2\cdot}$
3	n_{31}	n_{32}	n_{33}	$\cdots$	n_{3c}	$n_{3\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
r	n_{r1}	n_{r2}	n_{r3}	$\cdots$	n_{rc}	$n_{r\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	$\cdots$	$n_{\cdot c}$	n

Events are independent if the probability of their intersection is the product of their individual probabilities. Therefore, assuming independence of classification criteria, one can find the probability that an observation will be classified into a particular cell of the contingency table. From that, the expected number, i.e. frequency, of observations to be classified into a cell can be computed for a given sample size (n). For $j = 1, \ldots, r$ and $k = 1, \ldots c$, define event R_j and C_k to be the events that an observation is counted in j^{th} level of classification for the row criterion and that an observation is counted in k^{th} level of classification for the column criterion, respectively. Then, under independence, we have

$$\widehat{P(R_j \cap C_k)} = \widehat{P(R_j)}\widehat{P(C_k)} = \left(\frac{n_{j\cdot}}{n}\right)\left(\frac{n_{\cdot k}}{n}\right),$$

where the hat (ˆ) signifies that it is an estimator for the true value. With a sample size of n, we expect there to be

$$n\left(\frac{n_{j.}}{n}\right)\left(\frac{n_{.k}}{n}\right) = \frac{n_{j.}n_{.k}}{n} \tag{2.2}$$

observations classified into cell j, k. In words, the expected frequency of cell is the product of the corresponding row and column total frequencies divided by the total sample size.

Now, we want to compare the observed and expected frequencies for each of the $r \cdot c$ cells. If the values are quite similar for all cells, then the assumption of criteria independence is plausible. On the other hand, if the difference between the observed and expected frequencies is sufficiently large we would want to reject the null hypothesis and conclude that the variables are not independent. The method to measure this discrepancy involves scaling the square difference of the observed and expected frequencies by the expected frequency and then taking the sum of those values for all cells. In order to use the notation most commonly seen for the test statistic, we will consider cell 1,1 to be the first cell and cell r, c as the m^{th} cell; ordering will go across an entire row and then move to the next row.

Thus the statistic of interest, called Pearson's chi-square statistic, is given by

$$\chi_P^2 = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i} \tag{2.3}$$

where O_i is the observed frequency of the i^{th} cell and E_i is the expected frequency in the i^{th} cell under the null hypothesis; E_i is equal to the expression in (2.2). Note that when the classification criteria are in fact independent, χ_P^2 is approximately distributed as a χ^2 random variable with $(r-1)(c-1)$ degrees of freedom [20]. For a proof of this fact, refer to Panchenko [21]. As intended, we reject the null hypothesis when the total of all scaled

deviations of the expected and observed frequencies is sufficiently large, i.e. when the p-value of χ^2_P is below the chosen significance level.

Here are some rules of thumb on when the use of the approximate distribution for χ^2_P is inappropriate [20, 22]:

- $n < 20$
- $n = 20, \ldots, 40$ and $E_i < 5$ for at least one i
- $E_i < 1$ for at least one i
- For more than 20% of the cells, $E_i < 5$

Fisher's Exact Test is a better option if any of the above are true, because the test statistic is highly sensitive to small expected frequencies. Others have even more strict criterion; e.g. McDonald recommends to use the exact test if $n < 1000$ [19].

A commonly seen case for the $r \times c$ contingency table is where $r = c = 2$; both classification criteria have only two levels. Other ways to analyze the information contained in this type of table will be considered in Section 2.3.2. For a general table of this form with simplified notation, see Table 2.8. A shortcut to calculating χ^2_P in this case is given by the following formula [20]

$$\chi^2_P = \frac{n(ad - bc)^2}{(a + c)(b + d)(a + d)(c + d)}. \tag{2.4}$$

This shortcut is utilized by many, including Spielman et al., whose test we consider in Section 3.2 [13]. Note that χ^2_P in (2.4) has only one degree of freedom.

2.3.2. McNemar's Test. In certain contexts for the 2×2 contingency table, McNemar's test should be employed in place of the chi-square test of independence. These circumstances involve matched pairs of observations/subjects. The pairs are created by matching according to characteristics that are associated with the response being stud-

Table 2.8. The 2×2 Contingency Table.

Classification Criterion 2	Classification Criterion 1		
	1	2	Total
1	a	b	$a + b$
2	c	d	$c + d$
Total	$a + c$	$b + d$	n

ied [23]. For example, matched pairs occur in case-control studies, comparing new and old treatments, opinions of two experts, and before-and-after data [23, 24, 25]. The same table that is shown in Table 2.8 can be used to organize the data when a dichotomous measurement/outcome is considered. However, the frequencies represent numbers of pairs; hence, n is the total number of pairs considered [23]. Note that in the case of before-and-after data, a "pair" refers to the pair of measurements on a single individual at two different time points [26]. It each of these scenarios, we do not have independence of classification criteria. For example, repeated classification on the same observation/individual would not be independent [25]. We will reason towards McNemar's test in the context of a case-control study.

Consider the row criterion to be the presence (present or not) of a factor among the cases and the column criterion to be the presence of the same factor among the controls. The factor being present will correspond to the first level of classification in Table 2.8. Then the estimated proportion of the controls where the factor is present is

$$\hat{p}_1 = \frac{a + c}{n},$$

and the estimated proportion of the cases with the presence of the factor is

$$\hat{p}_2 = \frac{a+b}{n}.$$

We are interested in the difference of these proportions, i.e

$$\hat{p}_2 - \hat{p}_1 = \frac{b-c}{n}.$$

The null hypothesis of the true proportion having the factor among the cases being the same as that among the controls can be stated in a variety of equivalent ways [23]. One that is commonly used is

$$H_0 : p_1 - p_2 = 0.$$

It can be shown that the standard error of the estimated proportion difference is $\frac{\sqrt{b+c}}{n}$ under the null hypothesis that the true proportions are equal. Taking the ratio of the estimated difference and its standard error and then squaring it yields the test statistic:

$$\chi^2_M = \frac{(b-c)^2}{b+c} \tag{2.5}$$

which approximately follows a chi-square distribution with one degree of freedom under the null for sufficiently large samples [23]. Note that only pairs with different classifications, called the discordant pairs, are used in the calculation of this statistic.

This test can be derived using other perspectives as well. Keeping in mind that only b and c contribute to the test statistic, let $n_* = b + c$, which is the total number of discordant pairs. Under the null hypothesis, b, and thus c as well, follows a binomial distribution with

$p_* = \frac{1}{2}$. So we compare b with $BIN(n_*, p_*)$. Recall that the mean and variance of a binomial random variable are $n_* p_* = \frac{b+c}{2}$ and $n_* p_*(1 - p_*) = \frac{b+c}{4}$. For sufficiently large n_*, greater than 20 is widely agreed upon, $b \overset{approx.}{\sim} N(\frac{b+c}{2}, \frac{b+c}{4})$. Hence,

$$\frac{b - (b+c)/2}{\sqrt{(b+c)/4}} \overset{approx.}{\sim} N(0,1) \Rightarrow \left(\frac{b - (b+c)/2}{\sqrt{(b+c)/4}} \right)^2 = \frac{(b-c)^2}{b+c} \overset{approx.}{\sim} \chi_1^2.$$

The right-most expression is exactly the test statistic above, χ_M^2. If $n_* \leq 20$, one needs to use the binomial test (or sign test) [22]. Others say that the chi-square approximate test statistic can be used as long as $n_* > 10$ provided that the significance level is not much less than 0.05 [9].

We remark that other more standard tests for difference in proportions, e.g. the z-test, cannot be used in this situation, because matched pairs are employed which results in the populations being related/connected.

2.3.3. The Multiple-Testing Problem. In statistical hypothesis testing, an acceptable maximum probability of rejecting a true null hypothesis is specified; i.e. the maximum probability of Type I error is fixed. This selected probability is called the significance level and is typically denoted with α. Then, one wishes to choose a test statistic that maximizes the power (the probability of rejecting a false null hypothesis), or equivalently, that minimizes the probability of Type II error (failing to reject a false null hypothesis). When a Type I or Type II error is committed, the incorrect decision is referred to as a false positive or false negative, respectively. Testing more than a single hypothesis at a given time is referred to as multiple testing. In this setting, the probability that more than one Type I error occurs increases with the number of hypotheses, and this increase is often sharp [27, p. 562]. Therefore, the maximum probability of Type I error will not be maintained across the numerous tests when each is conducted at the stated significance level. One wants to

control the false positives while still preserving decent power, so the significance level must be corrected. The multiple-testing problem occurs in genetic association studies that involve statistical hypothesis tests at numerous genotyped SNPs and is most extreme in the context of candidate gene studies and genome-wide association studies (GWAS). One approach to account for this issue involves adjusting the significance level, or equivalently the p-value, of each hypothesis test.

Consider a general multiple-testing scenario where m hypotheses are being tested. In the context of family-based association studies, we can think of this as testing for linkage and association between a marker and the DSL for m different markers; see Section 3.1 for more on the hypotheses in this setting. Therefore, the m hypotheses and corresponding decisions can be broken down using the following terms

$$
\begin{aligned}
m_0 &= \text{number of true null hypotheses} \\
m - m_0 &= \text{number of false null hypotheses} \\
R &= \text{number of rejected null hypotheses} \\
m - R &= \text{number of null hypotheses that we failed to reject} \\
V &= \text{number of rejected true null hypotheses (Type I errors)} \\
S &= \text{number of rejected false null hypotheses} \\
T &= \text{number of false null hypothesis that we failed to reject (Type II errors)} \\
U &= \text{number of true null hypotheses that we failed to reject.}
\end{aligned}
$$

These classifications are summarized in Table 2.9. Note that m_0 is considered to be fixed but unknown while R, and therefore V, S, T, and U, are regarded as random.

There are a variety of ways to measure Type I errors in this setting; we will discuss two of the most fundamental approaches. The family-wise error rate (FWER) is defined

Table 2.9. Outcomes of Testing m Hypotheses.

	Fail to reject H_0	Reject H_0	Total
True H_0	U	V	m_0
False H_0	T	S	$m - m_0$
Total	$m - R$	R	m

to be the probability that at least one Type I error occurs, i.e. FWER = $P(V \geq 1)$ [28, p. 349]. Essentially, the FWER considers the error rate given that all the null hypotheses are true, which is clearly not always the case in practice [9, p. 163]. Thus, a measure that evaluates the error rate regardless of the number of true null hypotheses is desirable. The false discovery rate (FDR) is defined to be the expected proportion of rejections that are incorrect, i.e. FDR = $E[V/R]$ [27, p. 567]. Note that when $R = 0$, V/R is defined to be 0. The FDR can never be larger than the FWER (FDR $\leq$ FWER). [28, p. 349-350]. An alternative motivation for the FDR is a common misconception. The overall significance level, α, is not the proportion of rejected hypotheses that are actually true; for example, if all m null hypotheses are true then the all rejected hypotheses are errors. The FDR addresses the expected proportion [27, p. 567]. In other words, FWER is concerned with the number of Type I errors, whereas FDR is concerned with the balance between Type I and Type II errors [28, p. 382].

In general, multiple-testing procedures involving adjusted significance levels can be categorized in three ways: one-step, step-up, and step-down. The overall significance level, α, is the upper bound for the Type I error rate under consideration. One-step procedures compare all p-values to a predetermined cutoff that is typically a function of only α and m. In step-up procedures, p-values are first ranked from largest to smallest and then, starting

with the highest, compared to a cutoff dependent upon its rank. If it is larger than the cutoff, the subsequent ranked p-value is considered. This continues until a significant p-value is reached and the null hypotheses corresponding to that p-value and all those with smaller p-values are rejected. Note that "step-up" is referring to the movement toward p-values that are more likely to be significant; this is upward when one thinks of the smallest p-values as being higher. Conversely, step-down procedures move from the smallest to largest ranked p-values and stop once a p-value larger than the cutoff is reached; the corresponding null hypothesis and all those for the larger p-values are not rejected [28, p. 353]. Thus, both step-up and step-down procedures utilize cutoffs dependent on the data while one-step procedure cutoffs are independent of the data.

The unweighted Bonferroni method (also called correction) is a one-step procedure where the cutoff is α/m [28, p. 360]. It can be generalized so that the i^{th} cutoff is α_i; the sum of all α_i's must equal α. This method is based on the first-order Bonferroni inequality which states that the probability of the union of m events is less than or equal to the sum of their individual probabilities. The Bonferroni correction ensures the FWER is no larger than α [27, p. 569]. Additionally, there is no assumption regarding the independence of the hypothesis tests [9, p. 162]. On the other hand, provided the tests are independent, the Benjamini-Hochberg approach is a step-up procedure that controls the FDR at α with a cutoff for the i^{th} ranked p-value being $\alpha_i = \frac{i}{m}\alpha$ [28, p. 362]. Controlling the FDR is more powerful than the Bonferroni correction, especially for large m [27, p. 572].

Even though FDR approaches are more powerful, the Bonferroni correction for FWER is implemented the most frequently in association studies because of theses studies' nature and history. In these investigations, false positives have been commonplace, and hence one wishes to avoid them. Thus, typically investigators forgo maximizing statistical power via FDR control that aims to reduce the number of false negatives. Other approaches

to manage the multiple-testing problem, which will not be discussed in detail here, include permutation/re-sampling techniques, using haplotypes, and using simultaneous test strategies; the last two are only implemented for a small number of SNPs in a defined region [9].

3. HYPOTHESIS TESING IN FAMILY-BASED DESIGNS

In this chapter, we examine the common statistical hypothesis testing framework for family-based designs. We present the corresponding hypotheses and two statistical tests, the transmission disequilibrium test (TDT) and the family-based association test (FBAT). We conclude by discussing the conditional power of these tests.

3.1. HYPOTHESES FOR FAMILY-BASED DESIGNS

As mentioned in Section 2.1.2, association mapping is able to utilize data from unrelated individuals. When the null hypothesis is rejected, evidence suggests that the marker and DSL are associated. However, when testing for association using families, which is done in family-based designs, the complexity of the hypotheses changes by incorporating linkage. First, we will discuss the motivation for including linkage. Then, the specific hypotheses that result will be examined.

Recall that association and linkage cannot be inferred from each other. In family-based designs, we are concerned with association and linkage between a marker and a DSL being passed from the parental generation to the offspring. We will illustrate the necessity for both linkage and association by considering two cases. First, suppose association is present between a DSL and the marker in the parental population but that linkage is absent. Since there is no linkage, $\theta = 0.5$; i.e. recombination will occur between the DSL and marker with a probability of 0.5. Hence, the association will not be consistently passed from the parental generation to the offspring. Second, suppose linkage is present between a DSL and the marker in the parental population but that association is absent (no LD). "Then any observed pattern of association between the marker and the DSL will differ in

every family and their offspring" [9]. In other words, "the marker and the DSL will tend to be transmitted together, but different marker alleles will be transmitted with the DSL in different families" [11]. Thus, no systematic association will be present collectively among the offspring [9].

These two cases indicate that family-based designs "have no power to detect association unless linkage is present" [29]. Therefore, there is only a single alternative hypothesis:

H_A : presence of both association and linkage between a DSL and the marker.

The null hypotheses involve the other combinations of the presence of association and linkage. The possible null hypotheses, in words, are

1. H_0 : presence of association without linkage between a DSL and the marker
2. H_0 : presence of linkage without association between a DSL and the marker
3. H_0 : absence of both linkage and association between a DSL and the marker

The types of studies that correspond to the various null hypotheses are, respectively, [9]

1. follow up of case-control association studies that showed association
2. "follow up of linkage signals"
3. genome-wide association studies or "candidate gene studies without prior linkage"

For all the above null hypotheses, the allele transmission to the offspring from the parents is governed by Mendel's laws. Similarly, it does not matter which null is used in trio designs, which consider only the proband and his/her parents. In that case, the test statistic, which is discussed in Section 3.2.1, has the same distribution under all the null hypotheses. However, for family-based designs that involve siblings, the presence of linkage changes the distribution of the test statistic. Hence, when incorporating siblings, the first and

third null hypotheses will result in the same distribution, but it will differ for the second hypothesis [29].

Therefore, one must be explicit about which null hypothesis is assumed. Even though we are using markers from high confidence and strong candidate genes that already exhibited association, the individual SNPs that are being tested have not necessarily been shown to be associated with the DSL. Therefore, we assume the third null from above: no association nor linkage.

One issue with GWAS in population-based designs is that it is relatively easy to establish significant associations between many markers and a DSL. However, the presence of association between a marker and the disease via the DSL without linkage is actually misleading. The association could be due to other reasons previously discussed in Section 2.1.2. The advantage of family-based designs is that these misleading conclusions are avoided because rejection of the null hypothesis requires both association and linkage [11].

3.2. THE TRANSMISSION DISEQUILIBRIUM TEST (TDT)

In this section, we introduce the pioneering statistical test, the transmission disequilibrium test (TDT), for association and linkage in family-based designs. The theory of the test statistic is presented first. Then, we illustrate the differences resulting from two distributions that can be used to compute the p-values and interpret the hypothesis testing results utilizing the ASD genetic data.

3.2.1. Theory of the TDT. When trios involving the proband and both parents are utilized in a family-based design for genetic linkage and association studies, allele transmission can be regarded as a case-control study where the untransmitted alleles from parents are deemed the controls. This can be displayed in a 2×2 contingency table, Table 3.1, and McNemar's test in this framework is referred to as the transmission disequilibrium test

(TDT) [25, 30]. The test statistic is then

$$\chi^2_{TDT} = \frac{(b-c)^2}{b+c}, \tag{3.1}$$

which approximately follows a chi-square distribution with one degree of freedom under the null hypothesis for sufficiently large sample sizes [23]. Note that half the classifications, those corresponding to a and d, are not utilized.

A single count in a cell of the contingency table represents the *pair* of alleles from a single parent. Therefore, the classification criteria are not independent, because a parent must transmit exactly one allele and not transmit exactly one allele. In other words, if one is transmitted, the other must be untransmitted. This is why Pearson's chi-square test of independence is not appropriate here.

There are numerous assumptions behind the use of the TDT. The disease under consideration is assumed to be recessive, but this does not imply the recessive genetic model is assumed [13]. In reality, since each parental allele transmission is considered separately, an additive genetic model is assumed; see Section 2.1.3. Moreover, only bi-allelic markers are considered, because the standard McNemar's test involves 2×2 tables [9]. Finally, there is an implicit assumption that segregation distortion is absent at the allele locus [13].

Given that McNemar's test only depends on information from pairs with different classification, i.e. b and c in Table 2.8, one would want most, if not all, of the observations to fall in these categorizations. By example, we will demonstrate that no information is lost if only heterozygous parents are considered. In order to maintain the notation in Table 2.8, we will denote the minor and normal alleles as "F" and "f," respectively. Moreover, we restate the contingency table in terms of cases and controls for the transmitted and untransmitted alleles, respectively, in Table 3.1. We consider four families, three

with a single offspring and one with three offspring. The genotypes of the parents and each child are displayed in Table 3.2; note that the coded genotypes assume an additive mode of inheritance. Table 3.3 presents the allele transmission from each parent for all offspring considered; it also classifies the transmission-untransmission pair to a quadrant of the contingency table. One can clearly see that heterozygous parents only contribute to b or c. It is impossible for him/her to contribute to a or d. Since each count represents a pair of alleles and each pair corresponds to a single parent, $n_* = b + c$ is the number allele transmissions from heterozygous parents [30]. In the case where only one affected offspring is under consideration, n_* will also be the number heterozygous parents. We can also think of n_* as the number of heterozygous parent-child pairs. Note that this is different than n in the contingency table, which denotes the total number of allele transmission considered. In our example where all parental and offspring genotypes were known, $n = 12$ is also double the number of offspring included in the study; see Table 3.4.

Table 3.1. The General Contingency Table for the TDT. The minor and normal alleles are "F" and "f," respectively.

Cases (Transmitted Alleles)	Controls (Untransmitted Alleles) F	f	Total
F	a	b	$a + b$
f	c	d	$c + d$
Total	$a + c$	$b + d$	n

The test statistic was created to test for linkage when association is present, which corresponds to the first null hypothesis in Section 3.1. Nonetheless, the usage of the TDT has changed since its inception. Currently, it is often utilized to test association. Recall that the TDT test statistic is for a family-based design, so in reality it is testing for both

Table 3.2. Example Set of Families for the TDT. The minor and normal alleles are "F" and "f," respectively, and the additive mode of inheritance is used for coding the genotypes.

Family	Parental Genotypes	Offspring Genotype	Offspring Coded Genotype
1	FF, FF	FF	$X_{11} = 2$
2	FF, Ff	Ff	$X_{21} = 1$
3	ff, Ff	Ff	$X_{31} = 1$
4	Ff, Ff	FF	$X_{41} = 2$
	Ff, Ff	Ff	$X_{42} = 1$
	Ff, Ff	ff	$X_{43} = 0$

Table 3.3. Classifying Parental Contributions for the Contingency Table for the TDT. The minor and normal alleles are "F" and "f," respectively, and the additive mode of inheritance is used for coding the genotypes. Additionally, the parents are classified as being homozygous (Hom) or heterozygous (Het).

Coded Genotype	Parent Genotype	Hom or Het	Transmitted Allele	Untransmitted Allele	Contributes to Quadrant
$X_{11} = 2$	FF	Hom	F	F	a
	FF	Hom	F	F	a
$X_{21} = 1$	FF	Hom	F	F	a
	Ff	Het	f	F	c
$X_{31} = 1$	ff	Hom	f	f	d
	Ff	Het	F	f	b
$X_{41} = 2$	Ff	Het	F	f	b
	Ff	Het	F	f	b
$X_{42} = 1$	Ff	Het	F	f	b
	Ff	Het	f	F	c
$X_{43} = 0$	Ff	Het	f	F	c
	Ff	Het	f	F	c

Table 3.4. The Contingency Table for the TDT Example. Classifications of all allele transmissions that occurred for the example families are shown. The minor and normal alleles are "F" and "f," respectively.

	Controls		
Cases	F	f	Total
F	3	4	7
f	4	1	5
Total	7	5	12

linkage and association. Moreover, it is currently used for rare diseases, because in this case, using only affected offspring information is usually sufficient; refer to Section 3.3.3 for more details [9]. Most importantly, the TDT has motivated the development of other family-based tests.

Because χ^2_{TDT} approximately follows a chi-square distribution with one degree of freedom for large n_*, it should not be used when the number of heterozygous parents is too small. There are various recommendations on what constitutes too small. One of the more conservative criteria states that if $n_* \leq 20$, one needs to use an exact test, which in this case is the binomial test [22].

The exact test statistic for the TDT is

$$B_{TDT} = b, \tag{3.2}$$

which, under the null hypothesis ($H_0 : p = 0.5$), follows a binomial distribution with a probability of success of $p = 0.5$ and the total number of trials is $n_* = b + c$ [22]. Note

that the choice of b instead of c as the test statistic was arbitrary, because the distribution is symmetric under the two-sided null. Moreover, this is a specific case of the binomial test.

3.2.2. Comparing the Use of the Exact and Approximate Distributions of the TDT. For our data set, we are dealing with small sample sizes, especially when considering a single cluster at a time. Recall from Section 2.2.3, there are 10, 13, and 21 families in Cluster 1, 2, and 3, respectively. Therefore, we are concerned with the appropriateness of using the approximate distribution of the test statistic for the TDT. With this as a motivation, we wish to better understand the effect of small sample sizes and the distribution utilized on the results of the hypothesis tests.

Two analyses are carried out to examine differences in the test statistics χ^2_{TDT} and B_{TDT} for large and small n_*. In reality, only a single family-based test for genetic association and linkage is performed; the difference is just in the calculation of the test statistic and its associated p-value. We will utilize the ASD genetic data introduced in Section 2.2.3. To illustrate differences between the exact and approximate distributions of the TDT test statistic, we are not yet attempting to answer the research question concerning the facial features of the boys with ASD. After quality control, we are able to conduct tests for 2828 of the 3335 markers; no individuals were removed.

The TDT test using the approximate chi-square test statistic was carried out in PLINK; we will refer to the p-values of this test as the approximate p-values. The output of this procedure also reports the number of transmitted and untransmitted alleles used to calculate the test statistic, i.e. b and c. We utilized this information to conduct the exact binomial test using R, and its p-values will be reported as the exact p-values [31].

In this analysis, the families are not divided based on the facial-feature cluster to which the affected offspring belongs but instead are grouped together; the plots denote this with "Cluster All." This corresponds to testing for linkage and association with an autism

DSL in the entire dataset. First, we will consider the appropriateness of the approximate test for this data. Recall that there are 44 families and the TDT only utilizes trio information. Thus, we use 132 individuals of which 88 are parents. The criterion for determining the validity of the approximate test depends on the number of heterozygous parents for the marker (SNP) under consideration; the exact test is recommended when this is 20 or less, i.e. $n_* \leq 20$. We found that 622 of the 2828 markers, which is approximately 22%, require the exact binomial test. Hence, the approximate p-values for those markers should not be interpreted. Also, no single SNP had more than 58 heterozygous parents out of maximum of 88. See Figure 3.1 for a breakdown of the number of heterozygous parents.

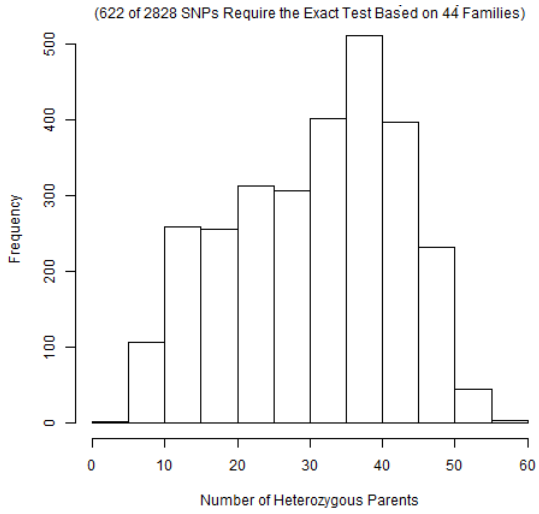


Figure 3.1. Histogram of the Number of Heterozygous Parents. The exact binomial test is required when the number of heterozygous parents is 20 or less ($n_* \leq 20$) for a given marker.

The exact binomial p-values were calculated for all markers even if the number of heterozygous parents utilized exceeded 20. These exact p-values were never less than the approximate ones, so the TDT using the approximate chi-square test statistic never overestimated the p-value for a marker being linked and associated with an autism DSL;

refer to Figure 3.2. In other words, any SNP that is declared to be linked and associated with a DSL using the exact distribution must also be significant when using the approximate distribution, but the converse is not necessarily true. In fact, the approximate and exact p-values were only equal when they both were one, which occurs when the number of transmitted and untransmitted alleles are identical, i.e. $b = c$. There were some markers whose approximate p-value was less than one but exact p-value equaled one. This was due to the number of transmitted and untransmitted alleles differing by only one. In this instance, the continuous chi-square statistic resulted in a relatively high p-value, but the discrete binomial test statistic was able to correctly assign a p-value of one by taking the discrete nature of the numbers into account. This can be seen in Figure 3.2 as the line of dots occurring where the exact p-value is one.

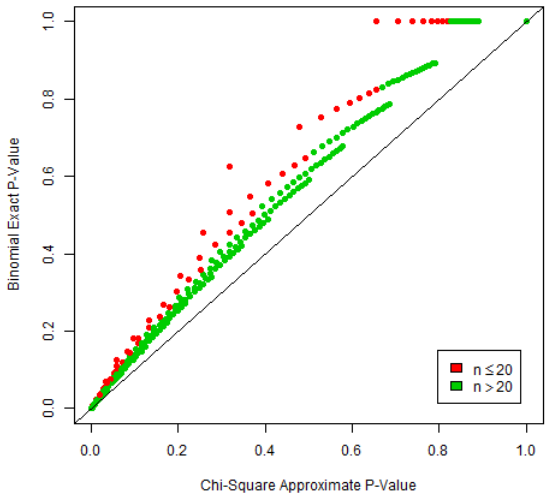


Figure 3.2. Exact vs. Approximate p-Values. The exact p-values are never less than the approximate p-values. The difference between the two is generally larger when the approximate p-value is larger. The colors indicate that there is greater agreement between the p-values when the approximation is typically deemed acceptable ($n_* > 20$).

In the same vein, the difference between the exact and approximate p-values for markers with larger approximate p-values tended to be higher even when $n_* > 20$, because

the closer the values are for the number of transmitted and untransmitted alleles, the greater the impact of a single discrete transmission. Hence, the employment of the discrete binomial test statistic will change the p-value to a greater extent in this case. However, within this trend the approximate p-values were closer to the exact p-values for $n_* > 20$. This is represented in Figure 3.2 by the fact that the points corresponding to markers where the approximation is considered appropriate fall closer to the 45 degree line. Figure 3.3 illustrates this point in another way; the difference in the p-values is more exaggerated when the number of heterozygous parents is smaller, especially for $n_* \leq 20$, and when the approximate p-value is larger.

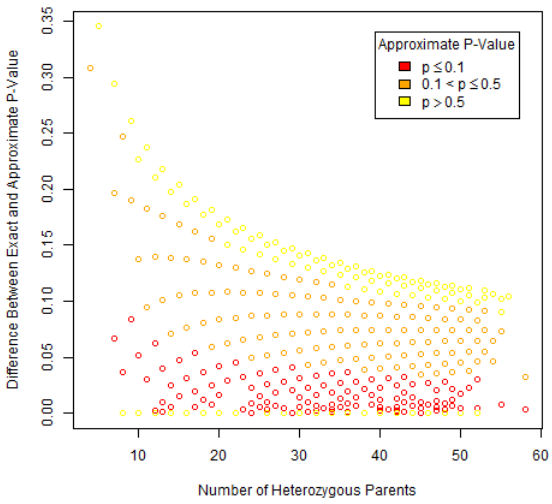


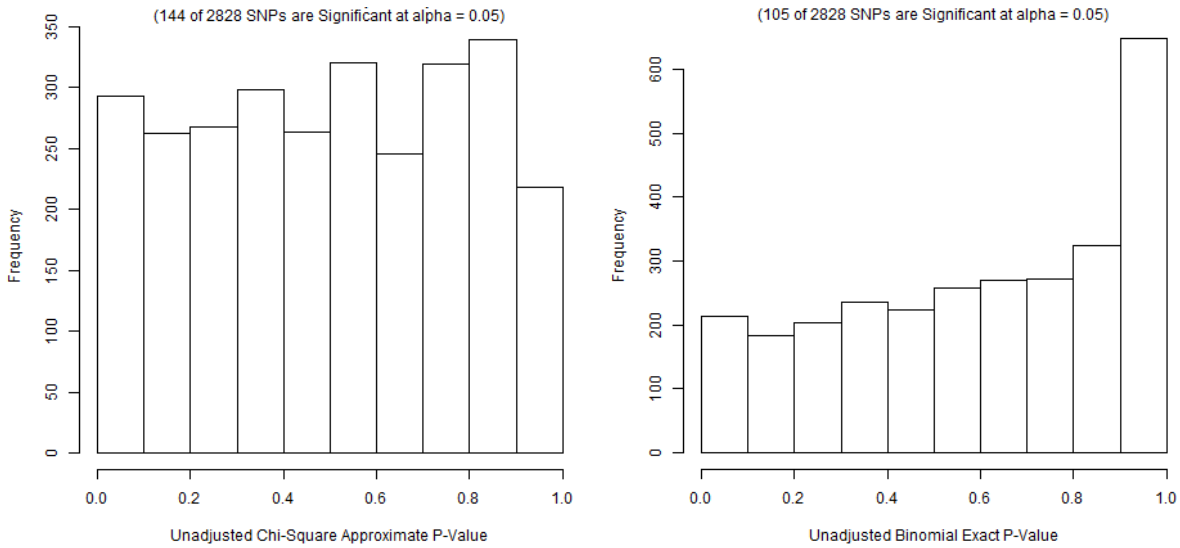
Figure 3.3. Change in p-Value vs. the Number of Heterozygous Parents. The difference in the p-values is more exaggerated when the number of heterozygous parents is smaller, especially for $n_* \leq 20$, and when the approximate p-value is larger.

As one would expect, the distribution of p-values differs based on which test statistic is used. Figure 3.4 compares these distributions. The histogram of the approximate p-values appears to be somewhat uniform with a slight dip near one. At a significance level of 0.05, 144 of the markers would be declared significant without any adjustment for multiple testing.

Contrastingly, we see a large peak at one for the histogram of the exact p-values. This is expected since any p-value that is one using the chi-square test statistic will remain one, and for any marker where the number of transmitted and untransmitted alleles differs by only one, the exact p-value will be one as well. Also, matching the fact that the exact p-values are never less than the approximate ones, fewer markers are declared significant; 105 SNPs are significant at the 0.05 significance level without any adjustment for multiple testing.

At the end of Section 2.2.2, it was mentioned that a method of assessing the overall genotyping quality utilizes the distribution of the p-values. For the chi-square version of the TDT, the distribution under the null is uniform; the corresponding histogram, which is Figure 3.4a, suggests that this is approximately the case. On the other hand, uniformity of the p-values is not expected under the null for the discrete test statistic using the exact binomial distribution [32]. Therefore, the peak near one is not a cause for concern. In fact, using simulation, it can be shown that the distribution under the null for 2828 binomial tests has the same general shape as that shown in Figure 3.4b. As should be the case, the distribution of p-values for both versions of the test demonstrate the same assessment of the overall genotyping quality: there are no issues, because the observed distribution matches that expected under the null.

In summary, there can be substantial differences between the p-values for the TDT computed using the approximate chi-square distribution and those using the exact binomial distribution. These differences are more exaggerated for smaller sample sizes and larger p-values. There is little disparity between the p-values when there is much evidence against the null hypothesis, i.e. when the p-values are small. We mainly care about the markers for which that is the case, so utilizing the approximate distribution will give similar declarations of significance. However, it is possible for a marker's approximate p-value to fall below the significance level while its exact p-value does not, which is undesirable. Hence, to



(a) Histogram of approximate p-values for the TDT. (b) Histogram of exact p-values for the TDT.

Figure 3.4. Histograms of Approximate and Exact p-Values for the TDT. Figure 3.4a is for the p-values computed using the approximate distribution, while those in Figure 3.4b were computed using the exact distribution. We see a large peak at one for the histogram of the exact p-values. This is due to the facts that any p-value that is one using the chi-square test statistic will remain one, and for any marker where the number of transmitted and untransmitted alleles differs by only one, the exact p-values will be one as well.

be prudent, one should work with the exact binomial distribution in the computation of p-values for the TDT.

3.2.3. Interpreting the Results of the TDT for the ASD Data. Next, we will properly interpret the results of these tests for illustrative purposes. Recall from Section 2.3.3, that in order to address the multiple-testing problem the per-test significance level, or equivalently the p-values, must be adjusted. For the autism data, using the Bonferroni correction to control the FWER at 0.05 results in a corrected significance level of $0.05/2828 = 1.768 \times 10^{-5}$. Therefore, zero markers were found to be significantly linked and associated with an autism DSL when using both the approximate and exact TDT. Refer to Table 3.5 for the SNPs closest to attaining family-wise significance for the exact TDT. We do not consider controlling the FDR, because the hypothesis tests considered are presumably not independent. Moreover, association studies have historically tried to minimize the number of Type I errors by utilizing the Bonferroni correction. Additionally, because approximate p-values are unnecessary when exact p-values are available, only the exact ones are reported here.

3.3. THE FAMILY-BASED ASSOCIATION TEST (FBAT)

Here, we discuss the generalization of the TDT, the family-based association test (FBAT). We introduce its theory, demonstrate its relationship to the TDT, and consider its extensions to some additional settings.

3.3.1. FBAT Theory. Family-based designs include a wide variety of setups of which the TDT's setting is just one. The TDT dealt with known parental genotypes and only affected offspring under the assumptions of no linkage, bi-allelic markers, and an underlying additive model. Therefore, there is a need for a more general test statistic that may be extended and modified to apply to diverse scenarios.

Table 3.5. SNPs Closest to Significance via Exact TDT without Considering Facial-Feature Clusters. An arbitrary level of 0.005 was used to filter the markers whose p-values are smallest when using the exact TDT. For each, information on the SNP's location, minor allele, minor allele frequency (MAF), and the number of heterozygous parents (NHZP, i.e. n_*) is displayed. None fall below the family-wise significance level of 1.768×10^{-5}. *High-confidence ASD gene

Chr.	Gene	Marker	Allele	MAF	NHZP	p-Value
3	CACNA2D3	rs10510773	A	0.3636	34	0.000195
21	DSCAM	rs6517607	G	0.1932	24	0.001544
3	CACNA2D3	rs3906509	A	0.3466	29	0.002316
3	CACNA2D3	rs13068008	A	0.3506	29	0.002316
3	CNTN4	rs925820	A	0.392	45	0.002459
12	GRIN2B*	rs12829455	G	0.4886	42	0.002887
2	NRXN1	rs11125326	C	0.2356	34	0.002935
5	CTNND2	rs10474912	G	0.2898	31	0.003327
21	DSCAM	rs8132311	G	0.2784	31	0.003327
3	CNTN4	rs7637377	A	0.08523	13	0.003418
2	NRXN1	rs10490175	A	0.2045	36	0.003933
12	GRIN2B*	rs12821108	A	0.3977	36	0.003933

It is important to retain the desirable characteristics of the TDT when generalizing. The robustness to population substructure is the chief advantage attained by using families, so its preservation is crucial. In order to accomplish that, three principles related to the null distribution need to be maintained through the generalization. First, one must condition on the parental genotypes when computing the null distribution to ensure the robustness to population substructure. Second, while generalizing to arbitrary phenotypes, the offspring genotype under the null needs to be conditioned on the offspring's trait. Third, in order to guarantee the validity of the test, the null distribution must be the "distribution of offspring genotypes, conditional on parental genotypes and offspring traits." Note that some of the specific implementations of these principles may need to be handled differently in certain contexts, e.g. including multiple offspring in the presence of linkage [9].

The test statistic for the general family-based association test (FBAT) can based on a score statistic. First, we will introduce the notation needed and then provide the expression for the test and score statistic.

$$
\begin{aligned}
n &= \text{number of families} \\
n_i &= \text{number of offspring in family } i \\
i &= 1, \ldots, n \quad \text{(family index)} \\
j &= 1, \ldots, n_i \quad \text{(offspring index)} \\
X_{ij} &= \text{coded genotype of } j^{\text{th}} \text{ offspring in } i^{\text{th}} \text{ family} \\
T_{ij} &= \text{coded trait of } j^{\text{th}} \text{ offspring in } i^{\text{th}} \text{ family} \\
P_i &= \text{parental genotypes for the } i^{\text{th}} \text{ family}
\end{aligned}
$$

The coding of the genotype (X_{ij}) depends on the genetic model assumed; see Table 3.6. It has been shown that even when the actual genetic model is not additive, using the additive model assumption to code X_{ij} has good power. Hence, the FBAT software's default is the additive model [14]. In this case, X_{ij} is the number of A alleles possessed by the j^{th} offspring in i^{th} family [29].

When a marker is sex-linked, i.e. it is on the X sex chromosome, the coding may differ slightly from that in Table 3.6 depending on the genetic model and the gender of the individual. For the additive model, the coding of sex-linked markers for females is identical to that for autosomal markers, but that for males is modified. Since males only possess a single X chromosome, it is impossible for them to have two alleles of a sex-linked marker. Therefore, X_{ij} is 0 if the A allele is absent or 1 if the A allele is present [14, p. 10]. The contributions to the test statistic for this modified genetic coding is the same as if the

Table 3.6. Coding the Genotype based on the Mode of Inheritance. One typically uses a coded genotype, X, to represent the actual genotype, G, where "A" and "a" are the minor and normal alleles, respectively, of the marker considered. This table specifies the codings for each mode of inheritance; note that a coded genotype vector is used in the codominant case.

Recessive		Dominant		Additive		Codominant		
X	G	X	G	X	G	X_1	X_2	G
1	AA	1	AA or Aa	2	AA	1	0	AA
				1	Aa	0	1	Aa
				0	aa	0	0	aa

standard additive coding is utilized while forcing males to be homozygous for the allele they possess for sex-linked markers.

Normally, the coded trait is taken to be the phenotypic residual, which is defined as $T_{ij} = Y_{ij} - \mu$, where

$$Y_{ij} = \text{coded phenotype of } j^{\text{th}} \text{ offspring in } i^{\text{th}} \text{ family}$$

$$\mu = \text{offset parameter (chosen by investigator)}$$

The optimal choice of μ is approximately the prevalence of the disease, i.e. $\mu \approx E[Y]$ [9]. Choosing the value of the offset will be discussed in more detail in Section 3.3.3.

This information can be combined into vectors, which gives

$$\mathbf{X} = [X_{11}, X_{12}, \ldots, X_{n\,n_i}]' \tag{3.3}$$

$$\mathbf{T} = [T_{11}, T_{12}, \ldots, T_{n\,n_i}]' \tag{3.4}$$

$$\mathbf{Y} = [Y_{11}, Y_{12}, \ldots, Y_{n\,n_i}]' \tag{3.5}$$

$$\boldsymbol{\mu} = \mu \mathbf{j}_N' \tag{3.6}$$

$$\mathbf{P} = [P_1, P_2, \ldots, P_n]'. \tag{3.7}$$

These represent the vector of coded offspring genotypes ($\mathbf{X}$), coded offspring traits ($\mathbf{T}$), coded offspring phenotype ($\mathbf{Y}$), the offset parameter ($\boldsymbol{\mu}$), and the parental genotypes ($\mathbf{P}$). The vectors in (3.3)-(3.6) are of length $N = \sum_{i=1}^{n} n_i$, which is the total number of offspring considered, but $\mathbf{P}$ in (3.7) is of length n. Also, the phenotypic residual can be represented as $\mathbf{T} = \mathbf{Y} - \boldsymbol{\mu}$.

Note that the only variable that will be considered random is $\mathbf{X}$. The traits, and consequently the phenotypes and offset, along with the parental genotypes are regarded as observed. Thus, the FBAT can be considered a conditional test where one conditions on the parental genotypes and the traits [33]. With this in mind, we denote the conditional expectation and variance under the null hypothesis as $E_0[\mathbf{X}|\mathbf{P}]$ and $\text{Var}_0(\mathbf{X}|\mathbf{P})$, respectively. Those under the alternative hypothesis will be referred to as $E_A[\mathbf{X}|\mathbf{P}]$ and $\text{Var}_A(\mathbf{X}|\mathbf{P})$ and will be utilized for conditional power calculations in Section 3.4.

The FBAT test statistic is

$$\begin{aligned}\chi^2_{FBAT} &= \frac{[(\mathbf{Y}-\boldsymbol{\mu})'(\mathbf{X}-E_0[\mathbf{X}|\mathbf{P}])]^2}{(\mathbf{Y}-\boldsymbol{\mu})'\mathrm{Var}_0(\mathbf{X}|\mathbf{P})(\mathbf{Y}-\boldsymbol{\mu})} \\ &= \frac{[\mathbf{T}'(\mathbf{X}-E_0[\mathbf{X}|\mathbf{P}])]^2}{\mathbf{T}'\mathrm{Var}_0(\mathbf{X}|\mathbf{P})\mathbf{T}} \\ &= (\mathbf{X}-E_0[\mathbf{X}|\mathbf{P}])'\frac{\mathbf{T}\mathbf{T}'}{\mathbf{T}'\mathrm{Var}_0(\mathbf{X}|\mathbf{P})\mathbf{T}}(\mathbf{X}-E_0[\mathbf{X}|\mathbf{P}]) \\ &= (\mathbf{X}-E_0[\mathbf{X}|\mathbf{P}])'\mathbf{A}(\mathbf{P},\mathbf{T})(\mathbf{X}-E_0[\mathbf{X}|\mathbf{P}]).\end{aligned} \tag{3.8}$$

The matrix of this quadratic form, $\mathbf{A}$, is "a $N \times N$ symmetric weight matrix whose elements depend on the parental genotypic information [($\mathbf{P}$)] and on the trait [($\mathbf{T}$)]" [33, p. 167]. The test statistic only corresponds to testing a single marker but does so using all families.

We can break down the test statistic into a score statistic and its associated variance to gain further insight into its structure and motivate its distribution. The expression for the FBAT score statistic, which can be viewed of as the covariance between the trait and genotype [29, p. 230], is

$$U = \mathbf{T}'(\mathbf{X}-E_0[\mathbf{X}|\mathbf{P}]) \tag{3.9}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n_i} T_{ij}\left(X_{ij}-E_0[X_{ij}|P_i]\right) \tag{3.10}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n_i}(Y_{ij}-\mu)\left(X_{ij}-E_0[X_{ij}|P_i]\right). \tag{3.11}$$

The FBAT software computes this for each allele of each SNP, but in the bi-allelic setting, the two p-values for a marker will be identical. The conditional expectation, $E[X_{ij}|P_i]$, is what maintains the robustness to population substructure. In fact, under the null hypothesis, it can be calculated using Mendel's first law. Hence, the score statistic, U, has an expected value of zero under any of the three null hypotheses mentioned in Section 3.1 [9].

The variance of the score statistic is

$$\begin{aligned} \text{Var}_0(U) &= \mathbf{T}'\text{Var}_0(\mathbf{X}|\mathbf{P})\mathbf{T} \\ &= \sum_i \sum_{j,j'} T_{ij}T_{ij'}\text{Cov}_0(X_{ij}, X_{ij'}|P_i). \end{aligned} \tag{3.12}$$

If no linkage is assumed to be present, i.e. the first and third null hypotheses from Section 3.1, or only a single sibling is used, $n_i = 1$ for $i = 1, \ldots, n$, then

$$\text{Var}_0(U) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} T_{ij}^2 \text{Var}_0(X_{ij}|P_i). \tag{3.13}$$

However, if linkage is present, i.e. the second null hypothesis is assumed, and more than a single offspring in a family is included, then within a family, allele transmissions are correlated. In this case, an approximate empirical variance is utilized for the score statistic and is given by

$$\text{Var}_0(U) \approx \sum_{i=1}^{n} \left[\sum_{j=1}^{n_i} T_{ij}(X_{ij} - E_0[X_{ij}|P_i]) \right]^2. \tag{3.14}$$

Note that this approximation only holds under the null hypothesis [9].

Standardizing U leads us to the distribution of the test statistic previously introduced. Two equivalent expressions for the test statistic are

$$Z_{FBAT} = \frac{U}{\sqrt{\text{Var}_0(U)}} \tag{3.15}$$

$$\chi^2_{FBAT} = (Z_{FBAT})^2 = \frac{U^2}{\text{Var}_0(U)}. \tag{3.16}$$

For large samples and under the null hypothesis, Z_{FBAT} approximately follows the standard normal distribution, $N(0,1)$, and thus, χ^2_{FBAT} approximately follows a central chi-square distribution with one degree of freedom [29].

Technically, for this hypothesis test, the size of the sample is determined by the number of informative families (NIF). This is related to but different from the measure of sample size used with the TDT, i.e. the number of transmissions from heterozygous parents. An informative family is one whose contribution to the score statistic is nonzero. In other words, a family is non-informative if there is no randomness in regards to the offsprings genotype. More formally, the i^{th} family is informative if

$$\text{Var}_0\left(\sum_{j=1}^{n_i} T_{ij}(X_{ij} - E_0[X_{ij}|P_i])\right) > 0.$$

Thus, informativeness depends on the coding of the trait (T_{ij}), the parental genotypes (P_i), the genetic model (i.e. the coding of the genotypes), and the family configuration [9, p.150-151]. For the additive mode of inheritance, informative families have at least one heterozygous parent or "sibships with at least two distinct genotypes" [11, p. 389]. In the trio setting, clearly, the NIF will never be larger than the number of heterozygous parents. It is recommended to use these approximate distributions only when there are at least ten informative families [14].

3.3.2. Obtaining the TDT from the FBAT. Recall that in the TDT setting, we assume all offspring are affected ($T_{ij} = 1$ for all i, j combinations), there is no segregation distortion, and the additive mode of inheritance is utilized. In its original derivation, the null hypothesis assumes no linkage [13, p. 510]. The FBAT test statistic, χ^2_{FBAT}, under these conditions will reduce to the test statistic for the TDT, χ^2_{TDT} [9, 29]. We will now demonstrate this.

First, consider the values of the terms in the FBAT test statistic under the additive genetic model; they are summarized in Table 3.7. Note that these results are only specific to the additive mode of inheritance and not the TDT setting, and we will return to referring to the minor and normal alleles as "A" and "a," respectively. It is important to observe that families with only homozygous parents do not contribute to the test statistic, which was also the case for the TDT. In general, the expected value of the coded genotype is half of the combined number of A alleles possessed by the parents. Additionally, the variance is the number of heterozygous parents divided by four.

Table 3.7. Contributions to FBAT Test Statistic with Additive Mode of Inheritance. The number of heterozygous parents present in the family is referred to as NHZP.

P_i	NHZP	$E[X_{ij}\|P_i]$	$\text{Var}(X_{ij}\|P_i)$	$X_{ij} - E[X_{ij}\|P_i]$
aa, aa	0	0	0	$0 - 0 = 0$
AA, aa	0	1	0	$1 - 1 = 0$
AA, AA	0	2	0	$2 - 2 = 0$
Aa, aa	1	0.5	0.25	$1 - 0.5 = 0.5$ A transmitted $0 - 0.5 = -0.5$ A untransmitted
Aa, AA	1	1.5	0.25	$2 - 1.5 = 0.5$ A transmitted $1 - 1.5 = -0.5$ A untransmitted
Aa, Aa	2	1	0.5	$2 - 1 = 1$ 2 A's transmitted $1 - 1 = 0$ 1 A transmitted $0 - 1 = -1$ 0 A's transmitted

Now, we find the value of the score statistic U in the TDT setting. Without loss of generality, we will assume that there is only one offspring per family, i.e. $n_i = 1$ for

$i = 1, \ldots, n$. Then, we have

$$\begin{aligned} U &= \sum_{i=1}^{n} \sum_{j=1}^{n_i} T_{ij} \left(X_{ij} - E_0[X_{ij}|P_i] \right) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{1} (1) \left(X_{ij} - E_0[X_{ij}|P_i] \right) \\ &= \left(\sum_{i=1}^{n} X_{ij} \right) - \left(\sum_{i=1}^{n} E_0[X_{ij}|P_i] \right). \end{aligned}$$

Since only families with at least one heterozygous parent contribute to U, we only need to consider the following, where the subscript indicates the genotype of the parent beyond the initial heterozygous parent,

$$\begin{aligned} n_{aa} &= \text{number of families where parent genotypes are Aa, aa} \\ n_{AA} &= \text{number of families where parent genotypes are Aa, AA} \\ n_{Aa} &= \text{number of families where parent genotypes are Aa, Aa.} \end{aligned}$$

We will also use the notation introduced in Section 3.2.1 for the TDT. Thus, b is the number of heterozygous parents that transmitted A, c is the number of heterozygous parents that transmitted a, and $n_* = b + c$ is the total number of transmissions from heterozygous parents. Note that n_* is also the number of heterozygous parents when there is only one offspring per family, i.e. $n_* = n_{aa} + n_{AA} + 2n_{Aa}$. Additionally, $\sum_{i=1}^{n} X_{ij} = b + n_{AA}$ for three reasons: only families with at least one heterozygous parent are considered, we have assumed only one offspring per family, and the parent with genotype aa cannot transmit an A allele. The

second part of U becomes

$$\begin{aligned}\sum_{i=1}^{n} E_0[X_{ij}|P_i] &= (0.5)n_{aa} + (1.5)n_{AA} + (1)n_{Aa} \\ &= 0.5(n_{aa} + n_{AA} + 2n_{Aa}) + n_{AA} \\ &= 0.5n_* + n_{AA},\end{aligned}$$

which is the sum of the expected number of A transmissions from heterozygous parents and that of the single homozygous parent considered. Hence, the score statistic is $U = b - 0.5n_* = b - 0.5(b+c) = 0.5(b-c)$, which holds for any number of affected offspring in a given family. This implies that only heterozygous parents contribute to the score statistic, one of the properties of the TDT.

In the absence of segregation distortion, the probability that the A allele is transmitted by a heterozygous parent is 0.5. We can think of this as a single Bernoulli trial whose corresponding random variable has a variance of $(0.5)(1-0.5) = 0.25$. The transmissions from the heterozygous parents are independent between families in general. When no linkage is assumed, as in the TDT, transmissions between parents within the same family are also independent. Therefore, $\text{Var}_0(U) = (0.25)n_* = (0.5)^2(b+c)$ and

$$\chi^2_{FBAT} = \frac{U^2}{\text{Var}_0(U)} = \frac{[0.5(b-c)]^2}{(0.5)^2(b+c)} = \frac{(b-c)^2}{b+c} = \chi^2_{TDT}.$$

In words, the FBAT simplifies to the TDT in this setting, or alternatively, the FBAT is a generalization of the TDT.

3.3.3. Extensions of the FBAT. The FBAT was constructed generally so that additional modifications and extensions could arise naturally. There are countless possible scenarios to which the FBAT could be applied, but we will examine three for brevity. The

first two extensions involve moving from a univariate to a multivariate test statistic. The final extension incorporates unaffected offspring information, which necessitates the precise choice of the null hypothesis.

For multi-allelic markers (or a codominant genetic model), the coded genotype, X_{ij}, will be a vector where each of its elements is a coding for a specific allele. Another modification concerns simultaneously testing for association between a marker and multiple traits. This results in the coded trait term, T_{ij}, being a vector. For both of these expansions, the score statistic, $\mathbf{U}$, will also be a vector and the variance, $\text{Var}_0(\mathbf{U}|\mathbf{P})$, is a variance-covariance matrix. Note that $\mathbf{P}$ is the collection of all parental genotypes. Thus, the test statistic becomes the quadratic form $\mathbf{U}'\text{Var}_0(\mathbf{U}|\mathbf{P})\mathbf{U}$ which approximately follows a chi-square distribution with $\text{rank}[\text{Var}_0(\mathbf{U}|\mathbf{P})]$ degrees of freedom [9]. Although, generally useful, for this work, these particular extensions do not address our data type or do not lead to desirable interpretations of the results for the research question, respectively.

The remaining application incorporates information from unaffected offspring; these individuals can be siblings of affected offspring or be a member of a family of only unaffected offspring. This is especially advantageous for common diseases, because it increases the power of the test when compared with using only affected offspring. Including families with only unaffected offspring does not result in substantial power increases for rare diseases, because the allele distribution among the unaffected, their parents, and the overall population will be analogous (unlikely to possess disease alleles) [9]. However, inclusion of any unaffected offspring is better than using the TDT [33, p. 580]. But it may be the case that the extra resources, both time and money, required to genotype unaffected offspring may not be worth the minor gain in power for rare diseases.

Incorporation of these individuals in the test statistic requires an appropriate formulation of the coded trait, T_{ij}, via the phenotype, Y_{ij}. For dichotomous traits, let

$$Y_{ij} = \begin{cases} 1 & \text{if the } j^{\text{th}} \text{ offspring in } i^{\text{th}} \text{ family has the disease (affected)} \\ 0 & \text{if the } j^{\text{th}} \text{ offspring in } i^{\text{th}} \text{ family does not have the disease (unaffected)} \end{cases}$$

which implies that

$$T_{ij} = \begin{cases} 1-\mu & \text{if the } j^{\text{th}} \text{ offspring in } i^{\text{th}} \text{ family has the disease (affected)} \\ -\mu & \text{if the } j^{\text{th}} \text{ offspring in } i^{\text{th}} \text{ family does not have the disease (unaffected)} \end{cases}$$

where the offset parameter, μ, is chosen by the investigator. Hence, the score statistic, U, can be viewed as a contrast between affected and unaffected offspring with weights $(1-\mu)$ and μ, respectively. This contrast is needed, because over/undertransmission of an allele to the affected offspring should coincide with under/overtransmission to the unaffected. Finally, recall that the inclusion of siblings will necessitate the use of the approximate variance, Equation (3.14), in the calculation of the test statistic when linkage is assumed to be present [29].

Typically, μ is chosen to be the prevalence of the disease in the overall population. Nevertheless, the misspecification of the offset does not affect the validity of the test statistic, because it does not depend on the genotype X [34, p. 613]. This fact has motivated the development of approaches that optimally choose the value of the offset. For instance, the variance of the score statistic U is minimized by selecting $\mu = n_a/(n_a+n_u)$. where n_a and n_u are the total number of transmissions to affected and unaffected offspring, respectively. In essence, this choice is the "sample estimate of the prevalence, where offspring are weighted by the number of heterozygous parents" [34, p. 609]. Often it is similar to the sample

prevalence [29, p. 231]. This optimal offset can be computed using the "-o" option in the FBAT software [14, p. 10]. Note that a different variance-minimizing offset is chosen for each hypothesis, i.e. for each marker that is tested.

However, the truly optimal offset choice should take the power of the test into account. Note that minimizing variance does not necessarily maximize the power of a test [29, p. 231]. The power is maximized when μ is approximately the disease prevalence [29, p. 226]. In fact, a value slightly larger than the prevalence generally results in the highest power [9, p. 153], but the power is relatively unaffected by different offset choices within a reasonable range around the disease prevalence [33, p. 580]. Yet an offset substantially larger than the disease prevalence will yield a power smaller than that attained by using affected offspring only, i.e. the TDT [29, p. 231]. In that case, one assigns too much weight to the unaffected offspring [29, p. 226]. To combat this issue, some suggest to limit the variance-minimizing offset to a maximum of 0.5, but this is not guaranteed to preserve high power [29, p. 231]. Therefore, it is not desirable to use the variance-minimizing offset if the sample prevalence and population prevalence differ drastically. Instead choosing the largest reasonable estimate of the population prevalence should be optimal in terms of power.

3.4. CONDITIONAL POWER FOR FAMILY-BASED ASSOCIATION TESTS

The conditional power of the FBAT test statistic can be computed, and these power calculations are the foundation for a modified testing method called screening. The goal of the modified screening method is to increase the power when testing many markers through the utilization of estimates of the conditional power for each marker considered. The motivation for its use and the specifics of the method are covered in Chapter 5.

In order to obtain the conditional power formula, we need to consider the distribution of the test statistic whose two common forms are (3.8) and (3.16). Under the alternative

hypothesis, the expected value of the score statistic is

$$\begin{aligned}\boldsymbol{\mu}_{H_A} &= E_A[U] \\ &= E_A[\mathbf{T}'(\mathbf{X} - E_0[\mathbf{X}|\mathbf{P}])|\mathbf{P}] \\ &= \mathbf{T}'(E_A[\mathbf{X}|\mathbf{P}] - E_0[\mathbf{X}|\mathbf{P}]),\end{aligned}$$

and its variance is

$$\begin{aligned}\sigma^2_{H_A} &= Var_A(U) \\ &= \mathbf{T}'\mathrm{Var}_A(\mathbf{X}|\mathbf{P})\mathbf{T}.\end{aligned}$$

Now, we can rewrite the test statistic as follows

$$\begin{aligned}\chi^2_{FBAT} &= \frac{[\mathbf{T}'(\mathbf{X} - E_0[\mathbf{X}|\mathbf{P}])]^2}{\mathbf{T}'\mathrm{Var}_0(\mathbf{X}|\mathbf{P})\mathbf{T}} \\ &= \frac{\mathbf{T}'\mathrm{Var}_A(\mathbf{X}|\mathbf{P})\mathbf{T}}{\mathbf{T}'\mathrm{Var}_0(\mathbf{X}|\mathbf{P})\mathbf{T}}\left(\frac{\mathbf{T}'(\mathbf{X} - E_0[\mathbf{X}|\mathbf{P}])}{\sqrt{\mathbf{T}'\mathrm{Var}_A(\mathbf{X}|\mathbf{P})\mathbf{T}}}\right)^2 \\ &= aW^2.\end{aligned}$$

Hence, from above, W will asymptotically follow a normal distribution with unit variance and mean μ_{H_A}/σ_{H_A}, i.e. $W \sim N(\mu_{H_A}/\sigma_{H_A}, 1)$. If follows that W^2 will be asymptotically distributed as a noncentral chi-square distribution with one degree of freedom with noncentrality parameter $\gamma = (\mu_{H_A}/\sigma_{H_A})^2$, i.e. $W^2 \sim \chi^2_1(\gamma)$ [33].

At a significance level of α, the conditional power is

$$\begin{aligned} \mathcal{CP} &= P(\text{reject } H_0 \mid H_A \text{ is true}) \\ &= P\{\chi^2_{\text{FBAT}} \geq \chi^2_{1,\,1-\alpha}(0)\} \\ &= P\{aW^2 \geq \chi^2_{1,\,1-\alpha}(0)\} \\ &= P\{W^2 \geq (1/a)\chi^2_{1,\,1-\alpha}(0)\} \\ &= P\{\chi^2_1(\gamma) \geq \omega\chi^2_{1,\,1-\alpha}(0)\} \end{aligned} \tag{3.17}$$

where

$$\omega = \frac{1}{a} = \frac{\mathbf{T}'\text{Var}_0(\mathbf{X}|\mathbf{P})\mathbf{T}}{\mathbf{T}'\text{Var}_A(\mathbf{X}|\mathbf{P})\mathbf{T}}, \tag{3.18}$$

$$\gamma = \frac{\mu^2_{H_A}}{\sigma^2_{H_A}} = \frac{\{\mathbf{T}'(E_A[\mathbf{X}|\mathbf{P}] - E_0[\mathbf{X}|\mathbf{P}])\}^2}{\mathbf{T}'\text{Var}_A(\mathbf{X}|\mathbf{P})\mathbf{T}}, \tag{3.19}$$

and $\chi^2_{1,\,1-\alpha}(0)$ denotes the $1-\alpha$ percentile of a central chi-square distribution with one degree of freedom [35, p. 171].

In order to compute the first and second moments of the coded genotype under the alternative hypothesis, that is $E_A[\mathbf{X}|\mathbf{P}]$ and $\text{Var}_A(\mathbf{X}|\mathbf{P})$, one needs the conditional marker distribution under H_A. Under the assumption of the additive model, these can be computed with [35]

$$P_{H_A}\left[X_i = x_i \middle| (P_i, T_i)\right] = \frac{f(T_i|x_i) \cdot P(X_i = x_i|P_i)}{\sum_{X_i=0}^{2} f(T_i|X_i) \cdot P(X_i|P_i)}. \tag{3.20}$$

We will discuss the computation in more detail in Section 5.1.2, where we outline how one can estimate this quantity from sample information.

4. FAMILY-BASED ASSOCIATION STUDIES FOR FACIAL-FEATURE CLUSTERS AND ASD

In this chapter, we perform two sets of family-based association studies for the facial-feature clusters detected in boys with ASD. Before discussing the tests and their results, we describe how the facial-feature clusters are incorporated into the phenotype.

4.1. THE CLUSTER MEMBERSHIP PHENOTYPES

When considering the autism spectrum disorder (ASD), the standard phenotype would be whether or not an individual has been diagnosed with ASD. That is, the phenotype is the affection status of ASD where 1 is assigned to individuals with ASD and 0 to those who do not have it; we will refer to this as the ASD Phenotype (ASDP). This phenotype was utilized in Sections 3.2.2 and 3.2.3.

However, including the clinically meaningful facial morphology information is what differentiates this study from previous genetic association studies involving ASD. To accomplish this incorporation, we will consider the facial-feature clusters as three different phenotypes; one for each cluster from Obafemi-Ajayi et al [8]. As a whole, we will call them the Cluster Membership Phenotypes (CMPs). The phenotype for the first cluster will be called the Cluster 1 Membership Phenotype (C1MP) Similarly for the second and third clusters, we will have C2MP and C3MP, respectively. Each of the phenotypes will be affection statuses for the corresponding facial-feature cluster. A summary of the phenotypes, their abbreviations, and their interpretations are displayed in Table 4.1.

These four binary phenotypes can be combined in a variety of ways. We will consider the product of ASDP and a particular CMP, that is the phenotype where the individual has

Table 4.1. Summary of Phenotypes. We consider one phenotype for autism spectrum disorder (ASD) and three phenotypes for facial-feature cluster membership. All four phenotypes are binary, i.e. affection statuses, where 1 is assigned to affected individuals and 0 is assigned to the unaffected.

Phenotype	Abbreviation	Affected (1)	Unaffected (0)
ASD	ASDP	Diagnosed with ASD	Not diagnosed with ASD
Cluster 1	C1MP	Facial features match those of Cluster 1.	Facial features do not match those of Cluster 1.
Cluster 2	C2MP	Facial features match those of Cluster 2.	Facial features do not match those of Cluster 2.
Cluster 3	C3MP	Facial features match those of Cluster 3.	Facial features do not match those of Cluster 3.

ASD *and* whose facial features align with those of the given cluster. Note that the product of binary variables is interpreted as "and." For example, all probands who belong to the second facial-feature cluster would be assigned a 1 for C2MP but would be assigned a 0 for the C1MP and C3MP. By definition, all probands are affected with ASD, i.e. are assigned a 1 for ASDP. For example, the phenotype of being affected with ASD *and* Cluster 2 is the product of the ASDP and C2MP affection statuses, ASDP$*$C2MP. When this is considered for all individuals, only probands that belong to the second cluster are affected with this combined phenotype, while all unaffected siblings and parents are unaffected. Hence, only the probands of Cluster 2 would contribute to the test statistic for the TDT, where only affected offspring are utilized.

Note that the facial features of the unaffected siblings and parents have not formally been clustered, so in reality their CMPs are missing. Thus, any combined phenotype for them will be missing as well. This is not an issue for the parents' information, because

only their genotypes are utilized in the test statistics we have introduced. On the other hand, the sibling information would not contribute to the test statistic due to their missing phenotype. Provided an extension of the FBAT that incorporates siblings is not being implemented, this is of no concern, because sibling information is not used with the TDT. In Section 4.2, we are only concerned with the TDT, so the above distinction is not necessary. However, in Section 4.3, we will incorporate sibling information for combined phenotypes; its motivation and implementation will appear there.

4.2. WITHIN-CLUSTER ANALYSIS VIA THE EXACT TDT

The incorporation of the CMPs involves taking the product of each individual cluster phenotype with the ASD phenotype. In terms of the phenotypic variables, that is ASDP$*$CxMP for $x = 1, 2, 3$. Due to the small sample sizes within a given cluster, we use the exact TDT to avoid eliminating SNPs due to the approximate distributions of the TDT and FBAT not being applicable. Recall that the exact TDT utilizes information from probands only and not unaffected individuals for the phenotype under consideration. In this setting, conducting the exact TDT for the ASDP$*$CxMP for $x = 1, 2, 3$ is identical to using solely the ASD phenotype within each cluster separately. We will refer to this as the "within-cluster TDT analysis" instead of the less intuitive language associated with the phenotypic products.

We performed the exact TDT within each of the three clusters, using the filtered data described in Section 2.2.3. In Section 3.2.3, the exact TDT was performed with the ASD phenotype without considering facial-feature cluster information, i.e. for ASDP alone. The adjusted significance level was 1.768×10^{-5}, while the smallest p-value in the overall analysis was 1.951×10^{-4} for the marker rs10510773. The number of SNPs tested does not change when considering the combined phenotype ASDP$*$CxMP for $x = 1, 2, 3$, so the

adjusted significance level remains the same. For each within-cluster analysis, the lowest p-value is 1.953×10^{-3}, 1.953×10^{-3}, and 4.883×10^{-4} for Clusters 1, 2, and 3, respectively; they correspond to markers rs10263964, rs10490175, and rs1025768, respectively. Hence, we do not have sufficient evidence to conclude that any markers are linked and associated with an ASDP$*$CxMP DSL. An ASDP$*$CxMP DSL is one that causes both ASD and facial features that align with those characteristic of Cluster x, for $x = 1, 2, 3$. In other words, there is not enough evidence to conclude that any markers are linked and associated with an autism DSL within a given cluster. See Table 4.2 for more SNPs that were closest to obtaining significance.

Table 4.2. SNPs Closest to Significance via Exact TDT within Facial-Feature Clusters. An arbitrary level of 0.005 was used to filter the SNPs whose p-values are smallest when using the exact TDT for ASDP$*$CxMP for $x = 1, 2, 3$. For each marker, information on the SNP's location, minor allele, minor allele frequency (MAF), and the number of heterozygous parents (NHZP, i.e. n_*) is displayed. None fall below the family-wise significance level of 1.768×10^{-5}.

Cluster	Chr.	Gene	Marker	Allele	MAF	NHZP	p-Value
1	7	CNTNAP2	rs10263964	A	0.3750	10	0.001953
	2	NRXN1	rs17480512	G	0.3125	9	0.003906
	3	CNTN4	rs11718833	A	0.3678	9	0.003906
	7	CNTNAP2	rs10230132	G	0.2784	9	0.003906
	7	CNTNAP2	rs1997530	A	0.3977	9	0.003906
2	2	NRXN1	rs10490175	A	0.2045	10	0.001953
3	15	CHD2	rs1025768	G	0.1875	12	0.000488
	15	CHD2	rs4777755	A	0.1839	11	0.000977
	3	CACNA2D3	rs2048809	A	0.3506	18	0.001312
	15	CHD2	rs2272458	G	0.2765	14	0.001831
	3	CACNA2D3	rs10510773	A	0.3636	20	0.002577
	3	CACNA2D3	rs551114	A	0.1534	13	0.003418
	15	CHD2	rs1371390	G	0.1932	13	0.003418

The extremely small sample size utilized translates into a proportionally low power to detect markers that are linked and associated with ASD. This occurred even when cluster

information was ignored and all 44 families were utilized in the testing, which was done in Section 3.2.2. The low power is more severe when the 44 families are broken down into smaller subsets in order to incorporate the facial-feature cluster membership into the phenotype. In order to further analyze these data, we need a modified approach in order to increase the statistical power.

4.3. INCORPORATION OF UNAFFECTED OFFSPRING WITH THE C2MP

In this section, we restrict the focus to Cluster 2. We start by motivating this restriction and then present the testing procedure and results.

4.3.1. Motivation for Focusing on the C2MP. Recall from Chapter 1, that both Aldridge et al. and Obafemi et al., using different methods, identified a single common subgroup of boys with ASD based on their facial features. It is unique in terms of the facial features and its characteristic clinical phenotype [5, 8]. This subgroup corresponds to Cluster 2 in the data utilized in this report. With respect to the cluster's facial features, Cluster 2 is both compact and distinct from the other subjects with ASD. In particular, it is defined by "the most exaggerated facial features" [8, p. 7-8]. Moreover, the facial features of the control group of typically developing boys are disparate from Cluster 2's features. The control group's facial features, however, did overlap with those of Clusters 1 and 3 [8, p. 7]. From a clinical perspective, Cluster 2 possesses the most consistent clinical phenotype [8, p. 13] which corresponds to a more severe form of ASD [5, p. 10]. Later, it was also described to be "reflective of the more classic [ASD] diagnosis" [6, p. 16]. Additionally, the fact that the mothers of Cluster 2 individuals possessed more severe measurements of subclinical ASD traits is further motivation for investigating the genetic underpinnings associated with this phenotype [8, p. 10]. In summary, focusing on the Cluster 2 Membership Phenotype (C2MP) is a natural choice, because it is unique to

boys with ASD, represents a coherent sub-phenotype, and may be associated with maternal characteristics.

Using the FBAT to incorporate unaffected offspring will likely enable a larger number of families to contribute to the test statistic for a given marker. We hope this results in increased statistical power compared to the exact TDT test that only incorporated the affected offspring. However, it will most likely be minor, because the ASDP∗C2MP is not a common disease. Nonetheless, recall that Lange et al. claim that including unaffected offspring is always better than not doing so [33, p. 580]. Here we will discuss the assumptions, test statistic considerations, and results of the corresponding analysis.

4.3.2. Phenotypic Assumptions. Our focus on Cluster 2 here differs from its within-cluster analysis carried out in Section 4.2. There, we conducted the exact TDT for the ASDP∗C2MP, which involves only probands who belong to Cluster 2. Technically, this combined phenotype for subjects who were not clustered based on their facial features, i.e. parents and unaffected siblings, was considered missing due to the unknown C2MP affection status. Eliminating missing ASDP∗C2MP affection statuses will enable more information to contribute to the test statistic via the extension of the FBAT that incorporates unaffected offspring; see Section 3.3.3.

Now, given the fact that the defining facial features of Cluster 2 are not present in prepubescent boys without ASD, we will safely assume that all unaffected prepubescent male siblings are not members of Cluster 2, that is, for them, C2MP = 0. We do not need to make any assumptions concerning the phenotypes of parents, because only parental genotypes, not phenotypes, contribute to the test statistic. That leaves all unaffected female, postpubescent male, and adult male siblings with missing C2MP affection statuses. However, because none of them have ASD (ASDP = 0), the choice of the affection status for C2MP is irrelevant when the phenotype of interest is ASDP∗C2MP. No matter their facial features, the siblings

without ASD will have ASDP$*$C2MP = 0. Therefore, deeming C2MP as non-missing for unaffected female, postpubescent male, and adult male siblings allows their ASDP$*$C2MP to be zero without assuming anything about the applicability of the clustering results beyond prepubescent males. Therefore, all unaffected siblings of probands that belong to Cluster 2 and all offspring of families whose proband belongs to Cluster 1 or 3 may contribute to the FBAT test statistic.

4.3.3. Test-Statistic Considerations. Two additional considerations accompany the use of the modified test statistic. First, the FBAT, unlike the exact TDT, uses the approximate distribution of the test statistic. Recall, it is recommended to have at least 10 informative families in order for the approximation to be appropriate. This may result in a fewer number of SNPs that can be tested when compared to the exact TDT. Second, the investigator must specify the value of the offset, μ, utilized in the contrast between affected and unaffected offspring. Note that the ascertainment condition of the proband being a prepubescent male with ASD will inflate the sample prevalence of the ASDP greatly above that of the entire population. Following the reasoning at the end of Section 3.3.3, we want to select the largest reasonable estimate of the population prevalence as the value of the offset.

To determine our choice for μ, we need to estimate the population prevalence of ASDP$*$C2MP. Given the information available to us, the most accurate method to accomplish this involves taking the product of the prevalence of ASD with the prevalence of Cluster 2 membership among those with ASD. Since the clusters based on facial features are unique to the utilized data set, the sample prevalence of the C2MP among those with ASD, $13/44 = 0.29\overline{54}$, must suffice. For ASD, the CDC reports official prevalence values that are updated on a periodic basis, but other large-scale studies have estimated this value as well. Recently, the National Health Statistics Reports published ASD prevalence values based on

43,283 interviews from a National Health Interview Survey (NHIS) [36, p. 2]. They reported that the estimated prevalence of ASD is 0.0224 (2.24%) in the overall population [36, p. 4]. We focus on this estimate, because it is larger than the CDC's official prevalence estimate while remaining reasonable. Moreover, we will utilize the 95% upper confidence limit of the true ASD prevalence as the largest reasonable estimate, i.e. as the value contributing to the offset μ. Using the reported standard errors, the approximate 95% upper confidence limit is 0.0273 (2.73%) [36, p. 8]. Therefore, the prevalence estimate, and thus the value of the offset μ, for ASDP$*$C2MP will be $0.0273 \times 13/44 = 0.008066$ (0.8066%).

4.3.4. Results from Incorporating Unaffected Offspring. In this setting, a SNP whose p-value falls below the significance level after controlling the family-wise error (FWER) rate would be considered linked and associated with an ASDP$*$C2MP disease susceptibility locus (DSL). An ASDP$*$C2MP DSL, by definition, is a genetic locus that has a causal relationship with the phenotype of having both ASD and facial features aligning with those of Cluster 2. When restricting this conclusion to prepubescent boys, one may say that the DSL causes a "subtype" of ASD characterized by the facial features of Cluster 2, because those facial features are not present in normally developing boys. Without the restriction, we would have to assume that the aspects of the facial feature clustering results extend to other populations, i.e. females and males of all ages, in order to make the "subtype" statement. Without the assumption concerning facial features in the overall population, one arrives back to the generally worded interpretation involving an ASDP$*$C2MP DSL.

Due to only testing markers with at least 10 informative families, there was a total 2730 SNPs tested instead of the 2828 when performing the exact TDT in Section 3.2.2 and Section 4.2. Therefore, when controlling the FWER with the Bonferroni method, the per-test significance level is $0.05/2730 = 1.832 \times 10^{-5}$. With this, no markers could be declared significantly linked and associated with the ASDP$*$C2MP DSL with the amount

of evidence available. Table 4.3 shows the SNPs that were closest to obtaining family-wise significance; there are now five markers below the 0.005 cutoff as opposed to only one when using the exact TDT. This difference is likely due to the slightly increased power that accompanies the inclusion of unaffected offspring. Notice that the number of informative families, a measure of the sample size, for each marker is substantially larger than that seen in the previous within-Cluster-2 analysis. For marker rs10490175, the only one that also appears in Table 4.2, the p-value (0.001569) using FBAT to incorporate unaffected offspring is slightly smaller than the p-value (0.001953) from the exact TDT. In summary, the marginal power improvement from including unaffected offspring did not change the overall conclusion: there is insufficient evidence to conclude that any marker is linked and associated with an ASDP∗C2MP DSL.

Table 4.3. SNPs Closest to Significance via FBAT with Unaffected Offspring for ASDP∗C2MP. An arbitrary level of 0.005 was used to filter the SNPs whose p-values are smallest when using the FBAT for the ASDP∗C2MP; the offset μ was taken to be 0.008066. For each marker, information on the SNP's location, minor allele, minor allele frequency (MAF), and the number of informative families (NIF) is displayed. None fall below the family-wise significance level of 1.832×10^{-5}.

Chr.	Gene	Marker	Allele	MAF	NIF	p-Value
2	NRXN1	rs10490175	A	0.205	30	0.001569
3	CNTN4	rs13093332	A	0.335	35	0.003871
5	CTNND2	rs441973	A	0.318	25	0.004045
12	GRIP1	rs11176241	A	0.148	16	0.004558
5	CTNND2	rs1638362	A	0.415	38	0.004729

5. SCREENING FOR FAMILY-BASED DESIGNS

When controlling the family-wise error rate (FWER) via the uncorrected Bonferroni method, we were unable to detect statistically significant association and linkage between any SNP and a DSL corresponding to the phenotype tested. The small sample size, both within clusters and overall, results in low statistical power for the tests. However, the primary ways for experimenters to increase power, reducing variability or increasing sample size, are not currently options. Therefore, for this analysis, increasing power amounts to addressing the multiple comparison problem. Essentially, we need to increase the significance level for the hypothesis test for a given SNP while maintaining the overall significance level of 0.05. This can be accomplished through reducing the number of tests performed or systematically assigning unequal portions of the significance level instead of allotting an equal fraction to each test. After reviewing the literature, we consider the latter, because many of the options corresponding to the former are not applicable to family-based designs with dichotomous traits.

In this chapter, we will delve into the theory behind a screening method developed by Murphy et al. to increase power and then apply the method to the ASD data [37]. The strategy was developed for screening in family-based studies where all the probands are affected, but has been extended to more general settings for categorical phenotypes. At a high level, the method uses information about the offspring phenotype and parental genotypes to estimate the conditional power of the TDT/FBAT for each SNP. Then, the weight for assigning a significance level to a given SNP depends on the rank of its conditional power estimate. The information used in the power assessment is independent of the test statistic, so the overall significance level will not be biased [37]. This results in performing the

same number of tests, but a larger proportion of the overall significance level is used with SNPs for which the power to detect association and linkage is high. Once the individual significance levels have been determined, we compare them to the p-values obtained in Sections 4.2 and 4.3. Note that the test itself is not changed for this new method; only the allotment of the significance level is different.

5.1. CONDITIONAL POWER ESTIMATION WHERE ALL PROBANDS ARE AFFECTED

Here we will outline the theory behind the screening approach; we follow Murphy et al. while providing additional details and explanation [37]. We are interested in computing the power of detecting association and linkage, i.e. rejecting the null when it is false, conditional on information we have about the sample. The key is to use only sample information that is independent of the test statistic (TDT/FBAT). We will state the test statistic for this setting and then discuss the method of estimating the conditional marker density to be used in the conditional power computation.

5.1.1. Setting-Specific Test Statistic. First, we will be explicit in stating the test statistic in the setting where all probands are affected with a disease. The same notation used for the FBAT's general setup in Section 3.3.1 will be utilized here as well. Recall that the FBAT, in its univariate form using summations, is

$$\chi^2_{FBAT} = \frac{\left[\sum_{i=1}^{n}\sum_{j=1}^{n_i} T_{ij}\left(X_{ij} - E_0[X_{ij}|P_i]\right)\right]^2}{\sum_{i=1}^{n}\sum_{j=1}^{n_i} T_{ij}^2 \mathrm{Var}_0(X_{ij}|P_i)}.$$

There are a certain set of characteristics to consider for the specific setting of only utilizing data from trios, i.e. from the proband and both parents. First, $n_i = 1$ for $i = 1, \ldots, n$, because no siblings are included. Therefore, the j subscript can be dropped. Second, the trait describes affection status, so

$$T_{ij} = Y_{ij} = \begin{cases} 1, & \text{if the } j^{\text{th}} \text{ offspring in } i^{\text{th}} \text{ family has the disease (affected)} \\ 0, & \text{if the } j^{\text{th}} \text{ offspring in } i^{\text{th}} \text{ family does not have the disease (unaffected).} \end{cases}$$

However, because all probands are affected, $T_i = 1$ for all $i = 1, \ldots, n$. Lastly, we assume an additive genetic model, so the coded genotype is

$$X_{ij} = \begin{cases} 0, & \text{if the proband of the } i^{\text{th}} \text{ family has 0 minor alleles} \\ 1, & \text{if the proband of the } i^{\text{th}} \text{ family has 1 minor allele} \\ 2, & \text{if the proband of the } i^{\text{th}} \text{ family has 2 minor alleles.} \end{cases}$$

Taking into account these specifics, the test statistic becomes

$$\chi^2_{FBAT} = \frac{\left[\sum_{i=1}^{n} (X_i - E_0[X_i|P_i])\right]^2}{\sum_{i=1}^{n} \text{Var}_0(X_i|P_i)}. \tag{5.1}$$

In fact, this is equivalent to the approximate test statistic for the TDT given by (3.1). Note that P_i contains information about the genotypes of both parents (P_{1i} and P_{2i}). We will continue to use the concise notation, because we will never consider an individual parent's genotype on its own.

5.1.2. Estimating the Conditional Marker Density for the Conditional Power.

In this section, we will consider the conditional marker density under the alternative hypothesis, which will be used for conditional power calculations; the conditional power was introduced in Section 3.4. The conditional marker density is the probability of a particular coded genotype occurring given the parental genotypes and affection status provided that the marker under consideration is linked and associated with a DSL. Notationally, we have $P_{H_A}\left[X_i = x_i \middle| (P_i, T_i = 1)\right]$, which is a specific case of (3.20). Before we derive the expression for this, we need to introduce additional terminology and its corresponding notation. The penetrance probability for each genotype is

$$f_{X_i=x_i} = P(T_i = 1|X_i = x_i) \text{ for } x_i = 0, 1, 2.$$

From this, we can construct the genotype relative risk

$$\Psi_{X_i=x_i} = \frac{f_{X_i=x_i}}{f_0} = \frac{P(T_i = 1|X_i = x_i)}{P(T_i = 1|X_i = 0)}.$$

Therefore, by definition, Ψ_0 is 1. Additionally, let's note that $P[T_i = 1|(X_i = x_i, P_i)] = P(T_i = 1|X_i = x_i)$, because affection ($T_i = 1$) is independent of parenteral genotypes (P_i) when the proband's genotype (X_i) is known (x_i) Finally, the events $X_i = 0$, $X_i = 1$, and $X_i = 2$ are mutually exclusive and exhaustive, so

$$P(P_i, T_i = 1) = \sum_{X_i=0}^{2} P(X_i, P_i, T_i = 1)$$

by the law of total probability. Therefore, the conditional marker density under the alternative hypothesis is

$$\begin{aligned}
P_{H_A}\left[X_i = x_i \middle| (P_i, T_i = 1)\right] &= \frac{P(X_i = x_i, P_i, T_i = 1)}{P(P_i, T_i = 1)} \\
&= \frac{P(X_i = x_i, P_i, T_i = 1)}{\sum_{x_i=0}^{2} P(X_i = x_i, P_i, T_i = 1)} \\
&= \frac{P[T_i = 1|(X_i = x_i, P_i)] \cdot P(X_i = x_i|P_i) \cdot P(P_i)}{\sum_{X_i=0}^{2} P[T_i = 1|(X_i, P_i)] \cdot P(X_i|P_i) \cdot P(P_i)} \\
&= \frac{P(T_i = 1|X_i = x_i) \cdot P(X_i = x_i|P_i)}{\sum_{X_i=0}^{2} P(T_i = 1|X_i) \cdot P(X_i|P_i)} \\
&= \frac{f_{X_i=x_i} \cdot P(X_i = x_i|P_i)}{\sum_{X_i=0}^{2} f_{X_i} \cdot P(X_i|P_i)} \\
&= \frac{\Psi_{X_i=x_i} \cdot P(X_i = x_i|P_i)}{\sum_{X_i=0}^{2} \Psi_{X_i} \cdot P(X_i|P_i)}.
\end{aligned} \tag{5.2}$$

Note that $P(X_i|P_i)$ is determined by Mendel's Laws. However, the penetrance probabilities (f_{x_i}), and thus the relative risk probabilities (Ψ_{x_i}), must be estimated using information that is statistically independent of the testing step [37].

In order to determine what choice will not bias the statistical test, we must breakdown the information used in family-based designs. One can think of family-based association tests, e.g. the TDT and FBAT, as conditional tests. In them, we condition upon the parental genotypes P_i and the offspring phenotype T_i while the offspring genotype, X_i, is considered random. Then, "the evidence for SNP-trait association is evaluated by comparing the observed offspring genotype with the expected offspring genotype, which are computed by conditioning upon the parental genotypes, assuming Mendelian transmissions." The joint density function for all the information can be decomposed as

$$f(X_i, T_i, P_i) = f[X_i|(T_i, P_i)] \cdot f(T_i, P_i), \tag{5.3}$$

where the two components are statistically independent. The first component, $f[X_i|(T_i, P_i)]$, and second component, $f(T_i, P_i)$, correspond to the testing step and the conditional power estimation, respectively. Notice that the nonrandom components of the test, i.e. the parental genotypes and offspring phenotypes, may be utilized for the conditional power computation. In the current setting, there is no variability in the offspring phenotypes, because all probands are affected. In terms of the density function, we have $f(T_i = 1, P_i)$. Hence, we can only use information about the parental genotypes to estimate the relative risk probabilities [37, p. 1].

With three coded genotype possibilities and two parents, there are a total of nine possible parental mating types. However, it does not matter if the genotypes of the parents are swapped, so we are only interested in the distinct mating types. There are six distinct mating types

$$\{P_i\} = \{(0, 0), (0, 1), (0, 2), (1, 1), (1, 2), (2, 2)\}$$

where a given value is the number of minor alleles that a parent possesses. We are interested in the frequency of each of these parental mating types in our sample where $T_i = 1$. Symbolically, we want

$$p_{kl} = P(P_i = (k, l)|T_i = 1) = \frac{P(P_i = (k, l), T_i = 1)}{P(T_i = 1)}.$$

Now, we will denote the minor allele frequency (MAF) for the marker in the general population with p. When Hardy-Weinberg equilibrium holds, the coded offspring genotype follows a binomial distribution with two trials and a probability of success of p, i.e

$X_i \sim BIN(2, p)$. Therefore, by the law of total probability and HWE,

$$\begin{aligned} P(T_i = 1) &= \sum_{X_i=0}^{2} P(T_i = 1, X_i) \\ &= \sum_{X_i=0}^{2} P(X_i) \cdot P(T_i = 1|X_i) \\ &= (1-p)^2 P(T_i = 1|X_i = 0) + 2p(1-p)P(T_i = 1|X_i = 1) + p^2 P(T_i = 1|X_i = 2) \\ &= (1-p)^2 f_0 + 2p(1-p) f_1 + p^2 f_2 \end{aligned} \tag{5.4}$$

Combining this with some of the steps taken to obtain (5.2), we get

$$\begin{aligned} p_{kl} &= \frac{\sum_{X_i=0}^{2} f_{X_i} \cdot P[X_i|P_i = (k,l)] \cdot P[P_i = (k,l)]}{(1-p)^2 f_0 + 2p(1-p) f_1 + p^2 f_2} \\ &= \frac{P[P_i = (k,l)] \sum_{X_i=0}^{2} \Psi_{X_i} \cdot P[X_i|P_i = (k,l)]}{(1-p)^2 + 2p(1-p)\Psi_1 + p^2 \Psi_2}. \end{aligned} \tag{5.5}$$

Here, $P[P_i = (k,l)]$ is determined by the minor allele frequency, p, and the actual mating type "[u]nder the assumption of random mating and Hardy-Weinberg equilibrium at the marker locus in the general population" [37, p. 3].

In this setting, the likelihood of observing the parental mating types is given by

$$\begin{aligned} \ell(\Psi_1, \Psi_2, p) &= \prod_{i=1}^{n} p_{(kl),i} \\ &= p_{00}^{n_{00}} p_{01}^{n_{01}} p_{02}^{n_{02}} p_{11}^{n_{11}} p_{12}^{n_{12}} p_{22}^{n_{22}}, \end{aligned} \tag{5.6}$$

where n_{kl} is the observed number of mating types of $P_i = (k, l)$ and

$$\begin{aligned} n &= n_{00} + n_{01} + n_{02} + n_{11} + n_{12} + n_{22} \\ &= \text{total number of observed mating types} \\ &= \text{number of trios.} \end{aligned}$$

This is because of an independently and identically distributed (i.i.d.) assumption; in other words we have a random sample of size n parental genotypes from the population. That is, the families are independent. This function would need to be maximized over Ψ_1, Ψ_2, p in order to find the maximum likelihood function of the genotype relative risks. Unfortunately, this is impossible to do directly, because the "Fisher information matrix is ill conditioned" and estimating the minor allele frequency (p) is "problematic in the presence of population admixture" [37, p. 3].

The next step involves deriving estimators for the genotype relative risks, Ψ_1 and Ψ_2. We want these estimators to be independent of the minor allele frequency, p, so it does not depend on an additional estimated value. Therefore, we need ratios of the frequency of the parental mating types that do not depend on the minor allele frequency. When we assume that Hardy-Weinberg equilibrium (HWE) holds in the general population, only four ratios will satisfy this requirement [37, p. 4]. They are

$$R_1 = \frac{p_{22} \cdot p_{00}}{p_{02}^2} \tag{5.7}$$

$$R_2 = \frac{p_{01} \cdot p_{11}}{p_{00} \cdot p_{12}} \tag{5.8}$$

$$R_3 = \frac{p_{00} \cdot p_{12}}{p_{02} \cdot p_{01}} \tag{5.9}$$

$$R_4 = \frac{p_{12} \cdot p_{01}}{p_{02} \cdot p_{11}}. \tag{5.10}$$

We will show this is the case for each of the four ratios but will not demonstrate that they are the only ratios where this occurs.

Recall that $P(X_i|P_i)$ is determined by Mendel's Laws. Table 5.1 displays this probability for each possible parental mating type when the additive mode of inheritance is assumed, i.e. X_i is the number of minor alleles. Additionally, when we assume that mating occurs randomly on top of the assumption that HWE holds in the general population, we can specify $P[P_i = (k, l)]$ in terms of the minor allele frequency. As an example, the probability that one parent possesses one minor allele while the other possesses none is

$$\begin{aligned} P[P_i = (0,1)] &= P[(0,1) \text{ or } (1,0)] \\ &= P[(0,1)] + P[(1,0)] \\ &= [2p(1-p)] \cdot [(1-p)^2] + [(1-p)^2] \cdot [2p(1-p)] \\ &= 4p(1-p)^3. \end{aligned}$$

This and the probabilities for all other parental mating types are displayed in Table 5.2.

Table 5.1. Genetic Probabilities for Parental Mating Types Under the Additive Mode of Inheritance.

$P_i = (k, l)$	$P[X_i = 0\|P_i = (k, l)]$	$P[X_i = 1\|P_i = (k, l)]$	$P[X_i = 2\|P_i = (k, l)]$
(0,0)	1	0	0
(0,1)	1/2	1/2	0
(0,2)	0	1	0
(1,1)	1/4	1/2	1/4
(1,2)	0	1/2	1/2
(2,2)	0	0	1

Now that we have information about $P[X_i|P_i = (k, l)]$ and $P[P_i = (k, l)]$, we can find the expressions for the ratios in (5.7)-(5.10). Notice that the denominator of p_{kl} cancels

Table 5.2. Probabilities of Parental Mating Types Under Random Mating and HWE.

$P_i = (k, l)$	$P[P_i = (k, l)]$
(0,0)	$(1-p)^4$
(0,1)	$4p(1-p)^3$
(0,2)	$2p^2(1-p)^2$
(1,1)	$4p^2(1-p)^2$
(1,2)	$4p^3(1-p)$
(2,2)	p^4

out in all the ratios and that $\Psi_0 = 1$. Starting with R_1, we have

$$\begin{aligned} R_1 &= \frac{p_{22} \cdot p_{00}}{p_{02}^2} \\ &= \frac{\{P[P_i = (2,2)] \cdot [1(0) + \Psi_1(0) + \Psi_2(1)]\} \cdot \{P[P_i = (0,0)] \cdot [1(1) + \Psi_1(0) + \Psi_2(0)]\}}{\{P[P_i = (0,2)] \cdot [1(0) + \Psi_1(1) + \Psi_2(0)]\}^2} \\ &= \frac{p^4(1-p)^4\Psi_2}{[2p^2(1-p)^2\Psi_1]^2} \\ &= \frac{\Psi_2}{4\Psi_1^2}. \end{aligned} \tag{5.11}$$

Similarly,

$$
\begin{aligned}
R_2 &= \frac{p_{01} \cdot p_{11}}{p_{00} \cdot p_{12}} \\
&= \frac{\{4p(1-p)^3[1(1/2) + \Psi_1(1/2) + \Psi_2(0)]\} \cdot \{4p^2(1-p)^2[1(1/4) + \Psi_1(1/2) + \Psi_2(1/4)]\}}{\{(1-p)^4[1(1) + \Psi_1(0) + \Psi_2(0)]\} \cdot \{4p^3(1-p)[1(0) + \Psi_1(1/2) + \Psi_2(1/2)]\}} \\
&= \frac{(1+\Psi_1)(1+2\Psi_1+\Psi_2)}{\Psi_1+\Psi_2}, \qquad (5.12) \\
R_3 &= \frac{p_{00} \cdot p_{12}}{p_{02} \cdot p_{01}} \\
&= \frac{\{(1-p)^4[1(1) + \Psi_1(0) + \Psi_2(0)]\} \cdot \{4p^3(1-p)[1(0) + \Psi_1(1/2) + \Psi_2(1/2)]\}}{\{2p^2(1-p)^2[1(0) + \Psi_1(1) + \Psi_2(0)]\} \cdot \{4p(1-p)^3[1(1/2) + \Psi_1(1/2) + \Psi_2(0)]\}} \\
&= \frac{\Psi_1+\Psi_2}{2\Psi_1(1+\Psi_1)}, \qquad (5.13) \\
R_4 &= \frac{p_{12} \cdot p_{01}}{p_{02} \cdot p_{11}} \\
&= \frac{\{4p^3(1-p)[1(0) + \Psi_1(1/2) + \Psi_2(1/2)]\} \cdot \{4p(1-p)^3[1(1/2) + \Psi_1(1/2) + \Psi_2(0)]\}}{\{2p^2(1-p)^2[1(0) + \Psi_1(1) + \Psi_2(0)]\} \cdot \{4p^2(1-p)^2[1(1/4) + \Psi_1(1/2) + \Psi_2(1/4)]\}} \\
&= \frac{2(\Psi_1+\Psi_2)(1+\Psi_1)}{\Psi_1(1+2\Psi_1+\Psi_2)}. \qquad (5.14)
\end{aligned}
$$

As previously stated, these ratios, (5.11)-(5.14), do not depend on the minor allele frequency, p, which is unknown. Therefore, they can be estimated based on the observed frequency of mating types in the sample [37, p. 4]. For example,

$$\hat{R}_1 = \frac{\hat{p}_{22} \cdot \hat{p}_{00}}{\hat{p}_{02}^2} = \frac{\left(\frac{n_{22}}{n}\right)\left(\frac{n_{00}}{n}\right)}{\left(\frac{n_{02}}{n}\right)^2} = \frac{n_{22}n_{00}}{n_{02}^2}.$$

From the definition of the additive mode of inheritance on the linear scale from Section 2.1.3, we have the following equivalent equations

$$\begin{aligned} P(T_i = 1|X_i = 1) &= \frac{1}{2}[P(T_i = 1|X_i = 2) + P(T_i = 1|X_i = 0)] \\ f_1 &= \frac{1}{2}(f_2 + f_0) \\ \Psi_1 &= \frac{1}{2}(\Psi_2 + 1). \end{aligned}$$

Hence, we can express each ratio in terms of a single genotype relative risk, which enables us to construct four different estimators of the genotype relative risk. For instance, $R_1 = (2\Psi_1 - 1)/(4\Psi_1^2)$ implies $\hat{\Psi}_1 = \left(-1 \pm \sqrt{1 - 4\hat{R}_1}\right)/4\hat{R}_1$, which is one estimator for Ψ_1. One could also take the average of the four estimators to estimate the overall effect size [37, p. 4]. However, simulation shows that R_4 is the best ratio to use for the estimators [37, p. 5].

The estimator for the genotype relative risk may not be unique. In fact, it is only unique when HWE holds exactly in the sample. If its solutions are not unique but real, the two solutions will likely be similar due to the fact that HWE is assumed in the overall population. In a communication with the corresponding author, Lange, we were informed that "it was sufficient to select the solution that was range consistent" but were not provided any additional details. We acknowledge that non-uniqueness is not theoretically satisfying, but it appears that, computationally, it does not affect the validity of the method.

The estimators for genetic relative risk ($\hat{\Psi}_1$ and $\hat{\Psi}_2$) can then be used to estimate the conditional marker density in (5.2) for $X_i = 0, 1, 2$. Those three estimated probabilities are utilized in computing the conditional power of the a family-based association test, (3.17) from Section 3.4 for a given marker. It is important to note that the estimators only depend parental genotypes which corresponds to the $f(T_i = 1, P_i)$ term in (5.3). Thus, the

significance level of the the hypothesis test will not be biased. It is possible to determine analogous ratios and estimators for other nuclear family structures as well [37, p. 4].

HWE need not be present in the ascertained sample, because it is only assumed to hold for the given marker in the overall population. However, being out of HWE can impact the power. When SNPs not associated with affection status are out of HWE, the power of this screening strategy may be reduced. On the other hand, the power may increase or decrease when the DSL violates HWE. In other words, departures from HWE may impact the conditional power estimation but they will not affect the testing step. Hence, the validity of the overall strategy is robust to violations of HWE [37, p. 4].

5.2. WEIGHTING THE SIGNIFICANCE LEVEL

Next, we need to utilize the calculated conditional power estimates to allot fractions of the overall significance level, α, to the various markers. Low conditional power for a given SNP implies a low minor allele frequency and/or small genetic effect size may be present. Because the power to detect whether the marker is significantly linked and associated with a DSL is limited, it is reasonable that its corresponding hypothesis test would deserve a smaller fraction of the overall significance level [38, p. 608]. Murphy et al. accomplishes this through a weighting method introduced by Ionita et al. [37, 38]. We will call this the rank-weighting step. It is important to note that this approach may be employed in any setting where conditional power estimates are available; it is not specific to family-based designs.

In order to discuss the method, we first must introduce new notation. Consider

$$
\begin{aligned}
m &= \text{total number of markers considered} \\
i &= \text{index for the power rankings of the markers } (i = 1, \ldots, m) \\
w_i &= \text{weight for marker of rank } i \ (w_i \geq 0) \\
\alpha_i &= w_i \alpha = \text{significance level allotted to the marker of rank } i.
\end{aligned}
$$

Note that the sum of the weights must be one, i.e. $\sum_{i=1}^{m} w_i = 1$, in order to maintain the stated control over the the overall Type I error rate. Also, different weighting choices will result in different power.

Other common multiple-testing corrections can be viewed in this weighting framework. The Bonferroni correction equally weights all markers; for all $i = 1, \ldots, m$, $w_i = 1/m$. A common screening method for family-based association studies involving quantitative traits, sometimes called the top-R method, considers a power assessment and weighting. After power estimates are calculated, only the markers with the top R conditional power values are then tested. A Bonferroni correction is made on these R hypothesis tests instead of those for all the markers considered in the power estimation step [39]. Therefore, for the top-R approach, one chooses $w_1 = \cdots = w_R = 1/R$ and $w_{R+1} = \cdots = w_m = 0$. The advantage to using the Bonferroni and top-R approaches are that all markers are tested and conditional power information is incorporated, respectively. The method we will introduce combines both of these advantages [38, p. 68].

Most markers considered in genetic association studies will not be linked or associated with the DSL, so only a small fraction of the markers should receive a large proportion of the significance level. Therefore, in order to maintain robustness on this front, we will partition the markers into groups of similar estimated conditional power values; these

groups will be referred to as partitions. All markers belonging to the same partition will be assigned the same weight, and hence the same portion of α. Moreover, this approach is robust to small alterations in the ranking value for a marker. Note that only the ranking, i, is used for partitioning and not the actual value of the conditional power estimate [38].

The following notation will be used for the partitions:

K = number of partitions

j = the contitional power rank for the partitions ($j = 1, \ldots, K$)

k_j = number of markers in the partition with the j^{th} highest power

w^j = weight for all SNPs in the partition with the j^{th} highest power.

For example, the first partition will contain the k_1 markers with the highest conditional power estimates while the K^{th} partition contains the k_K markers with the lowest conditional power estimates. We must have $\sum_{j=1}^{K} w^j k_j \leq 1$ in order to control Type I error at the specified level; although, $\sum_{j=1}^{K} w^j k_j = 1$ is preferred. Also, we must account for all the markers under consideration $\sum_{j=1}^{K} k_j \geq m$. The Bonferroni method can then be described as choosing $K = 1$, $k_1 = m$, and $w^1 = 1/m$. On the other hand, the top-R approach utilizes $K = 2$, $k_1 = R$, $k_2 = m - R$, $w^1 = 1/R$, and $w^2 = 0$ [38].

All that remains is to specify a strategy for assigning partition sizes and weights. Intuitively, the partitions containing the markers with the highest conditional power should be the smallest and received the largest weight, because they involve the most promising SNPs. In terms of our notation, as j increases, k_j should increase while w^j decreases. There are many different weighting functions that will satisfy this requirement; both Ionita et al. and Murphy et al. employed an exponential weighting scheme [37, 38].

In general, one needs to specify two parameters for the exponential weighting: the size of the first partition, k_1, and the base, r. Then, the number of markers in a partition is defined recursively as

$$k_j = r^{j-1} k_1. \tag{5.15}$$

The weights are chosen in a similar fashion,

$$w^j = \frac{r-1}{r^{2j-1} k_1}. \tag{5.16}$$

Therefore, the significance level assigned to the j^{th} partition is $\alpha_j = w^j \alpha$. Note that

$$\sum_{j=1}^{K} k_j w^j = \sum_{j=1}^{K} (r-1) \left(\frac{1}{r}\right)^j = 1 - \left(\frac{1}{r}\right)^K \approx 1$$

approximately satisfies the requirement that the sum of all the weights equals 1 [37]. The fact that it will never exceed one is important; that means we will always control the Type I error rate to be at most the specified level, α.

According to simulation studies involving 500,000 markers, the $k_1 = 7$ and $r = 2$ are the optimal choices of the parameters in terms of power. However, $k_1 = 5, \ldots, 10$ and $r = 2, 3$ resulted in similarly high power levels; only $k_1 = 3, \ldots, 10$ and $r = 2, 3, 4, 5$ were considered [37, p. 7]. When examining 100,000 markers, $k_1 = 5$ and $r = 2$ were found to be the best selections. Moreover, for 100,000 markers, simulations indicate that this weighting scheme is more powerful than both the Bonferroni and top-R approaches and even attains power near that of population-based designs [38, p. 609].

Clearly, provided the number of markers are divided evenly with the partitioning method, there is $1/r^K$ of the significance level that goes unused. There will be an additional

portion of the significance level that is not utilized if the number of markers is not divided evenly with the partitioning method, which will almost always be the case. For instance, when considering 100,000 markers with $k_1 = 5$ and $r = 2$, the number of SNPs allotted to the last partition needed (the fifteenth, i.e. $K = 15$) would be 81,920. However, only 18,085 SNPs remain unpartitioned after creating the fourteenth grouping. Thus, the true value of k_{15} will be much smaller than that prescribed by the exponential weighting method. When the number of markers considered is relatively high, the combined unused significance level from these two sources is negligible. Including it in the approach would only make a difference if a p-value for a given SNP were extremely close to the allotted significance level.

5.3. SCREENING FOR THE WITHIN-CLUSTER ANALYSES VIA THE EXACT TDT

The screening method discussed involves two parts: the conditional power estimation and the rank-weighting step. The conditional power calculations were carried out using PBAT v3.61, and we implemented the rank-weighting step ourselves in R v3.2.3 [31, 40]. In this work, we consider 2828 markers, which is lower than those considered in the simulation studies that established optimal choices of the weighting parameters. We specified the size of the group of most promising SNPs to be 5, i.e. $k_1 = 5$, because that was the optimal choice for the simulation involving fewer markers. Since the base of $r = 2$ was superior in both simulations, we will also set it to two. More details about each of the partitions with these weighting parameters is displayed in Table 5.3. Notice that the significance level assigned to the partition with the most power happens to coincide with the arbitrary cutoff we previously utilized to see which markers were closest to significance. Recall that the adjusted significance level using the Bonferroni correction is 1.768×10^{-5}. Thus, partitions

six through ten are actually assigned a more stringent significance level than without the screening method.

Table 5.3. Partition Details for Exponential Weighting with $k_1 = 5$, $r = 2$, and $\alpha = 0.05$. For the first ten partitions, we display the size of the partition (k_j), the last rank value that would be a member of the partition, and the significance level assigned to all SNPs in the partition (α_j). Ten partitions are needed for the 2828 markers considered, but the tenth partition will contain 273 markers in this setting instead of the 2560 prescribed in general.

Partition Num (j)	Partition Size (k_j)	Last Rank	$\alpha_j = w^j\alpha$
1	5	5	5.00×10^{-3}
2	10	15	1.25×10^{-3}
3	20	35	3.13×10^{-4}
4	40	75	7.81×10^{-5}
5	80	155	1.95×10^{-5}
6	160	315	4.88×10^{-6}
7	320	635	1.22×10^{-6}
8	640	1275	3.05×10^{-7}
9	1280	2555	7.63×10^{-8}
10	2560	5115	1.91×10^{-8}

None of the markers considered attained statistical significance using the screening approach. Due to the fact that the largest significance level from the exponential weighting is 0.005, the only markers that had a chance of being significantly linked and associated with an ASDP$*$CxMP DSL, for $x = 1, 2, 3$, were those closest to significance shown in Table 4.2. Hence, we only discuss those same markers in the screening setting; refer to Table 5.4. Most of those markers did not have high ranking conditional power estimates. The exception is rs2272458 for the Cluster 3 analysis that was grouped into the third partition with a rank of 19. Nonetheless, its p-value was an order of magnitude larger than the assigned significance level.

The amount of the overall significance level, $\alpha = 0.05$, that goes unused due to $1/r^K$ and k_K, where $K = 10$, not being the size prescribed is 9.77×10^{-4} and 4.37×10^{-5},

Table 5.4. Screening Results for Within-Cluster Analyses via the Exact TDT. The SNPs whose p-value from the exact TDT was less than 0.005 are shown; they also appeared in Table 4.2. For each, we display the conditional power estimate, the rank of the conditional power among all markers in the study, and the significance level (α_j) assigned to the SNP based on the partition to which its rank belongs. No marker is below its assigned significance level. (The p-value is shown in scientific notation this time to make the comparison to α_j more conducive.)

Cluster	Marker	p-Value	Power	Rank	α_j
1	rs10263964	1.953×10^{-3}	0.1726	1013	3.05×10^{-7}
	rs17480512	3.906×10^{-3}	0.4958	117	1.95×10^{-5}
	rs11718833	3.906×10^{-3}	0.3298	342	1.22×10^{-6}
	rs1997530	3.906×10^{-3}	0.2524	602	1.22×10^{-6}
	rs10230132	3.906×10^{-3}	0.1087	1515	7.63×10^{-8}
2	rs10490175	1.953×10^{-3}	0.0696	1917	7.63×10^{-8}
3	rs1025768	4.883×10^{-4}	0.1645	1627	7.63×10^{-8}
	rs4777755	9.766×10^{-4}	0.3007	1031	3.05×10^{-7}
	rs2048809	1.312×10^{-3}	0.2384	1287	7.63×10^{-8}
	rs2272458	1.831×10^{-3}	0.9739	19	3.13×10^{-4}
	rs10510773	2.577×10^{-3}	0.3819	776	3.05×10^{-7}
	rs551114	3.418×10^{-3}	0.2798	1113	3.05×10^{-7}
	rs1371390	3.418×10^{-3}	0.0530	2500	7.63×10^{-8}

respectively. Thus, the overall unused portion of α is 0.00102. This is substantially higher than the example considered in Section 5.2 for two reasons. First, there are considerably fewer SNPs in this study, and second, the final partition only uses about 10% of the combined significance level assigned to it. However, one cannot simply assign this extra portion of the significance level to a marker of one's choosing. For example, we should not add this to α_j for rs1025768 for the Cluster 3 analysis in order to make the new assigned significance level 0.00102, which results in the SNP being significantly linked and associated with an ASDP$*$C3MP DSL. Doing so would be a form of data snooping, which would not control the Type I error at the specified level. In other words, we cannot assign significance levels after seeing the results of hypothesis testing. Therefore, if the unused portion of α is to be reassigned, it must be done so in a meaningful fashion that is planned before any hypothesis testing is carried out.

Overall, we cannot say that any of the 2828 markers considered are linked and associated with an ASDP$*$CxMP DSL, for $x = 1, 2, 3$, using the exact TDT, while controlling the FWER at 0.05. That does not mean that none of the markers in our study truly exhibit that relationship. It does mean that, given the amount of evidence present in the data set utilized, we cannot conclude any linkage and association with a DSL with sufficiently high confidence. Hence, it may be the case that some of those markers are linked and associated with a DSL, or it could be that none are. Additionally it is possible, that some markers are linked and associated with the DSL but were not included in the markers utilized here. As was the case without the screening method, the inability to detect statistically significant linkage and association is most likely due to the small sample size, i.e. the small number of families contributing to the test statistic has resulted in low statistical power.

5.4. SCREENING FOR THE C2MP FOCUS VIA THE FBAT

The screening method developed in Sections 5.1 and 5.2 applies to affection status phenotypes where all offspring are affected. In this section, we consider the FBAT for the ASDP∗C2MP that includes unaffected offspring. However, PBAT v3.61 has implemented a generalized version of the approach in Murphy et al. that works for any categorical phenotype [37, 39]. Therefore, we employ this feature to compute the conditional power estimates when the unaffected offspring information is included. We will use the same parameter values ($k_1 = 5$ and $r = 2$) for the rank-weighting step as those utilized in Section 5.3. Thus, Table 5.3 is applicable here as well.

As before, the largest significance level from the exponential weighting approach is 0.005, so we restrict the discussion to those markers whose p-value is smaller than that, i.e. those previously shown in Table 4.3. The results of the screening method for those SNPs are displayed in Table 5.5. All the p-values are at least two orders of magnitude larger than the corresponding assigned portion of the significance level. Therefore, we have insufficient evidence to conclude that any marker is linked and associated with an ASDP∗C2MP DSL. When restricting this conclusion to prepubescent boys, the DSL that causes a "subtype" of ASD characterized by the facial features of Cluster 2. As mentioned at the end of Section 5.3, this solely means that the data we have does not contain enough evidence to conclude that any of the markers are linked and associated with sufficiently high confidence. We have not proven that every marker considered is not related to the DSL.

5.5. EXAMINING THE CONDITIONAL POWER ESTIMATES

Neither including unaffected offspring information to increase the effective sample size for a given marker nor allotting greater portions of the significance level to the most

Table 5.5. Screening Results for the C2MP Focus. The SNPs whose p-value from the FBAT was less than 0.005 are shown; they appeared in Table 4.3 too. For each, we display the conditional power estimate, the rank of the conditional power among all markers in the study, and the significance level (α_j) assigned to the SNP based on the partition to which its rank belongs. No marker is below its assigned significance level. (The p-value is shown in scientific notation this time to make the comparison to α_j more conducive.)

Marker	p-Value	Power	Rank	α_j
rs10490175	1.569×10^{-3}	0.0635	2000	7.63×10^{-8}
rs13093332	3.871×10^{-3}	0.1916	1050	3.05×10^{-7}
rs441973	4.045×10^{-3}	0.5334	156	4.88×10^{-6}
rs11176241	4.558×10^{-3}	0.4632	230	4.88×10^{-6}
rs1638362	4.729×10^{-3}	0.6305	75	7.81×10^{-5}

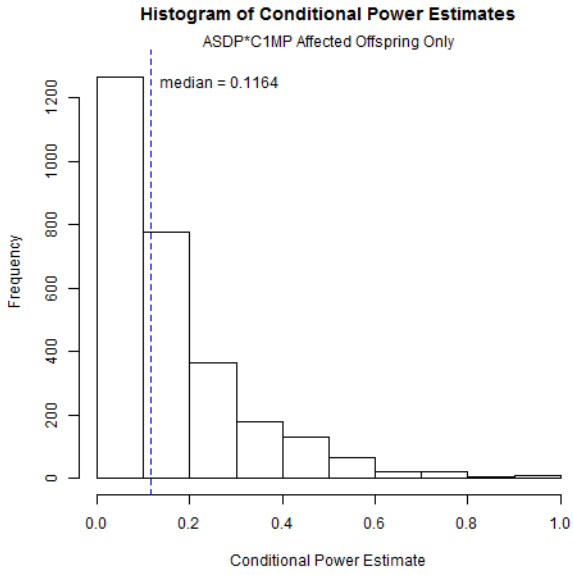
promising SNPs resulted in detecting any markers that are significantly linked and associated with a DSL. Theoretically, those adjustments should lead to a larger statistical power; albeit, in the first case the increase is marginal due to the infrequency of the phenotype in the overall population. Therefore, either the increase in power was not sufficient to detect markers in the study that are truly linked and associated with the DSL given the data set, or there are no markers in the 2828 considered that are truly linked and associated with the DSL. We cannot say which is the case, because we assume from the beginning that linkage and association are absent. However, an examination of the conditional power estimates will provide some insight into the status of the power for the analyses conducted.

The conditional power estimates for each of the analyses are graphically summarized via histograms in Figure 5.1. In every case, at least half of the markers have conditional power estimates less than 0.2051; this can be seen quickly, because the median is represented by the blue dashed vertical line. We utilize the median as the measure of center, because the estimates are highly skewed to the right. At the best, the sample median is about a fourth of the popular value of the minimum power recommended, which is 0.8 [41]. There is an

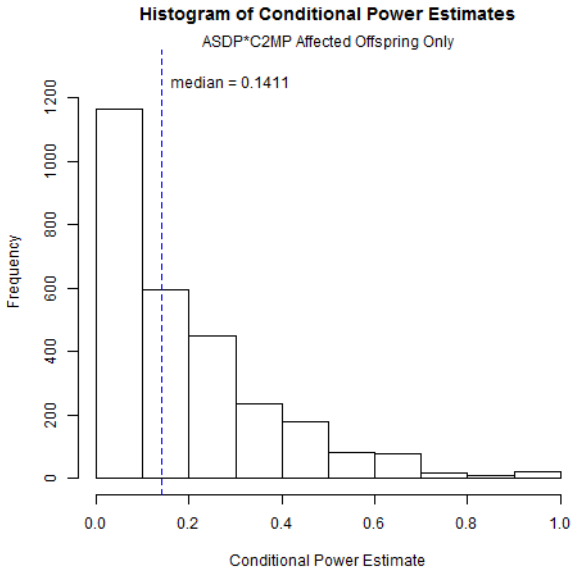
important distinction that must be made here. The recommended power of 0.8 is the true statistical power of the test, so it involves the true effect size, which is unknown in practice. The conditional power estimates are not the true power of the tests. The estimate is an attempt to approximate the true value conditioning on, i.e. using, the parental genotypes and the offspring phenotypes.

For the analyses that do not utilize unaffected offspring, shown in Figures 5.1a - 5.1c, the conditional power estimate are generally larger as the cluster number increases. This corresponds to an increasing number of affected offspring for the ASDP$*$CxMP; recall from Table 2.4, Clusters 1, 2, and 3 have 10, 13, and 21 probands, respectively. As expected, the marginal difference in the number of affected offspring between Clusters 1 and 2 resulted in negligible changes in the distribution of the conditional power estimates. Whereas, doubling, or nearly so, the number of probands when going from Cluster 1 or 2 to Cluster 3 causes the distribution and sample median to change more dramatically. The sample median conditional power estimate for Clusters 1, 2, and 3 are 0.1164, 0.1411, and 0.2051, respectively. Nonetheless, the values of the conditional power estimates are quite low for many markers, so it will be difficult to detect a true linkage and association with a DSL.

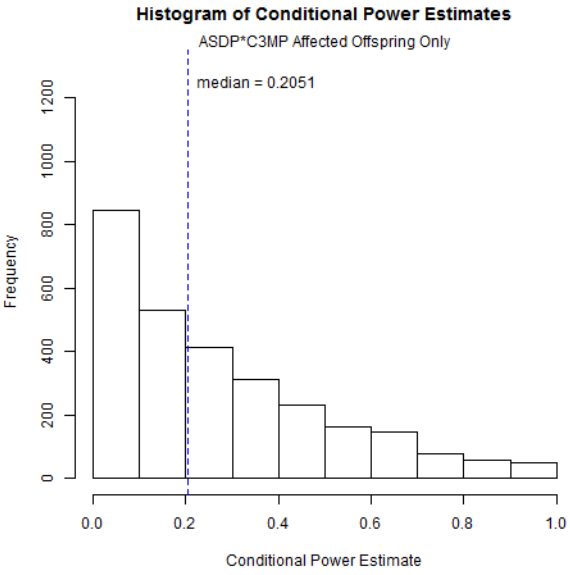
Now for the incorporation of unaffected offspring within the restriction to the C2MP, we can compare Figure 5.1b and Figure 5.1d. The distributions appear nearly identical and their sample medians are approximately the same as well. Hence, incorporating the unaffected offspring had little impact on the distribution of the conditional power estimates. This is not completely unexpected, because the offset chosen, $\mu = 0.008066$, is quite small, which results in the unaffected offspring's contribution to be nonzero but still minor. Recall, including families with only unaffected offspring does not result in substantial power increases for rare diseases, because the allele distribution among the unaffected, their parents, and the overall population will be analogous (unlikely to possess disease alleles) [9].



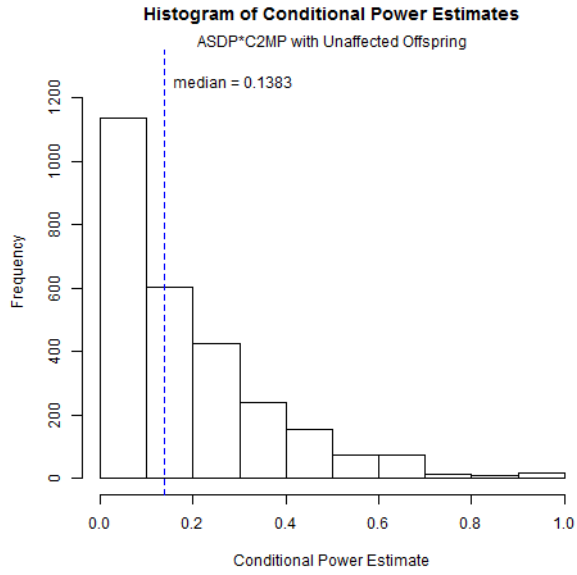
(a) Histogram of conditional power estimates for the ASDP∗C1MP via the exact TDT.



(b) Histogram of conditional power estimates for the ASDP∗C2MP via the exact TDT.



(c) Histogram of conditional power estimates for the ASDP∗C3MP via the exact TDT.

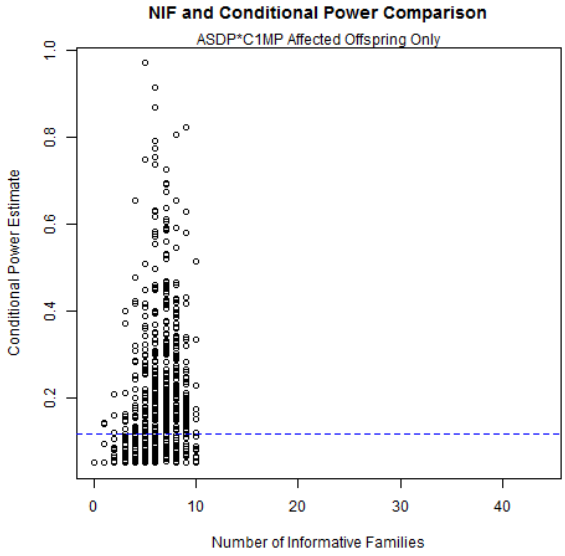


(d) Histogram of conditional power estimates for the ASDP∗C2MP via the FBAT.

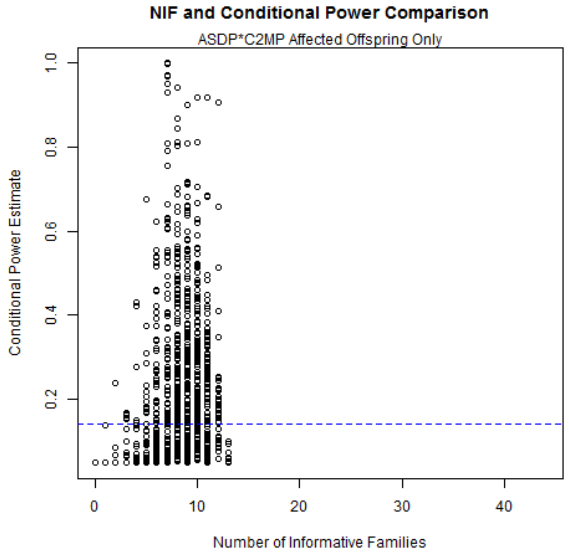
Figure 5.1. Histograms of Conditional Power Estimates. The conditional power estimates were computed using PBAT and the method introduced by Murphy et al. [37, 39]. The sample median conditional power is denoted with the blue dashed vertical line.

One may notice that the inclusion of unaffected offspring actually decreased the sample median of conditional power estimates from 0.1411 to 0.1383. This is not a cause for concern, because the change is minimal. The apparent invariance of the conditional power distribution to the incorporation of unaffected offspring does not imply that the modification was not helpful. In general, one always increases power by including information from unaffected offspring. In this case, the increase was marginal due to the low prevalence of the disease considered, and this change was imperceptible in the distribution.

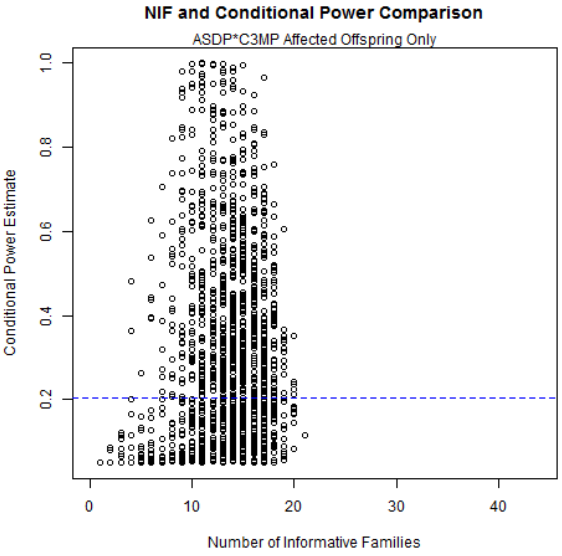
When considering the distribution of the conditional power estimates, it appears that increasing the number of probands is a far more efficient way to improve the conditional power for rare diseases like ASD, and through that the facial-feature clusters. The change in the distribution when moving from 10 or 13 probands to 21 is prominent, while the change due to including 71 unaffected offspring is negligible. This can also be seen in Figure 5.2, where the conditional power estimate is plotted against the number of informative families for each marker. Incorporating unaffected offspring when focusing on the ASDP$*$C2MP resulted in a median increase of 19 informative families which corresponds to a median percent increase of 233%. One can notice the density of markers whose conditional power estimate is larger than 0.7 is greater for the ASDP$*$C3MP than the others. Thus, solely from the perspective of enhancing statistical power through enlarging the sample size, clustering the facial features of and genotyping additional affected offspring will be most effective. Note that the meaning of affected offspring depends on which phenotype, ASDP$*$CxMP for $x = 1, 2, 3$, is being considered.



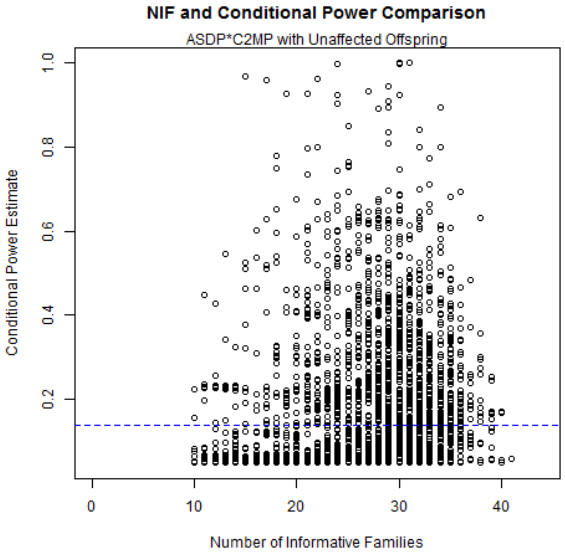
(a) Conditional Power Estimate vs. the NIF for the ASDP∗C1MP via the exact TDT.



(b) Conditional Power Estimate vs. the NIF for the ASDP∗C2MP via the exact TDT.



(c) Conditional Power Estimate vs. the NIF for the ASDP∗C3MP via the exact TDT.



(d) Conditional Power Estimate vs. the NIF for the ASDP∗C2MP via the FBAT.

Figure 5.2. Conditional Power Estimate vs. Number of Informative Families (NIF). The conditional power estimates were computed using PBAT and the method introduced by Murphy et al. [37, 39]. The sample median conditional power estimate is denoted with the blue dashed horizontal line.

6. DISCUSSION AND CONCLUSIONS

We conclude by summarizing our findings and then outlining our recommendations for future work.

6.1. SUMMARY

Because of the heterogeneous nature of phenotypes for individuals with autism spectrum disorder (ASD), researchers have attempted to identify homogenous subgroups with statistically different behavioral and clinical phenotypes between groups. One method to identify subgroups has been to cluster individuals based on their facial features, noting that facial and brain development are intertwined. This, combined with the fact that ASD is inherently a genetic disease, provides motivation to identify the genetic underpinnings specific to the facial-feature clusters of boys with ASD.

We attempted to identify genetic markers, in particular SNPs, that align with the combined ASD and Cluster Membership Phenotypes (CMPs) using the family-based association test (FBAT). In a specific setting that we consider, the FBAT is often referred to as the transmission disequilibrium test (TDT). The FBAT is an advantageous option: it avoids misleading associations by requiring the presence of both association and linkage, and it is more powerful than its population-based counterpart, especially in the trio setting, for relatively uncommon diseases like ASD [9, p. 7]. We conducted "within-cluster" analyses that utilized affected offspring and parents for each of the three facial-feature clusters identified by Obafemi et al. [8] using 2828 SNPs from high-confidence and strong-candidate ASD genes. Then, we argued that focusing on Cluster 2 alone is natural and carried out its analysis using a modified test statistic that incorporated information from unaffected

offspring. In all four analyses, there was not enough evidence to conclude that any marker was linked and associated with a causal DSL of the corresponding phenotype.

The data available included information on only 44 prepubescent boys with ASD and their immediate family, which was reduced when incorporating CMPs. Therefore, we implemented a screening method in order to address the low power resulting from this small sample size. Screening first involved estimating the conditional power of being able to detect linkage and association with a DSL when it is truly present. Then we systematically allotted larger portions of the overall significance level to the markers with higher conditional power estimates. This method has been shown to increase the power of the overall testing procedure. However, for our data, the overall conclusion remained unchanged compared to implementing the standard multiple-testing correction for these studies. Hence, we were unable to say any of the 2828 SNPs considered are linked and associated with a DSL involving ASD and the CMPs. Next, we discuss possible future work taking into account these results.

6.2. FUTURE WORK

The drawback specific to this study is low power due to small sample sizes, but let's consider what is the best measure of sample size for this research setting. We have mentioned two different ways that sample size can be measured in family-based association studies. For the TDT, which includes only affected offspring, it is taken to be the number of transmissions from heterozygous parents; this becomes the number of heterozygous parents when trios are considered. For the more general FBAT, the sample size was recorded as the number of informative families (NIF). When parental genotypes are known, the NIF is the number of families that have at least one heterozygous parent. However, as seen in Section 5.5, increasing the NIF for rare diseases essentially does not increase power when

many of the families consist of only unaffected offspring. Hence, we argue that the best measure of sample size for the ASD facial-feature cluster setting is actually that under the TDT setting: the number of transmissions from heterozygous parents to affected offspring.

With this precise choice for the meaning of sample size, we can say that increasing the sample size should result in more substantial gains in power. Therefore, in order to have more acceptable power levels, one should recruit more probands (prepubescent boys with ASD), determine the facial-feature cluster to which they belong, and genotype the proband and both his parents. We do not feel that genotyping unaffected offspring is worth the extra cost and time given that their contribution does not substantially impact the power of the FBAT in this setting. If one is interested in restricting the focus to Cluster 2 as we did, then "affected" means that the boy must have autism *and* facial features that align with those characteristic of Cluster 2. Note, we are not implying that increasing the sample size, and through that the power, will guarantee that a marker considered would will be declared significantly linked and associated with a DSL while controlling the FWER at 0.05; it will only raise the probability of detecting this when it is actually the case.

Given that the TDT is more appropriate for the ASD-facial-cluster setting, we can refer to the literature on sample size calculations for the TDT, which is nonexistent for the more general FBAT. These calculations depend on the characteristics of the disease and marker, the desired power (typically 0.8), and significance level, which changes with the number of markers considered. Some examples of disease and marker characteristics include the true risk ratio, minor allele frequency, and mode of inheritance. In the best case scenario, there is a large risk ratio, a MAF near 0.5, and an additive mode of inheritance. For this, with a power of 0.8 and a significance level typical of GWAS, the number of trios required is on the scale of hundreds [42]. But for realistic risk ratio values, more than a thousand trios are needed [43]. However, in order to obtain a more specific estimate

of the required number of trios, one would need to consult with medical practitioners for reasonable estimates of values related to the disease and marker. Note that these ranges were computed assuming a GWAS, which includes many more markers than the 2828 considered here. Nonetheless, realistically, we would still need hundreds of trios. When focusing on the C2MP, we only had thirteen families, so we would need to increase the number of trios by at least an order of magnitude.

One potential alternative to the TDT for the current data set is a sequential probability ratio test (SPRT). There are three possible conclusions for a sequential test: accept H_0, accept H_A, or continue sampling. The third conclusion means that there is not enough evidence to accept one of the hypotheses given the sample size, which is treated as a random variable. Its form analogous to the TDT requires fewer trios than the TDT; e.g. a minimum of 70 are required to obtain a power of at least 0.8 for "highly associated SNPs" [44, p. 918]. However, the investigator must choose a specific value for the alternative hypothesis and the desired levels of Type I and II error control. Therefore, the conclusions only apply to those choices, which is quite limiting. Additionally, the alternative hypothesis for SPRT differs from that of the TDT, and the TDT is more powerful if the linkage and association between the marker and a DSL is moderate. Moreover, there is no software implementation of this method [44]. For the ASD-facial-cluster setting, the SPRT could help with power, but the current sample sizes may still be deemed insufficient even if the implementation difficulty and excessive specificity of the test are overcome.

The most promising future research path that moves away from the TDT involves testing for association at a higher level. There exist population-based tests that utilize the gene as the "primary unit of analysis" instead of a single marker. This higher-level approach is typically more powerful, because it combines possibly weak associations from multiple markers within a gene. These gene-based tests are particularly useful for complex

diseases [45, p. 343]. However, in the published literature, we could not find a test for family-based association studies that utilizes gene-level information. Thus, the development of such an approach would be valuable for the complex disease involving ASD and facial-feature cluster membership.

BIBLIOGRAPHY

[1] National Institute of Mental Health. Autism spectrum disorder. Online, March 2016.

[2] Jeremy Veenstra-VanderWeele, Susan L Christian, and Edwin H Cook, Jr. Autism as a paradigmatic complex genetic disorder. *Annu. Rev. Genomics Hum. Genet.*, 5:379–405, 2004.

[3] Nicholas Wade. In the genome race, the sequel is personal. *The New York Times*, September 2007.

[4] Cathryn M Lewis and Jo Knight. Introduction to genetic association studies. *Cold Spring Harbor Protocols*, 2012(3):pdb–top068163, 2012.

[5] Kristina Aldridge, Ian D George, Kimberly K Cole, Jordan R Austin, T Nicole Takahashi, Ye Duan, and Judith H Miles. Facial phenotypes in subgroups of prepubertal boys with autism spectrum disorders are correlated with clinical phenotypes. *Molecular Autism*, 2(1):1, 2011.

[6] Tayo Obafemi-Ajayi, Gayla Olbricht, Cynthia Germeroth, Luke Settles, T Nicole Takahashi, Judith Miles, and Donald Wunsch. Genetic variant analysis of facially delineated clusters of boys with autism spectrum disorders using family-based association testing. 2016. Manuscript submitted for publication in Molecular Autism.

[7] Margaret L Bauman and Thomas L Kemper. Neuroanatomic observations of the brain in autism: a review and future directions. *International journal of developmental neuroscience*, 23(2):183–187, 2005.

[8] Tayo Obafemi-Ajayi, Judith H Miles, T Nicole Takahashi, Wenchuan Qi, Kristina Aldridge, Minqi Zhang, Shi-Qing Xin, Ying He, and Ye Duan. Facial structure analysis separates autism spectrum disorders into meaningful clinical subgroups. *Journal of autism and developmental disorders*, 45(5):1302–1317, 2015.

[9] Nan M Laird and Christoph Lange. *The fundamentals of modern statistical genetics*. Springer Science & Business Media, 2011.

[10] 23andMe. What is the difference between genotyping and sequencing?, 2016.

[11] Nan M Laird and Christoph Lange. Family-based designs in the age of large-scale gene-association studies. *Nature Reviews Genetics*, 7(5):385–394, 2006.

[12] Wessler Griffiths, Anthony J.F. *Introduction to Genetic Analysis*. W.H. Freeman and Company, New York, 8 edition, 2005.

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

[14] Gourab De, Steve Horvath, Nan Laird, Steve Lake, Christoph Lange, Kristel Van Steen, Lin Wang, Wai-Ki Yip, Xin Xu, and Jin Zhou. *FBAT User's Manual*, July 2013.

[15] Affymetrix. How affymetrix genechip dna microarrays work, 2004.

[16] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

[17] Carl A Anderson, Fredrik H Pettersson, Geraldine M Clarke, Lon R Cardon, Andrew P Morris, and Krina T Zondervan. Data quality control in genetic case-control association studies. *Nature protocols*, 5(9):1564–1573, 2010.

[18] Simons Foundation Autism Research Initiative. Gene scoring criteria v3.0, October 2013.

[19] John H McDonald. *Handbook of Biological Statistics*. Sparky House Publishing, Baltimore, Maryland, U.S.A., third edition, 2014.

[20] Wayne W. Daniel and Chad L. Cross. *Biostatistics: A Foundation for Analysis in the Health Sciences*. Wiley, 10 edition, 2013.

[21] Dmitry Panchenko. Pearson's theorem, 2003.

[22] W.J. Conover. *Practical Nonparametric Statistics*. John Wiley & Sons, 3rd edition, 1999.

[23] Joseph L. Fleiss. *Statistical Methods for Rates and Proportions*. John Wiley & Sons, second edition, 1981.

[24] Susan Telke. Chi-square and mcnemar's test. Lecture Notes, November 2011. PubH6414: Biostatistical Methods I.

[25] Sylvia Wassertheil-Smoller and Jordan Smoller. *Biostatistics and epidemiology: a primer for health and biomedical professionals*. Springer, fourth edition, 2015.

[26] NCSS. Two correlated proportions (mcnemar test), 2014.

[27] Juliet Popper Shaffer. Multiple hypothesis testing. *Annual review of psychology*, 46:561, 1995.

[28] Alessio Farcomeni. A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical Methods in Medical Research*, 2007.

[29] Nan M Laird and Christoph Lange. Family-based methods for linkage and association analysis. *Advances in genetics*, 60:219–252, 2008.

[30] Steven J Schrodi and Hywel B Jones. Calculating exact p-values from the mcnemar transmission/disequilibrium test statistic. *Journal of Investigative Genomics*, 2, September 2015.

[31] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[32] Duncan J Murdoch, Yu-Ling Tsai, and James Adcock. P-values are random variables. *The American Statistician*, 2012.

[33] Christoph Lange and Nan M Laird. Power calculations for a general class of family-based association tests: dichotomous traits. *The American Journal of Human Genetics*, 71(3):575–584, 2002.

[34] Kathryn L Lunetta, Stephen V Faraone, Joseph Biederman, and Nan M Laird. Family-based tests of association and linkage that use unaffected sibs, covariates, and interactions. *The American Journal of Human Genetics*, 66(2):605–614, 2000.

[35] Christoph Lange and Nan M Laird. On a general class of conditional tests for family-based association studies in genetics: The asymptotic distribution, the conditional power, and optimality considerations. *Genetic epidemiology*, 23(2):165–180, 2002.

[36] Benjamin Zablotsky, Lindsey I Black, Matthew J Maenner, Laura A Schieve, and Stephen J Blumberg. Estimated prevalence of autism and other developmental disabilities following questionnaire changes in the 2014 national health interview survey. *National health statistics reports*, (87):1–21, 2015.

[37] Amy Murphy, Scott T Weiss, and Christoph Lange. Screening and replication using the same data set: testing strategies for family-based studies in which all probands are affected. *PLoS Genet*, 4(9):e1000197, 2008.

[38] Iuliana Ionita-Laza, Matthew B McQueen, Nan M Laird, and Christoph Lange. Genomewide weighted hypothesis testing in family-based association studies, with an application to a 100k scan. *The American Journal of Human Genetics*, 81(3):607–614, 2007.

[39] Kristel Van Steen, Matthew B McQueen, Alan Herbert, Benjamin Raby, Helen Lyon, Dawn L DeMeo, Amy Murphy, Jessica Su, Soma Datta, Carsten Rosenow, et al. Genomic screening and replication using the same data set in family-based association testing. *Nature genetics*, 37(7):683–691, 2005.

[40] Kristel Van Steen and Christoph Lange. Pbat: a comprehensive software package for genome-wide association analysis of complex family-based studies. *Human genomics*, 2(1):1, 2005.

[41] Jacob Cohen. A power primer. *Psychological bulletin*, 112(1):155, 1992.

[42] Michael Knapp. A note on power approximations for the transmission/disequilibrium test. *The American Journal of Human Genetics*, 64(4):1177–1185, 1999.

[43] Christoph Neumann, Margaret A Taub, Samuel G Younkin, Terri H Beaty, Ingo Ruczinski, and Holger Schwender. Analytic power and sample size calculation for the genotypic transmission/disequilibrium test in case-parent trio studies. *Biometrical Journal*, 56(6):1076–1092, 2014.

[44] Ozlem Ilk, Farid Rajabli, Dilay Ciglidag Dungul, Hilal Ozdag, and Hakki Gokhan Ilk. A novel approach for small sample size family-based association studies: sequential tests. *European Journal of Human Genetics*, 19(8):915–920, 2011.

[45] Pak C Sham and Shaun M Purcell. Statistical power and significance testing in large-scale genetic studies. *Nature Reviews Genetics*, 15(5):335–346, 2014.

VITA

In May 2015, Luke Settles graduated from Southern Illinois University Edwardsville with two degrees: B.S. in Mathematical Studies with an Applied Mathematics Specialization and B.A. in Foreign Languages and Literature with a Spanish Specialization. He earned a Master of Science in Applied Mathematics with a Statistics Emphasis from Missouri University of Science and Technology in May 2017.