STATISTICAL ISSUES IN FAMILY-BASED GENETIC ASSOCIATION STUDIES WITH APPLICATION TO CONGENITAL HEART DEFECTS IN DOWN SYNDROME

by

Yan Lin

MS in Biostatistics, University of Pittsburgh, 2003

PhD in Biology,University of Michigan, 2001

BS in Biology, Beijing Normal University, 1994

Submitted to the Graduate Faculty of

the Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2007

UNIVERSITY OF PITTSBURGH

Graduate School of Public Health

This dissertation was presented

by

Yan Lin

It was defended on

May 4th , 2007

and approved by

Dissertation Advisor:
Eleanor Feingold, Ph.D.
Associate Professor
Department of Human Genetics
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Committee Member:
Chien-Cheng (George) Tseng, Sc.D.
Assistant Professor
Department of Biostatistics
Department of Human Genetics
Graduate School of Public Health
University of Pittsburgh

Committee Member:
Daniel E. Weeks, Ph.D.
Professor
Department of Human Genetics
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Committee Member:
Lisa Weissfeld, Ph.D.
Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

STATISTICAL ISSUES IN FAMILY-BASED GENETIC ASSOCIATION STUDIES WITH APPLICATION TO CONGENITAL HEART DEFECTS IN DOWN SYNDROME

Yan Lin, PhD

University of Pittsburgh, 2007

This dissertation is motivated by data generated from a genetic association study of congenital heart defects in Down syndrome (DS). Congenital heart defects are among the most common abnormalities seen at birth. The genetic basis for most congenital heart defects is unknown. One severe form of congenital heart defect, atrioventricular septal defect (AVSD), is highly associated with DS. This makes the DS population a useful tool for discovering of genes that are associated with this specific form of congenital heart defect. Discovering genes that influence risk of AVSD will lead to a better understanding of heart development and of the etiology of these defects. This in turn can lead eventually to improved public health through better screening, prevention, and treatment strategies.

Family trios were collected for the Down syndrome heart study. This dissertation discusses statistical issues raised in genetic association studies using family trio data, including the genotype calling problem (i.e. how to generate genotype data from the raw data produced by high-throughput SNP arrays) and analysis strategies. Although the motivating dataset involves trisomic individuals, we developed statistical methods both for disomic and trisomic data.

For the genotype-calling problem, we generated two genotype calling methods specifically for disomic family trio data. The first method is an ad-hoc modification of the K-means clustering algorithm that incorporates family information. The second is a likelihood-based method that combines the mixture model approach with a pedigree likelihood. These two

methods out-performed existing methods, which ignore the family information, both in simulation studies and a real data analysis. We also extended these two methods to trisomic trio data.

With regard to analysis strategies, we discussed alternative analysis methods for trio designs, particularly for the combination of case trios and control trios that we have in the Down syndrome data. We derived likelihood models that help explain the differences among some published methods. We also proposed an extension of a combined likelihood-based method proposed by Epstein and others for analysis of case trios plus independent controls to our design of case and control trios.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

First I would like to express my profound gratitude to my advisor Professor Eleanor Feingold for her guidance, persistent support, encouragement, and patience during past few years. She introduced me to this exciting project and made tremendous effort in guiding me through my thesis research. Without her invaluable guidance and support, I would not have completed my thesis.

Also I would like to thank Professors George Tseng, Daniel Weeks, and Lisa Weissfeld for precious instruction on statistics and genetics during courses, fruitful discussion and valuable inputs on my research.

Id like to thank Professors Lora Bean and Stephanie Sherman at Emory University for providing the AVSD dataset and valuable input to my project.

I am grateful of the Departments of Biostatistics and Human Genetics for the education and generous support during the past few years. Also I would like to thank the students, post docs and faculties of the statistical genetics group. The journal club and the seminars organized by this group are very interesting and helpful to me.

I am grateful for my parents and family. Thank you for being with me all the time.

1.0 INTRODUCTION

1.1 GENETIC ASSOCIATION STUDIES

There are two major approaches for mapping genes that are associated with human disease, linkage analysis and association analysis. Allelic association refers to the increase or decrease of a specific marker allele frequency in individuals with a disease trait. There are two types of association studies: (1) population-based case-control studies and (2) family-based association studies. Case-control studies compare allelic frequencies in cases and controls. The control population needs to be matched to case population with respect to all factors (such as ethnicity, age and sex) that might have an effect with the outcome. An unique problem of the case-control genetic association studies is that spurious association may occur because of population stratification. Population stratification refers to the situation in which multiple population subtypes are hidden in a population that appears to be homogeneous. Family-based association studies use the parents' (or other family members') genotypes as surrogate controls. Therefore, they are robust to the population stratification. The classic family-based association design is the trio design. A trio is a nuclear family with one offspring. The transmission disequilibrium test (TDT) is designed for analysis of trio designs (see Chapter 4 for a detailed description of the TDT test). My dissertation deals with problems raised in the family-based association studies, with application to a study of congenital heart defects (CHD) in Down syndrome (DS).

1.2 DOWN SYNDROME AND CONGENITAL HEART DEFECTS

A normal human cell has 46 chromosomes. That is, it has two copies of each chromosome (*disomic*). However, occasionally, there is a mishap, called nondisjunction, in which the members of a pair of homologous chromosomes do not move apart properly during meiosis I, or in which sister chromatids fail to separate during meiosis II. In these cases, one of the gametes then receives two copies of that chromosome instead of one (Figure 1.1).

If this gamete unites with a normal one at fertilization, the offspring will have an extra copy of that chromosome. This offspring then becomes *trisomic* for that chromosome. Over 95% of the DS cases are caused by trisomy 21 (*i.e.*, the presence of an extra copy of chromosome 21). In addition to mental retardation, 44% of all DS individuals also have some form of CHD (Freeman et al., 1998). The most severe form of CHD observed is atrioventricular septal defect (AVSD). There are two forms of AVSD, complete and partial. The complete form of AVSD is most often associated with DS. To further understand the etiology of CHD, and gain insights into aspects of human heart development, an association study of CHD in DS population is being conducted by our collaborators at Emory University in Atlanta. The cases are defined as trisomy 21 individuals with confirmed cases of complete AVSD. The controls are trisomy 21 individuals with no major associated heart defect. The cases and controls for this study were ascertained from two primary sources. The majority were selected from the larger study of live births with trisomy born and ascertained in the five county metropolitan area of Atlanta, Georgia (Freeman et al., 1998). Others were ascertained from individuals who attended the DS clinic at Kennedy Krieger in Baltimore, MD. Blood samples were collected from the mother, the father and the DS offspring. Currently, we have data available for 136 case trios (nuclear families with one offspring) and 126 control trios. Genotype data generated using the Sequenom platform are available for 25 single nucleotide polymorphisms (SNPs). We also have received genotype data of approximately 400 SNPs generated by the Illumina platform. This dissertation contains statistical methods we developed to address several problems raised in this study. We also discuss practical strategies for analysis of family trio designs.

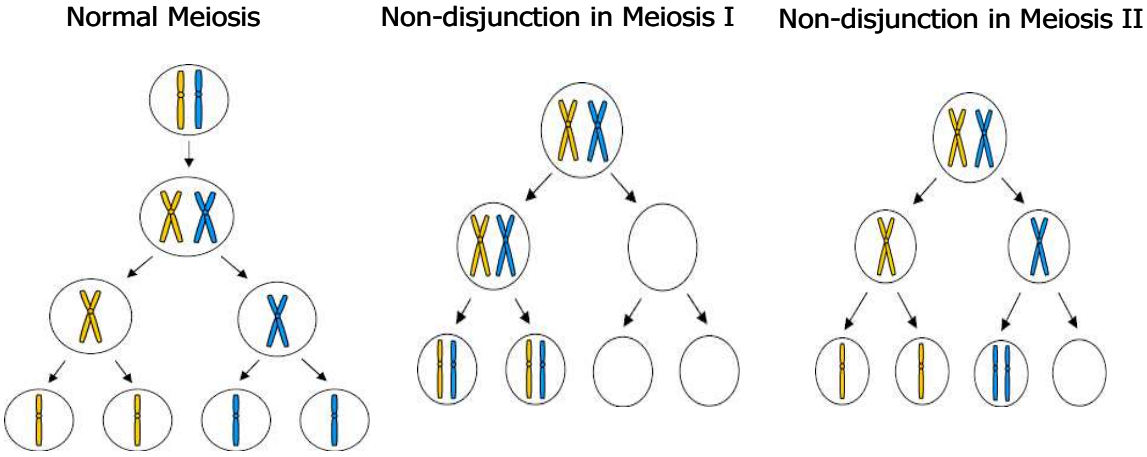

Figure 1.1: Schematic Drawing of Meiosis

1.3 A MODEL FOR TRISOMIC TRIOS

Xu and others proposed a basic model for genotype data of trisomic trios (Xu et al., 2004). We base our likelihood-based methods for trisomic data, including the model-based genotype calling methods (see Chapter 3) and the two-marker trisomic TDT test(see Chapter 5), on this model. The following is a brief description of the model.

Assume a SNP marker with two alleles marked A and B. There are nine possible mating types (*i.e.*, different combinations of parental genotypes) as shown in Table 3.1 in Chapter 3. The nondisjoining parent (NDJP) is the parent that contributes two copies of the chromosomes and the correctly dsijoining parent (CDJP) is the parent that contributes only one copy of the chromosome. Because only a small portion of the trisomic conceptuses survive to term, we can only observe the disease status of these trisomic individuals. Hence it is impossible to separate the two events, survival to term and affected with the disease. Therefore, the association parameters in the model are defined as the following,

$$
\begin{aligned}
w_0 &= \text{probability of survival and affectedness of a conceptus with genotype } AAA, \\
w_1 &= \text{probability of survival and affectedness of a conceptus with genotype } AAB, \\
w_2 &= \text{probability of survival and affectedness of a conceptus with genotype } ABB, \\
w_3 &= \text{probability of survival and affectedness of a conceptus with genotype } BBB.
\end{aligned}
$$

The map parameter used in this model is h, which is defined as the probability that the two chromosomes contributed by the NDJP are reduced to homozygosity (duplicates of the same parental chromosome). Given the parental data, the probability of a living diseased trisomic offspring's genotype depends only on the h and the w's. For example, for mating type AB (NDJP) $\times$ AA (CDJP), the CDJP must contribute an A. If h=0, *i.e.*, the two chromosomes are not reduced to homozygosity, then the NDJP contribute AB; if $h = 1$, *i.e.*, the two chromosomes are reduced to homozygosity, then NDJP contributes either AA

or BB, with half of a chance each. Therefore, given the parental genotypes shown above,

$$\begin{aligned}
&Pr(\text{the diseased child is } AAA\text{—NDJP is } AB \text{ and CDJP is } AA)\\
&= Pr(\text{NDJP contributes } AA) \times Pr(\text{the child survives to term and is diseased}|\text{the child is } AAA)\\
&= \frac{h}{2} \times w_0.
\end{aligned}$$

Similarly,

$$\begin{aligned}
&Pr(\text{the diseased child is } AAB\text{—NDJP is } AB \text{ and CDJP is } AA)\\
&= Pr(\text{NDJP contributes } AB) \times Pr(\text{the child survives to term and is diseased}|\text{the child is } AAB)\\
&= \frac{1-h}{2} \times w_1,
\end{aligned}$$

and

$$\begin{aligned}
&Pr(\text{the diseased child is } ABB\text{—NDJP is } AB \text{ and CDJP is } AA)\\
&= Pr(\text{NDJP contributes } BB) \times Pr(\text{the child survives to term and is diseased}|\text{the child is } ABB)\\
&= \frac{h}{2} \times w_2.
\end{aligned}$$

These probabilities are normalized so that they add up to 1 for each mating type, and listed in the 5^{th} column of Table 3.1. This example is the fourth mating type shown in Table 3.1.

1.4 OVERVIEW OF PROBLEMS CONSIDERED IN THIS DISSERTATION.

As described above, the motivating dataset for this dissertation is the combined sample of case and control trios generated from the CHD study. The current goal of this study is to test the association of AVSD with candidate genes in these trisomic trios using SNP markers. The offspring in our dataset are trisomy 21 individuals. However, this dissertation addresses the problems for both disomic and trisomic family trio data.

After the data are collected, the following steps are followed in the analysis of family-based association studies.

1. Genotype calling, i.e. to generate genotype data from the SNP array raw data. This is usually done by using a clustering method. Here we have not only compared several different commonly-used genotype-calling methods but also developed two family-based genotype calling methods for both disomic and trisomic trio data. These results are presented in Chapters 2 and 3.
2. Test for association. Several questions arise related to the analysis of the family-based association studies.
   a. What are the choices that we have for the analysis of our family trio design?
   b. What questions do we have about how the methods compare?
   c. What new methods need to be developed?

   We discuss these questions in Chapters 4 and 5.

1.4.1 Genotype calling methods for family trio data

In Chapter 2, we develop two new family-based genotype calling methods for SNP array data of disomic trios. Furthermore, we discuss the impacts of certain features of the genotype calling problem on the performance as compared to the methods. Our family-based methods showed much improved performances than other methods that ignore the family information. In Chapter 3, we extend these methods to the trisomic trios.

1.4.2 Alternative analysis of family trio designs

In Chapters 4 and 5 we discuss the available methods for analysis of trio designs. We focus on the combined designs in which both case trios and controls (either independent controls or control trios) are available. We compare different analysis strategies under each trio design. We also discuss the question of whether it is sensible to collect trios at all any more.

2.0 SMARTER CLUSTERING METHODS FOR HIGH-THROUGHPUT SNP GENOTYPE CALLING

Manuscript submitted to Biostatistics.

Yan Lin[1], George C. Tseng[1,2], Lora J.H. Bean[3],

Stephanie L. Sherman[3], Eleanor Feingold[1,2,]

1. Department of Biostatistics, University of Pittsburgh

2. Department of Human Genetics, University of Pittsburgh

3. Department of Human Genetics, Emory University

Email: feingold@pitt.edu

Telephone: (412)383-8599

Fax: (412)624-3020

2.1 ABSTRACT

Many high-throughput genotyping technologies for single nucleotide polymorphism (SNP) markers have been developed. Most use clustering methods to "call" the SNP genotypes, but standard clustering methods are not optimal in distinguishing the genotype clusters of a SNP because they do not take advantage of a number of specific features of the genotype calling problem. In particular, when family data are available, pedigree information is ignored. Furthermore, prior information about the distribution of the measurements for each cluster can be used to choose an appropriate model-based clustering method and can significantly improve the genotype calls. In this paper, we discuss the impact of incorporating external information into clustering algorithms to call the genotypes. We also propose two new methods to call genotypes using family data. The first method is a modification of the K-means method which uses the family information by updating all members of a family together. The second method is a likelihood-based method that combines the Gaussian or beta mixture model with pedigree information. We compare the performance of these two methods and some other existing methods using simulation studies. We also compare the performance of these methods on a real dataset generated by the Illumina platform (www.illumina.com).

2.2 INTRODUCTION

Single nucleotide polymorphisms (SNPs) are DNA sequence variations that occur when a single nucleotide (A, T, C, or G) in the genome sequence is altered. They are becoming the most popular type of marker used in genome-wide linkage and association studies to discover genes relevant to disease. The vast majority of SNPs are biallelic. Consider a SNP marker with two alleles A and B. There are three possible genotypes for a disomic individual, AA, AB and BB. Many high-throughput technologies have been developed to genotype the SNPs efficiently, including the GeneChip Human Mapping Array from Affymetrix, the Illumina platform, the Sequenom platform, and the Invader assay. Each platform uses a different

technology, and they give somewhat different forms of data. In general, they all give certain quantitative measures of allelic abundance for the two alleles, y_A and y_B. The abundance measures can either be scalars or vectors. Individuals with genotype AA are expected to have high y_A value and low y_B value. The opposite is expected for individuals with genotype BB. Those with genotype AB are expected to have similar y_A and y_B values. Figure 2.1A gives an example of data generated for a SNP marker using the Illumina platform. Each dot on the plot represents one individual. In SNP genotyping, we seek to identify genotype clusters based on these measurements and "call" each person's genotype by assigning them to a cluster. Normally we expect to find 3 clusters, but if one allele is rare in the population, a particular dataset might only have 2 clusters (genotypes).

Different platforms generate data of different dimension. For example, the Affymetrix SNP array generates raw data of very high dimension. Each SNP is assessed by 20 probe pairs. The Illumina platform generates data of 2 dimensions. For data generated by the platforms which produce high dimensional data, dimension reduction is typically done before the genotype calling procedure, usually reducing the data to 2 dimensions, as shown in Figure 2.1A, or 1 dimension, as shown in Figure 2.1B. The method used for dimension reduction is platform specific. We use the transformation (Intensity of allele B)/(Intensity of allele A+Intensity of allele B) to achieve a 1-dimensional summary from the 2-dimensional data generated by the Illumina platform. The second dimension (distance from the origin) mostly contains information on data quality. It is common to exclude points very close to the origin before clustering.

Genotypes are typically assigned ("called") from raw data using clustering methods. But most of the commonly-used clustering methods, such as the K-means clustering method, do not incorporate any special information about the genotyping problem. However, there are several features of genotyping problem that could potentially facilitate the process. First of all, the number of clusters is limited (1, 2, or 3 for standard disomic data). Second, there is typically known prior knowledge of the distribution of the data points from previous use of the technology. These distributions are platform-specific, however, in general, the heterozygote cluster almost always has higher variation than the homozygote clusters and the homozygote clusters are sometimes highly skewed (Figure 2.1). Finally, when working with family data,

prior information or constraints about the genotypes are also available. Transmission of alleles from the parents to the offspring must follow Mendelian rules. Proper use of these various types of prior knowledge can greatly increase the accuracy of the genotype calls.

Both supervised and unsupervised methods have been used for genotype calling. When training datasets are available, supervised clustering algorithms can be used. This is the case for the modified partitioning around the medoid (MPAM) algorithm developed for the GeneChip Human Mapping 10K array by Affymetrix (Liu et al., 2003) and the robust linear model with the Mahalanobis distance classifier (RLMM) developed by Rabbee and Speed (Rabbee and Speed 2005). In most cases, appropriate training datasets are not available. An unsupervised clustering algorithm is then needed in genotype calling. Many platforms use the K-means clustering algorithm to call the genotypes. When the clusters are reasonably separated, the K-means clustering algorithm can give satisfactory results. However, the K-means method is not always effective, especially when the variances for each group differ (Fujisawa et al., 2004). The dynamic model-based algorithm (DM) is an ad-hoc method developed by Affymetrix for their 100k array (Di et al., 2005). DM is generally accurate, but exhibits higher error rates for heterozygous bases than for homozygous bases (Rabbee and Speed 2005). Fujisawa et al. proposed a Gaussian-mixture model for data generated with the Invader assay (Fujisawa et al., 2004). Unlike the K-means clustering algorithm, which requires knowing the number of clusters beforehand, the use of a penalized likelihood in their method performs well in selecting the number of clusters. More importantly, the K-means algorithm is equivalent to maximizing the classification likelihood under a mixture of Gaussian models with equal variances for all clusters (Celeux and Govaert, 1992). The Gaussian-mixture model estimates variances for each cluster separately, which should be better for genotyping, since different genotype clusters usually have dramatically different variances.

All the methods described above are designed for independent samples. When family data are available, the prior constraints on the genotype can play an important role in genotype calling. Sabatti and Lange developed a method for family data collected for linkage studies (Sabatti and Lange, 2005). They combined a Gaussian-mixture penetrance model with the pedigree likelihood. The empirical Bayesian method was used so that information across all

SNPs could be borrowed in parameter estimation. The general idea of this method could be applied to any platform. However, they developed this method for the high-dimensional data of the Affymetrix GeneChip Human Mapping Array. In addition, this method is designed for data collected from linkage studies, and makes assumptions about allele and genotype frequencies (e.g. Hardy-Weinberg Equilibrium (HWE)) that may not be appropriate for all applications.

In this paper, we illustrate the improvement that can be made by taking better advantage of the special features of the genotype calling problem, such as prior information about cluster distributions and family genotype constraints. We apply our methods to both simulated and real data. The real data are genotypes on 262 trios (parents and a child). The data were generated by the Illumina platform (www.illumina.com). Subjects were recruited from the Atlanta, Georgia metropolitan area and from the Down syndrome clinic at Kennedy Kreiger in Baltimore, MD as described in detail by Kerstann et al.(2004). Additional families were recruited from the Sibley Heart Center, Cardiology, Children's Healthcare of Atlanta.

2.3 METHODS

2.3.1 K-means methods for trio data

The K-means clustering method (Hartigan and Wong 1979) is one of the popular methods used in genotype calling. It is fast, straightforward, and fairly effective. Here we propose to modify the K-means algorithm so that family information can be used to improve the accuracy of the genotypes called in trios. Our method could be extended to nuclear families, but would probably not work well for large pedigrees. We refer to this modified K-means method as "trio K-means" in this paper. The method uses the following iterative procedure,

Step 1: Start with a set of initial centroids.

The initial centroids are $\{C_{AA}^{(0)}, C_{AB}^{(0)}, C_{BB}^{(0)}\}$.

Step 2: At the k+1 step, update all three observations in a family as described in the following.

Assume we have two alleles, A and B. For disomic family trios, there are 15 possible genotype combinations that agree with Mendelian segregation rules (Table 2.1).

Let g_1, g_2 and g_3 be the possible genotypes for parent 1, parent 2 and their child. For all combinations of g_1, g_2 and g_3 shown in Table 2.1, we calculate $D_{g_1,g_2,g_3} = d(x_1, C_{g_1}^{(k)}) + d(x_2, C_{g_2}^{(k)}) + d(x_3, C_{g_3}^{(k)})$, where $C_g^{(k)}$'s are the estimated group centers from the k^{th} step and $d(x_i, C_j^{(k)})$ is the squared Euclidean distance between the observed value x_i and the center for j^{th} genotype group, $C_j^{(k)}$. Family members are then assigned to the genotypes $\tilde{g_1}$, $\tilde{g_2}$ and $\tilde{g_3}$ that minimize D_{g_1,g_2,g_3}.

Step 3: Iterate until convergence.

Note that our trio K-means procedure assumes that all the family information is correct, and no Mendelian errors are acceptable (e.g. no sample switches or non-paternity or mutations). We discuss this assumption further in the discussion section of this paper. The trio K-means procedure is straightforward and does not make any assumption about the distribution of each genotype cluster, which makes it more robust to outliers than the model-based methods (see discussion). On the other hand, the trio K-means method does not use the information on the shape of the clusters. This makes it less efficient than the model-based methods. In the next section, we introduce a model-based method that integrates the pedigree information and the distribution information together to call the genotypes.

2.3.2 Model-based methods for trio data

In order to incorporate the pedigree information into the model-based clustering methods, we propose a genotype calling method that combines the pedigree likelihood and a parametric mixture model approach. This method is easily applicable to pedigrees of almost any size and configuration. Our likelihood is similar to that of Sabatti and Lange, but we do not use it in a Bayesian context, so we do not make any assumptions about allele or genotype frequencies. Moreover, we work only with 1-dimensional data, so our likelihood is applicable to data from any platform.

2.3.2.1 Likelihood Let y be the observed 1-dimensional value for an individual. We assume the following parametric penetrance model.

$$y|g = \lambda \sim f(y, \xi_\lambda) \tag{2.3.1}$$

where

$$\lambda \in \Lambda = \{AA, AB, BB\},$$

and ξ_λ is the parameter vector for genotype λ. $f(\xi_\lambda)$ could be any parametric model that fits the data well. In this paper, we illustrate the use of the Gaussian-mixture model and the beta-mixture model. We will refer these two methods as the trio Gaussian-mixture model and the trio beta-mixture model respectively.

Let $\mathbf{Y_i} = (y_{fi}, y_{mi}, y_{ki})$ be the observed data for the father, the mother, and the child of the i^{th} trio. Let $\mathbf{G_i} = (g_{fi}, g_{mi}, g_{ki})$ be the corresponding genotype vector. First, let us assume that we can observe the genotype vector $\mathbf{G_i}$. Then the likelihood for the i^{th} trio is:

$$\begin{aligned}
&L_i(Y_i, G_i, \theta) \\
&= Pr(g_{fi})Pr(g_{mi})Pr(g_{ki}|g_{fi}, g_{mi})Pr(y_{fi}|g_{fi})Pr(y_{mi}|g_{mi})Pr(y_{ki}|g_{ki}) \qquad (2.3.2)\\
&= \prod_{\lambda\in\Lambda} p_\lambda^{1\{g_{fi}=\lambda\}} p_\lambda^{1\{g_{mi}=\lambda\}} Pr(g_{ki}|g_{fi}, g_{mi}) f(y_{fi}, \xi_\lambda)^{1\{g_{fi}=\lambda\}} f(y_{mi}, \xi_\lambda)^{1\{g_{mi}=\lambda\}} f(y_{ki}, \xi_\lambda)^{1\{g_{ki}=\lambda\}},
\end{aligned}$$

where

$$\lambda \in \Lambda = \{AA, AB, BB\},$$

and

$$\theta = (p_\lambda\text{'s}, \xi_\lambda\text{'s})^T. \tag{2.3.3}$$

If we have a total of n trios, the full likelihood is

$$L(Y, \theta) = \prod_{i=1}^{n} L_i(Y_i, G_i, \theta). \tag{2.3.4}$$

If the parameters are known, then we can determine the genotypes of all three members of a family using Bayes' rule. The posterior probability of the family genotype vector $\mathbf{G} = (g_f, g_m, g_k)$ given the observed values $\mathbf{Y} = (y_f, y_m, y_k)$ is

$$p(G|Y) = \frac{p_{\lambda=g_f} p_{\lambda=gm} Pr(g_k|g_f, g_m) f(y_f, \xi_{\lambda=g_f}) f(y_m, \xi_{\lambda=g_m}) f(y_k, \xi_{\lambda=g_k})}{\sum_{j=1:15} p_{\lambda=g_{fj}} p_{\lambda=g_{mj}} Pr(g_{kj}|g_{fj}, g_{mj}) f(y_f, \xi_{\lambda=g_{fj}}) f(y_m, \xi_{\lambda=g_{mj}}) f(y_k, \xi_{\lambda=g_{kj}})} \quad (2.3.5)$$

where g_{fj}, g_{mj} and g_{kj} are the genotypes of the father, the mother and the child for the j^{th} family type listed in Table 2.1.

2.3.2.2 Estimation method If the Gaussian-mixture model is assumed for the penetrance term of the model, $f(y|g = \lambda)$, a convenient EM algorithm can be constructed to estimate the parameters. Here the parameter vector is

$$\theta_\lambda = (p_\lambda\text{'s}, \mu_\lambda\text{'s}, \sigma^2_\lambda\text{'s})^T.$$

The update algorithm is:

$$\begin{aligned}
p_\lambda^{(t+1)} &= \frac{E(S_{1,\lambda}|Y, \theta^{(t)})}{2n} \\
\mu_\lambda^{(t+1)} &= \frac{E(S_{2,\lambda}|Y, \theta^{(t)})}{E(S_{1,\lambda}|Y, \theta^{(t)}) + E(S_{4,\lambda}|Y, \theta^{(t)})} \\
\sigma_\lambda^{2(t+1)} &= \frac{E(S_{3,\lambda}|Y, \theta^{(t)})}{E(S_{1,\lambda}|Y, \theta^{(t)}) + E(S_{4,\lambda}|Y, \theta^{(t)})} - \left(\mu_\lambda^{(t+1)}\right)^2
\end{aligned}$$

where

$$\begin{aligned}
S_{1,\lambda} &= \sum_{i=1}^{n} (1\{g_{fi} = \lambda\} + 1\{g_{mi} = \lambda\}), \\
S_{2,\lambda} &= \sum_{i=1}^{n} \left[1\{g_{fi} = \lambda\} y_{fi} + 1\{g_{mi} = \lambda\} y_{mi} + 1\{g_{ki} = \lambda\} y_{ki}\right], \\
S_{3,\lambda} &= \sum_{i=1}^{n} \left[1\{g_{fi} = \lambda\} y_{fi}^2 + 1\{g_{mi} = \lambda\} y_{mi}^2 + 1\{g_{ki} = \lambda\} y_{ki}^2\right],
\end{aligned}$$

and

$$S_{4,\lambda} = \sum_{i=1}^{n} 1\{g_{ki} = \lambda\}.$$

If we assume a beta-mixture model for the penetrance term, the parameter vector becomes

$$\theta_\lambda = (p_\lambda\text{'s}, \alpha_\lambda\text{'s}, \beta_\lambda\text{'s})^T.$$

We can still use the same update algorithm for p_λ. For the estimation of α_λ's and β_λ's, we use the *nlm* package in *R* to maximize the $E(\log(L(Y, \theta)))$ at the M-step. The *nlm* algorithm converges fairly fast. Details on the estimation methods are shown in Appendix A.

2.3.2.3 Determination of the cluster number Fujisawa et al. proposed a mixture Gaussian approach in combination with the penalized likelihood, which performs well in selecting the number of clusters (Fujisawa et al., 2004). Here we took advantage of the fact that the number of the clusters is limited, and there are a limited number of configurations for missing clusters (e.g. we do not expect to have a missing middle cluster). We modified the EM algorithm so that when $p_\lambda^{(t)}$ is smaller than a preset small number x, then we consider the cluster empty from step t and up.

2.4 RESULTS

2.4.1 Simulation studies

We performed two simulation studies to compare the performance of different clustering algorithms. For each simulation study, we simulated 1000 datasets. Each dataset consisted of 150 trios. A beta distribution was used in simulating the observations in different genotype groups because we observed that the distribution of the homozygote clusters is highly skewed in most of the platforms with which we have experience. We applied six different clustering methods to these datasets. Three methods treat each individual independently: the K-means clustering method, the Gaussian-mixture model for independent data (we will refer to this method as the Gaussian-mixture model throughout the rest of this chapter) and the beta-mixture model for independent data (we will refer to this method as the beta-mixture model throughout the rest of this chapter). Three corresponding methods treat the family

as a group: the trio K-means method, the trio Gaussian-mixture model and the trio beta-mixture model.

The datasets we simulated in the first simulation study represent "good" data, because the genotype clusters are reasonably well separated. The distributions of the AA and the BB genotype clusters are highly skewed and the distribution of the AB genotype cluster is relatively more symmetric following our experience with SNP array data in general. Figure 2.2A is a plot of one of the "good" datasets simulated. We then simulated datasets that represent "bad" data. The genotype clusters are less well defined and the distributions of each cluster are wider compared to the "good" data. Figure 2.2B is an example of the "bad" datasets simulated.

The results for simulation study 1 are summarized in Table 2.2 and Figure 2.3. In general, when the three clusters are well separated, all six methods give reasonably good results. As expected, the methods that incorporate the family information consistently perform better than their counterpart methods that ignore the family information. On average, we made $20\% - 50\%$ fewer mistakes when we utilized the family information in our genotype calling process. In general, we would expect model-based methods to perform better than the k-means related methods, since they allow different variances to be estimated for each genotype cluster. However, the Gaussian models seem to perform worse than corresponding K-means method in simulation study 1. This is because in the good data, the distribution of the homozygote cluster is extremely skewed (Figure 2.2A). The Gaussian-mixture model simply does not fit the data well. The beta-mixture model seems to fit the data well, and performs the best.

Table 2.3 and Figure 2.4 summarize the results for simulation study 2. When the data quality is not good, external information can improve the genotype calls significantly. On average, 1/3 to 1/2 of the mistakes were avoided when using the family information. In this simulation study, the clusters are less well separated, and the homozygote clusters are less skewed. The Gaussian-mixture models perform better than the K-means methods on the bad data. Again, the trio beta-mixture model performs the best among the six methods compared. It may in some sense seem obvious that the beta model would perform better since our data were generated using the beta distribution. However, the beta model has not

previously been used for clustering of genotype data to our knowledge, despite the fact that most genotyping technologies produce a skewed intensity distribution for the homozygote clusters.

It is also worth noticing that in both simulation studies, the K-means and the trio K-means methods have a higher error rate in heterozygote individuals than in homozygote individuals. That is, they tend to call a true heterozygote individual as a homozygote. The opposite is true for the Gaussian-mixture models. The beta-mixture models, on the other hand, have similar error rates in both heterozygous and homozygous individuals (Tables 2.2 and 2.3). We also saw a similar pattern in the analysis of a real dataset generated by the Illumina platform (see below).

2.4.2 Real data example

We applied all of the methods described above to our real dataset. The subjects were genotyped using the BeadStation from Illumina Inc. (www.illumina.com). The dataset includes a total of 178 trios.

The clustering was done with one-dimensional data. The clustered data are restored into 2-dimensional data and shown in Figure 2.5. The data shown here are for one SNP that represents typical "good" data. Therefore, all six methods agree on the calls for most of the data points. However, we still observed improvements due to our methods. As we can see from Figure 2.5, the variances for the heterozygote clusters are much bigger than those for the homozygote clusters. As a result, the K-means clustering method tends to mistakenly assign some of the heterozygous individuals to the homozygous genotype cluster, as seen in Figure 2.5A. The presumed misclassified heterozygous individuals are circled. Two of these individuals' genotypes do not follow Mendelian rules, and were corrected by the trio K-means method. They are circled in Figure 2.5D. The homozygote clusters are highly skewed. The Gaussian-mixture model did not fit the homozygote clusters well, and tends to misclassify some of the homozygote individuals as heterozygote individuals (Figure 2.5B circled points). Since these seemingly misclassified individuals still follow Mendelian rules, the trio Gaussian-mixture model failed to correct these mistakes (Figure 2.5E). The

beta-mixture model and the trio beta-mixture model gave identical genotype calls for this SNP. The results of these two methods seem to be more reasonable than those of the other four methods. This suggests that the beta-mixture model is a better fit than the Gaussian-mixture model. Although we do not see a difference between the calls of the beta-mixture model and the trio beta-mixture model, we would expect a better performance of the trio beta-mixture model for data with poorer quality.

We also compared the calls made by the trio beta-mixture model to the calls made by Illumina's software (Shen et al., 2005). They are almost identical except for one individual, which is marked as 5 in Figure 2.6A. This data point is at the margin of the genotype clusters. It is not called by the company's software. We know that it most likely should be a heterozygous AB individual, since the genotypes of the parents of this child are AA and BB.

2.5 DISCUSSION

The goal of this paper is to show that we can use some specific features of the genotyping problem to improve clustering methods for making genotype calls from SNP array data. The specific features we discussed in this paper include: a) differences of variance structures and shapes of the distributions for different genotype clusters; b) constraints on genotype calls based on family structure; and c) limited number of clusters. We also proposed two new genotype calling methods for family data (demonstrated for trios). We studied the performance of the various methods by simulation. We also compared the results of these methods on a real dataset. We found that, when the quality of the data is good, all methods compared can give satisfactory results, though improvement is still possible. However, when the data quality is low, those methods that use additional information improved the genotype calls significantly. The trio beta-mixture model can be easily extended to incorporate larger pedigrees or individuals not in families. However, the trio K-means method, though very simple and straightforward, cannot be extended to handle larger family data.

Our results suggest that when calling genotypes for data of reasonably good quality, we

might want to choose the K-means clustering method (or trio K-means clustering method when family data is available) for simplicity of calculation. When the data quality is not good, it will be worth the effort to choose a model-based method that fits the data as well as possible. Our results suggest that the beta-mixture model fits the Illumina data better than the Gaussian-mixture model, since the homozygote clusters are very skewed. The Gaussian-mixture model tends to assign some of the homozygous individuals heterozygous genotypes. This might not necessarily be true for other platforms. Figure 2.7 shows two examples of datasets generated from other platforms and datasets. The data in Figure 2.7A and 2.7B look very similar to our Illumina data. However, the data in Figure 2.7C and 2.7D are quite different. In addition to the different platform, DNA preparation is another important factor that affects the quality and shape of the data. The dataset shown in 7C and 7D was prepared using whole-genome amplification. The quality of the data is much worse, and the distributions of all three genotype clusters are much more flat and symmetric as compared to our Illumina dataset. The Gaussian-mixture model seems to be a better fit for this dataset (results not shown).

One important issue in the family-based methods is that they force all genotypes to follow Mendelian inheritance rules within each family. But genetic studies often have a few errors in reported family structure, most often due to sample swaps, non-paternity, or unreported adoption. If the genotypes are called using a family-based method without first finding these family structure errors, there will be two problems. The most obvious problem is that some genotypes will be mis-called. The other problem, however, is that there will be outliers in the clusters, which may distort all of the estimation. For example, suppose a true AB is called as AA in order to enforce Mendelian rules. Then the AA cluster will include a point that may be far beyond its natural boundaries, which will affect both mean and variance estimates for that cluster and thus potentially affect other genotype calls. We recommend that genotype calling be done first with non-family-based methods in order to identify families with an excess of non-Mendelian calls. Then the family-based methods can be applied after the reasons for the non-Mendelian calls have been identified. Another way to deal with this problem (for the model-based methods only) is to examine the posterior probabilities of the genotype calls and set up a no-call cutoff value. This solution should also help maintain the

stability of the genotype calls if there are true technical outliers (e.g. a true AA point that falls in the AB cluster because of pure technical aberration). As a final note, we would like to suggest (see also Sabatti and Lange 2005) that using posterior probabilities of genotypes rather than absolute genotype calls might improve almost all statistical genetic analyses. The model-based methods that we have proposed here are of course easily adaptable to generate such probabilistic data.

Table 2.1: Fifteen Family Types of a SNP Marker for a Nuclear Family with One Disomic Offspring

Family Type	Parent 1	Parent 2	Child
1	AA	AA	AA
2	AA	AB	AA
3			AB
4	AA	BB	AB
5	AB	AA	AA
6			AB
7	AB	AB	AA
8			AB
9			BB
10	AB	BB	AB
11			BB
12	BB	AA	AB
13	BB	AB	AB
14			BB
15	BB	BB	BB

Table 2.2: Simulation Study 1.

Methods	K-means	Gaussian-mixture model	beta-mixture model	trio K-means	trio Gaussian-mixture model	trio beta-mixture model
Average number of mistakes						
Total	0.253	0.699	0.037	0.127	0.549	0.024
Misscalled heterozygotes	0.25	0	0.019	0.125	0	0.011
Misscalled homozygotes	0.003	0.699	0.018	0.002	0.549	0.013
Number of simulations with						
0 miscall	775	514	964	883	589	976
1 miscalls	200	320	35	107	295	24
2 miscalls	22	122	1	10	94	0
3 miscalls	3	41	0	0	22	0
4 miscalls	0	3	0	0	0	0

A total of 1000 datasets were simulated. Each dataset consisted 150 disomic trios. Population genotype frequencies were set at $p_{AA} = 0.2$, $p_{AB} = 0.35$ and $p_{BB} = 0.45$. The beta parameters used in the simulations for the three genotype clusters were $\alpha_{AA} = 1$, $\beta_{AA} = 40$, $\alpha_{AB} = 20$, $\beta_{AB} = 20$, $\alpha_{BB} = 40$, $\beta_{BB} = 1$.

Table 2.3: Simulation Study 2.

Methods	K-means	Gaussian-mixture model	beta-mixture model	trio K-means	trio Gaussian-mixture model	trio beta-mixture model
Average number of mistakes						
Total	14.62	6.65	5.31	7.13	4.42	3.47
Misscalled heterozygotes	14.46	1.38	2.89	6.93	1.09	1.90
Misscalled homozygotes	0.16	5.27	2.42	0.20	3.33	1.57
Number of simulations with						
0 miscalls	0	1	3	3	15	35
1-5 miscalls	8	349	572	285	705	819
6-10 miscalls	158	574	394	599	274	145
10-15 miscalls	449	73	31	108	6	1
>15 miscalls	385	3	0	5	0	0

A total of 1000 datasets were simulated. Each dataset consisted 150 disomic trios. Population genotype frequencies were set at $p_{AA} = 0.2$, $p_{AB} = 0.35$ and $p_{BB} = 0.45$. The beta parameters used in the simulations for the three genotype clusters were $\alpha_{AA} = 5$, $\beta_{AA} = 40$, $\alpha_{AB} = 10$, $\beta_{AB} = 10$, $\alpha_{BB} = 40$, $\beta_{BB} = 5$.

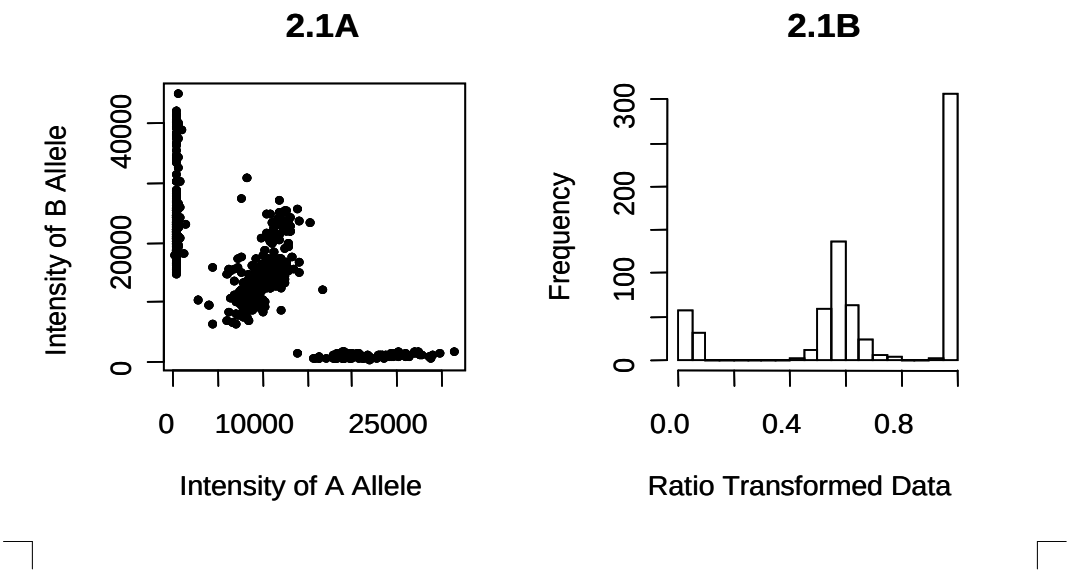

Figure 2.1: Plots of 2-dimensional (2.1A) and transformed 1-dimensional (2.1B) Illumina data.

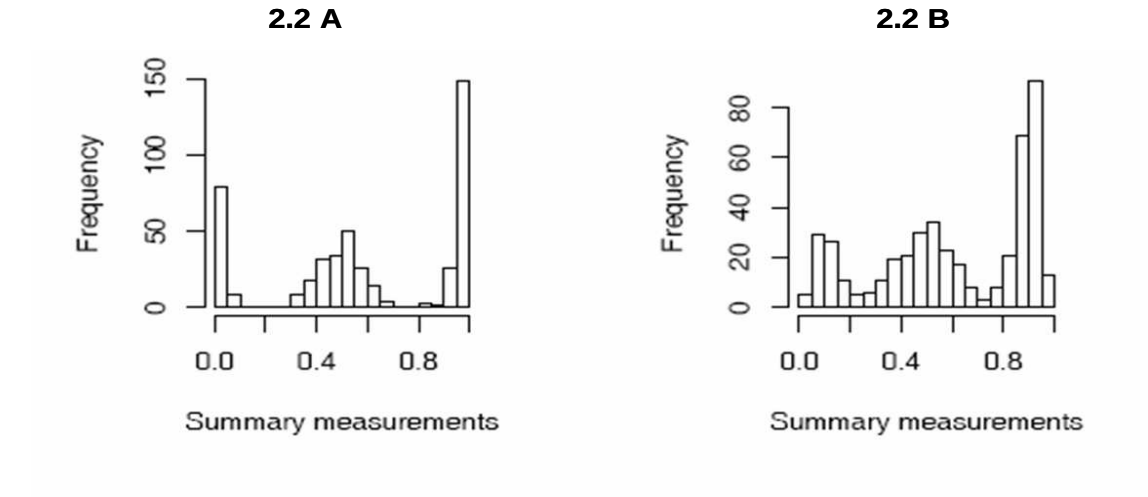

Figure 2.2: Histograms of examples of simulated data.2.2A: "Good" data; 2.2B: "Bad" data.

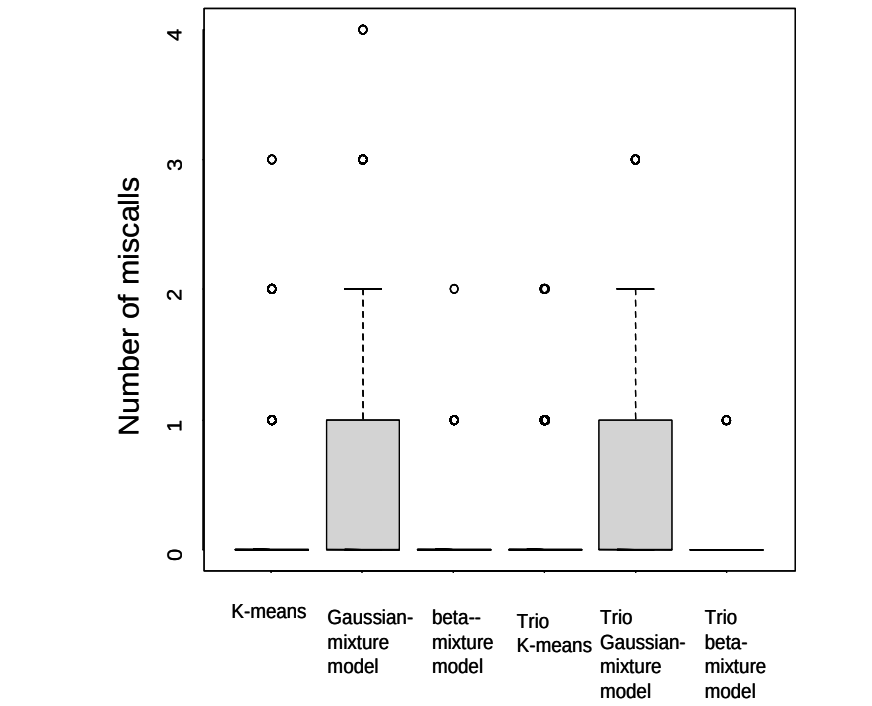

Figure 2.3: Boxplots for the results of simulation study 1.

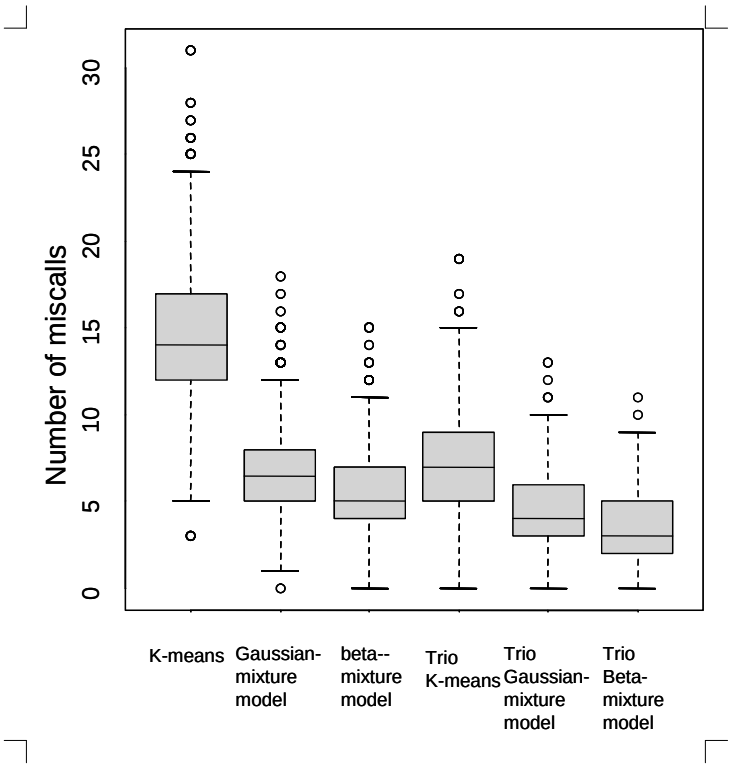

Figure 2.4: Boxplots for the results of disomic simulation study 2.

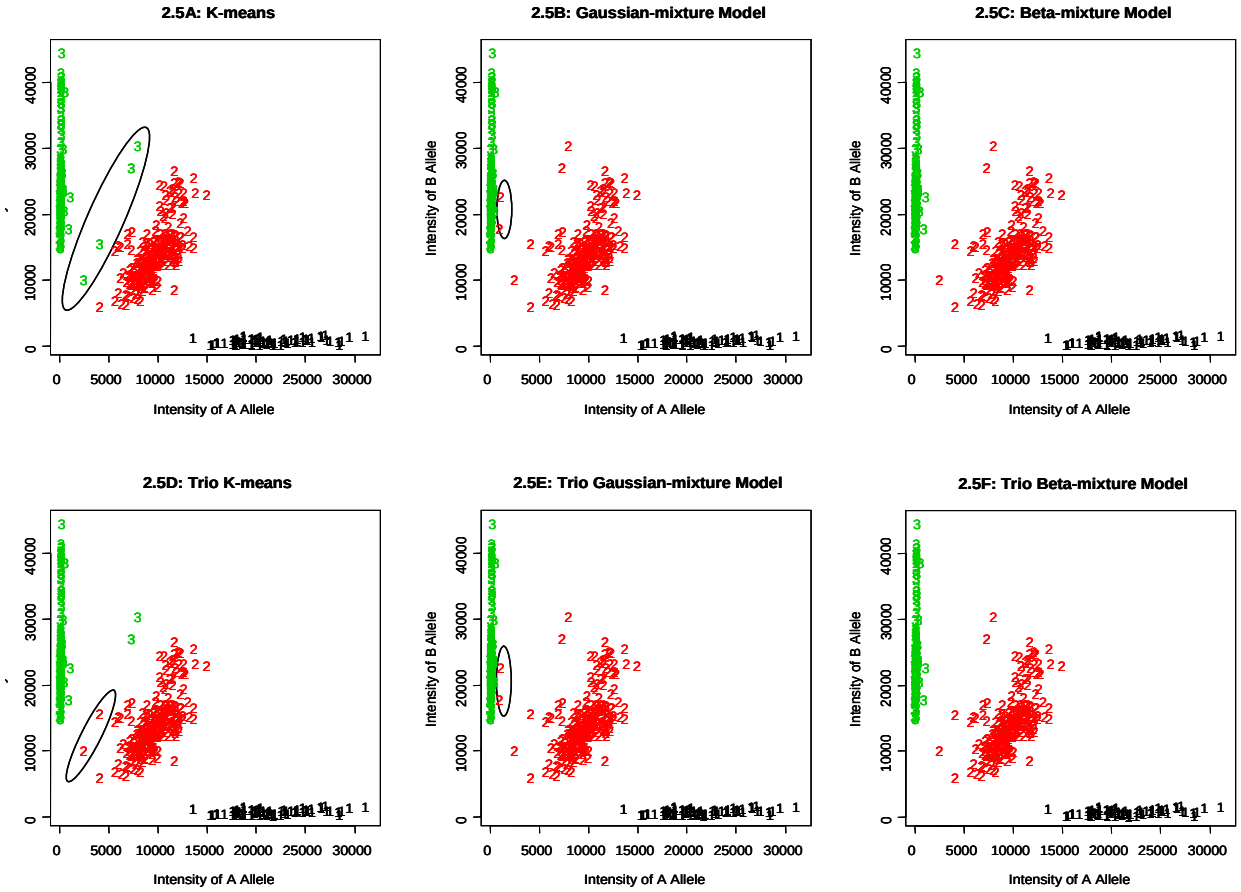

Figure 2.5: Restored clustering results for the real dataset.

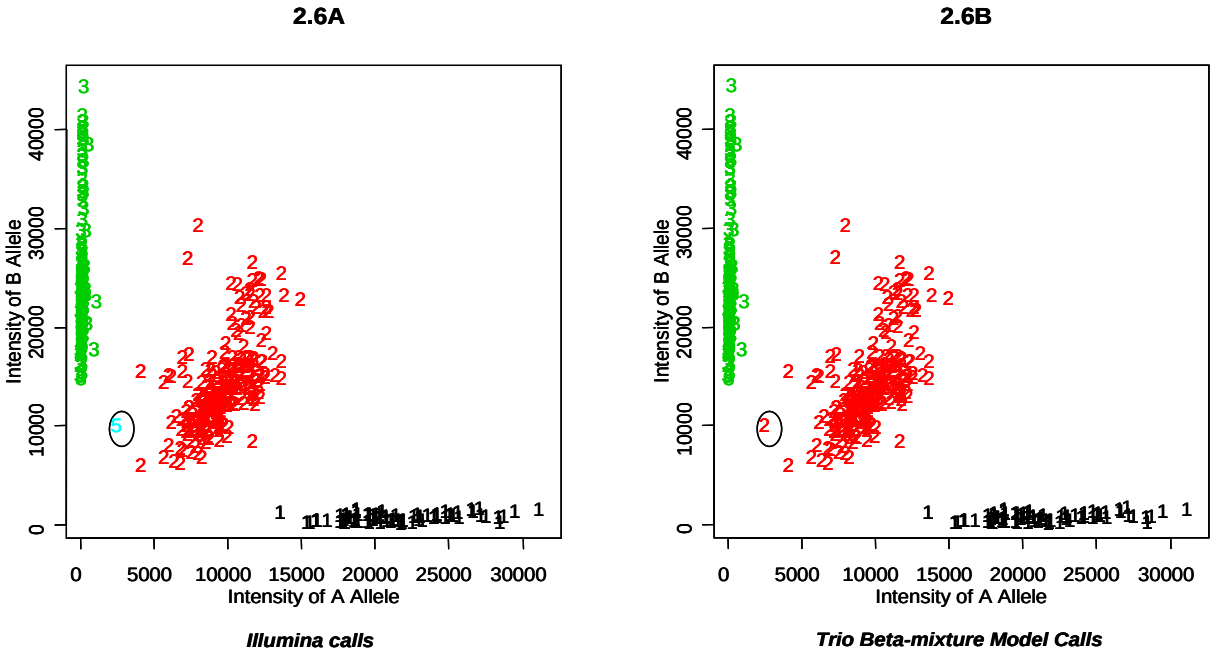

Figure 2.6: Comparison of Illumina calls (2.6A) and trio beta-mixture model calls (2.6B). Genotype cluster 1=AA genotype group, genotype cluster 2=AB genotype group, genotype cluster 3=BB genotype group, and 5=no call. The circled points are the "problematic" calls.

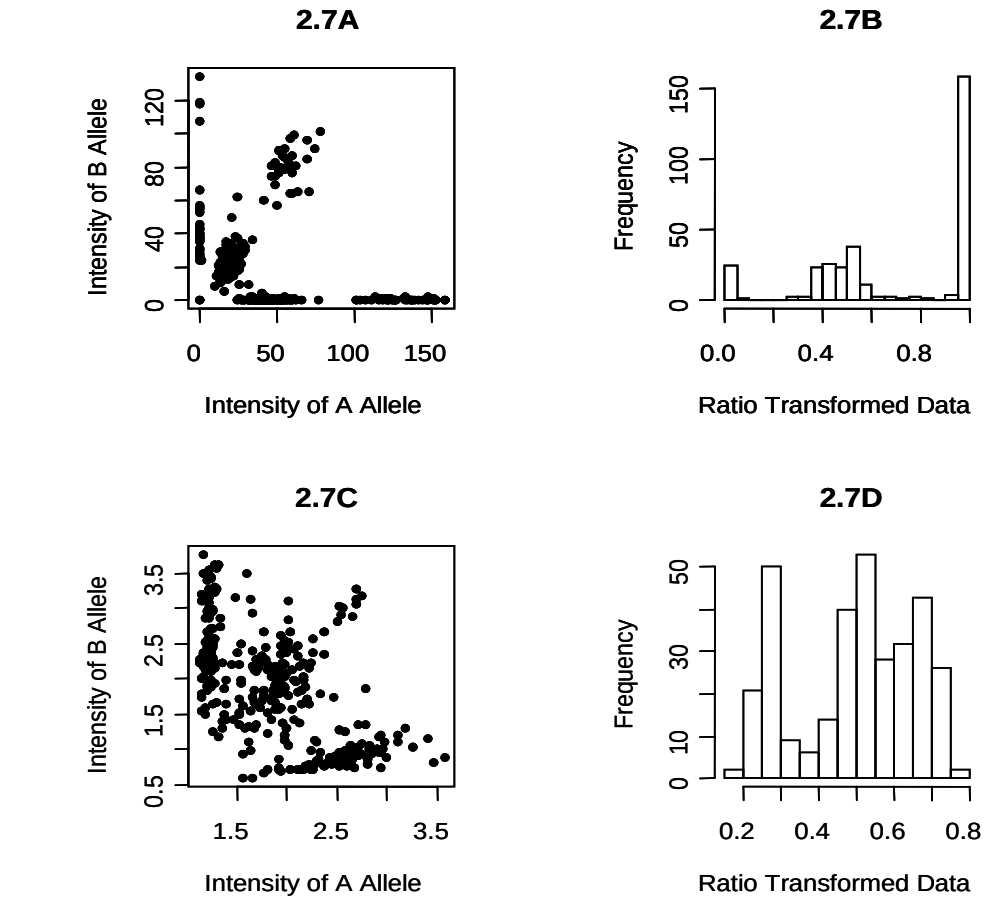

Figure 2.7: Examples of data generated by two other platforms.

3.0 GENOTYPE CALLING METHODS FOR HIGH-THROUGHPUT SNP GENOTYPING OF TRISOMIC INDIVIDUALS

Manuscript in preparation.

Yan Lin[1], Eleanor Feingold[1,2,]

1. Department of Biostatistics, University of Pittsburgh

2. Department of Human Genetics, University of Pittsburgh

Email: yal14@pitt.edu

Telephone: (412)624-7178

Fax: (412)624-3020

3.1 ABSTRACT

Genotyping of trisomic individuals has never been addressed formally to date. It is not clear which, if any, standard genotyping methods have the ability to distinguish the two heterozygous genotype clusters. In this paper we showed that when carefully conducted, the trisomic individuals could be successfully genotyped by existing techniques. We also extended two family-based methods developed by our previous paper to trisomic trios. We compared the performance of these two methods and related clustering methods by both simulation study and the analysis of a real dataset. Our results suggested that the family-based methods perform significantly better than the methods that ignore the family information.

3.2 INTRODUCTION

Single nucleotide polymorphysms (SNPs) are the most common type of genetic variants in the human population. It has been estimated that there are about 10 million SNPs with minor allele frequency $> 1\%$ in the human genome (The International HapMap Consortium, 2003). SNPs are a popular choice of genetic markers for genome-wide linkage and association studies to discover genes relevant to disease. In addition to the fact that they have high density throughout the genome, it is also relatively cheaper to genotype SNPs than other type of genetic markers. Many high-throughput SNP genotyping techonologies have been developed by different companies, for example the GeneChip Human Mapping Array from Affymetrix (www.affymetrix.com), the BeadStation from Illumina (www.illumina.com), and the iPLEX assay from Sequenom (www.sequenom.com).

The vast majority of the SNPs are biallelic. If we denote the two alleles of a SNP marker A and B, a disomic individual then will have three possible genotypes for the marker, AA, AB and BB. Different platforms produce data of different forms and dimensions. In general, they all give quantitative measures for each of the two alleles of a single SNP. It is common to represent the genotyping data in a two-dimensional space, as illustrated in Figure 3.1A with the horizontal axis for the observed intensity of one allele and the vertical axis for the

other allele. It is an example of the raw data for a single SNP generated by the Illumina platform. Each point in the figure represents an individual. The genotype calling procedure generates genotype data from the raw data produced by these high-throughput technologies, using a clustering algorithm. Each individual is assigned to one of the genotype clusters, AA, AB and BB, based on his/her quantitative measurements of A and B alleles. In practice, often the two-dimensional data are reduced to one-dimension before genotype calling. The transformation that was used in this paper is y =(intensity of B allele)/(intensity of A allele+intensity of B allele). Therefore, for individuals with AA genotype, the observed values of y are expected to be close to 0. For those with BB genotype, the observed values of y are expected to be close to 1. The heterozygous individuals will have y values close to 0.5.

Both supervised (Rabbee and Speed 2005, Liu et al., 2003) and unsupervised (Sabatti and Lange, 2005, Di et al., 2005, Fujisawa et al., 2004) clustering methods have been used in genotype calling for SNP array data from disomic individuals. Mostly training data are not available in practice, and an unsupervised clustering method is then required to call the genotypes for the study subjects. In a previous paper, we developed two family-based clustering methods that incorporate the family information for genotype calling of family trio data (Lin et al., 2007). We compared the performance of these two methods and several popular unsupervising clustering methods on genotype calling of SNP data for disomic cases. We found that better use of prior knowledge of distribution and pedigree structure can significantly improve the genotype calls.

Genotyping trisomic individuals, those with an extra copy of one of the chromosomes, has not been discussed by any published literature to date. Trisomic individuals have four possible genotypes for a biallelic marker, AAA, AAB, ABB and BBB. This makes the genotype calling procedure more difficult since the two heterozygous groups are close together (Figure 3.1B). It is not clear which, if any, standard genotyping methods have the ability to distingish the two heterozygous genotype clusters. The object of this paper is to show that with proper statistical methods, we can successfully call the genotypes of the trisomic individuals. As discussed later in a real data example, the unique family structure for trisomic data makes family-based methods particularly suitable here.

The genotype calling problem in trisomic individuals is some times confused with the copy number variation (CNV) problem. In the genotype-calling problem we know there are 3 alleles (4 genotype clusters) and we want to look at one marker at a time and determine what alleles it has. In the CNV problem, we are looking at markers over a whole genomic region at a time, and looking at the total intensity of all alleles. We try to classify individuals in terms of their total intensity (copy number) on average over lots of markers. The differences are single marker versus multiple markers, and using a different dimension of the dataset (Komura et al., 2007).

In this chapter, we extend the two family-based methods that were proposed in the previous chapter to trisomic trios. We compare the performance of these two methods and other popular clustering methods for genotype calling on both simulated trisomic trio data and on a real dataset. The motivating dataset for this paper is from an association study of atrioventricular septal defects (AVSD) in a Down syndrome (DS) population. It is known that over 95% of the DS cases are caused by trisomy 21. These individuals have three copies of chromosome 21. In addition to mental retardation, 44% of all DS individuals also have some form of congenital heart defects (CHD, Freeman et al., 1998), and the severe form of CHD is AVSD. One hundred and sixty family trios were collected. A trio is a nuclear family with one child. In our case, the child is a trisomy 21 child. Typically we know which parent is the source of the extra chromosome, and whether that parent passed two different (not reduced to homozygosity) or identical (reduced to homozygosity) chromosomes to the child (Xu et al., 2004). This parent is denoted as the non-disjoining parent (NDJP). The other parent, who contributes only one copy of chromosome 21, is denoted as the correctly disjoining parent (CDJP). The parents are disomic. In these families, there are more constraints on the genotypes of the children than in disomic case. For example, if the genotype of the NDJP is AA and the genotype of the CDJP is AB, then we know that child's genotype can only be either AAA or AAB but not ABB or BBB. Individual's genotyping data were acquired by the Illumina platform.

3.3 METHODS

3.3.1 Trio beta-mixture model for trisomic data

Previously, we developed a model-based clustering method that incorporates the pedigree likelihood and a beta-mixture model for disomic trios (Lin et al., 2007). Here this trio beta-mixture model was extended to the trisomic family data.

3.3.1.1 Likelihood for complete data Let $\mathbf{Y_i} = (y_{Ni}, y_{Ci}, y_{Ki})$ denotes the observed one-dimensional data for the NDJP, CDJP and the child of the i^{th}trio; $\mathbf{G_i} = (g_{Ni}, g_{Ci}, g_{Ki})$ the corresponding genotype vector, where $\mathbf{G_i}$ is unknown. The contribution to the complete-data likelihood function from the i^{th} trio is:

$$\begin{aligned} & L_i(\theta, Y_i, G_i, h_i) \\ & = \{Pr(g_{Ni})Pr(g_{Ci})Pr(g_{Ki}|g_{Ni}, g_{Ci})\}\{Pr(y_{Ni}|g_{Ni})Pr(y_{Ci}|g_{Ci})Pr(y_{Ki}|g_{Ki})\}, \end{aligned} \quad (3.3.1)$$

where the first component, $Pr(g_{Ni})Pr(g_{Ci})Pr(g_{Ki}|g_{Ni}, g_{Ci})$, is the pedigree likelihood; the second component, $Pr(y_{Ni}|g_{Ni})Pr(y_{Ci}|g_{Ci})Pr(y_{Ki}|g_{Ki})$, is the penetrance term. h_i is the probability that the two alleles contributed by the NDJP are reduced to homozygousity (see Chapter 1 section 1.3).

3.3.1.2 A beta-mixture model We assume a beta-mixture model for the penetrance term. In a trisomic trio, the parents are disomic and the child is trisomic. Therefore, we need to set up two mixture models for the data, one for the parents, and one for the children. Let y be the obseved value for an individual, we assume the following beta mixture-model for the parents

$$y \sim \sum_{\lambda 1 \in \Lambda 1} \nu_{\lambda 1} f(y, \alpha_{\lambda 1}, \beta_{\lambda 1}), \quad (3.3.2)$$

where $\nu_{\lambda 1}$ is the probability of a person having genotype $\lambda 1 \in \Lambda 1 = \{AA, AB, BB\}$, and

$$\begin{aligned} f(y, \alpha, \beta) &= \frac{1}{B(\alpha, \beta)} y^{\alpha-1}(1-y)^{\beta-1} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}, 0 < y < 1, \alpha, \beta > 0. \end{aligned}$$

For the childeren, we assume

$$y \sim \sum_{\lambda 2 \in \Lambda 2} \nu_{\lambda 2} f(y, \alpha_{\lambda 2}, \beta_{\lambda 2}), \tag{3.3.3}$$

where $\nu_{\lambda 2}$ is the probability of a person having genotype $\lambda 2 \in \Lambda 2 = \{AAA, AAB, ABB, BBB\}$,

$$\begin{aligned} f(y, \alpha, \beta) &= \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, 0 < y < 1, \alpha, \beta > 0. \end{aligned}$$

$\nu_{\lambda 2}$'s are functions of population genotype frequencies and the probability that the two alleles contributed by the NDJP are reduced to homozygousity (h_i's).

3.3.1.3 Pedigree likelihood We follow the model proposed by Xu et al., 2004, as discussed in Chapter 1.

Let us define h as the probability that the two chromosomes contributed by the NDJP are reduced to homozygosity, that is they are replicates of the same allele from the NDJP parent. In our real data example, this probability is estimated based on the microsatelite marker map already established on this dataset (Feingold, et al. 2000). $h = 1$ when we are sure that the two alleles from the NDJP are reduced to homozygosity, and $h = 0$ when we are sure that the two alleles from the NDJP are not reduced to homozygosity. For the purpose of genotype calling, all w's are set to 1. The conditional probabilities for each outcome given the parental genotype are listed in the fifth column of Table 3.1. They are functions of h. The marginal probabilities for each mating type are listed in the last column of Table 3.1. The products of the corrsponding values from these two columns are the marginal probabilities for each family type.

3.3.1.4 Estimation The expectation-maximization (EM) algorithm was applied to estimate the model parameters, $\theta = (\nu_{\lambda 1}\text{'s}, \alpha_{\lambda 1}\text{'s}, \beta_{\lambda 1}\text{'s}, \alpha_{\lambda 2}\text{'s}, \beta_{\lambda 2}\text{'s})^T$. Assuming that $\theta^{(t)}$ is the current estimate, the E-step is

$$\nu_{\lambda 1}^{(t+1)} = \frac{E(S_{1,\lambda 1}|Y, \theta^{(t)})}{2n} \tag{3.3.4}$$

where

$$S_{1,\lambda 1} = \sum_{i=1}^{n}[1\{g_{Ni} = \lambda 1 + 1\{g_{Ci} = \lambda 1\}], \quad \lambda 1 \in \Lambda 1 = \{AA, AB, BB\},$$

and n is the number of family trios. In order to estimate the α's and β's for each genotype cluster, we used *nlm* function included in *R* to maximize the conditional expectation of the log complete-data likelihood in the M-step. Details for estimation are given in Appendix B.

3.3.1.5 Genotype prediction using Bayes rule Once the parameters are estimated, we can use the Bayes rule to determine the genotypes of the family members. The posterior probability of the family genotype vector $G = (g_N, g_C, g_K)$ given the observed values $Y = (y_N, y_C, y_K)$ is

$$\begin{aligned} &p(G|Y) \\ &= \frac{\nu_{\lambda 1=g_N}\nu_{\lambda 1=g_C} Pr(g_K|g_N, g_C) f(y_N, \xi_{\lambda 1=g_N}) f(y_C, \xi_{\lambda 1=g_C}) f(y_K, \xi_{\lambda 2=g_K})}{\sum_{j=1:18} \nu_{\lambda 1=g_{Nj}}\nu_{\lambda 1=g_C} Pr(g_{Kj}|g_{Nj}, g_{Cj}) f(y_{Nj}, \xi_{\lambda 1=g_{Nj}}) f(y_{Cj}, \xi_{\lambda 1=g_{Cj}}) f(y_{Kj}, \xi_{\lambda 2=g_{Kj}})} \end{aligned} \tag{3.3.5}$$

Here the $\xi_\lambda = (\alpha_\lambda, \beta_\lambda)^T$ are the parameters for the beta distribution of genotype cluster λ.

3.3.2 Trio K-means algorithm for trisomic data

In a previous paper, we proposed an ad-hoc modification of the K-means algorithm, the trio K-means algorithm, that incoporates the family information for disomic family trio data (Lin et al., 2007). It is fairly straightforward to extend it to trisomic trios. The iterative procedure is described as the following:

Step1: Start with a set of initials

The initial centroids are $\{C_{AA}^{(0)}, C_{AB}^{(0)}, C_{BB}^{(0)}, C_{AAA}^{(0)}, C_{AAB}^{(0)}, C_{ABB}^{(0)}, C_{BBB}^{(0)}\}$. Note that in a trisomic trio, the parents are disomic and the child is trisomic. Therefore, we have to consider 3 genotype clusters for the parents and 4 genotype clusters for the children.

Step2: At the k+1 step, update the genotype of family members as described in the following

For a biallelic marker with alleles A and B, there are 18 possible genotype combinations that agree with Mendelian segregation rules (Table 3.1). Let y_N, y_C, and y_K be the observed one-dimensional values for the NDJP, the CDJP and the child in a family. Let us denote g_{Nj}, g_{Cj} and g_{Kj} the genotypes for the NDJP, the CDJP and the child in the j^{th} family type in Table 3.1. For all j, we calculate $D_{g_{Nj},g_{Cj},g_{Kj}} = d(y_N, C^{(k)}_{g_{Nj}})^2 + d(y_C, C^{(k)}_{g_{Cj}})^2 + d(y_K, C^{(k)}_{g_{Kj}})^2$, where $C^{(k)}_g$'s are the estimated group centers from the k^{th} step and $d(y, C^{(k)}_g)$ is the Euclidiean distance between the observed value y and the centroid for genotype cluster $C^{(k)}_g$. $\tilde{g_{Nj}}$, $\tilde{g_{Cj}}$ and $\tilde{g_{Kj}}$ that minimize $D_{g_{Nj},g_{Cj},g_{Kj}}$ then will be assigned to the NDJP, the CDJP and the child of the family.

Step3: Iterate until convergence

3.4 RESULTS

3.4.1 Simulation study

We compared the performance of the regular K-means clustering algorithm, the regular beta-mixture model, the trio K-means clustering method and the trio beta-mixture model on genotype calling of simulated trisomic trio data. In order to apply the K-means clustering method and the beta-mixture model to trisomic trio data, we need to separate the dataset into two datasets. One contains disomic individuals (parents), and the other one contains trisomic individuals (offspring). We then applied the above two methods to these two datasets separately. A total of 1000 datasets were simulated and 150 family trios were simulated in each dataset. We used a beta distribution to simulate the observed values because we observed that the distributions of the homozygote genotype clusters are quite skewed in real datasets. Figure 3.2 is an example of the simulated dataset, which looks similar to the real dataset (Figure 3.1). We applied the four methods mentioned above to these datasets and compared the number of mistakes made in each dataset by these methods. The results of

the simulation study is summarized in Table 3.2 and Figure 3.3. The improvements to the genotype calls after the incorporation of the family information and controlling for variance structure are apparent, consistent with what we observed in the disomic case (Lin et al. 2007). However, the family information seems to make a bigger contribution in the trisomic genotype calling procedure than does the control of variance structure. This is especially obvious when comparing the results between the regular beta-mixture model and the trio beta-mixture model. On average, 2/3 of the miscalls were avoided when applying the trio beta-mixture model instead of the regular beta-mixture model to the data. The trio K-means algorithm made 1/3 fewer mistakes then the regular K-means algorithm. It is worth noticing that the performance of the regular beta-mixture model approach is quite unstable. Although on average it performs better than the K-means method, there are quite a few cases in which the beta-mixture model makes more miscalls than then K-means method. The trio beta-mixture model, on the other hand, is much more stable (Figure 3.3).

3.4.2 Real data analysis

We applied the regular K-means clustering algorithm, the regular beta-mixture model, and their corresponding family-based methods, the trio K-means algorithm and the trio beta-mixture model to a real dataset generated by the AVSD study. The reconstructed two-dimensional results for the parents were shown in Figure 3.4, and those for the children were shown in Figure 3.5. As expected, the K-means method seemed to have misclassified some heterozygous individuals as homozygous (Figures 3.4A and 3.5A, the circled individuals). The trio K-means method corrected some but not all of these "miscalls" (Figures 3.4C and 3.5C). The regular beta-mixture model also seems to have misclassified some individuals (Figures 3.4B and 3.5B), with seemingly homozygous individuals misclassified as heterozygous. To our surprise, the trio beta-mixture model seem to produce the "worst" results, with many obviously misclassified trisomic individuals (Figure 3.5D). We then realized that there are some mixup for a small portion of the family data, *i.e.*, who is the NJDP and the h estimates. Table 3.3 gives out an example of how wrong family information can "force" genotype called for the trisomic individual to the wrong cluster. As shown in Table 3.3, either

the wrong identity of the NJDP or the wrong h can cause trouble in the process. However, if we set $h = 0.5$, i.e. assume non-informative microsatelite marker data, the genotype of the child will be correctly called. This is because that both genotypes ABB and BBB become possible but the observed value for the child is very close to the center of the BBB cluster. Before we can straighten out the family data with the investigators, we set h for all families to 0.5. The corresponding results for this are shown in Figure 3.6. We believe that if correct family information is used, the trio beta-mixture model should produces the best results. This is also an example of how the family-based methods can be misled by the wrong family information.

3.5 DISCUSSION

Genotype calling of trisomic individuals is more complicated than that for the disomic individuals. There are four genotype clusters, the two heterozygous genotype clusters are close to each other. In this paper, we have shown that with caution, standard clustering methods could be used for genotype calling in trisomic individuals. We also extended two family-based genotype calling methods we previously developed for disomic trios to trisomic trios. The simulation results are similar to those presented for disomic trios. The family-based methods perform significantly better than the corresponding methods that ignore the family information. As in the disomic case, the model-based methods perform better than the K-means methods. This is because that the K-means methods assume equal variances for all clusters, but the mixture-model approach allows different variances for each genotype cluster. In our case, the heterozygous genotype clusters have a quite different variance structure than the homozygous genotype cluster. On the other hand, the performance of the regular beta-mixture model is quite unstable as compared to the other methods. After incorporating the family information, the results of the model-based methods look stable.

The fact that the family-based methods performed better in both the simulation study and the real data analysis suggests that family information can improve genotype calls significantly. However, we are surprised to see how well the trio beta-mixture model performed

when all h's were set to 0.5 in the real data example. That is, we assumed that all the microsatelite markers were uninformative. This suggests that the model fitting is mostly driven by the observed genotyping data and the basic family structure, i.e. which parent is the NJDP. We plan to evaluate the impact of h values in the future by simulation study.

To simplify calculation, we also assume all w's to be 1 for the trio beta-mixture model. As mentioned above, we suspect that the model fitting is mostly driven by the observed genotyping data and the basic family structure. The difference of the w's in cases and controls may not affect the clustering results dramatically. We also plan to evaluate this hypothesis in the future.

Table 3.1: Eighteen Family Types of a SNP Marker for a Nuclear Family with One Trisomic Offspring

Family Type	NDJP	CDJP	Child	Probability	
				p	q
1	AA	AA	AAA	1	p_{aa}^2
2	AA	AB	AAA	$(\frac{w_0}{w_0+w_1})$	$p_{aa}p_{ab}$
3			AAB	$(\frac{w_1}{w_0+w_1})$	$p_{aa}p_{ab}$
4	AA	BB	AAB	1	$p_{aa}p_{bb}$
5	AB	AA	AAA	$(\frac{w_0 h}{w_0 h+2w_1(1-h)+w_2 h})$	$p_{ab}p_{aa}$
6			AAB	$(\frac{2w_1(1-h)}{w_0 h+2w_1(1-h)+w_2 h})$	$p_{ab}p_{aa}$
7			ABB	$(\frac{w_2 h}{w_0 h+2w_1(1-h)+w_2 h})$	$p_{ab}p_{aa}$
8	AB	AB	AAA	$(\frac{w_0 h}{w_0 h+(w_1+w_2)(2-h)+w_3 h})$	$p_{ab}p_{ab}$
9			AAB	$(\frac{w_1(2-h)}{w_0 h+(w_1+w_2)(2-h)+w_3 h})$	$p_{ab}p_{ab}$
10			ABB	$(\frac{w_2(2-h)}{w_0 h+(w_1+w_2)(2-h)+w_3 h})$	$p_{ab}p_{ab}$
11			BBB	$(\frac{w_3 h}{w_0 h+(w_1+w_2)(2-h)+w_3 h})$	$p_{ab}p_{ab}$
12	AB	BB	AAB	$(\frac{w_1 h}{w_1 h+2w_2(1-h)+w_3 h})$	$p_{ab}p_{bb}$
13			ABB	$(\frac{2w_2(1-h)}{w_1 h+2w_2(1-h)+w_3 h})$	$p_{ab}p_{bb}$
14			BBB	$(\frac{w_3 h}{w_1 h+2w_2(1-h)+w_3 h})$	$p_{ab}p_{bb}$
15	BB	AA	ABB	1	$p_{bb}p_{aa}$
16	BB	AB	ABB	$(\frac{w_2}{w_2+w_3})$	$p_{bb}p_{ab}$
17			BBB	$(\frac{w_3}{w_2+w_3})$	$p_{bb}p_{ab}$
18	BB	BB	BBB	1	p_{bb}^2

p=conditional probability of the child's genotype given the parents' genotype. q=marginal probability of the family type.

Table 3.2: Simulation Study Results

Methods	K-means	beta-mixture model	trio K-means	trio beta-mixture model
Average number of mistakes				
	4.41	4.18	3.23	1.44
Number of simulations with				
0 miscalls	14	32	35	233
1-5 miscalls	714	762	825	761
6-10 miscalls	265	162	138	6
10-15 miscalls	7	27	2	0
>15 miscalls	0	15	0	0

Table 3.3: Example of Wrong Family Data

	Mother (CDJP)	Father (NDJP)	Child
True Genotypes	AB	BB	BBB
Observed Values	0.6	0.95	0.98
Observed $h = 0$			
Genotype called by trio beta-mixture model assuming father is the NDJP			
	AB	BB	BBB
Genotype called by trio beta-mixture model assuming mother is the NDJP			
	AB	BB	ABB
Genotype called by trio beta-mixture model assuming mother is the NDJP & $h = 0.5$			
	AB	BB	BBB

An example of potential impact of wrong family structure. In this example, the true NDJP is the father.

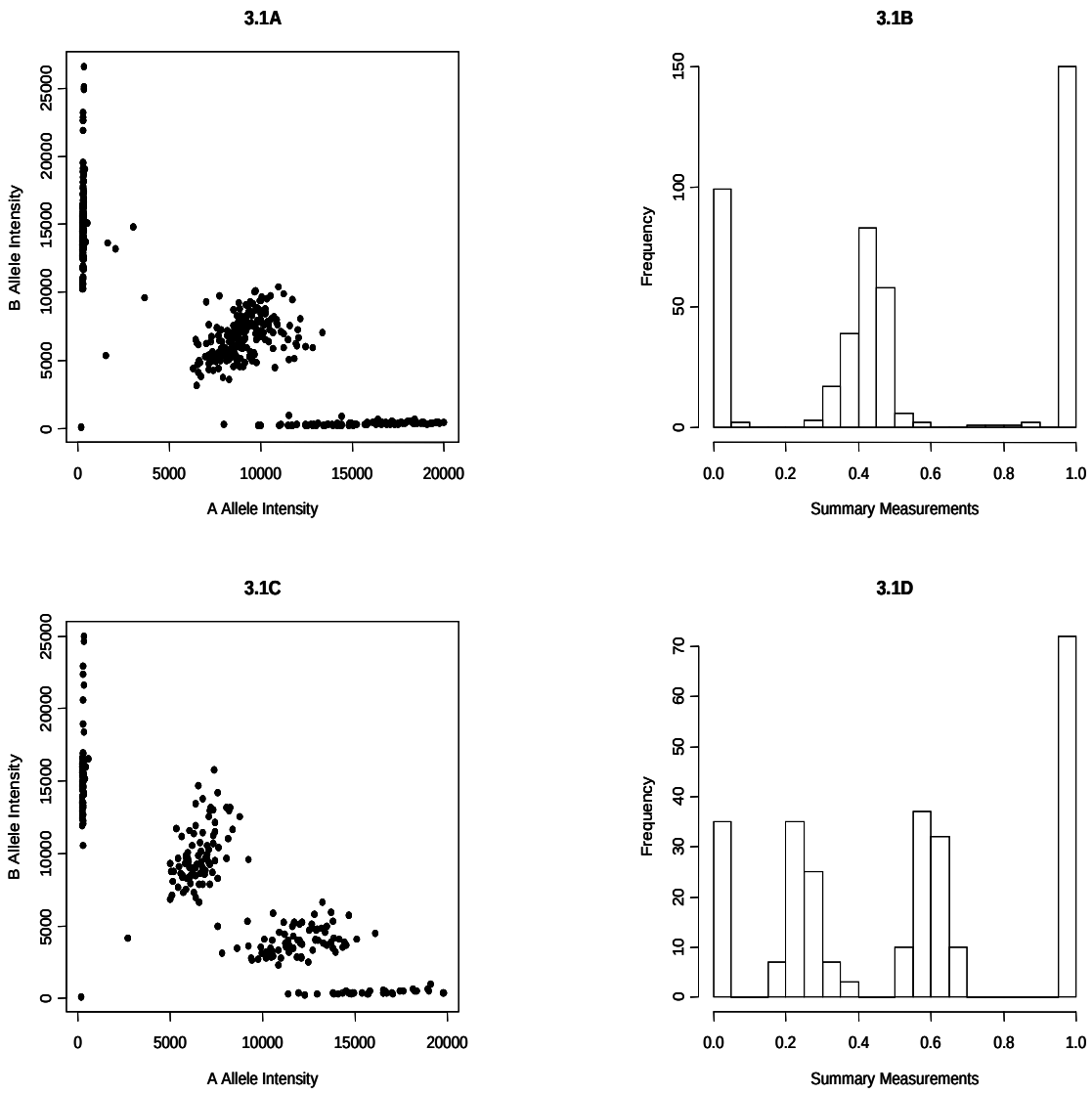

Figure 3.1: Plots of a example Illumina data.

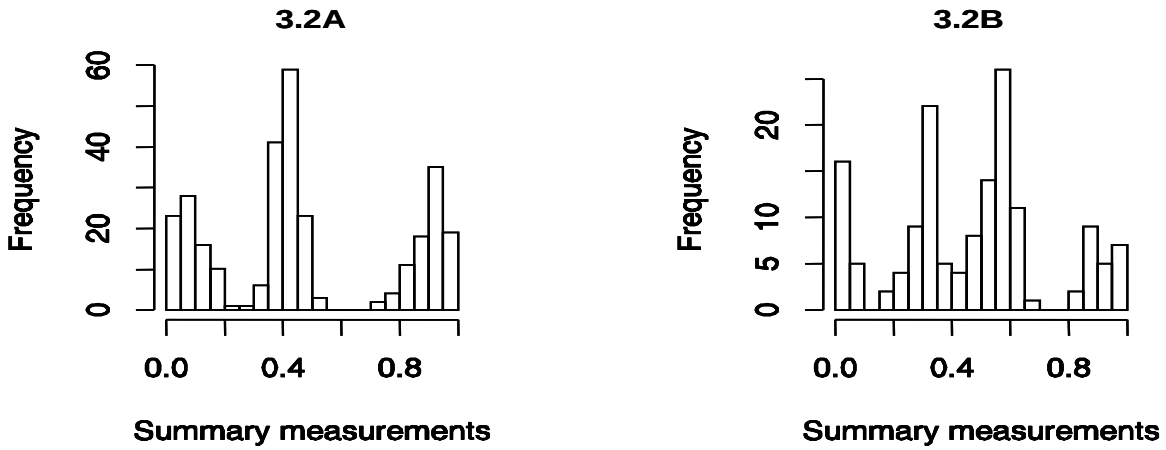

Figure 3.2: Histograms of an example of simulated disomic (3.2A) and trisomic (3.2B) data.

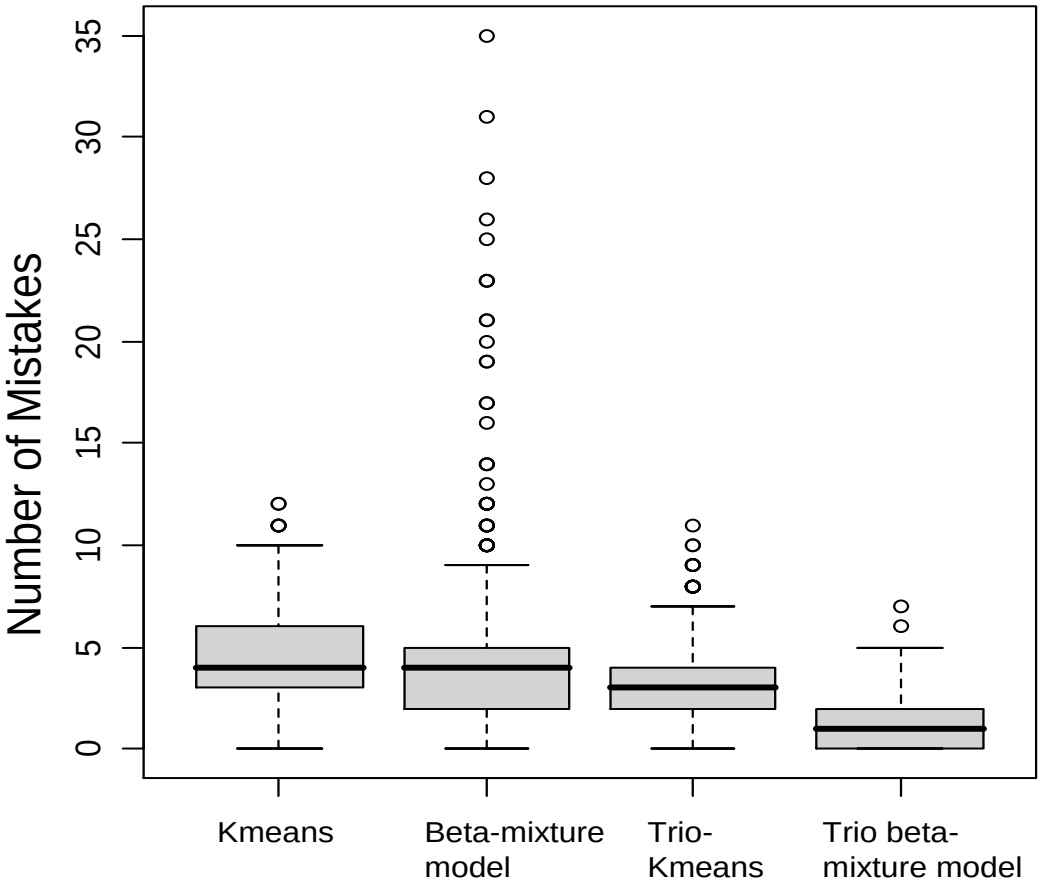

Figure 3.3: Boxplots for the results of simulation study.

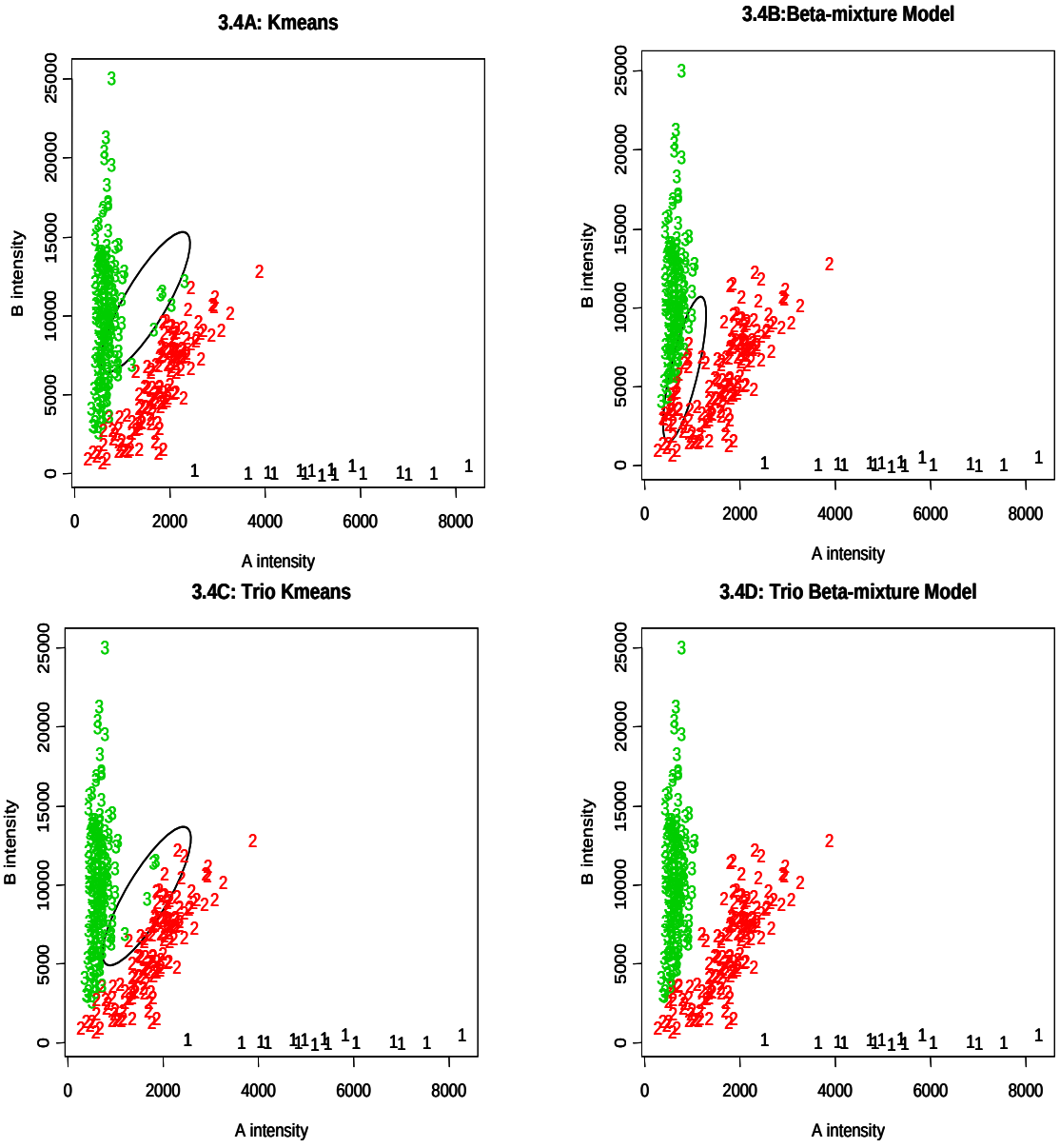

Figure 3.4: Restored clustering results for the disomic individuals (the parents) in the real dataset.

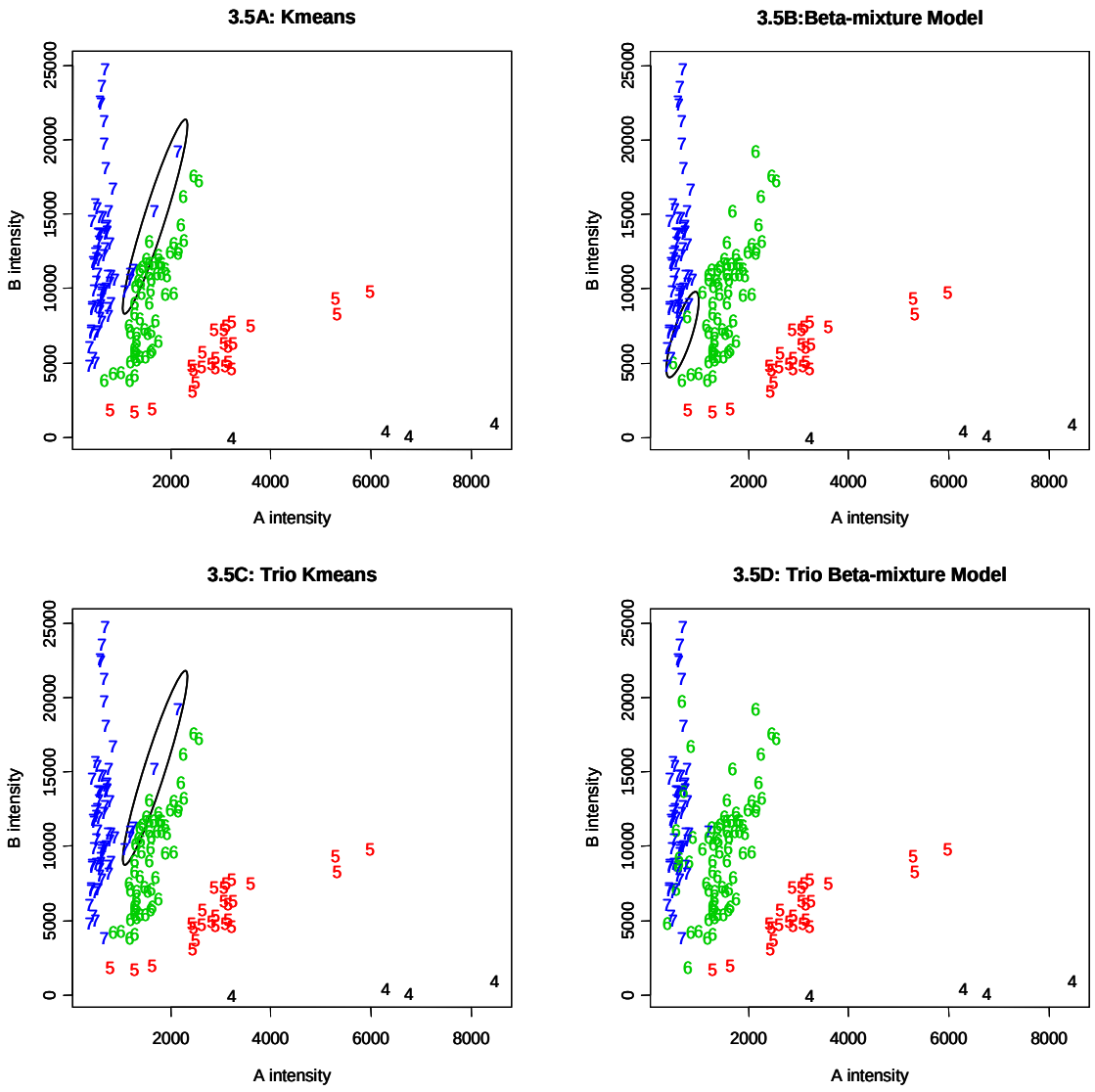

Figure 3.5: Restored clustering results for the trisomic individuals (the children) in the real dataset.

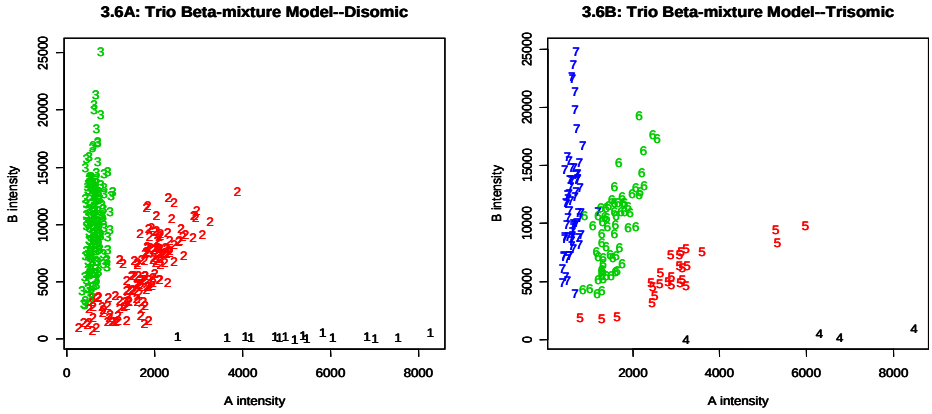

Figure 3.6: Restored clustering results for the second analysis using the trio beta-mixture model.

4.0 ALTERNATIVE ANALYSES FOR TRIO DESIGNS

Manuscript in preparation.

Yan Lin[1], Eleanor Feingold[1,2,]

1. Department of Biostatistics, University of Pittsburgh

2. Department of Human Genetics, University of Pittsburgh

Email: yal14@pitt.edu

Telephone: (412)624-7178

Fax: (412)624-3020

4.1 ABSTRACT

In this chapter, we discussed and compared the available methods on the analysis of trio designs under the factorizations of one unified conditional likelihood. We also discussed the problem of whether it is sensible to collect trio data any more.

4.2 INTRODUCTION

There are two major approaches for mapping genes that are associated with human disease, the linkage analysis and the association analysis. Recently, association studies have become more and more important in the search for genes that contribute to complex disease susceptibility. The simplest design for a genetic association with a binary trait is a case-control study. In a case-control study, we compare the allele frequencies or genotype frequencies in a group of independent cases and a set of controls. The controls can be either true controls (i.e. disease-free individuals) or population controls.

One problem with association studies is that spurious association may occur as a result of population stratification. Population stratification refers to the phenomenon in which multiple population subtypes are hidden in a population that appear to be homogeneous. If these subtypes are associated with different risks of the disease, the population composition then becomes a confounder (Kleibaum et al., 1982). To perform an association test in the presence of population stratification, subpopulation structure should be properly adjusted for. To make the issue more complicated, the memberships in the subpopulations are often not observed. Genomic control (GC) and structured association (SA) are two popular methods in dealing with the stratification problem in case-control genetic association studies (Delvlin and Roder, 1999, Pritchard et al., 2000, Reich and Goldstein, 2001). GC and SA have proven to be useful over the years, but they do have limitations. GC uses information from extra "null" markers that are not related to the disease of interest, and adjust the chi-square statistic using the median of the null-marker statistics. This method adjusts uniformly on all tested markers. However, different markers might have different population

allele frequencies across ancestral populations. A uniform adjustment might not be appropriate for all the markers (Price et al., 2006). The SA method also uses information from the genotypes of extra markers to estimate whether there are subpopulations in the study population. Then it uses an association test that takes subgroup membership into account. It is computationally intensive and not suitable for a whole genome scan. It also requires the specification of the number of subgroups beforehand, which could be tricky.

The only foolproof solution for spurious association due to population stratification is a family-based association study, in which the test is conditioned on parental genotypes. The classic design for family-based association study is case trios. The data are then analyzed using the transmission disequilibrium test (TDT). In order to perform a TDT test, at least one of the parents must be heterozygous for the marker of interest. In a TDT test, we test whether the transmission rates of the two alleles from the heterozygous parent to the child are equal. Larger families, unrelated controls, or other variations can also be used in a family-based association study. The disadvantage of family-based association studies is that they have much lower power than case-control studies.

A few new methods have very recently emerged that can handle the stratification problem in a case-control study more effectively than previous methods. Price *and others* proposed a method called EIGENSTRAT that uses principal components analysis to infer continuous axes of genetic variation (Price et al., 2006). Their method performs very well, but requires a very large number of markers. Others have proposed to use a logistic regression to account for the population stratification in case-control studies (Epstein et al., 2006, Setakis et al., 2007). These methods appear to perform at least as well as the EIGENSTRAT, yet they only require about 50 markers.

Because of these methods, we are more and more confident that the population stratification problem in case-control studies can be handled well. However, many investigators still have trio data around. If only case trios are available, the TDT test is the only analysis option. However, when there are controls of some kind, various analysis options exist. The purpose of this paper is to discuss alternative analyses of trio data, especially in designs that include controls. We are oriented towards what to do with existing datasets. In the discussion we return to the question of whether it is sensible to collect trios at all any more.

4.3 FACTORIZATION OF THE LIKELIHOOD FOR FAMILY TRIOS

The first question that we are trying to answer is whether the parental genotypes add any information on the inference of the parameters of interest, namely relative risks to develop disease for different genotype groups. In most family trio designs, genotypes (but not phenotypes) of the parents are collected along with the genotype and the disease status of the child in a family trio. The study designs that we consider here are 1) case trios + independent controls; 2) case trios + control trios. Let us denote D_o the disease status of the child, $D_o = 1$ for cases, and $D_o = 0$ for controls. Let G_p and G_o be the observed genotypes of the parents and the child respectively. In general the conditional likelihood of a trio (either a case trio or a control trio) given the disease status of the offspring can be written as

$$L = Pr(G_p, G_o|D_o) \tag{4.3.1}$$

$$= \frac{Pr(D_o|G_p, G_o)Pr(G_p, G_o)}{Pr(D_o)} \tag{4.3.2}$$

If we assume that the child's disease status is independent of the parents' genotypes conditional on his/her own genotype, then equation (4.2.2) becomes

$$\frac{Pr(D_o|G_o)Pr(G_p, G_o)}{Pr(D_o)}. \tag{4.3.3}$$

With simple algebra, we can rewrite it as

$$Pr(G_o|D_o)Pr(G_p|G_o) \tag{4.3.4}$$

$$= L1 \times L2. \tag{4.3.5}$$

That is, the whole conditional likelihood $Pr(G_p, G_o|D_o)$ (denoted as L), can be factored into two independent parts, $Pr(G_o|D_o)$ (denoted as $L1$) and $Pr(G_p|G_o)$ (denoted as $L2$).

Let us define the relative risk ψ_g as

$$\psi_g = \frac{Pr(D_o = 1|G_o = g)}{Pr(D_o = 1|G_o = g_0)},$$

where g_0 is the baseline genotype group. Let p_g be the population frequency for genotype g. It is easy to show that $L1 = f1(\psi_g\text{'s}, p)$ is a function of both the ψ_g's and the p_g's, while $L2$ is only determined by the p_g's and the Mendelian segregation rules. Therefore, only $L1$

contains the parameters of interest, the ψ_g's. In real life, the p_g's are sometimes known (e.g. estimated from larger population sample). Under this situation, the parental genotype distribution itself does not contribute anything to the inference of the ψ_g's. If all we want to do is the overall test of whether the genotype of the marker (gene) is associated with the disease status, the most efficient test is the independent chi-square test using the children's genotype and phenotype data, i.e. the test based on $L1$ above.

When the p_g's need to be estimated from the current dataset, there are two ways that it could be done. If the controls are population controls, the p_g's can be estimated using the controls (for study design of case trios + independent controls) or the parents of the controls (for study design of case trios + control trios) if the sample size is big enough. Even if the controls are true controls, under the assumption of a rare disease, we can still estimate the p_g's from the controls. The p_g's cannot be well estimated from the controls directly if the sample size is small or if the disease is common and the controls are true controls. A joint likelihood approach that combines the likelihoods of the case trios with the available controls to estimate the p_g's and the ψ_g's simultaneously can be used in these situations. A better estimate of the p_g's can potentially improve the efficiency of the estimate of the ψ_g's. We will discuss this approach in detail for different study designs in later sections.

4.4 ALTERNATIVE FACTORIZATION OF THE LIKELIHOOD

An alternative factorization of the likelihood shown in equation (4.2.1) is

$$\begin{aligned} L = \quad & Pr(G_p, G_o|D_o) \\ = \quad & Pr(G_o|G_p, D_o)Pr(G_p|D_o) \qquad (4.4.6) \\ = \quad & L3 \times L4. \qquad (4.4.7) \end{aligned}$$

$L3$ contains transmission information, that is how different alleles of the heterozygous parents are transmitted to the child. $L4$ contains founder information, namely the distribution of the parental genotypes for cases or controls. It is easy to show that both parts are functions of the ψ_g's and the p_g's. Moreover, if we estimate the ψ_g's using $L3$ and test the hypothesis

of $\psi_g = 0$, it should be equivalent to the transmission tests (e.g. the TDT test). This explains why the TDT test has low power to test for association between marker alleles and disease status. It uses only part of the information that involves the ψ_g's. A test based on $L4$ should be equivalent to a case-control test (e.g. chi-square test) of the founders' (the parents) genotypes by the non-founders' (the children's) disease status.

Furthermore, this factorization of the likelihood suggests another analysis strategy for designs that involve family trios – do both a stratification-proof test (e.g. the TDT test) and a founder test (e.g. a chi-square test of the founders' genotype). The PBAT is an example of a quantitative-trait TDT that makes use of this factorization. They screen on the founder test and then do a final test on the transmission test (VanSteen et al., 2005). The basic idea of the PBAT is to increase the power of the analysis by reducing the number of the final transmission tests. However, under severe population stratification, the list produced by the founder test in the first stage is questionable. We will discuss this in detail in the discussion section.

To simplify the presentation, we will continue our discussion under the assumption that the controls are true controls (i.e. disease-free individuals). However, the basic principles discussed in this paper can be applied to studies that use either kind of controls.

4.5 COMBINED ANALYSIS OF CASE TRIOS AND EXTERNAL CONTROLS

In many studies, both trios and unrelated controls are collected. In some cases, the trios were collected to confirm some association found by a previous case-control study. In other cases, unrelated controls were collected to adjust for potential confounders (Epstein et al., 2005). In such situations, a combined analysis that uses all the data is usually preferred. Several methods have been developed to address this type of analysis (Epstein et al., 2005, Nagelkerke et al., 2004, Whittemore et al., 2000). These likelihood-based approaches combined founder information and the transmission information to increase the power of the association test. Although these methods differ in the way the tests are constructed and whether they make

some certain population assumptions (e.g. HWE), the overall conditional likelihood for the trios is the same for all of them. Assuming a sample of I trios and J independent controls, the whole conditional likelihood for both the case trios and the independent controls is then

$$\begin{aligned} L &= L_{\text{case trios}} \times L_{\text{control}} \\ &= \prod_{i=1}^{I} Pr(G_{pi}, G_{oi}|D_{oi}=1) \times \prod_{j=1}^{J} Pr(G_j|D_j=0). \end{aligned} \tag{4.5.8}$$

From equation (4.3.6), the likelihood of the trios can be separated into two parts, one contains founder informations the other contains non-founder information:

$$L = \prod_{i=1}^{I} Pr(G_{oi}|D_{oi}=1)Pr(G_{pi}|G_{oi}) \times \prod_{j=1}^{J} Pr(G_j|D_j=0). \tag{4.5.9}$$

As discussed above, when the p_g estimates are available from external information or when the sample size is large so that the p_g's can be estimated accurately from the controls, the only parts of the likelihood that are involved in the estimation of the ψ_g's are $\prod_{i=1}^{I} Pr(G_{pi}, G_{oi}|D_{oi}=1)$ and $\prod_{j=1}^{J} Pr(G_j|D_j=0)$, so an overall chi-square test for the children's and the independent controls' genotypes and phenotypes should be the most efficient test. To avoid spurious association results, the test statistics should be adjusted using one of the methods that account for the population stratification. It is only when we don't have a good estimate of the p_g's that, the combined likelihood methods are needed.

Epstein et al. demonstrated in their papers that their method is more powerful than the traditional TDT test (Epstein et al., 2005) by simulation studies. This is also suggested by the alternative factorization of the likelihood shown in equation (4.3.6). They did not compare the power of their methods to the overall chi-square test of the cases and controls. It will be interesting to see how much, if any, power can be gained by including the parents' genotype data in the analysis.

One important issue in this type of analysis is whether the trios and the independent controls *should* be combined. Both Epstein et al. and Nagelkerke et al. proposed tests for testing whether the trios and the unrelated controls can be combined. Whittemore et al. also developed a score statistic that decomposes nicely into two components, the NFS (non-founder statistic) and the FS (founder statistic). The FS is related to $L4$ in equation

(4.3.7). It compares the founder's (the parents') genotype distribution to that of the general population to which they belong. It is not robust to population stratification or inappropriate assumptions (e.g. random mating or HWE).The NFS tests for the deviation of the observed and the expected marker alleles of the non-founders (the children), conditional on the founder genotypes. The NFS is related to the TDT statistic, which uses information contained in $L3$ in equation (7), and thus is robust to population stratification. Therefore, we can use both the FS and the NFS to test for the association and compare the results. At the presence of population stratification, the NFS should be used in testing the association between the marker genotype and the disease. The drawback of this method is that the transmission test is still low in power to detect association between the marker and the disease of interests.

4.6 COMBINED ANALYSIS OF CASE TRIOS AND CONTROL TRIOS

In some cases, both case trios and control trios are available. This is particularly common in studies of early developmental disease or birth defects. Assume that there are I case trios and J control trios. The combined likelihood for the whole data now becomes

$$\begin{aligned} L &= L_{\text{case trios}} \times L_{\text{control trios}} \\ &= \prod_{i=1}^{I} Pr(G_{pi}, G_{oi}|D_{oi}=1) \times \prod_{j=1}^{J} Pr(G_{pj}, G_{oj}|D_{oj}=0) \\ &= \prod_{i=1}^{I} Pr(G_{oi}|D_{oi}=1)Pr(G_{pi}|G_{oi}) \times \prod_{j=1}^{J} Pr(G_{oj}|D_{oj}=0)Pr(G_{pj}|G_{oj}). \quad (4.6.10) \end{aligned}$$

As before, when the p_g's can be estimated independently, the portions of the likelihood that contribute to the ψ_g's are $\prod_{i=1}^{I} Pr(G_{oi}|D_{oi}=1)$ and $\prod_{j=1}^{J} Pr(G_{oj}|D_{oj}=0)$. Therefore, to test the overall association of the marker genotype and the disease status, we just need to do a chi-square test using the case children and the control children with proper adjustment for population stratification. This test does not use parental information at all.

When the p_g estimates cannot be obtained easily from the controls, a combined likelihood approach might be more appropriate. The method proposed by Epstein *and others* can be

easily extended to be applied to the data that contain case trios and control trios (Epstein et al., 2005, Nagelkerke et al., 2004, Whittemore et al., 2000). Details of an example of the extension of their method are given in the Appendix C. Again, case and control trios are assumed to be from the same population. When this assumption is violated, we then should use a transmission disequilibrium test, which has low power.

Another option is to use an add-hoc combination (e.g. a weighted average) of the TDT statistics and the chi-square statistic of the parents' genotype in case trios and control trios. Similar idea worked well in QTL analysis of sib pairs (Forrest and Feingold, 2000). Recently, Kazeem and Farall also presented a similar idea for combined analysis of independent case-control and TDT studies (Kazeem and Farrall, 2005). However, it is hard to determine on the optimal weights for the two types of statistics.

Alternatively, an overall chi-square tests can be applied to test the association of the genotypes and phenotypes of the children with proper adjustment for subpopulation structure. This test should be much more powerful than the transmission disequilibrium test. It accounts for the population stratification, yet it is straightforward to apply.

4.7 ANALYSIS OF TRISOMIC CASE TRIOS AND CONTROL TRIOS

4.7.1 TDT analysis

Our trisomic case-control data consists of trios, case trios plus control trios. One important issue that is not considered in any of the TDT tests or the likelihood-based combined analysis methods is transmission in controls. This is a more severe problem when dealing with trisomic data. A good portion of trisomic fetuses do not survive to term. For example, only 20% of the clinically recognized trisomy 21 conceptuses survived to term (Hassold and Jacobs, 1984). Therefore, the cases and the controls that we observe are not really random samples from the population. They are from the subset of the population that survive to term. It is very possible that we will observe segregation distortion in the case trios on a marker due to gene-specific selection effects that have nothing to do with the disease (Xu et al., 2004,

Kerstann et al., 2004). Xu *and others* proposed a TDT test for trisomic trios (Xu et al., 2004). The TDT test can be applied to either the case trios or the control trios. A positive test result in the control trios suggests that a locus near this marker is associated with the survival of the trisomic embryo. A positive test result in the case trios can be explained by confounded effect of the marker on selection and/or on the trait of interest. Some genes may be associated with susceptibility to the trait only, but not survival. In that case, positive results will only be observed in the case trios. Some genes might be involved in the survival of the embryo only. If this is true, similar results will be seen in both case and control trios. However, it is also possible that there are genes that are involved in both processes. Therefore, to fully understand the genetic mechanism of the disease, it is recommanded that the TDT should be applied separately to the case trios and the control trios.

4.7.2 Combined analysis

To do a combined analysis of the trisomic trio data (case trios + control trios), the strategies mentioned in the last section are also applicable. To date, no combined likelihood method has been developed for trisomic data, although most of the methods developed for dismic trio data can be extended to deal with trisomic trios. The most efficient test for overall association is still the chi-square test using the trisomic case children and control children. However, this analysis will not yield details of whether the gene is associated with survival of the embryos and/or disease status as the separate TDT analyzes do.

4.8 DISCUSSION

When case trio data with either independent controls or control trios are available, the following strategies can be used to analyze the data.

1. Apply a TDT analysis of the case trios and control trios (when available). This is equivalent to do the inference using $L3$ only. The TDT is robust for population stratification. However, it has low power to detect association. On the other hand, this strategy pro-

vides important additional information in the analysis of trisomic trio data.

2. Use a combined likelihood approach, which is currently available for disomic case trios and controls. These methods can be easily extended to trisomic trio data. This approach uses the complete likelihood L. When the assumptions are correct, i.e. the cases and controls are from the same population so that the two type of data can be combined, this should be the most powerful test. When this assumption is not right, the test is biased.
3. Use an ad-hoc statistic that takes the form of the weighted mean of a regular chi-square statistic for the parents' genotype and the TDT statistic. This method uses information contains in both $L3$ and $L4$ simultaneously. This is an appealing approach except for the fact that it is hard to define the optimal weight for each type of statistics.
4. Use a PBAT-type approach. This approach uses $L3$ and $L4$ separately in two different stages of the analysis. PBAT increases the overall power of the analysis by reducing the number of transmission tests in the final stage. However, it is questionable whether it is appropriate to use a case-control statistic to select the list in the first stage. Under the presence of severe population stratification, genes that make the list might be the ones that distributed differently in different subpopulation, and has nothing to do with the disease of interest.
5. Do an overall chi-square test on the cases and controls (ignoring parental information), adjust for subpopulation structure using one of the newly developed methods that accounts for stratification. This method uses information contained in $L1$ only. As discussed before, this is the most efficient test when the sample size is large enough. This approach requires genotyping of extra "null" markers. With the advances of high throughput genotyping technologies, this should not be a big problem for most the studies. We believe that this is the best strategy for this type of analysis, even when the sample size is small. Potentially we may lose some power due to less efficient estimates of the ψ_g's. We suspect the loss will be small though.

So, are trio designs obsolete? Probably, except in special situations, as for the trisomic studies. The main cost of including the parental genotype data is recruiting. For studies of the birth defects, the parents are often registered automatically. Hence it won't be too much more costly to collect their genotype data. For other studies that need more effort to

recruit the parents, such as studies for late onset disease, it is probably not worth collecting parents. It is more efficient to do a population case-control study and control for potential population stratification using one of the methods disscussed earlier.

So far, we discuss the situation where only the parental genotypes but not phenotypes are collected along with the genotypes and phenotypes of the children. If the parental phenotypes are available, it is an entirely different situation. The likelihood for a trio becomes

$$L = \quad Pr(G_p, G_o | D_o, D_p),$$

where D_p is the disease status of the parents. It is easily seen that the parents data will also contribute to estimation of the ψ_g's. That is, collecting parents *only* for their genotype data is almost certainly wasteful, but if parental phenotype are also available and scientifically relevant, then trios may indeed be useful.

5.0 MULTIPOINT EXTENSION OF TRISOMIC TDT

5.1 MOTIVATION

As explained earlier, the TDT statistic is robust to population stratification, assortive mating, and other factors that can distort the parental distribution. It has been the most popular test used in family-based control studies. A trisomic TDT test was developed by Xu et al. 2004. It is a likelihood ratio test comparing the likelihood of the data under random segregation model to the likelihood of the data under a model that allows non-random segregation (Xu et al. 2004). Assume N_k families are of the k^{th} mating type. Mating types are defined by different combinations of the parental genotypes of a family. Let n_{0k}, n_{1k}, n_{2k} and n_{3k} denote the number of families with the offspring of genotype AAA, AAB, ABB and BBB respectively. The multinomial likelihood is expressed as

$$L = \prod_{k=1}^{5} \frac{N_k!}{n_{0k}!n_{1k}!n_{2k}!n_{3k}!} P_{0k}^{n_{0k}} P_{1k}^{n_{1k}} P_{2k}^{n_{2k}} P_{3k}^{n_{3k}},$$

where P_{0k}, P_{1k}, P_{2k} and P_{3k} are the probabilities of surviving to term with phenotype of interest (CHD in our case) conditioning on the parents' genotypes of the k^{th} mating type for the four offspring genotypes. P_{ik}'s are functions of allelic association parameters (w's, as described in section 2.3.2), as well as genetic map parameters (h's, as described in section 2.3.2). The map parameter, h, could estimated simultaneously with the w's, but the test could be adapted to use map parameters that are known from other sources.

The trisomic TDT test is applicable for transmission disequilibrium of the two alleles for a single marker. When multiple markers exist, multiple comparison problems emerge. The tests for different markers are not independent, since markers that are close together are closely related to each other. Therefore, if we use traditional Bonferroni correction to correct for multiple comparisons, the power of the test will be very low. We attempted to extend the trisomic TDT to a multi-marker test by applying GEE to account for the correlation across different markers. The general idea for a test for two markers is described in the next section. However, this method turns out to be impractical, as discussed in section 4.2.2. We will come up with more practical solutions for our data analysis.

5.2 TWO MARKER TRISOMIC TDT TEST

Assume we have two biallelic markers. We adapted the notation from Xu et al. 2004. For the marker locus g, we denote

$$\begin{aligned} h_g &= \text{probability before selection of disomic homozygosity at locus } g, \\ w_{0g} &= \text{probability of survival with disease phenotype of genotype AAA at locus } g, \\ w_{1g} &= \text{probability of survival with disease phenotype of genotype AAB at locus } g, \\ w_{2g} &= \text{probability of survival with disease phenotype of genotype ABB at locus } g, \\ w_{3g} &= \text{probability of survival with disease phenotype of genotype BBB at locus } g. \end{aligned}$$

5.2.1 Set up of the test

Only the families with at least one heterozygous parent for the marker are informative for TDT test. A single family may be informative for some markers but non-informative for other markers. We consider three types of families in our two marker trisomic TDT test:

Type I family: informative for both markers. Assume that we have m such families.

Type II family: only 1^{st} marker is informative. Assume that we have n_1 such families.

Type III family: only 2^{nd} marker is informative. Assume that we have n_2 such families.

We want to construct a test for $H_0 : w_{1g} = w_{2g} = w_{3g} = 1$. There are 5 different mating types for a informative marker. Our final marginal likelihood is a product of 5 multinomial likelihoods. The mating types and their corresponding probability of each offspring genotype conditioning on the mating type are listed in Table 3.1. The marginal likelihood for marker g is

$$L_g = \prod_{k=1}^{5} \frac{N_{kg}}{n_{0kg}!n_{1kg}!n_{2kg}!n_{3kg}!} P_{0kg}^{n_{0kg}} P_{1kg}^{n_{1kg}} P_{2kg}^{n_{2kg}} P_{3kg}^{n_{3kg}}.$$

For each family i, the two scores (derived from the marginal likelihood for each marker) are:

$$S_{1i} = \frac{\partial f(x_i, \theta_1)}{\partial \theta_1} \qquad \text{and} \qquad S_{2i} = \frac{\partial f(x_i, \theta_2)}{\partial \theta_2}$$

$$\text{where } \theta_1 = (h_1, w_{01}, w_{11}, w_{21}, w_{31})^T \text{ and } \theta_2 = (h_2, w_{02}, w_{12}, w_{22}, w_{32})^T$$

For type I family, the contribution to the score is

$$T_1 = (\sum_{i=1}^{m} S_{1i}^T, \sum_{i=1}^{m} S_{2i}^T)\Sigma^{-1}(\sum_{i=1}^{m} S_{1i}, \sum_{i=1}^{m} S_{2i})^T$$

For type II family, the contribution to the score is

$$T_2 = (\sum_{i=1}^{n1} S_{1i}^T)\Sigma_{11}^{-1}(\sum_{i=1}^{n1} S_{1i})^T$$

For type III family, the contribution to the score is

$$T_3 = (\sum_{i=1}^{n2} S_{2i}^T)\Sigma_{22}^{-1}(\sum_{i=1}^{n2} S_{2i})^T$$

The final score statistic becomes

$$T = T_1 + T_2 + T_3.$$

Here

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

where

$$\Sigma_{11} = \frac{\sum_{i=1}^{m+n_1}\{\left(S_1(x_i,\theta_{10}) - \bar{S}_1(x_i,\theta_{10})\right)\left(S_1(x_i,\theta_{10}) - \bar{S}_1(x_i,\theta_{10})\right)^T\}}{m+n_1-1}$$

$$\Sigma_{12} = \frac{\sum_{i=1}^{m}\{\left(S_1(x_i,\theta_{10}) - \bar{S}_1(x_i,\theta_{10})\right)\left(S_2(x_i,\theta_{20}) - \bar{S}_2(x_i,\theta_{20})\right)^T\}}{m-1}$$

$$\Sigma_{22} = \frac{\sum_{i=1}^{m+n_2}\{\left(S_2(x_i,\theta_{20}) - \bar{S}_2(x_i,\theta_{20})\right)\left(S_2(x_i,\theta_{20}) - \bar{S}_2(x_i,\theta_{20})\right)^T\}}{m+n_2-1}$$

and

$$\bar{S}_1 = \text{sample mean of } S_1,\ \bar{S}_2 = \text{sample mean of } S_2.$$

Under the null hypothesis that none of the markers is associated with the outcome,

$$H_0 : \theta_{10} = (\widehat{h}_1, 1, 1, 1, 1)^T \text{ and } \theta_{20} = (\widehat{h}_2, 1, 1, 1, 1)^T.$$

$$T \sim \chi^2(6).$$

In addition, we can test association for each marker separately. To test whether marker1 is associated with the outcome, the null hypothesis is,

$$H_0 : \theta_{10} = (\widehat{h}_1, 1, 1, 1, 1)^T \text{ and } \theta_{20} = (\widehat{h}_2, 1, \widehat{w}_{12}, \widehat{w}_{22}, \widehat{w}_{32})^T.$$

Under the null hypothesis,

$$T \sim \chi^2(3).$$

Similarly, to test whether marker 2 is associated with the outcome, the null hypothesis is

$$H_0 : \theta_{10} = (\widehat{h}_1, 1, \widehat{w}_{11}, \widehat{w}_{21}, \widehat{w}_{31})^T \text{ and } \theta_{20} = (\widehat{h}_2, 1, 1, 1, 1)^T.$$

Under the null hypothesis,

$$T \sim \chi^2(3).$$

The details for derivation of the score statistic and parameter estimation are listed in Appendix D.

5.2.2 Practical issues for multi-marker trisomic TDT

The idea described above can be extended to deal with n markers, $n > 2$. However, it will not work in practice. Here is why: for 2 markers, we have 3 types of families to deal with based on the informativeness pattern of the markers. Therefore, the total score is the summation of three independent scores. In general, for n markers, we will have $C_n^1 + C_n^2 + ... + C_n^n$ types of families. That is, for 3 markers, there will be 7 types of families, for 4 markers, there will be 15 types of families... As we can see, the number of different types of families increase exponentially with the number of markers. That is, we divide the families into so many subgroups, even for decent sized data, there would be certain subgroups that are either empty or have very few observations. Therefore, we cannot estimate the variance-covariance matrix correctly.

Another issue is the h parameter. It is defined as the probability of homozygosity for a marker. In practice, for a single marker, the h parameter would be different for MI cases and MII cases, it would also be different for maternal and paternal non-disjoining events. That is, for each marker, we should have to estimate 4 h's. There would be 7 parameters for each marker. In most cases, the nondisjoining event happened in maternal meiosis and majority of the data we have are MI cases. We can chose to assume that there is only one h for each marker. However, this could lead to biased estimates. Some time (like in our case), the h parameter could be estimated by some external information. For the CHD study, the h parameters are estimated by existing microsatalite maps. We can apply the test to the data treating hs as known quantities.

As discussed above, it is hard in practice to apply the multi-marker trisomic TDT to real data. We decided that we will not continue working on this topic further in the future.

6.0 CONCLUSIONS AND DISCUSSION

Family-based genetic association studies have been proven important in studying gene associated with complex diseases. My dissertation is focused on statistical issues raised in family-based genetic association studies. The motivating data for this dissertation are generated by the atrioventricular septal defect (AVSD) study of Down sydrome (DS). Family trios were collected in this study. Therefore, my dissertation mainly dealt with analysis of trio designs. The analysis of a family-based association study consists of two main steps, genotype calling and testing for association. We have made progress in statistical issues raised in each step, as concluded in the following sections.

6.1 IMPROVED GENOTYPE CALLING METHODS OF SNP ARRAY DATA FOR FAMILY TRIOS

Many high throughput SNP genotyping technologies have been developed recently. In Chapter 2, we discussed the question of how to better call the genotypes for disomic family trio data. We felt that existing methods for genotype calling can be improved by better use of some specific features of genotype calling problem. The specific features that we considered in this chapter include:

1. Prior information about the distribution of the data, including the variance structures and the shpaes of the genotype clusters.
2. Family constraints due to Mendelian rules.
3. Limited number of clusters (1, 2 or 3 for disomic data).

We also developed two family-based genotype calling methods, namely the trio K-means method, and the family-based mixture model approaches. We compared these two methods with some other commonly used clustering methods for genotype calling using both simulation studies and a real data analysis. Our results suggest that when the data quality is not good, external information, including prior knowledge of the distribution of the data and family structure, can improve the genotype calls for family trio data significantly.

In Chapter 3, we extended the two family-based methods to trisomic trio data. We saw similar results as we did in the disomic case when we compared these two methods to other genotype calling methods. The trio beta-mixture model performed the best among the four methods compared.

One important issue in the family-based methods is how to deal with potential family errors. The family-based methods force all genotypes to follow Mendelian inheritance rules within each family. However, genetic studies often have a few errors in reported family information. Genotypes called by the family-based method using the wrong family information will cause some trouble not only for this family but also for the whole dataset. We discussed this issue in the discussion of Chapter 2, and we also suggested some practical strategies of dealing with such situation. Interestingly, when we applied the trio beta-mixture model to a real dataset generated by the AVSD study, we encountered the problem of family data mixup ourselves. The trio beta-mixture model assigned a few individuals to the obviously wrong genotype clusters. This is a good example of how the family errors can affect the results of family-based genotyping.

In Chapter 3, we used families with all three members to illustrate the application of our family-based methods. Currently, the trio beta-mixture model deals with only families with no missing memeber. However, the EM algorithm can be modified to accomondate missing parents. We plan to fix the codes for the trio beta-mixture model so that it can deal with families with missing parents in the near future.

6.2 ALTERNATIVE ANALYSIS OF TRIO DESIGNS

In Chapter 4 we discussed and compared the available methods on the analysis of trio designs. The two types of trio designs we focused on are case trios plus independent controls and case trios plus control trios. We factorized the conditional likelihood for the trios in two different ways (equations 4.2.4 and 4.3.6)and related different analysis strategies to the part(s) of the likelihood they use. We believe that, when the sample size is large, the best strategy for this type of analysis is an case-control test of the children's genotypes and phenotypes. To avoid spurious association due to population structure, the subpopulation structure should be properly adjusted. We can use one of the several new methodologies developed recently, which can handle the stratification fairly effectively, to accomplish this task.

Trisomic association studies are unique in that the samples are not really random samples from the population. Since only a small portion of the triosmy fetuses (e.g. about 20% for trisomy 21) survive to term, we are dealing with a subpopulation that survived. Therefore, when a positive result turns up in a traditional TDT test in case trios, we don't know whether the marker is associated with the disease or survival of the fetus or both. When the TDT test is applied to the case trios and the control trios separately, the combined results of these two tests provides important details regarding where the positive result is coming from. We discussed this in details in Chapter 4. Hence, for trisomic trio designs, although the more powerful test is still the overall case-control test of the children's genotypes and phenotypes, separate TDT tests are still recommanded because they provides important details in the analysis of trisomic trio data.

In the discussion of Chapter 4, we returned to the question of whether it is sensible to collect trio data any more. We concluded that the trio designs that collect parents only for their genotypes are probably obsolete now. Because armed with the recent development of methodologies that deal with population stratification, we can taken into the account of subpopulation structure in population-based association studies fairly efficiently. However for special situations, such as in trisomic case, trio design is still useful.

There is still one important questions related to the analysis of trio designs that have not been addressed in the literature or in this dissertation. The potential drawback of the

overall case-control test is that by throwing out parents' data, we lose part of the information related to the p_g's. This can potentially cause some loss in the efficiency of the estimation of the ψ_g's in a likelihood frame work. Although we suspect that the loss is small, exactly how much power might be lost has never been tested. A related question is how the combined-likelihood methods compare to the overall case-control test of children in the analysis of a trio design. None of the authors did the comparison between these two strategies. It will be interesting to see how they compare to each other under different situations, maybe by simulation studies.

APPENDIX A

EM ALGORITHM FOR TRIO GAUSSIAN-MIXTURE MODEL AND TRIO BETA-MIXTURE MODEL

Let $\mathbf{Y_i} = (y_{fi}, y_{mi}, y_{ki})$ be the observed data for the father, the mother, and the child of the i^{th} trio. Let $\mathbf{G_i} = (g_{fi}, g_{mi}, g_{ki})$ be the corresponding genotype vector. First, let us assume that we can observe the genotype vector $\mathbf{G_i}$. Then the log likelihood for the i^{th} trio is

$$\begin{aligned} l_i(Y_i, G_i, \theta) &= \log p_{\lambda=g_{fi}} + \log p_{\lambda=g_{mi}} + \log Pr(g_{ki}|g_{fi}, g_{mi}) \\ &\quad + \log f(y_{fi}, \xi_{\lambda=g_{fi}}) + \log f(y_{mi}, \xi_{\lambda=g_{mi}}) + \log f(y_{ki}, \xi_{\lambda=g_{ki}}) \end{aligned} \quad \text{(A.0.1)}$$

where

$$f(y, \xi_\lambda) = \phi(y, \mu_\lambda, \sigma^2_\lambda)$$

for trio Gaussian-mixture model and

$$\begin{aligned} f(y, \xi_\lambda) &= f(y, \alpha_\lambda, \beta_\lambda) \\ &= \frac{1}{B(\alpha, \beta)} y^{\alpha-1}(1-y)^{\beta-1} \end{aligned}$$

for trio beta-mixture model. The parameter vector is

$$\theta = (p_\lambda\text{'s},\ \xi_\lambda\text{'s})^T \quad \text{(A.0.2)}$$

A.1 ESTIMATION OF P_λ'S.

The sufficient statistic for p_λ is

$$S1_\alpha = \sum_{i=1}^{n} [1\{g_{fi} = \lambda\} + 1\{g_{mi} = \lambda\}] \tag{A.1.3}$$

E-step

$$Pr(g_{fi} = \lambda | Y_i, \theta^{(t)}) = \frac{Pr(Y_i | g_{fi} = \lambda, \theta^{(t)}) Pr(g_{fi} = \lambda, \theta^{(t)})}{\sum_{\lambda \in \Lambda = \{AA,AB,BB\}} Pr(Y_i | g_{fi} = \lambda, \theta^{(t)}) Pr(g_{fi} = \lambda, \theta^{(t)})}$$
$$= \frac{p_\lambda f(y_{fi}, \xi_\lambda) \sum_{g_{mi}} \sum_{g_{ki}} p_{\lambda = g_{mi}} Pr(g_{ki} | g_{mi}, g_{fi} = \lambda, \theta^{(t)}) \prod_{\gamma \in \{mi,ki\}} f(y_\gamma, \xi_{g_\gamma})}{\sum_{g_{fi}} \sum_{g_{mi}} \sum_{g_{ki}} p_{\lambda = g_{fi}} p_{\lambda = g_{mi}} Pr(g_{ki} | g_{mi}, g_{fi}, \theta^{(t)}) \prod_{\gamma \in \{fi,mi,ki\}} f(y_\gamma, \xi_{g_\gamma})}$$

Similarly

$$Pr(g_{mi} = \lambda | Y_i, \theta^{(t)})$$
$$= \frac{p_\lambda f(y_{mi}, \mu_\lambda) \sum_{g_{fi}} \sum_{g_{ki}} p_{\lambda = g_{fi}} Pr(g_{ki} | g_{mi} = \lambda, g_{fi}, \theta^{(t)}) \prod_{\gamma \in \{fi,ki\}} f(y_\gamma, \mu_{g_\gamma})}{\sum_{g_{fi}} \sum_{g_{mi}} \sum_{g_{ki}} p_{\lambda = g_{fi}} p_{\lambda = g_{mi}} Pr(g_{ki} | g_{mi}, g_{fi}, \theta^{(t)}) \prod_{\gamma \in \{fi,mi,ki\}} f(y_\gamma, \mu_{g_\gamma})}$$

$$E(S_{1,\lambda} | Y, \theta^{(t)})$$
$$= \sum_{i=1}^{n} \left[Pr(g_{fi} = \lambda | Y_i, \theta^{(t)}) + Pr(g_{mi} = \lambda | Y_i, \theta^{(t)}) \right] \tag{A.1.4}$$

M-step

$$p_\lambda^{(t+1)} = \frac{E(S_{1,\lambda} | Y, \theta^{(t)})}{2n}. \tag{A.1.5}$$

A.2 ESTIMATION OF THE NORMAL COMPONENTS.

For the trio Gaussian-mixture model, the sufficient statistics for μ_λ and σ^2_λ are:

$$S_{1,\lambda} = \sum_{i=1}^{n} \left[1\{g_{fi} = \lambda\} + 1\{g_{mi} = \lambda\}\right], \tag{A.2.6}$$

$$S_{2,\lambda} = \sum_{i=1}^{n} \left[1\{g_{fi} = \lambda\}y_{fi} + 1\{g_{mi} = \lambda\}y_{mi} + 1\{g_{ki} = \lambda\}y_{ki}\right], \tag{A.2.7}$$

$$S_{3,\lambda} = \sum_{i=1}^{n} \left[1\{g_{fi} = \lambda\}y^2_{fi} + 1\{g_{mi} = \lambda\}y^2_{mi} + 1\{g_{ki} = \lambda\}y^2_{ki}\right], \tag{A.2.8}$$

$$S_{4,\lambda} = \sum_{i=1}^{n} \mathbf{1}\{g_{ki} = \lambda\}. \tag{A.2.9}$$

E-step

At E-step, we calculate $E(S_{1,\lambda}|Y, \theta^{(t)})$, $E(S_{2,\lambda}|Y, \theta^{(t)})$, $E(S_{3,\lambda}|Y, \theta^{(t)})$, and $E(S_{4,\lambda}|Y, \theta^{(t)})$. As shown above,

$$\begin{aligned}
&E(S_{1,\lambda}|Y, \theta^{(t)}) \\
&= \sum_{i=1}^{n} \left[\frac{p_\lambda \phi(y_{fi}, \mu_\lambda, \sigma^2_\lambda) \sum_{g_{mi}} \sum_{g_{ki}} p_{\lambda=g_{mi}} Pr(g_{ki}|g_{mi}, g_{fi} = \lambda, \theta^{(t)}) \prod_{\gamma\in\{mi,ki\}} \phi(y_\gamma, \mu_{g_\gamma}, \sigma^2_{g_\gamma})}{\sum_{g_{fi}} \sum_{g_{mi}} \sum_{g_{ki}} p_{\lambda=g_{fi}} p_{\lambda=g_{mi}} Pr(g_{ki}|g_{mi}, g_{fi}, \theta^{(t)}) \prod_{\gamma\in\{fi,mi,ki\}} \phi(y_\gamma, \mu_{g_\gamma}, \sigma^2_{g_\gamma})} \right] \\
&+ \sum_{i=1}^{n} \left[\frac{p_\lambda \phi(y_{mi}, \mu_\lambda, \sigma^2_\lambda) \sum_{g_{fi}} \sum_{g_{ki}} p_{\lambda=g_{fi}} Pr(g_{ki}|g_{mi} = \lambda, g_{fi}, \theta^{(t)}) \prod_{\gamma\in\{fi,ki\}} \phi(y_\gamma, \mu_{g_\gamma}, \sigma^2_{g_\gamma})}{\sum_{g_{fi}} \sum_{g_{mi}} \sum_{g_{ki}} p_{\lambda=g_{fi}} p_{\lambda=g_{mi}} Pr(g_{ki}|g_{mi}, g_{fi}, \theta^{(t)}) \prod_{\gamma\in\{fi,mi,ki\}} \phi(y_\gamma, \mu_{g_\gamma}, \sigma^2_{g_\gamma})} \right].
\end{aligned}$$

Similar to the derivation shown from last section,

$$\begin{aligned}
&Pr(g_{ki} = \lambda|Y_i, \theta^{(t)}) \\
&= \frac{\phi(y_{ki}, \mu_\lambda, \sigma^2_\lambda) \sum_{g_{fi}} \sum_{g_{mi}} p_{\lambda=g_{fi}} p_{\lambda=g_{mi}} Pr(g_{ki} = \lambda|g_{mi}, g_{fi}, \theta^{(t)}) \prod_{\gamma\in fi,mi} \phi(y_\gamma, \mu_\gamma, \sigma^2_\gamma)}{\sum_{g_{fi}} \sum_{g_{mi}} \sum_{g_{ki}} p_{\lambda=g_{fi}} p_{\lambda=g_{mi}} Pr(g_{ki}|g_{mi}, g_{fi}, \theta^{(t)})\} \prod_{\gamma\in\{fi,mi,ki\}} \phi(y_\gamma, \mu_{g_\gamma}, \sigma^2_{g_\gamma})}.
\end{aligned}$$

Hence

$$\begin{aligned}
&E(S_{4,\lambda}|Y, \theta^{(t)}) = \sum_{i=1}^{n} Pr(g_{ki} = \lambda|Y_i, \theta^{(t)}) \\
&= \sum_{i=1}^{n} \frac{\sum_{g_{fi}} \sum_{g_{mi}} p_{\lambda=g_{fi}} p_{\lambda=g_{mi}} Pr(g_{ki}|g_{mi}, g_{fi}, \theta^{(t)}) \prod_{\gamma\in fi,mi} \phi(y_\gamma, \mu_\gamma, \sigma^2_\gamma)}{\sum_{g_{fi}} \sum_{g_{mi}} \sum_{g_{ki}} p_{\lambda=g_{fi}} p_{\lambda=g_{mi}} Pr(g_{ki}|g_{mi}, g_{fi}, \theta^{(t)})\} \prod_{\gamma\in\{fi,mi,ki\}} \phi(y_\gamma, \mu_{g_\gamma}, \sigma^2_{g_\gamma})},
\end{aligned}$$

and

$$
\begin{aligned}
&E(S_{2,\lambda}|Y,\theta^{(t)}) \\
&= \sum_{i=1}^{n}\left[Pr(g_{fi}=\lambda|Y_i,\theta^{(t)})y_{fi} + Pr(g_{mi}=\lambda|Y_i,\theta^{(t)})y_{mi} + Pr(g_{ki}=\lambda|Y_i,\theta^{(t)})y_{ki}\right], \\
&E(S_{3,\lambda}|Y,\theta^{(t)}) \\
&= \sum_{i=1}^{n}\left[Pr(g_{fi}=\lambda|Y_i,\theta^{(t)})y_{fi}^2 + Pr(g_{mi}=\lambda|Y_i,\theta^{(t)})y_{mi}^2 + Pr(g_{ki}=\lambda|Y_i,\theta^{(t)})y_{ki}^2\right].
\end{aligned}
$$

M-step

$$
\mu_{\lambda}^{(t+1)} = \frac{E(S_{2,\lambda}|Y,\theta^{(t)})}{E(S_{1,\lambda}|Y,\theta^{(t)}) + E(S_{4,\lambda}|Y,\theta^{(t)})} \tag{A.2.10}
$$

$$
\sigma_{\lambda}^{2(t+1)} = \frac{E(S_{3,\lambda}|Y,\theta^{(t)})}{E(S_{1,\lambda}|Y,\theta^{(t)}) + E(S_{4,\lambda}|Y,\theta^{(t)})} - (\mu_{\lambda}^{(t+1)})^2 \tag{A.2.11}
$$

A.3 ESTIMATION OF THE BETA PARAMETERS.

The part of the log likelihood that involves $alpha_\lambda$ and β_λ) is:

$$
\begin{aligned}
l_\lambda = &\sum_{i=1}^{n}[1\{g_{fi}=\lambda\}\log f(y_{fi},\alpha_\lambda,\beta_\lambda) + 1\{g_{mi}=\lambda\}\log f(y_{mi},\alpha_\lambda,\beta_\lambda) \\
&+1\{g_{ki}=\lambda\}\log f(y_{ki},\alpha_\lambda,\beta_\lambda)]
\end{aligned} \tag{A.3.12}
$$

E-step

At E-step, we calculate

$$
\begin{aligned}
E(l_\lambda|Y,\theta^{(t)}) = &\sum_{i=1}^{n} Pr(g_{fi}=\lambda|Y_i,\theta^{(t)})\log f(y_{fi},\alpha_\lambda,\beta_\lambda) \\
&+\sum_{i=1}^{n} Pr(g_{mi}=\lambda|Y_i,\theta^{(t)})\log f(y_{mi},\alpha_\lambda,\beta_\lambda) \\
&+\sum_{i=1}^{n} Pr(g_{ki}=\lambda|Y_i,\theta^{(t)})\log f(y_{ki},\alpha_\lambda,\beta_\lambda)
\end{aligned} \tag{A.3.13}
$$

As shown above,

$$
\begin{aligned}
&Pr(g_{fi} = \lambda | Y_i, \theta^{(t)}) \\
&= \frac{p_\lambda f(y_{fi}, \alpha_\lambda, \beta_\lambda) \sum_{g_{mi}} \sum_{g_{ki}} p_{\lambda=g_{mi}} Pr(g_{ki} | g_{mi}, g_{fi} = \lambda, \theta^{(t)}) \prod_{\gamma \in \{mi,ki\}} f(y_\gamma, \alpha_{g_\gamma}, \beta_{g_\gamma})}{\sum_{g_{fi}} \sum_{g_{mi}} \sum_{g_{ki}} p_{\lambda=g_{fi}} p_{\lambda=g_{mi}} Pr(g_{ki} | g_{mi}, g_{fi}, \theta^{(t)}) \prod_{\gamma \in \{fi,mi,ki\}} f(y_\gamma, \alpha_{g_\gamma}, \beta_{g_\gamma})} \\
&Pr(g_{mi} = \lambda | Y_i, \theta^{(t)}) \\
&= \frac{p_\lambda f(y_{mi}, \alpha_\lambda, \beta_\lambda) \sum_{g_{fi}} \sum_{g_{ki}} p_{\lambda=g_{fi}} Pr(g_{ki} | g_{mi} = \lambda, g_{fi}, \theta^{(t)}) \prod_{\gamma \in \{fi,ki\}} f(y_\gamma, \alpha_{g_\gamma}, \beta_{g_\gamma})}{\sum_{g_{fi}} \sum_{g_{mi}} \sum_{g_{ki}} p_{\lambda=g_{fi}} p_{\lambda=g_{mi}} Pr(g_{ki} | g_{mi}, g_{fi}, \theta^{(t)}) \prod_{\gamma \in \{fi,mi,ki\}} f(y_\gamma, \alpha_{g_\gamma}, \beta_{g_\gamma})}.
\end{aligned}
$$

Similar to the derivation shown from last section,

$$
\begin{aligned}
&Pr(g_{ki} = \lambda | Y_i, \theta^{(t)}) \\
&= \frac{f(y_{ki}, \alpha_\lambda, \beta_\lambda) \sum_{g_{fi}} \sum_{g_{mi}} p_{\lambda=g_{fi}} p_{\lambda=g_{mi}} Pr(g_{ki} = \lambda | g_{mi}, g_{fi}, \theta^{(t)}) \prod_{\gamma \in fi,mi} f(y_\gamma, \alpha_\gamma, \beta_\gamma)}{\sum_{g_{fi}} \sum_{g_{mi}} \sum_{g_{ki}} p_{\lambda=g_{fi}} p_{\lambda=g_{mi}} Pr(g_{ki} | g_{mi}, g_{fi}, \theta^{(t)})\} \prod_{\gamma \in \{fi,mi,ki\}} f(y_\gamma, \alpha_{g_\gamma}, \beta_{g_\gamma})}.
\end{aligned}
$$

M-step

In M-step, we maximize $E(l_\lambda | Y, \theta^{(t)})$using the *nlm* procedure included in the *R*-package to get $\alpha_\lambda^{(t+1)}$ and $\beta_\lambda^{(t+1)}$.

APPENDIX B

ALGORITHM FOR PARAMETER ESTIMATION FOR TRISOMIC TRIO BETA-MIXTURE MODEL

B.1 COMPLETE DATA LIKELIHOOD

Let $\mathbf{Y_i} = (y_{Ni}, y_{Ci}, y_{Ki})$ be the observed data for the NDJP, CDJP and the child of the i^{th}trio. Let $\mathbf{G_i} = (g_{Ni}, g_{Ci}, g_{Ki})$ be the corresponding genotype vector. The likelihood for complete data of a trio is then:

$$L_i(Y_i, G_i, h_i, \theta) = Pr(g_{Ni})Pr(g_{Ci})Pr(g_{Ki}|g_{Ni}, g_{Ci})Pr(y_{Ni}|g_{Ni})Pr(y_{Ci}|g_{Ci}Pr(y_{Ki}|g_{Ki})$$

The parameter vector is

$$\theta = (p_{\lambda 1}s, \alpha_{\lambda 1}s, \alpha_{\lambda 2}s, \beta_{\lambda 1}s, \beta_{\lambda 2}s)^T$$

where

$$\lambda 1 \in \Lambda 1 = \{AA, AB, BB\}, \text{and } \lambda 2 \in \Lambda 2 = \{AAA, AAB, ABB, BBB\}.$$

Therefore,

$$\begin{aligned} L_i(Y_i, G_i, h_i, \theta) = \quad & p_{\lambda 1=g_{Ni}} p_{\lambda 1=g_{Ci}} Pr(g_{Ki}|g_{Ni}, g_{Ci}) \\ & f(y_{Ni}, \alpha_{\lambda 1=g_{Ni}}, \beta_{\lambda 1=g_{Ni}}) f(y_{Ci}, \alpha_{\lambda 1=g_{Ci}}, \beta_{\lambda 1=g_{Ci}}) f(y_{Ki}, \alpha_{\lambda 2=g_{Ki}}, \beta_{\lambda 2=g_{Ki}}) \end{aligned}$$

and the log likelihood is

$$\begin{aligned} l_i(Y_i, G_i, h_i, \theta) &= \log p_{\lambda 1 = g_{Ni}} + \log p_{\lambda 1 = g_{Ci}} + \log Pr(g_{Ki} | g_{Ni}, g_{Ci}) \\ &\quad + \log f(y_{Ni}, \alpha_{\lambda 1 = g_{Ni}}, \beta_{\lambda 1 = g_{Ni}}) \\ &\quad + \log f(y_{Ci}, \alpha_{\lambda 1 = g_{Ci}}, \beta_{\lambda 1 = g_{Ci}}) \\ &\quad + \log f(y_{Ki}, \alpha_{\lambda 2 = g_{Ki}}, \beta_{\lambda 2 = g_{Ki}}) \end{aligned} \tag{B.1.1}$$

B.2 ESTIMATION OF $\nu_{\lambda 1}$'S.

The sufficient statistic for $\nu_{\lambda 1}$ is

$$S_{1,\lambda 1} = \sum_{i=1}^{n} [1\{g_{Ni} = \lambda 1\} + 1\{g_{Ci} = \lambda 1\}] \tag{B.2.2}$$

E-step

At the E-step, we calculate $E(S_{1,\lambda 1} | Y, \theta^{(t)})$.

$$\begin{aligned} &Pr(g_{Ni} = \lambda 1 | Y_i, \theta^{(t)}) = \frac{Pr(Y_i | g_{Ni} = \lambda 1, \theta^{(t)}) Pr(g_{Ni} = \lambda 1, \theta^{(t)})}{\sum_{\lambda 1 \in \Lambda = \{AA, AB, BB\}} Pr(Y_i | g_{Ni} = \lambda 1, \theta^{(t)} Pr(g_{Ni} = \lambda 1, \theta^{(t)})} \\ &= \frac{\nu_{\lambda 1} f(y_{Ni}, \alpha_{\lambda 1}, \beta_{\lambda 1}) \sum_{g_{Ci}} \sum_{g_{Ki}} \nu_{g_{Ci}} Pr(g_{Ki} | g_{Ci}, g_{Ni} = \lambda 1, \theta^{(t)}) \prod_{\gamma \in \{Ci, Ki\}} f(y_\gamma, \alpha_{g_\gamma}, \beta_{g_\gamma})}{\sum_{g_{Ni}} \sum_{g_{Ci}} \sum_{g_{Ki}} \nu_{g_{Ni}} \nu_{g_{Ci}} Pr(g_{Ki} | g_{Ci}, g_{Ni}, \theta^{(t)})\} \prod_{\gamma \in \{Ni, Ci, Ki\}} f(y_\gamma, \alpha_{g_\gamma}, \beta_{g_\gamma})} \end{aligned}$$

Similarly

$$\begin{aligned} &Pr(g_{Ci} = \lambda 1 | Y_i, \theta^{(t)}) \\ &= \frac{\nu_{\lambda 1} f(y_{Ci}, \alpha_{\lambda 1}, \beta_{\lambda 1}) \sum_{g_{Ni}} \sum_{g_{Ki}} \nu_{g_{Ci}} Pr(g_{Ki} | g_{Ci} = \lambda 1, g_{Ni}, \theta^{(t)}) \prod_{\gamma \in \{Ni, Ki\}} f(y_\gamma, \alpha_{g_\gamma}, \beta_{g_\gamma})}{\sum_{g_{Ni}} \sum_{g_{Ci}} \sum_{g_{Ki}} \nu_{g_{Ni}} \nu_{g_{Ci}} Pr(g_{Ki} | g_{Ci}, g_{Ni}, \theta^{(t)})\} \prod_{\gamma \in \{Ni, Ci, Ki\}} f(y_\gamma, \alpha_{g_\gamma}, \beta_{g_\gamma})}. \end{aligned}$$

Therefore,

$$
\begin{aligned}
&E(S_{1,\lambda 1}|Y,\theta^{(t)}) \\
&= \sum_{i=1}^{n} \left[Pr(g_{Ni} = \lambda 1|Y_i,\theta^{(t)} + Pr(g_{Ci} = \lambda 1|Y_i,\theta^{(t)}) \right] \\
&= \sum_{i=1}^{n} \frac{\nu_{\lambda 1} f(y_{Ni},\alpha_{\lambda 1},\beta_{\lambda 1}) \sum_{g_{Ci}} \sum_{g_{Ki}} \nu_{g_{Ci}} Pr(g_{Ki}|g_{Ci},g_{Ni}=\lambda 1,\theta^{(t)}) \prod_{\gamma\in\{Ci,Ki\}} f(y_\gamma,\alpha_{g_\gamma},\beta_{g_\gamma})}{\sum_{g_{Ni}} \sum_{g_{Ci}} \sum_{g_{Ki}} \nu_{g_{Ni}}\nu_{g_{Ci}} Pr(g_{Ki}|g_{Ci},g_{Ni},\theta^{(t)})\} \prod_{\gamma\in\{Ni,Ci,Ki\}} f(y_\gamma,\alpha_{g_\gamma},\beta_{g_\gamma})} \\
&+ \sum_{i=1}^{n} \frac{\nu_{\lambda 1} f(y_{Ci},\alpha_{\lambda 1},\beta_{\lambda 1}) \sum_{g_{Ni}} \sum_{g_{Ki}} \nu_{g_{Ci}} Pr(g_{Ki}|g_{Ci}=\lambda 1,g_{Ni},\theta^{(t)}) \prod_{\gamma\in\{Ni,Ki\}} f(y_\gamma,\alpha_{g_\gamma},\beta_{g_\gamma})}{\sum_{g_{Ni}} \sum_{g_{Ci}} \sum_{g_{Ki}} \nu_{g_{Ni}}\nu_{g_{Ci}} Pr(g_{Ki}|g_{Ci},g_{Ni},\theta^{(t)})\} \prod_{\gamma\in\{Ni,Ci,Ki\}} f(y_\gamma,\alpha_{g_\gamma},\beta_{g_\gamma})}.
\end{aligned}
$$

M-step

At the M-step, we update $\nu_{\lambda 1}$ using the following formula,

$$\nu_{\lambda 1}^{(t+1)} = \frac{E(S1_\lambda|Y,\theta^{(t)})}{2n}, \tag{B.2.3}$$

where n is the number of family trios.

APPENDIX C

LIKELIHOOD-BASED METHOD FOR ASSOCIATION ANALYSIS OF CASE TRIOS AND CONTROL TRIOS

C.1 LIKELIHOOD DERIVATION

Let us denote the two alleles of the SNP of interest A and a. We code each genotype g, as the number of A allele carried by the individual of interest. The combined association test we consider are based on the likelihood proposed by Nagelkerke et al. (2004) and Epstein et al. (2005) for combining data on parental genotypes G_p, offspring genotypes G_o, and the disease outcome for the offspring D_o (1=affected, 0=unaffected). Assuming the dataset includes I case trios and J control trios, the likelihood can be written as:

$$L = \prod_{i=1}^{I} P(G_{pi}, G_{oi}|D_{oi} = 1) \times \prod_{j=1}^{J} P(G_{pj}, G_{oj}|D_{oj} = 0)$$

In order to construct L, $P(G_p, G_o|D_o = 1)$ and $P(G_p, G_o|D_o = 0)$ need to be specified. Let us define

$$P_0 = P(D_o = 1|G_o = 0),\ P_1 = P(D_o = 1|G_o = 1),\ \text{and}\ P_2 = P(D_o = 1|G_o = 2).$$

Accordingly, the relative risk (RR) for genotype group g then can be defined as

$$\psi_g = \frac{P_g}{P_0},\quad g = 1, 2.$$

- For case trios:

$$P(G_p, G_o|D_o = 1) = P(G_o = g|G_p = g_p, D_o = 1)P(G_p = g_p|D_o = 1)$$

where

$$P(G_o = g|G_p = g_p, D_o = 1) = \frac{\psi_g P(G_o = g|G_p = g_p)}{\sum_{g^*} \psi_{g^*} P(G_o = g^*|G_p = g_p)},$$

if we assume that the offspring disease risk is independent of parental genotype given offspring genotype.

Similarly,

$$P(G_p = g_p|D_o = 1) = \frac{\sum_g \psi_g P(G_o = g|G_p = g_p)P(G_p = g_p)}{\sum_{g_{p^*}} \sum_{g^*} \psi_{g^*} P(G_o = g^*|G_p = g_{p^*})P(G_p = g_{p^*})}.$$

- For control trios:

$$P(G_p, G_o|D_o = 0) = P(G_o = g|G_p = g_p, D_o = 0)P(G_p = g_p|D_o = 0)$$

where

$$P(G_o = g|G_p = g_p, D_o = 0) = \frac{(1 - \psi_g P_0)P(G_o = g|G_p = g_p)}{\sum_{g^*}(1 - \psi_{g^*} P_0)P(G_o = g^*|G_p = g_p)},$$

Again, we assume that the offspring disease risk is independent of parental genotype given offspring genotype, then

$$P(G_p = g_p|D_o = 0) = \frac{\sum_g (1 - \psi_g)P(G_o = g|G_p = g_p)P(G_p = g_p)}{\sum_{g_{p^*}} \sum_{g^*}(1 - \psi_{g^*})P(G_o = g^*|G_p = g_{p^*})P(G_p = g_{p^*})}.$$

$P(G_o|G_p, D_o = 1)$ values and $P(G_o|G_p, D_o = 0)$ values for all possible triad genotype combinations were listed in Table A1. Note that these values are different from the ones listed in Table A1. of Epstein et al. 2005. They assume that when the marker locus is not associated with the disease status (i.e. all ψ_g's equal to 1), then there is no segregation distortion. That is, the transmission rate of either allele to the offspring is 1/2. However, we can pick up segregation distortion that has nothing to do with the trait. Apparent segregation distortion can be caused by other reasons (Xu et al., 2004).

Here we generalize Epstein et al.'s likelihood by allowing segregation distortion when the locus is not associated with diseased status. Let us define

$$\begin{aligned} t_{A1} &= P(\text{A allele is transmitted}|D_o = 1), \\ t_{a1} &= P(\text{a allele is transmitted}|D_o = 1), \\ t_{A0} &= P(\text{A allele is transmitted}|D_o = 0), \\ t_{a0} &= P(\text{a allele is transmitted}|D_o = 0). \end{aligned}$$

where

$$t_{A1} + t_{a1} = 1, \quad \text{and } t_{A0} + t_{a0} = 1$$

When there is no segregation distortion,

$$t_{A1} = t_{a1} = t_{A0} = t_{a0} = 1/2.$$

When there is segregation distortion, but the gene is not associated with the disease status,

$$t_{A1} = t_{A0} \text{ and } t_{a1} = t_{a0},$$

but they are not equal to 1/2. When there is segregation distortion, and the gene is associated with the disease of interest,

$$t_{A1} \neq t_{A0} \text{ and } t_{a1} \neq t_{a0}.$$

We followed Epstein et al. and calculate $P(G_p)$ using the parental genotype distribution described by Weinberg et al. (1998). For a SNP , there are 6 possible mating types, {(2,2),(2,1), (2,0), (1,1), (1,0), (0,0)}. We define μ_l as the probability of the l^{th} mating type ($l = 1, 2, ...6$) in the population. μ_l's are functions of t_α's, $\alpha \in (A1, A0, a1, a0)$, and ψ_g's. The μ_l's may be any positive numbers that add up to 1.

C.2 TESTING HYPOTHESIS

The L specified above can be used to estimate the ψ_g's and the t_α's using standard maximum-likelihood procedures. LR test can be constructed to test the null hypothesis of $\psi_1 = \psi_2 = 1$. In addition, the null hypothesis $t_{A1} = t_{A0}$ or $t_{A1} = t_{A0} = 1/2$ can also be tested. The later one is equivalent to the regular TDT test.

Table C1: Evaluation of $P(G_o|G_p, D_o = 1)$ and $P(G_o|G_p, D_o = 0)$

G_p and G_o	$P(G_o\|G_p, D_o = 1)$	$P(G_o\|G_p, D_o = 0)$
$G_p = (2,2)$		
$G_o = 2$	1	1
$G_o = 1$	0	0
$G_o = 0$	0	0
$G_p = (2,1)$		
$G_o = 2$	$\frac{\psi_2 t_{A1}}{\psi_1 t_{a1}+\psi_2 t_{A1}}$	$\frac{(1-\psi_2 P_0)t_{A0}}{(1-\psi_2 P_0)t_{A0}+(1-\psi_1 P_0)t_{a0}}$
$G_o = 1$	$\frac{\psi_1 t_{a1}}{\psi_1 t_{a1}+\psi_2 t_{A1}}$	$\frac{((1-\psi_1 P_0)t_{a0}}{(1-\psi_2 P_0)t_{A0}+(1-\psi_1 P_0)ta0}$
$G_o = 0$	0	0
$G_p = (2,0)$		
$G_o = 2$	0	0
$G_o = 1$	1	1
$G_o = 0$	0	0
$G_p = (1,1)$		
$G_o = 2$	$\frac{\psi_2 t_{A1}^2}{\psi_2 t_{A1}^2+2\psi_1 t_{A1} t_{a1}+t_{a1}^2}$	$\frac{(1-\psi_2 P_0)t_{A0}^2}{(1-\psi_2 P_0)t_{A0}^2+2(1-\psi_1 P_0)t_{A0}t_{a0}+(1-P_0)t_{a0}^2}$
$G_o = 1$	$\frac{2\psi_1 t_{A1} t_{a1}}{\psi_2 t_{A1}^2+2\psi_1 t_{A1} t_{a1}+t_{a1}^2}$	$\frac{2(1-\psi_1 P_0)t_{A0}t_{a0}}{(1-\psi_2 P_0)t_{A0}^2+2(1-\psi_1 P_0)t_{A0}t_{a0}+(1-P_0)t_{a0}^2}$
$G_o = 0$	$\frac{t_{a1}^2}{\psi_2 t_{A1}^2+2\psi_1 t_{A1} t_{a1}+t_{a1}^2}$	$\frac{(1-P_0)t_{a0}^2}{(1-\psi_2 P_0)t_{A0}^2+2(1-\psi_1 P_0)t_{A0}t_{a0}+(1-P_0)t_{a0}^2}$
$G_p = (1,0)$		
$G_o = 2$	0	0
$G_o = 1$	$\frac{\psi_1 t_{A1}}{\psi_1 t_{A1}+t_{a1}}$	$\frac{(1-\psi_1 P_0)t_{A0}}{(1-\psi_1 P_0)t_{A0}+(1-P_0)t_{a0}}$
$G_o = 0$	$\frac{t_{a1}}{\psi_1 t_{A1}+t_{a1}}$	$\frac{(1-\psi_1 P_0)t_{A0}}{(1-P_0)t_{a0}+(1-P_0)t_{a0}}$
$G_p = (0,0)$		
$G_o = 2$	0	0
$G_o = 1$	0	0
$G_o = 0$	1	1

APPENDIX D

SUPPLEMENT MATERIAL FOR TWO-MARKER TRISOMIC TDT

D.1 DERIVATION OF SCORE FUNCTIONS FOR THE FIVE MATING TYPES

There are total of 5 possible mating types for the families that generate informative data for trisomic TDT test. The mating types and corresponding probabilities of each offspring are listed in Table D1.

Mating type I

Category 1: $P_{01} = \frac{1/2}{1/2+w_1/2} = \frac{1}{1+w_1}$.

Category 2: $P_{11} = \frac{w_1/2}{1/2+w_1/2} = \frac{w_1}{1+w_1}$

For each family in category 1,

$$S_i = \frac{\partial f(x_i,\theta)}{\partial \theta} = \begin{pmatrix} \frac{-\partial \log(1+w_1)}{\partial w_1} \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{-1}{1+w_1} \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{D.1.1}$$

Table D1: Fifteen Categories of a Informative SNP Marker for a Nuclear Family with One Trisomic Offspring

Mating Type	Category	NDJP	CDJP	Child	Probability
I	1	AA	AB	AAA	$\frac{w_0}{w_0+w_1}$
	2			AAB	$\frac{w_1}{w_0+w_1}$
II	3	AB	AA	AAA	$\frac{h}{h+2w_1(1-h)+w_2h}$
	4			AAB	$\frac{2w_1(1-h)}{h+2w_1(1-h)+w_2h}$
	5			ABB	$\frac{w_2h}{h+2w_1(1-h)+w_2h}$
III	6	AB	AB	AAA	$\frac{h}{h+(w_1+w_2)(2-h)+w_3h}$
	7			AAB	$\frac{w_1(2-h)}{h+(w_1+w_2)(2-h)+w_3h}$
	8			ABB	$\frac{w_2(2-h)}{h+(w_1+w_2)(2-h)+w_3h}$
	9			BBB	$\frac{w_3h}{h+(w_1+w_2)(2-h)+w_3h}$
IV	10	AB	BB	AAB	$\frac{w_1h}{w_1h+2w_2(1-h)+w_3h}$
	11			ABB	$\frac{2w_2(1-h)}{w_1h+2w_2(1-h)+w_3h}$
	13			BBB	$\frac{w_3h}{w_1h+2w_2(1-h)+w_3h}$
V	14	BB	AB	ABB	$\frac{w_2}{w_2+w_3}$
	15			BBB	$\frac{w_3}{w_2+w_3}$

and

$$\sum S_i = \begin{pmatrix} \frac{-n_{01}}{1+w_1} \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{D.1.2}$$

For each family in category 2,

$$S_i = \frac{\partial f(x_i,\theta)}{\partial \theta} = \begin{pmatrix} \frac{\partial[\log w_1 - \log(1+w_1)]}{\partial w_1} \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{w_1} - \frac{1}{1+w_1} \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

and

$$\sum S_i = \begin{pmatrix} \frac{n_{11}}{w_1} - \frac{n_{11}}{1+w_1} \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

Mating type II

Category 3: $P_{02} = \frac{h}{h+2w_1(1-h)+w_2h}$

Category 4: $P_{12} = \frac{2w_1(1-h)}{h+2w_1(1-h)+w_2h}$

Category 5: $P_{22} = \frac{w_2h}{h+2w_1(1-h)+w_2h}$

For each family in category 3,

$$S_i = \frac{\partial f(x_i,\theta)}{\partial \theta} = \begin{pmatrix} \frac{-\partial \log[h+2w_1(1-h)+w_2h]}{\partial w_1} \\ \frac{-\partial \log[h+2w_1(1-h)+w_2h]}{\partial w_2} \\ 0 \\ \frac{\partial[\log h - \log(h+2w_1(1-h)+w_2h)]}{\partial h} \end{pmatrix} = \begin{pmatrix} \frac{2(1-h)}{h+2w_1(1-h)+w_2h} \\ \frac{h}{h+2w_1(1-h)+w_2h} \\ 0 \\ \frac{1}{h} - \frac{1-2w_1+w_2}{h+2w_1(1-h)+w_2h} \end{pmatrix} \tag{D.1.3}$$

and

$$\sum S_i = \begin{pmatrix} \frac{2(1-h)n_{02}}{h+2w_1(1-h)+w_2h} \\ \frac{hn_{02}}{h+2w_1(1-h)+w_2h} \\ 0 \\ \frac{n_{02}}{h} - \frac{(1-2w_1+w_2)n_{02}}{h+2w_1(1-h)+w_2h} \end{pmatrix} \tag{D.1.4}$$

For each family in category 4,

$$\begin{aligned} S_i = \frac{\partial f(x_i,\theta)}{\partial \theta} &= \begin{pmatrix} \frac{\partial[\log 2w_1(1-h) - \log(h+2w_1(1-h)+w_2h)]}{\partial w_1} \\ \frac{-\partial \log[h+2w_1(1-h)+w_2h]}{\partial w_2} \\ 0 \\ \frac{\partial[\log(2w_1(1-h)) - \log(h+2w_1(1-h)+w_2h)]}{\partial h} \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{w_1} - \frac{2(1-h)}{h+2w_1(1-h)+w_2h} \\ \frac{h}{h+2w_1(1-h)+w_2h} \\ 0 \\ \frac{-1}{1-h} - \frac{1-2w_1+w_2}{h+2w_1(1-h)+w_2h} \end{pmatrix} \end{aligned} \tag{D.1.5}$$

and

$$\sum S_i = \begin{pmatrix} \frac{n_{12}}{w_1} - \frac{2(1-h)n_{12}}{h+2w_1(1-h)+w_2h} \\ \frac{hn_{12}}{h+2w_1(1-h)+w_2h} \\ 0 \\ \frac{-n_{12}}{1-h} - \frac{(1-2w_1+w_2)n_{12}}{h+2w_1(1-h)+w_2h} \end{pmatrix} \tag{D.1.6}$$

For each family in category 5,

$$S_i = \frac{\partial f(x_i,\theta)}{\partial \theta} = \begin{pmatrix} \frac{-\partial \log[h+2w_1(1-h)+w_2h]}{\partial w_1} \\ \frac{\partial[\log w_2h - \log(h+2w_1(1-h)+w_2h)]}{\partial w_2} \\ 0 \\ \frac{\partial[\log w_2h - \log(h+2w_1(1-h)+w_2h)]}{\partial h} \end{pmatrix} = \begin{pmatrix} \frac{-2(1-h)}{h+2w_1(1-h)+w_2h} \\ \frac{1}{w_2} - \frac{h}{h+2w_1(1-h)+w_2h} \\ 0 \\ \frac{1}{h} - \frac{1-2w_1+w_2}{h+2w_1(1-h)+w_2h} \end{pmatrix} \tag{D.1.7}$$

and

$$\sum S_i = \begin{pmatrix} \frac{-2(1-h)n_{22}}{h+2w_1(1-h)+w_2h} \\ \frac{n_{22}}{w_2} - \frac{hn_{22}}{h+2w_1(1-h)+w_2h} \\ 0 \\ \frac{n_{22}}{h} - \frac{(1-2w_1+w_2)n_{22}}{h+2w_1(1-h)+w_2h} \end{pmatrix} \tag{D.1.8}$$

Mating type III

Category 6: $P_{03} = \frac{h/4}{h/4+w_1(1/2-h/4)+w_2(1/2-h/4)+w_3(h/4)} = \frac{h}{h+w_1(2-h)+w_2(2-h)+w_3h}$.
Category 7: $P_{13} = \frac{w_1(1/2-h/4)}{h/4+w_1(1/2-h/4)+w_2(1/2-h/4)+w_3(h/4)} = \frac{w_1(2-h)}{h+w_1(2-h)+w_2(2-h)+w_3h}$.
Category 8: $P_{23} = \frac{w_2(1/2-h/4)}{h/4+w_1(1/2-h/4)+w_2(1/2-h/4)+w_3(h/4)} = \frac{w_2(2-h)}{h+w_1(2-h)+w_2(2-h)+w_3h}$.
Category 9: $P_{13} = \frac{w_3(h/4)}{h/4+w_1(1/2-h/4)+w_2(1/2-h/4)+w_3(h/4)} = \frac{w_3h}{h+w_1(2-h)+w_2(2-h)+w_3h}$.

For each family in category 6,

$$\begin{aligned} S_i = \frac{\partial f(x_i,\theta)}{\partial\theta} &= \begin{pmatrix} \frac{-\partial \log\left[h+w_1(2-h)+w_2(2-h)+w_3h\right]}{\partial w_1} \\ \frac{-\partial \log\left[h+w_1(2-h)+w_2(2-h)+w_3h\right]}{\partial w_2} \\ \frac{-\partial \log\left[h+w_1(2-h)+w_2(2-h)+w_3h\right]}{\partial w_3} \\ \frac{\partial[\log h - \log\left(h+w_1(2-h)+w_2(2-h)+w_3h\right)]}{\partial h} \end{pmatrix} \\ &= \begin{pmatrix} \frac{h-2}{h+w_1(2-h)+w_2(2-h)+w_3h} \\ \frac{h-2}{h+w_1(2-h)+w_2(2-h)+w_3h} \\ \frac{-h}{h+w_1(2-h)+w_2(2-h)+w_3h} \\ \frac{1}{h} - \frac{1-w_1-w_2+w_3}{h+w_1(2-h)+w_2(2-h)+w_3h} \end{pmatrix} \end{aligned} \tag{D.1.9}$$

and

$$\sum S_i = \begin{pmatrix} \frac{(h-2)n_{03}}{h+w_1(2-h)+w_2(2-h)+w_3h} \\ \frac{(h-2)n_{03}}{h+w_1(2-h)+w_2(2-h)+w_3h} \\ \frac{-hn_{03}}{h+w_1(2-h)+w_2(2-h)+w_3h} \\ \frac{n_{03}}{h} - \frac{(1-w_1-w_2+w_3)n_{03}}{h+w_1(2-h)+w_2(2-h)+w_3h} \end{pmatrix} \tag{D.1.10}$$

For each family in category 7,

$$S_i = \frac{\partial f(x_i,\theta)}{\partial \theta} = \begin{pmatrix} \frac{\partial[\log w_1(2-h) - \log(h+w_1(2-h)+w_2(2-h)+w_3h)]}{\partial w_1} \\ \frac{-\partial \log[h+w_1(2-h)+w_2(2-h)+w_3h]}{\partial w_2} \\ \frac{-\partial \log[h+w_1(2-h)+w_2(2-h)+w_3h]}{\partial w_3} \\ \frac{\partial[\log w_1(2-h) - \log(h+w_1(2-h)+w_2(2-h)+w_3h)]}{\partial h} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1}{w_1} - \frac{2-h}{h+w_1(2-h)+w_2(2-h)+w_3h} \\ \frac{h-2}{h+w_1(2-h)+w_2(2-h)+w_3h} \\ \frac{-h}{h+w_1(2-h)+w_2(2-h)+w_3h} \\ \frac{1}{h-2} - \frac{1-w_1-w_2+w_3}{h+w_1(2-h)+w_2(2-h)+w_3h} \end{pmatrix} \tag{D.1.11}$$

and

$$\sum S_i = \begin{pmatrix} \frac{n_{13}}{w_1} - \frac{(2-h)n_{13}}{h+w_1(2-h)+w_2(2-h)+w_3h} \\ \frac{(h-2)n_{13}}{h+w_1(2-h)+w_2(2-h)+w_3h} \\ \frac{-hn_{13}}{h+w_1(2-h)+w_2(2-h)+w_3h} \\ \frac{n_{13}}{h-2} - \frac{(1-w_1-w_2+w_3)n_{13}}{h+w_1(2-h)+w_2(2-h)+w_3h} \end{pmatrix} \tag{D.1.12}$$

For each family in category 8,

$$S_i = \frac{\partial f(x_i,\theta)}{\partial \theta} = \begin{pmatrix} \frac{-\partial \log[h+w_1(2-h)+w_2(2-h)+w_3h]}{\partial w_1} \\ \frac{\partial[\log w_2(2-h) - \log(h+w_1(2-h)+w_2(2-h)+w_3h)]}{\partial w_2} \\ \frac{-\partial \log[h+w_1(2-h)+w_2(2-h)+w_3h]}{\partial w_3} \\ \frac{\partial[\log w_2(2-h) - \log(h+w_1(2-h)+w_2(2-h)+w_3h)]}{\partial h} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{h-2}{h+w_1(2-h)+w_2(2-h)+w_3h} \\ \frac{1}{w_2} - \frac{2-h}{h+w_1(2-h)+w_2(2-h)+w_3h} \\ \frac{-h}{h+w_1(2-h)+w_2(2-h)+w_3h} \\ \frac{1}{h-2} - \frac{1-w_1-w_2+w_3}{h+w_1(2-h)+w_2(2-h)+w_3h} \end{pmatrix} \tag{D.1.13}$$

and

$$\sum S_i = \begin{pmatrix} \frac{(h-2)n_{23}}{h+w_1(2-h)+w_2(2-h)+w_3h} \\ \frac{n_{23}}{w_2} - \frac{(2-h)n_{23}}{h+w_1(2-h)+w_2(2-h)+w_3h} \\ \frac{-hn_{23}}{h+w_1(2-h)+w_2(2-h)+w_3h} \\ \frac{n_{23}}{h-2} - \frac{(1-w_1-w_2+w_3)n_{23}}{h+w_1(2-h)+w_2(2-h)+w_3h} \end{pmatrix} \tag{D.1.14}$$

For each family in category 9,

$$S_i = \frac{\partial f(x_i,\theta)}{\partial \theta} = \begin{pmatrix} \frac{-\partial \log\,[h+w_1(2-h)+w_2(2-h)+w_3h]}{\partial w_1} \\ \frac{-\partial \log\,[h+w_1(2-h)+w_2(2-h)+w_3h]}{\partial w_2} \\ \frac{\partial [log w_3 - \log\,(h+w_1(2-h)+w_2(2-h)+w_3h)]}{\partial w_3} \\ \frac{\partial [\log w_3 h - \log\,(h+w_1(2-h)+w_2(2-h)+w_3h)]}{\partial h} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{h-2}{h+w_1(2-h)+w_2(2-h)+w_3h} \\ \frac{h-2}{h+w_1(2-h)+w_2(2-h)+w_3h} \\ \frac{1}{w_3} - \frac{h}{h+w_1(2-h)+w_2(2-h)+w_3h} \\ \frac{1}{h} - \frac{1-w_1-w_2+w_3}{h+w_1(2-h)+w_2(2-h)+w_3h} \end{pmatrix} \tag{D.1.15}$$

and

$$\sum S_i = \begin{pmatrix} \frac{(h-2)n_{33}}{h+w_1(2-h)+w_2(2-h)+w_3h} \\ \frac{(h-2)n_{33}}{h+w_1(2-h)+w_2(2-h)+w_3h} \\ \frac{n_{33}}{w_3} - \frac{hn33}{h+w_1(2-h)+w_2(2-h)+w_3h} \\ \frac{n_{33}}{h} - \frac{(1-w_1-w_2+w_3)n_{33}}{h+w_1(2-h)+w_2(2-h)+w_3h} \end{pmatrix} \tag{D.1.16}$$

Mating type IV

Category 10: $P_{14} = \frac{w_1(h/2)}{w_1(h/2)+w_2(1-h)+w_3(h/2)} = \frac{w_1 h}{w_1 h+2w_2(1-h)+w_3 h}$
Category 11: $P_{24} = \frac{w_2(1-h)}{w_1(h/2)+w_2(1-h)+w_3(h/2)} = \frac{2w_2(1-h)}{w_1 h+2w_2(1-h)+w_3 h}$
Category 12: $P_{34} = \frac{w_3(h/2)}{w_1(h/2)+w_2(1-h)+w_3(h/2)} = \frac{w_3 h}{w_1 h+2w_2(1-h)+w_3 h}$

For each family in category 10,

$$S_i = \frac{\partial f(x_i,\theta)}{\partial \theta} = \begin{pmatrix} \frac{\partial[\log w_1 h - \log(w_1 h + 2w_2(1-h) + w_3 h)]}{\partial w_1} \\ \frac{-\partial \log[w_1 h + 2w_2(1-h) + w_3 h]}{\partial w_2} \\ \frac{-\partial \log[w_1 h + 2w_2(1-h) + w_3 h]}{\partial w_3} \\ \frac{\partial[\log w_1 h - \log(w_1 h + 2w_2(1-h) + w_3 h)]}{\partial h} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1}{w_1} - \frac{h}{w_1 h + 2w_2(1-h) + w_3 h} \\ \frac{-2(1-h)}{h + 2w_1(1-h) + w_2 h} \\ \frac{-h}{w_1 h + 2w_2(1-h) + w_3 h} \\ \frac{1}{h} - \frac{w_1 - 2w_2 + w_3}{w_1 h + 2w_2(1-h) + w_3 h} \end{pmatrix} \tag{D.1.17}$$

and

$$\sum S_i = \begin{pmatrix} \frac{n_{14}}{w_1} - \frac{h n_{14}}{w_1 h + 2w_2(1-h) + w_3 h} \\ \frac{-2(1-h) n_{14}}{h + 2w_1(1-h) + w_2 h} \\ \frac{-h n_{14}}{w_1 h + 2w_2(1-h) + w_3 h} \\ \frac{n_{14}}{h} - \frac{(w_1 - 2w_2 + w_3) n_{14}}{w_1 h + 2w_2(1-h) + w_3 h} \end{pmatrix} \tag{D.1.18}$$

For each family in category 11,

$$S_i = \frac{\partial f(x_i,\theta)}{\partial \theta} = \begin{pmatrix} \frac{-\partial \log[w_1 h + 2w_2(1-h) + w_3 h]}{\partial w_1} \\ \frac{\partial[\log w_2(1-h) - \log(w_1 h + 2w_2(1-h) + w_3 h)]}{\partial w_2} \\ \frac{-\partial \log[w_1 h + 2w_2(1-h) + w_3 h]}{\partial w_3} \\ \frac{\partial[\log w_2(1-h) - \log(w_1 h + 2w_2(1-h) + w_3 h)]}{\partial h} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{-h}{w_1 h + 2w_2(1-h) + w_3 h} \\ \frac{1}{w_2} - \frac{2(1-h)}{w_1 h + 2w_2(1-h) + w_3 h} \\ \frac{-h}{w_1 h + 2w_2(1-h) + w_3 h} \\ \frac{1}{h-1} - \frac{w_1 - 2w_2 + w_3}{w_1 h + 2w_2(1-h) + w_3 h} \end{pmatrix} \tag{D.1.19}$$

and

$$\sum S_i = \begin{pmatrix} \frac{-h n_{24}}{w_1 h + 2w_2(1-h) + w_3 h} \\ \frac{n_{24}}{w_2} - \frac{2(1-h) n_{24}}{w_1 h + 2w_2(1-h) + w_3 h} \\ \frac{-h n_{24}}{w_1 h + 2w_2(1-h) + w_3 h} \\ \frac{n_{24}}{h-1} - \frac{(w_1 - 2w_2 + w_3) n_{24}}{w_1 h + 2w_2(1-h) + w_3 h} \end{pmatrix} \tag{D.1.20}$$

For each family in category 12,

$$S_i = \frac{\partial f(x_i,\theta)}{\partial \theta} = \begin{pmatrix} \frac{-\partial \log[w_1 h + 2w_2(1-h) + w_3 h]}{\partial w_1} \\ \frac{-\partial \log[w_1 h + 2w_2(1-h) + w_3 h]}{\partial w_2} \\ \frac{\partial[\log w_3 h - \log(w_1 h + 2w_2(1-h) + w_3 h)]}{\partial w_3} \\ \frac{\partial[\log w_3 h - \log(w_1 h + 2w_2(1-h) + w_3 h)]}{\partial h} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{-h}{w_1 h + 2w_2(1-h) + w_3 h} \\ \frac{-2(1-h)}{w_1 h + 2w_2(1-h) + w_3 h} \\ \frac{1}{w_3} - \frac{h}{w_1 h + 2w_2(1-h) + w_3 h} \\ \frac{1}{h} - \frac{w_1 - 2w_2 + w_3}{w_1 h + 2w_2(1-h) + w_3 h} \end{pmatrix} \quad \text{(D.1.21)}$$

and

$$\sum S_i = \begin{pmatrix} \frac{-h n_{34}}{w_1 h + 2w_2(1-h) + w_3 h} \\ \frac{-2(1-h) n_{34}}{w_1 h + 2w_2(1-h) + w_3 h} \\ \frac{n_{34}}{w_3} - \frac{h n_{34}}{w_1 h + 2w_2(1-h) + w_3 h} \\ \frac{n_{34}}{h} - \frac{(w_1 - 2w_2 + w_3) n_{34}}{w_1 h + 2w_2(1-h) + w_3 h} \end{pmatrix} \quad \text{(D.1.22)}$$

Mating type V

Category 13: $P_{25} = \frac{w_2/2}{w_2/2 + w_3/2} = \frac{w_2}{w_2 + w_3}$.
Category 14: $P_{35} = \frac{w_3/2}{w_2/2 + w_3/2} = \frac{w_3}{w_2 + w_3}$

For each family in category 13,

$$S_i = \frac{\partial f(x_i,\theta)}{\partial \theta} = \begin{pmatrix} 0 \\ \frac{\partial[\log w_2 - \log(w_2 + w_3)]}{\partial w_2} \\ \frac{-\partial \log(w_2 + w_3)}{\partial w_3} \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{1}{w_2} - \frac{1}{w_2 + w_3} \\ \frac{-1}{w_2 + w_3} \\ 0 \end{pmatrix} \quad \text{(D.1.23)}$$

and

$$\sum S_i = \begin{pmatrix} 0 \\ \frac{n_{25}}{w_2} - \frac{n_{25}}{w_2+w_3} \\ \frac{-n_{25}}{w_2+w_3} \\ 0 \end{pmatrix} \tag{D.1.24}$$

For each family in category 14,

$$S_i = \frac{\partial f(x_i,\theta)}{\partial \theta} = \begin{pmatrix} 0 \\ \frac{-\partial \log(w_2+w_3)}{\partial w_2} \\ \frac{\partial[\log w_3 - \log(w_2+w_3)]}{\partial w_3} \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{-1}{w_2+w_3} \\ \frac{1}{w_3} - \frac{1}{w_2+w_3} \\ 0 \end{pmatrix} \tag{D.1.25}$$

and

$$\sum S_i = \begin{pmatrix} 0 \\ \frac{-n_{35}}{w_2+w_3} \\ \frac{n_{35}}{w_3} - \frac{n_{25}}{w_2+w_3} \\ 0 \end{pmatrix} \tag{D.1.26}$$

D.1.1 Estimation of The Parameters

For marker g:

1) Generate category index, t=1,2,...,14.

2) Sort data by sex and type of disjunction (MI/MII).

3) Count n_{FIt},n_{FIIt},n_{MIt}, and n_{MIIt} (a total of $14 \times 4 = 56$ numbers).

4) Then we can calculate $\sum_i S_{ig} = \sum_i \sum_t S_{igt}$s from the scores shown in the previous section.

5) Solve $\sum_i S_{ig} = 0$ to get estimates for θ_g by quasi-Newton method. Note that we have 7 parameters for each marker. $\theta^T = (w_{1g}, w_{2g}, w_{3g}, h_{FIg}, h_{FIIg}, h_{MIg}, h_{MIIg})$.

BIBLIOGRAPHY

[1] Celeux, G. and Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis* **14**, 315-332.

[2] Di, X., Matsuzaki, H., Webster, T.A., Hubbell, E., Liu, G., Dong, S., Bartell, D., Huang, J., Chiles, R., Yang, G., Shen, M., Kulp, D., Kennedy, G., Mei, R., Jones, K.W., and Cawley, S. (2005) Dynamic model based algorithms for screening and genotyping over 100K SNPs on oligonucleotide microarrays. *Bioinformatics*, **21**, 1958-1963.

[3] Devlin, B. and Roeder, K. (1999) Genomic control for association studies. *Biometrics* **55**, 997-1004.

[4] Epstein, M.P., Allen, A.S., and Satten, G.A. (2007). A simple and improved correction for population stratification in case-control studies. *Am. J. Hum. Genet.* **80**, 921-930.

[5] Epstein, M.P., Veal, D.D., Trembath, R.C., Barker, J.N.W.N., Li, C. and Satten, G.A. (2005). Genetic association analysis using data from trios and unrelated subjects. *Am. J. Hum. Genet.* **76**, 592-608.

[6] Feingold, E., Brown, A.S., and Sherman, S.L. (2000) Multipoint estimation of genetic maps for human trisomies with one parent or other partial data. *Am. J. Hum. Genet.* **66**, 958-968.

[7] Forest, W. and Feingold, E. (2000) Composite statistics for QTL mapping with moderately discordant sibling pairs. *Am. J. Hum. Genet.* **66**,1642-1660.

[8] Freeman, S.B., Faft, L.F., Dooley, K.J., Allran, K., Sherman, S.L., Hassold, M.J., and Saker, D.M. (1998) Population-based study of congenital heart defects in Down syndrome. *Am. J. Med. Genet.* **80**, 213-217.

[9] Fujisawa, H., Eguchi, S., Ushijima, M., Miyata, S., Miki, Y, Muto, T. and Matsuura, M. (2004) Genotyping of single nucleotide polymorphism using model-based clustering.*Bioinformatics*, **20**, 718-726.

[10] Hartigan, J. A. and Wong, M. A. (1979) A K-means clustering algorithm. *Applied Statistics*, **28**, 100-108.

[11] Hassold T.J. and Jacobs, P.A. (1984). Trisomy in man. *Annu. Rev. Genet.*, **18**, 69-97.

[12] Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U. and Speed, T.P. (2003) Exploration, normalization, and summaries of high density oligonuceotide array probe level data. *Biostatistics*, **4**, 249-264.

[13] Kazeem, G.R. and Farrall, M. (2005) Integrating case-control and TDT studies. *Ann. Hum. Genet.* **69**, 329-335.

[14] Kerstann, K.F., Feingold, E., Freeman, S.B., Bean, L.J.H., Pyatt, R., Tinker, S., Jewel, A.H., Capone, G., and Sherman, S.L.(2004) Linkage Disequilibrium Mapping in Trisomic Populations: Analytical Approaches and an Application to Congenital Heart Defects in Down Syndrome. *Genet. Epidemiol.* **27**, 240-251.

[15] Kelinbaum, D.G., Kupper, L.L., Morgenstern, H. (1982) Epidemiologic research: principles and quantitative methods. Van Nostrand Reinhold, New York.

[16] Komura et al., 2007Komura, d., Shen, F., Ishikawa, S., Fitch, K.R., Chen, W., Zhang, J., Liu,G., IharA, S., Nakamura, H., Hurles, M.E., lee, C., Scherer, S.W., Jones, K.W., Shapero, M.H., Huan, J. and Aburatani, H. (2007) Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays. *Genome Research***16**, 1575-1584.

[17] Liu, W., Di, X., Yang, G., Matsuzaki, H., Jing, H., Mei R., Ryder T., Webster, T.A., Dong, S., Liu, G., Jones, K., Kennedy, G., Kulp, D. (2003) Algorithms for large-scale genotyping microarrays. *Bioinformatics*, **19**, 2397-2403.

[18] Lin, Y.,Tseng, G.C., Lora J.H., Bean, L.J.H., Sherman, S.L. and Feingold, E. (2007) Smarter Clustering Methods for High-throughput SNP Genotype Calling. Submitted to *Biostatistics*.

[19] Nagelkerke, N.J.D., Hoebee, B., Teunis, P. and Kimman, T.G. (2004) Combining the transmission disequilibrium test and case-control methodology using generalized logistic regression. *Eur. J. Hum. Genet.* **12**, 964-970.

[20] Price, A. L., Patterson, N. J., Plenge, R. M., Weinblattt, M. E., Shadick, N.A. and Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904-909.

[21] Purcell, S., Sham, P. and Daly, M.J. (2005) Parental phenotypes in family-based association analysis. *Am. J. Hum. Genet.* **76**, 249-259.

[22] Rabbee, N. and Speed, T.P. (2005) A genotype calling algorithm for affymetirx SNP arrays. *Bioinformatics*,**22**, 7-12.

[23] Reich, D. and Goldstein, D. Detecting association in a case-control study while asllowing for population stratification. (2006) *Genet. Epidemiol.* **20**, 4-16.

[24] Sabatti, C. and Lange, K. (2005) Bayesian Gaussian Mixture Models for High Density Genotyping Arrays Department of Statistics Papers. Paper 2005040102.

[25] Setakis, E., Stirnadel, H. and Balding, D. J. (2007) Logistic regression protects against population structure in genetic asoociation studies. *Genome Res.* **16**, 290-296.

[26] Shen, R., Fan, J.B., Cambell, D., Dhang, W., Chen, J., Doucet, D., Yeakley, J., Bibikova, M., Garcia, E.W., Mcbride, C., Steemers, F, Garcia, F., Kermani, B.G., Gunderson, K and Oliphant, A. (2005) High-throughput SNP genotyping on universal bead arrays. *Mutation Research* **573**, 70-82.

[27] Steen, K.V., McQueen, M.B., Herbert, A., Raby, B., Lyon, H., DeMeo, D.L., Murphy, A., Su, J., Datta, S., Rosenow, C., Christman, M., Silverman, E.K., Laird, N.M., Weiss, S. and Lange, C. (2005) Genomic screening and replication using the same data set in family-based association testing. *Nature Genetics*, **37**, 683 691.

[28] Tu, I-P, Balise, R.R. and Whittemore, A.S. (2000) Detection of disease genes by using of family data II. Application to nuclear families. *Am. J. Hum. Genet.* **66**, 1341-1350.

[29] Weinberg, C.R., Wilcox, A.J. and Lie, R.T. (1998) A log-linear approach to case-parent-triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. *Am. J. Hum. Genet.* **62**, 969-978.

[30] Whittemore, A.S. and Tu, I-P (2000) Detection of disease genes by use of family data. I. Likelihood-based theory. *Am. J. Hum. Genet.* **66**, 1328-1348.

[31] Xu, Z., Kerstann, K.F., Sherman, S.L., Chakravarti, A. and Feingold, E. (2004) A Trisomic Transmission Disequilibrium Test *Genet. Epidemiol.* **26**, 125-131.