AFFECTED RELATIVE PAIR LINKAGE STATISTICS THAT MODEL RELATIONSHIP UNCERTAINTY

by

Amrita Ray

BSc, Presidency College, Calcutta, India, 2001

MStat, Indian Statistical Institute, Calcutta, India, 2003

Submitted to the Graduate Faculty of
the Department of Human Genetics in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2007

UNIVERSITY OF PITTSBURGH

HUMAN GENETICS DEPARTMENT

This dissertation was presented

by

Amrita Ray

It was defended on

April 11th 2007

and approved by

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

Michael Barmada, Ph.D., Associate Professor, Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh

Bernie Devlin, Ph.D., Associate Professor, Department of Psychiatry, School of Medicine, University of Pittsburgh

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

Sati Mazumdar, Ph.D., Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Dissertation Director: Daniel E. Weeks, Ph.D., Professor, Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh

AFFECTED RELATIVE PAIR LINKAGE STATISTICS THAT MODEL RELATIONSHIP UNCERTAINTY

Amrita Ray, PhD

University of Pittsburgh, 2007

In linkage analysis with affected related pairs (ARP), stated familial relationships are usually assumed to be correct, thus misspecified relationships can lead to either reduced power or false-positive evidence for linkage. In practice, studies either discard individuals with erroneous relationships or use the best possible alternative pedigree structure. We have developed several linkage statistics that model the relationship uncertainty by properly weighting over possible true relationships. We consider ARP data for a genome-wide linkage scan. A simulation study is performed to assess the proposed statistics, and to compare them to the maximum likelihood statistic (MLS) and S_{all} LOD score using true and discarded structures. We have simulated small and large pedigree datasets with different underlying true and apparent relationships, and typed for 367 microsatellite markers. The results show that two of our relationship uncertainty linkage statistics (RULS) have power almost as high as MLS and S_{all} using the true structure. Also, these two RULS have greater power to detect linkage than MLS and S_{all} using the discarded structure. Thus, our RULS provide a statistically sound and powerful approach for dealing with the commonly encountered problem of relationship errors. The RULS are relevant to public health because application of these RULS to complex human disease will facilitate the mapping and discovery of genes involved in the etiology of such diseases.

We attempted to apply RULS to Otitis Media with effusion (OME) data from Caucasian families. OME is an infection causing fluid in the middle ear, and is the most common cause of hearing loss among young children. We have recruited subjects (with history of tympanostomy tube insertion) and their families (parents and affected/unaffected siblings). Genotyping was done using Affymetrix 10K SNP chips, and out of 1,584 enrolled individuals (322 families), 1,191 (305 families) are genotyped at this date. We performed nonparametric multipoint linkage analysis using discarded structures. The preliminary results show suggestive linkage peaks on six chromosomes, the highest being at rs1345938 on chromosome 7 with S_{all} LOD score of 2.36 (p-value 0.0005).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I want to take this opportunity to thank my advisor, Dr. Daniel Weeks for all his support and encouragement throughout the course of my graduate studies. He has been an ideal mentor in my research as well as personal life. Whenever I would get very stressed with work or deadline, he would remind me that there is more to life than work. Then again, he would gently remind me that he expects more from me when I took his advice and tended to enjoy life a bit too much.

I also want to thank Dr. Eleanor Feingold, Dr. Bernie Devlin, Dr. Micheal Barmada and Dr. Sati Mazumdar for agreeing to serve on my dissertation committee. Their suggestions at various points helped me in my Ph.D. I have always admired Dr. Eleanor Feingold for her dynamicity, and she has gone out of way many times to help me, whether that be in reseach or when I was looking for the next step in my career. I cannot thank Dr. Sati Mazumdar enough for all the love and care she showed towards me.

It has been a pleasure to interact with many students in the Human Genetics department at University of Pittsburgh. My office mates were also nice to spend time beyond working. Special thanks goes to Nandita and Abigail Matthews who has over the period of two years have become close friends. I have also thoroughly enjoyed working with Jeesun and our friendship became deeper with time.

I would also like to thank all my friends here at Pittsburgh. Over the course of the last three and a half years, I have made many new friends, some I knew from India, but mostly I met new people from all over the world who were in the same boat as mine as we all landed in a foreign land looking lost. My roommate and myself have spent countless hours talking about silly things to get our minds off courses and homeworks. With my friends here, I have spent a few memorable moments in groups, be it just chatting, watching movies or plain old birthday parties.

But most importantly, I would like to thank my family for their love, support, and patience during these years while I focused on studies. Thanks to my mom, dad and sister for helping me aim high and provide support in every way possible. My dad has been my idol, my source of encouragement and my pillar of strength since when I remember. I want to credit him more than anyone for supporting me through ups and downs that has seen me through my PhD. My mom has always showered me with more love than I could ask for. And my little sister, being the sweetest person on earth, has been a source of pure joy to me. Last but certainly not the least, I would also like to thank Mainak for filling my life with happiness and I look forward to spending the rest of my life with him.

1.0 RELATIONSHIP UNCERTAINTY LINKAGE STATISTICS (RULS)

1.1 INTRODUCTION

Linkage analysis programs invariably assume that the relationships identified during pedigree collection are the true relationships, known without errors. Thus in linkage analysis, misspecified relationships can lead to either reduced power or false-positive evidence for linkage [16]. Reduced power might occur when an affected pair is actually more distantly related than assumed, for example, when a half sib pair is incorrectly analyzed as a full sib pair. Also, if an affected pair is actually more closely related than assumed, this might result in a false positive, for example when monozygotic twins are falsely coded as full sibs. So detection of such errors is useful prior to linkage analysis. Relationship testing [5] may allow one to detect and correct erroneous relationships resulting from cases of nonpaternity, unrecorded adoption or accidental sample swaps in the laboratory. Several studies show that relationship error is frequently present in real data; this motivated us to develop linkage statistics that properly model relationship uncertainty.

Here we consider the situation where we have collected affected relative pairs (ARPs) typed for a genome-wide set of markers, where the presumed relationship of the ARP actually might be different from the true relationship. There are two sources of information about relationships: the stated apparent relationship and the genome-wide marker data. We statistically model the relationship uncertainty by properly weighting over the possibilities to develop our relationship uncertainty linkage statistics (RULS). The relationship uncertainty is modeled via weights, the weights being the conditional probability of a true relationship type given the apparent one and the genome-wide marker data. All the computations are done at the affected pair level, using only data for the two individuals of each affected pair.

We perform a simulation study to explore the behavior of the RULS and to compare them to an MLS approach of Cordell et al. [9] and to the multipoint exponential S_{all} [21] non-parametric LOD score as implemented in Merlin [1]. S_{all} and MLS are computed using both the true pedigree structure and the discarded structure (where problematic individuals with erroneous relationships are removed from the pedigree). We evaluate genome-wide empirical significance thresholds and compute the power of our statistics under several genetic models.

1.2 BACKGROUND

Linkage studies assume that the relationships identified during pedigree collection are the true relationships, known without error. In linkage analysis, misspecified relationships can lead to either reduced power or false-positive evidence for linkage. So, detection of such errors is useful prior to linkage analysis. The errors in pedigree structure will often be uncovered through Mendelian inconsistencies. But these mistakes may go undetected when parental genotypes are not known. In these cases, the genome-screen data can be informative for imputation of relationships. This motivated the development of statistical methods for detecting misspecified relationships based on genotype data.

We now give brief introduction to linkage analysis and affected relative pair approaches to linkage analysis.

1.2.1 Linkage analysis

The goal of linkage analysis is to identify the location of a gene or set of genes in the genome, which cause a particular characteristic. Linkage analysis is used to narrow the possible region of a disease gene, so that molecular approaches or other analyses can be used to more precisely identify the disease gene. Linkage analysis is carried out several steps: a) defining the disease phenotype of interest, b) collecting a sample of patients along with their families, c) genotyping families at markers (either single marker study: specific candidate genes/regions chosen for genotyping, or whole genome studies, where genotyping is done at approximately

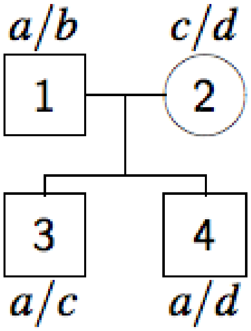


Figure 1: Example showing a sib pair sharing 1 allele IBD.

equally spaced marker throughout the genome), d) performing either parametric or nonparametric linkage analysis. Parametric linkage analysis requires one to specify the disease inheritance model, which is unknown in many real data scenarios, specially in the case of complex disease.

Nonparametric linkage analysis does not require disease model specification, and it compares the expected and observed similarity of identical by descent (IBD) allele sharing between individuals. When two alleles are inherited from a common ancester, then those alleles are identical by descent (IBD). For example, in Figure 1, individuals 3 and 4 share one allele IBD (the paternal allele, a). Similary, for a marker not linked to the disease, full sibs share 0, 1 and 2 alleles IBD with probability .25, .50, .25, and this sharing at an unlinked locus is called null sharing. Now, the underlying principle to mapping methods is that people who share traits should share genetic material more than expected near the genes that influence those traits. Nonparametric linkage analysis utilizes this principle, and compares IBD sharing at a marker with the null sharing. Thus, if the sharing at a marker is significantly higher than the null sharing, one concludes that the marker is linked to the disease.

1.2.2 Affected relative pair approaches

There are several affected relative pair (ARP) or affected sib pair (ASP) approaches in nonparametric linkage analysis. One is the likelihood ratio test statistic, where the observed

IBD sharing of the ARPs at a marker is compared to the null sharing. Risch [36] proposed a LOD score formulation for ARP linkage analysis, which we will describe in section 1.2.3. There have been several extensions and modifications of this basic ARP LOD score. The second approach is an allele sharing statistic that measures IBD sharing among affecteds within a pedigree. S_{all} and S_{pairs} [49] are the two score functions for allele sharing. Here we give a brief description of S_{pairs} and S_{all}, as discussed by Shih and Whittemore [50]. Briefly, S_{pairs} is the number of allele pairs from ARP that are IBD, and S_{all} puts extra weight on more than two affecteds sharing the same allele IBD [21]. At a given location on the genome, at marker x, the inheritance vector is $v(x) = (p_1, m_1, p_2, m_2, \cdots, p_n, m_n)$, where n is the number of non-founders, $p_i = 0 \,\text{or}\, 1$ according to whether the grandpaternal or grandmaternal allele is transmitted from father to child, and m_i has the same definition as p_i except that it shows transmission from mother to child. The inheritance vectors can thus be organized into IBD configurations. The score function S_{pairs} for IBD configuration ψ is given by

$$S_{pairs}(\psi) = \frac{2}{n(n-1)} \sum_{i,j} f_{ij}(v),$$

where f_{ij} is one-fourth the number of alleles shared IBD by relatives i, j. The second score function S_{all} is defined by

$$S_{all}(\psi) = \frac{1}{2^n} \sum_{h} [\Pi_i b_i(h)!],$$

where h is the collection of alleles taken from each affected individual, and $b_i(h)$ is the number of founder alleles i in collection h. S_{all} considers all affected relatives at the same time, unlike the score function S_{pairs} where pairwise comparisons are made. The normalized score for score function S for a pedigree at locus x is given by $\frac{S(x)-\mu(x)}{\sigma(x)}$ where μ and σ^2 are the mean and variance under H_0 that the marker at x is not linked to disease.

Kong and Cox [24] suggested two models, linear and exponential models to get the final non-parametric linkage scores. Taking IBD configuration probability for a pedigree to be a linear combination of the score functions gives the linear model, and taking the configuration as exponentially related to the score function, gives exponential model.

A third approach for ASPs, is the mean sharing statistic [4]. This compares the mean number of alleles shared IBD with expected sharing number under null hypothesis of no linkage, and is given by,

$$\begin{aligned} Z &= \frac{\sum_i \pi_i - n/2}{\sqrt{n/8}} \qquad (1.1) \\ &= \frac{(\#\text{pairs with IBD1})(1/2) + (\#\text{pairs with IBD2}) - n/2}{\sqrt{n/8}}. \end{aligned}$$

where n is number of ASPs, π_i is mean IBD sharing for ASP i.

1.2.3 Risch's approach

Risch [36] developed a LOD score formulation for ARP linkage analysis. Here we will give a brief introduction to his approach. Let f_i be the prior probability of sharing i alleles IBD, and w_{ij} be the probability of observed marker data given that the pair shares i alleles IBD at the marker. Risch [36] showed that for affected relative pairs (ARPs), the likelihood of the observed marker data for the jth pair is $\sum_{i=0}^{2} z_i w_{ij}$ and the LOD score considering all independent affected pairs is given by

$$\Lambda = \sum_j \log_{10}[\frac{\sum_i z_i w_{ij}}{\sum_i f_i w_{ij}}] \qquad (1.2)$$

where z_i is the probability that an ARP shares i alleles IBD at the marker. The maximum lod score (Λ) is obtained by maximizing equation (1.2) with respect to the parameters z_i. Considering only affected sib pairs (ASPs), under H_0 that the marker is not linked to the disease, the asymptotic distribution of $2\ln(10)\Lambda$ is χ^2_2. Holmans [19] and Faraway [15] showed that the power of Risch's method can be improved by restricting the parameter space as discussed below.

To improve the power of ASP linkage analysis, Holmans [19] introduced the "possible triangle" method by restricting maximization of parameters to the region of the triangle. In our approach also, for two of our statistics, we restrict maximization of the parameters to satisfy the "possible triangle" restriction. Holmans showed that for any genetic model, allele sharing estimates fall within a possible triangle as given in Figure 2. The likelihood ratio test satisfying the "possible triangle" constraints of $z_1 \leq .5$, $2z_0 \leq z_1$, and $z_0 \geq 0$ has higher power than an unrestricted likelihood ratio test, where z_i is the allele-sharing parameter among ARPs, $i = 0, 1, 2$.

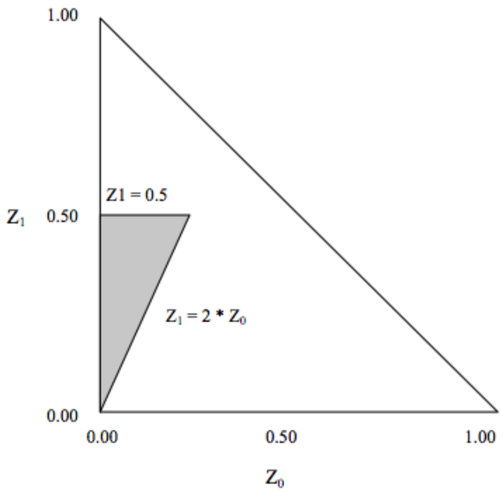


Figure 2: The possible triangle of allele sharing estimates.

1.2.4 Cordell et al.'s approach

Cordell et al. [9] have developed a maximum likelihood score (MLS) statistic for ARPs, starting from the statistics proposed by Risch [36], but maximizing only two parameters: additive and dominance variance. They have also considered more than one disease locus model in their approach. Here we give a brief description of their approach for a one disease locus model. z_i and f_i are defined as in the previous section. Let $\hat{f_{ij}}$ be the posterior probability of pair j sharing i alleles IBD given the observed genotype data of the pair. The likelihood of pair j with genotype data G_j is given as $L_j = \sum_{i=0}^{2} \frac{z_i \hat{f_{ij}} P(G_j)}{f_i}$ and MLS is given as

$$MLS = \sum_j \log_{10}(\sum_{i=0}^{2} \frac{z_i \hat{f_{ij}}}{f_i}).$$

The z_i is written in terms of f_i and two variance terms, the additive and dominance variances caused by the disease-causing locus, as $z_i = \frac{\lambda_i f_i}{\lambda_T}$, and λ_i, λ_T are functions of the variance terms (see discussion below). The likelihood is maximized with respect to the variance component parameters by restricting the variance components to nonnegative quantities, so that "possible triangle" [19] restriction holds. Under H_0, the asymptotic distribution of

$2\ln(10)MLS$ is a mixture of χ^2 distributions. For only affected sib pairs, MLS has asymptotic distribution as mixture of χ^2_1 and χ^2_2. For affected pairs with other relationship types, MLS has more complex mixture χ^2 distribution.

As derived by James [20], we can express λ_T in terms of the covariance of affection status of the affected pair and the population prevalence K:

$$\lambda_T = 1 + \frac{Cov(X_1, X_2)}{K^2}, \tag{1.3}$$

where X_i is the phenotype of person i defined to be 0 or 1, according to whether the person is affected or unaffected, and Cov denotes the covariance. James [20] showed that $Cov(X_1, X_2)$ can be expressed in terms of the additive and dominance variance (V_A, V_D) caused by disease locus, the kinship coefficient (r_A), and the prior probability of sharing 2 alleles IBD (f_2^T). The covariance is given by:

$$Cov = 2r_A V_A + u_R V_D \tag{1.4}$$

where $r_A = .5f_2^T + .25f_1^T$ and $u_R = f_2^T$. This expression and equation (1.33) show that both λ_T and λ_i are functions of V_A and V_D.

1.2.5 Olson's approach

Olson [33] developed a conditional-logistic representation of the ARP likelihood ratio. We give a brief description of her method for one disease locus when there is no covariate in her model. Let A be the event that both members of a relative pair are affected, and let A_1 and A_2 be the events that the first and second relative, respectively, are affected. The terms f_i and f_{ij} are defined as before. Then the likelihood ratio (LR_j) for an ARP of type T is

$$\begin{aligned} LR_j = \frac{P(G_j|A,T)}{P(G_j|T)} &= \frac{P(A|G_j,T)}{P(A|T)} \\ &= \frac{P(A_2|G_j,T,A_1)/P(A_2)}{P(A_2|T,A_1)/P(A_2)} \\ &= \frac{\sum_i [P(A_2|i,A_1)/P(A_2)]P(i|G_j,T)}{\sum_i [P(A_2|i,A_1)/P(A_2)]P(i|T)} \\ &= \frac{\sum_i \lambda_i \hat{f}_{ij}}{\sum_i \lambda_i f_i} \end{aligned} \tag{1.5}$$

where λ_i is the relative risk to an individual who shares i alleles IBD with an affected relative. Considering all ARP's, the LOD score (LR) is given by $LR = \sum_j \log_{10} LR_j$. Under H_0, the asymptotic distribution of $2\ln(10)LR$ is a mixture of χ^2 distributions.

Writing the z_i's (in equation 1.2) in terms of the λ_i's one can show that this model is equivalent to the Risch's [36] approach. First let's consider an ASP. Risch's likelihood ratio can be written as (see section 1.4 for details),

$$LR = \sum_i \frac{z_i \hat{f}_{ij}}{f_i}. \tag{1.6}$$

Now, for sib pairs, $z_0 = 1/(4\lambda_s)$, $z_1 = \lambda_o/(2\lambda_s)$, $z_2 = \lambda_m/(4\lambda_s)$, where $\lambda_s, \lambda_o, \lambda_m$ are risk ratios to sib, offspring and MZ twin, and $f_0 = 1/4$, $f_1 = 1/2$, $f_2 = 1/4$. So, the LR for pair j becomes,

$$LR = \hat{f}_{0j}\frac{1}{\lambda_s} + \hat{f}_{1j}\frac{\lambda_o}{\lambda_s} + \hat{f}_{2j}\frac{\lambda_m}{\lambda_s}, \tag{1.7}$$

and as the denominator of equation (1.5) can be written as $1/4 + \lambda_o/2 + \lambda_m/4 = \lambda_s$, LR_j in equation (1.5) takes the same form as LR in equation (1.7). Thus for ASPs, Olson's model is equivalent to Risch's model. The equivalence holds for a general ARP also. In case of an ARP of type r, denominator of Olson's model in equation (1.5) equals λ_r and numerator is a linear combination of λ_o, and λ_m. Though the models are equivalent, the statistic Λ by Risch [36] and $\sum_j \log_{10} LR_j$ by Olson [33] behave differently, because unlike the Risch's statistic, Olson's statistic has "possible triangle" constraints on the parameter space.

The ARP linkage analysis approaches discussed above assume that the true relationship of the ARP is known without error. This assumption might get violated in presence of relationship error, which is quite frequent in real data. This motivated us to develop ARP linkage statistics that model relationship uncertainty. We will discuss the details behind our approach after brief discussion of relationship error and its impacts on ARP linkage analysis.

1.3 RELATIONSHIP ERROR

Pedigree error, *i.e.* relationship error or misclassification of relationship between individuals is common issue in real data. Possible sources of relationship error include cases of nonpaternity, unrecorded adoption or accidental sample swaps in the laboratory. Using genome-wide marker data and the stated relationship, it is possible to identify the individuals with erroneous relationships in a pedigree. This method of identifying relationship error is called relationship testing. This plays an important role in checking a real dataset for any misspecified relationship, because not removing relationship errors from the pedigree might have a serious effect on linkage analysis. We now briefly discuss the frequency of relationship error in real data studies, and the different methods of relationship testing.

1.3.1 Examples

There are several real data studies that encounter relationship error- this shows relationship misspecification is quite a frequent problem in real data scenario. In a real study, one implements relationship testing using genome-wide marker data and the stated apparent relationship information and then one decides to either discard individuals with erroneous relationships or use the most likely alternative pedigree structure for linkage analysis. Many studies have found relationship errors in their data and have discarded individuals with erroneous relationships. Ehm et al. [13] found in their study of small pedigrees segregating for Type 2 diabetes in four American populations, that relationship testing revealed "24.4% of the families contained pedigree errors and 2.8% of the families contained errors in which an individual appeared to be unrelated to the rest of the members of the pedigree"; and pedigree errors were removed before performing the linkage analyses. Another linkage study by Shmulewitz et al. [42] made 55 modifications to a very large pedigree before analysis. Another study by Daly et al. [10] on chronic and recurrent otitis media found 9 families out of 133 (*i.e.*, 6.7%) families with "segregation problems that were not consistent with the reported family structure" and they changed the structures to the alternative structures as suggested by relationship testing. So, relationship errors are quite frequent in real data; this motivated us to develop linkage statistics that properly model relationship uncertainty.

1.3.2 Relationship testing

Here we will discuss on relationship testing, the method used to identify any relationship error among relative pairs. Relationship testing can be used to correct erroneous relationships, and thus might increase the power of linkage study. Marker information from a genome-wide scan can be highly informative for verifying relationships among individuals. The appropriate data for relationship testing will have relative pairs who have been typed for more than 50 unlinked microsatellite markers spread over at least 10 chromosomes and preferably, relative pairs who have been typed for markers throughout the genome [12]. There is an extensive literature addressing the statistical methods for detecting misspecified relationships based on genotype data. There are two methods for relationship testing, the likelihood-based method and identity-by-state (IBS) method.

Boehnke and Cox [5] used the likelihood ratio method to infer genetic relationships on the basis of genetic marker data. It compares the multipoint probability of the marker data, conditional on different genetic relationships and infers the relationship that makes the data most likely, *i.e.* computing the likelihood ratio

$$LR(R_1, R_2) = \frac{P(G|R_1)}{P(G|R_2)},$$

where R_1, R_2 are two relationships of a relative pair, G is observed genotype of the pair over all markers. $LR > 1$ supports relationship R_1 and $LR < 1$ supports relationship R_2. Likelihoods and posterior probabilities are calculated by assuming a IBD process to be Markov and no interference between markers.

For more distant relationships like avuncular and first-cousin relationships, the IBD process might not be Markov even under the non-interference assumption between markers. In such cases, McPeek and Sun [31] used an augmented IBD Markov process to calculate the likelihood of the marker data. The likelihood is calculated by applying the Baum [3] algorithm to the augmented process. Let D denotes the IBD process for an outbred pair. For cases where D is no longer a Markov process, an augmented process A is constructed. A is Markov under no interference assumption, and it contains all the information of process D. For any given relationship R of a relative pair, define $\alpha_1(j) = P(A_1 = j)$ and at marker

k, $\alpha_k(j) = P(G_1, G_2, \cdots, G_{k-1}, G_k, A_k = j)$ for $k > 1$, where G_m is genotype data for the pair at marker m. $\alpha_1(j)$ is the stationary distribution of A for relationship R. Using the recursion formula similar to Boehnke and Cox [5],

$$\alpha_{k+1} = \sum_i \alpha_k(i) P(A_{k+1} = j | A_k = i) P(G_k | A_k = i),$$

where $P(A_{k+1} = j | A_k = i)$ is the transition probability of A. Since A contains all information of IBD process denoted as $\{D\}$, $P(G_k | A_k = i) = P(G_k | D_k = IBD$ status associated with state i of $A)$. So, probabilities $P(G_k | A_k = i)$ are computed as $P(G_k | D_k = j)$ given by Thompson [46]. For cth chromosome, $P(G_1, \cdots, G_{n_c})$ is given by $\sum_j \alpha_{n_c}(j)$, and multiplying these terms over all chromosomes we get the likelihood of genotype data throughout the genome for the pair.

Göring and Ott [16] developed methods of computing on the basis of genetic marker data on the pair, the likelihoods of the sib, half-sib and unrelated relationships between pairs of individuals and calculated the posterior probabilities for alternate relationships by a Bayesian approach.

For the IBS method, one is interested to know how likely the observed identical by state sharing is conditional on the assumed relationships. To test the hypothesis that two individuals are sibs, Ehm and Wagner [12] proposed a test statistic based on the summation, over a large number of genetic markers, of the number of alleles shared identical by state by a pair of individuals. $S_k(G_k)$ denotes a score based on proportion of of alleles shared IBS by genotype G_k of a pair at marker k, and score over all markers is given by $S = \sum_k S_k(G_k)$. They calculate a test statistic $\frac{S - E(S|R)}{SD(S|R)}$, where $E(S|R)$ and $SD(S|R)$ denotes mean and standard deviation of S, conditional on relationship R. If the relationship R of the relative pair is correct, then this statistic is approximately distributed as standard normal for large samples.

1.3.3 Impact on Linkage analysis

The impact of misspecified relationships on linkage analysis is quite serious. Misspecified relationships can lead to either reduced power or false-positive evidence for linkage [16] (see

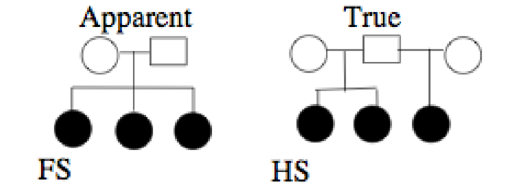
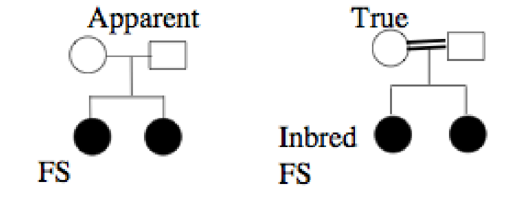
Reduced power	more distant relationship	
False positive	closer relationship	

Figure 3: In linkage analysis, misspecified relationships can lead to either reduced power or increased false positives in two different cases.

Figure 3). Reduced power might occur when an affected pair is actually more distantly related than assumed, for example, when a half sib pair is incorrectly analyzed as a full sib pair. If an affected pair is actually more closely related than assumed, this might result in a false positive, for example when monozygotic twins are falsely coded as full sibs. So detection of such errors is useful prior to linkage analysis.

1.3.4 Possible solutions

After relationship testing identifies an individual having a wrong relationship with other relatives, one usually proceeds either by discarding the individual from the pedigree, or by constructing an alternative pedigree structure. Discarding individuals leads to a conservative structure where individuals with erroneous relationships are removed. To construct an alternative structure, one has to first observe the p-value from the relationship testing and also the estimated IBD sharing probability, and then infer an alternative structure to the pedigree. These two options are usually followed by studies analyzing real data whenever they encounter relationship error.

In the first solution to deal with relationship error, *i.e.* for conservative structure, one loses information on the discarded individuals, and this might reduce power. In the second solution, replacing the stated apparent structure by an alternative structure might not be the best solution, as the 'most likely' pedigree structure might not be the 'most certain' structure. Also, it is hard to incorporate the choice of the 'most certain' structure in an automated code. In a situation where one has sparse markers, not enough to infer an alternative structure, this solution does not work well.

We thereby propose a third solution: statistically model the uncertainty by properly weighting over the possibilities. This leads to developing three new linkage statistics that model relationship uncertainty. In our approach, we have taken five true possible underlying relationships. Five true relationships are considered because together they give a good coverage of the outbred space of the relationship triangle, as discussed in the next section. Also, we do not take more than five true relationships as the number of parameters estimated in one of our statistics increases with the number of true relationships considered.

1.3.5 Relationship triangle

In our method we consider affected relative pair data and a genome-wide scan of the pairs, where the presumed relationship of the ARP actually might be different from the true relationship. In order to construct a linkage statistic that models relationship uncertainty via weights, the weights being the conditional probability of a true relationship type given the apparent one and the genome-wide marker data, we have to consider several true relationships. We take those relationships so that together they give good coverage of the space of IBD probabilities in the relationship triangle, which we now define.

The relationship triangle provides a way of diagramming the space of identity state probabilities between two noninbred individuals [47]. The identity by descent (IBD) probabilities are denoted as $k = (k_0, k_1, k_2)$ where k_i is the probability that the individuals share i genes IBD and $k_0 + k_1 + k_2 = 1$. Individuals are related if $k_0 < 1$. Each relationship may thus be represented by a point in an equilateral triangle of unit height, the vertices corresponding to unrelated pairs ($k_0 = 1$), parent-offspring ($k_1 = 1$), and the identity (monozygous twins) relationship ($k_2 = 1$) and k_i values being the perpendicular distances of the point from the three sides. The triangle representation is shown in Figure 4 and the values of k for some standard relationships are given in Table 1.

The kinship coefficient (ψ) is the probability that homologous genes segregating form two individuals are identical by descent and thus $\psi = (2k_2 + k_1)/4$. While each relationship determines a point k, the converse is not true. Several relationships give the same probabilities k, as for example, half sibs (HS), grandparent-grandchild (G), and avuncular (AV) all have $k = (0.50, 0.50, 0)$. We should note that some points in the triangle are not attainable by any non-inbred relationship, as there is a restriction on the parameters $k_1 \geq 4k_1k_2$. Here we will give the proof for the inequality condition. For non-inbred individuals

$$\psi = \frac{1}{4}(\psi_{MM} + \psi_{FF} + \psi_{MF}\psi_{FM})k_2 = (\psi_{MM}\psi_{FF} + \psi_{MF}\psi_{FM})$$

where the subscripted kinship coefficients are those between a parent (mother (M) or father (F)) of one individual, and a parent of the other. Now, as given by Thompson [47], the arithmatic-geometric mean inequality gives

$$\begin{aligned} 4k_2 &\leq (\psi_{MM} + \psi_{FF})^2 + (\psi_{MF} + \psi_{FM})^2 \\ &\leq (\psi_{MM} + \psi_{FF} + \psi_{MF} + \psi_{FM})^2 \\ &= (4\psi)^2 \\ &= (k_1 + 2k_2)^2 \qquad (1.8) \\ &= k_1^2 + 4k_2(k_1 + k_2), \qquad (1.9) \end{aligned}$$

and this implies that,

$$\begin{aligned} 4k_2k_0 &= 4k_2(1 - (k_1 + k_2)) \qquad (1.10) \\ &\leq k_1^2 \end{aligned}$$

For some relationships, such as full sibs ($\psi_{MM} = \psi_{FF} = 0.25, \psi_{MF} = \psi_{FM} = 0$) and double first cousins, equality of the above condition holds. In the relationship triangle, these relationships fall on the boundary parabola. For an inbred relationship, the IBD probabilities k_0, k_1, k_2 might not satisfy the inequality $k_1^2 \geq 4k_0k_2$. Thus an inbred relationship might not be represented in the possible region of the relationship triangle.

We now give example (Table 1) of k and ψ for several different outbred relationships.

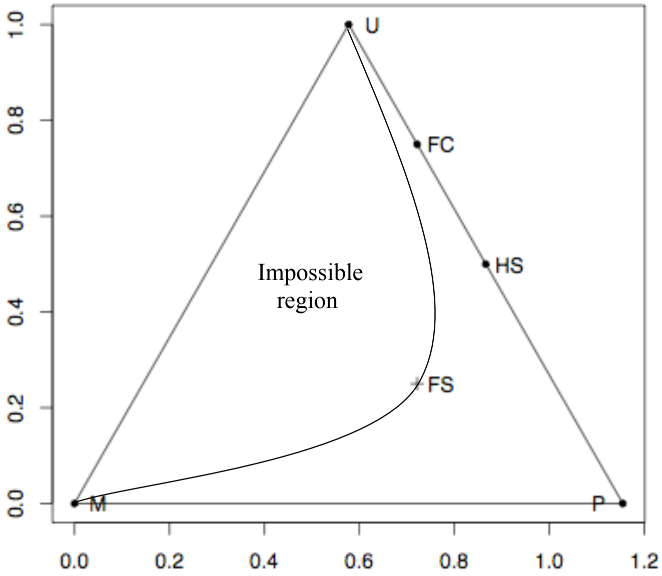


Figure 4: The Relationship triangle. The notations FS, HS, FC, U, P and M denote full sibs, half sibs, first cousins, unrelated, parent offspring and MZ twins. (This triangle is after Thompson [1986])

Table 1: Values of k and kinship coefficient ψ for some standard relationships between two non-inbred individuals

Pairwise relationship	k_0	k_1	k_2	ψ
Unrelated (U)	1.00	0	0	0
Parent-offspring (PO)	0	1.00	0	.25
Monozygous twin (M)	0	0	1.00	.50
Full sibs (FS)	.25	.50	.25	.25
Half sib (HS), Grandparent-grandchild (G), Avuncular (AV)	.50	.50	0	.125
First cousin (FC)	.75	.25	0	.0625

1.3.6 Summary

Relationship error is quite frequent in real data, and due to the assumption that true relationship is known without error, misspecified relationship can have potentially serious consequences on linkage analysis. The common practice to resolve the relationship error issue is either to discard erroneous individuals or to construct an alternative pedigree structure. In our approach, we statistically model the relationship uncertainty, and thus develop three new affected relative pair linkage analysis statistics. In section 1.4 we give the details of these statistics, and the derivations are given in the section 1.5.

1.4 METHODS

Affected relative pair (ARP) linkage analysis methods, and thus the software packages assume true relationship between the relative pairs is known without error. This assumption leads to reduced power or false positive evidence of linkage in presence of misspecified relationships. Thus, the fact that relationship error is quite frequent in real data motivated us to develop ARP linkage statistics that model relationship uncertainty. Consider the situation where we have collected affected relative pair (ARP) data and carried out a genome-wide scan for linkage. As the stated apparent relationships might be different from the true relationship, we have developed three linkage statistics that model relationship uncertainty. Our first relationship uncertainty linkage statistics (z-RULS) is an extension of the maximum lod score statistics of Risch [36]. Our second RULS (V-RULS), which is derived from our first one, is similar to the Cordell et al. [9] MLS, and our third RULS (L-RULS)is based on the conditional-logistic representation of Olson [33].

Extension of Risch's maximum lod score: Risch [36] showed that for ARPs, the likelihood for observed marker data for the jth pair is $\sum_{i=0}^{2} z_i w_{ij}$, where $w_{ij} = P(G_j|i)$. Now one can proceed as follows,

$$\begin{aligned}\sum_{i=0}^{2} z_i w_{ij} &= \sum_i z_i P(G_j|i)\\ &= \sum_i z_i \frac{P(i|G_j)P(G_j)}{P(i)}\\ &= \sum_i z_i \frac{\hat{f}_{ij}P(G_j)}{f_i}\end{aligned}$$

Under H_0 of no linkage, the likelihood of the observed marker data is $P(G_j)$, and thus the log likelihood ratio test statistic for testing H_0 becomes,

$$\sum_j \log_{10}(\sum_i \frac{\hat{f}_{ij} z_i}{f_i}). \tag{1.11}$$

If there is relationship error, then the stated apparent relationship might not be the same as the underlying true relationship. Thus, when one allows for relationship error in this statistical framework, one has to adjust how one computes f_i and $\hat{f}_{ij}$ as one can no longer use these for the observed relationship. Furthermore, one must also adjust how one computes the z_i.

1.4.1 z-RULS

Our RULS that estimates z, the identity by descent (IBD) sharing probabilities among affecteds, z-RULS, is an extension of the maximum likelihood statistic (MLS) developed by Risch [36] to the situation where true relationships might be different from the stated relationships. For possible true relationship types, we consider five outbred relationships: full sibs (FS), half sibs (HS), first cousins (FC), unrelated (U) and parent offspring (PO), as they cover the outbred relationship space of the relationship triangle [47].

Let Aff denote the event that both individuals in a pair are affected, and for affected pair j, let G_j be the genome-wide marker data, T the true relation type, and A_j the apparent relation type. Also, for a pair with apparent relation type A_j, let f_{ij}^{A} be the probability of sharing i alleles IBD at the marker given the genotype data, F_{ij}^{A} is the posterior probability of sharing i alleles IBD, and z_i^T be the probability that an affected pair shares i alleles IBD at the marker given the true relation type T.

The statistic for testing the null hypothesis that the marker is unlinked to the disease, $H_0:\ z_0^{FS} = .25,\ z_1^{FS} = .5,\ z_0^{HS} = .5,\ z_0^{FC} = .75$, is given as:

$$z\text{-}RULS = \sum_j \log_{10}[\sum_{i=0}^{2} \frac{f_{ij}^{A}}{F_{ij}^{A}} \sum_{T\in(FS,HS,FC,U,PO)} z_i^T P(T|\mathrm{Aff}, A_j)] \tag{1.12}$$

where $z_0^U = 1$, $z_1^P = 1$. As we can not estimate $P(T|\mathrm{Aff}, A_j)$, it is approximated by $P(T|A_j)$ (see section 1.5 for details behind the derivation of z-RULS).

We notice that our z-RULS (equation 1.12) is analogous to Risch's maximum lod score as given in equation (1.11), as each of the three terms (z_i,$\hat{f}_{ij}$,f_i) in equation (1.11) is replaced in z-RULS by the appropriately weighted average of their respective values for each of true relationships.

Similar to Risch's [36] approach for his maximum lod score, we maximize z-RULS over the parameters z_i^T. Under H_0, the asymptotic distribution of $2\ln(10)[z\text{-}RULS]$ is χ^2_4. This is because there is no genetic constraints over the parameter space for z-RULS.

The z-RULS ignores the correlations between the different z_i^T that are induced by a genetic model. The number of parameters we estimate (z_i^T) depends on the number of possible true relationships, thus increasing this number lowers the power of z-RULS. Considering these five true relationships (FS, HS, FC, PO, and U), we estimate a total of four parameters: z_0^{FS}, z_1^{FS}, z_0^{HS} and z_0^{FC}. One doesn't need parameters for U and PO, see details in section 1.5

The MLS proposed by Cordell et al. [9] is maximized with respect to the genetic variances; additive and dominance variances, subject to the constraint that the variance components are nonnegative. We adopt this technique to decrease the number of parameters in z-RULS and develop another RULS, the V-RULS. Our yet another RULS, L-RULS, based on the conditional-logistic representation of Olson [33], also estimates a smaller number of parameters. We now show that unlike z-RULS, both V-RULS and L-RULS model the correlations between the z_i^T, and thus use a lower number of parameters than is used by z-RULS.

1.4.2 V-RULS

The V-RULS estimates two parameters: $P_A = V_A/K^2$, $P_D = V_D/K^2$, where V_A is additive variance, V_D is dominance variance, and K is population prevalence. The V-RULS is derived from z-RULS following Cordell et al. [9]'s insight to reduce number of parameters. In their MLS, the parameter z_i is expressed in terms of P_A and P_D, here we denote it by $z_i(P_A, P_D)$. Thus, their MLS can be written as

$$MLS = \sum_j \log_{10}(\sum_{i=0}^{2} \frac{z_i(P_A, P_D)\hat{f}_{ij}}{f_{ij}}) \tag{1.13}$$

We extend z-RULS to V-RULS by writing z_i^T's in terms of P_A, P_D as $z_i^T(P_A, P_D)$. Thus V-RULS is given by:

$$V\text{-}RULS = \sum_j \log_{10}[\sum_{i=0}^{2} \frac{f_{ij}^A}{F_{ij}^A} \sum_{T\in(FS,HS,FC,U,PO)} z_i^T(P_A, P_D)P(T|\text{Aff}, A_j)] \tag{1.14}$$

Now, comparing Cordell's MLS and our z-RULS, we note that equation (1.14) is the analog of equation (1.13) where each of the three terms ($z_i(P_A, P_D)$,$\hat{f}_{ij}$,f_{ij}) in equation (1.13) is replaced by the appropriately weighted average of their respective values for each of true relationships (for details, see section 1.5).

In order to express z_i^T's as function of P_A, P_D, the parameters z_i^T are first written in terms of f_i^T, λ_i and λ_T where f_i^T is the prior probability of sharing i alleles IBD for true relationship T, λ_i are the relative risks to an individual sharing i alleles IBD with an affected individual, and λ_T is the risk ratio of an individual related as T with an affected individual. Both λ_i and λ_T being functions of the additive and dominance variances, equation (1.12) can be written as function of the P_A and P_D, and thus we named our statistic as V-RULS. So, the statistic V-RULS is given as,

$$V\text{-}RULS = \sum_j \log_{10} \sum_i \frac{f_{ij}^A \lambda_i(P_A, P_D)}{F_{ij}^A} \sum_{T \in \{FS,HS,FC,U,PO\}} f_i^T P(T|\text{Aff}, A_j)/\lambda_T(P_A, P_D) \tag{1.15}$$

The details behind the derivation is given in the section 1.5.

The V-RULS does not require specification of the population prevalence K, and thus is robust to K. Under H_0, the asymptotic distribution of $2\ln(10)[V\text{-}RULS]$ is a mixture of χ^2 distributions. Since it is difficult to obtain p-values from mixture of distributions, we will perform simulation study to calculate genome-wide threshold for V-RULS.

1.4.3 L-RULS

The L-RULS estimates two parameters: the relative risks to an individual who shares 1 or 2 alleles IBD with an affected individual, denoted as λ_1 and λ_2 respectively. The proposed likelihood ratio test statistic is derived similarly as conditional-logistic representation of the affected relative pair (ARP) likelihood ratio of Olson [33].

Let λ_i denote the relative risk of being affected to an individual who shares i alleles IBD with an affected relative for $i = 0$, 1, 2 and $\lambda_0 = 1$. The statistic L-RULS is given as,

$$L\text{-}RULS = \sum_j \log_{10}[\frac{\sum_i \lambda_i f_{ij}^A}{\sum_i \lambda_i F_{ij}^A}]. \tag{1.16}$$

Details are given in the section 1.5. Under H_0, the asymptotic distribution of $2\ln(10)[L\text{-}RULS]$ is a mixture of χ^2 distributions, and similar to V-RULS, genome-wide threshold for L-RULS is computed from the simulation study.

It can be shown analytically (as in section 1.5) that under certain assumptions, V-RULS becomes the same as L-RULS.

1.5 STATISTICAL DERIVATION OF THE RULS

Here we give the details behind the derivations of our three RULS and the details to compute each RULS (as shown in the flowchart Figure 5). For completeness, we reiterate our variable definitions here: Let Aff denote the event of both individuals in a pair being affected and for affected pair j, let G_j be the genome-wide marker data, T the true relationship type, and A_j the apparent relationship type. Also, let i be 0, 1, 2 alleles identity by descent (IBD) at a marker, f_{ij}^{A} be the probability of sharing i alleles IBD at the marker given the genotype data of the pair with apparent relationship A_j, F_{ij}^{A} is the posterior probability of sharing i alleles IBD, and z_i^T be the probability that an affected pair shares i alleles IBD at the marker given the true relation type T.

1.5.1 z-RULS

To develop z-RULS, the likelihood for the affected pair j with apparent relationship type A_j is given as:

$$L(z_i^T|G_j, \text{Aff}, A_j) \propto P(G_j|\text{Aff}, A_j)P(\text{Aff}, A_j)$$

where

$$\begin{aligned}
P(G_j|\text{Aff}, A_j) &= \sum_{i=0}^{2} P(G_j|\text{Aff}, A_j, i)P(i|\text{Aff}, A_j) \\
&\approx \sum_i P(G_j|\text{Aff}, A_j, i) \sum_{T\in(FS,HS,FC,U,PO)} P(i|\text{Aff}, A_j, T)P(T|\text{Aff}, A_j) \qquad (1.17)\\
&= \sum_i P(G_j|A_j, i) \sum_T P(i|\text{Aff}, A_j, T)P(T|\text{Aff}, A_j) \qquad (1.18)\\
&= \sum_i P(G_j|A_j, i) \sum_T P(i|T, \text{Aff})P(T|\text{Aff}, A_j) \\
&= \sum_i \frac{P(i|G_j, A_j)P(G_j|A_j)}{P(i|A_j)} \sum_T P(i|T, \text{Aff})P(T|\text{Aff}, A_j) \\
&= \sum_i \frac{f_{ij}^{A} P(G_j|A_j)}{F_{ij}^{A}} \sum_{T\in(FS,HS,FC,U,PO)} z_i^T\, P(T|\text{Aff}, A_j)
\end{aligned}$$

Step 1.17 is an approximation because we are summing over only 5 true relationships, which is not the full set of possible true relationships. Step 1.18 is true if there is no association of the genetic markers with the disease locus.

The likelihood under the null hypothesis of no linkage for pair j is given as

$$\begin{aligned} L_0 &\propto \sum_i \frac{f_{ij}^A P(G_j|A_j)}{F_{ij}^A} \sum_{T\in(FS,HS,FC,U,PO)} P(i|T)P(T|A_j)P(\text{Aff}, A_j) \\ &= \sum_i \frac{f_{ij}^A P(G_j|A_j)}{F_{ij}^A} F_{ij}^A P(\text{Aff}, A_j) \\ &= P(G_j|A_j) \sum_i f_{ij}^A P(\text{Aff}, A_j) \\ &= P(G_j|A_j)P(\text{Aff}, A_j) \end{aligned}$$

So, the likelihood ratio for affected pair j is

$$LR_j = \sum_i \frac{f_{ij}^A}{F_{ij}^A} \sum_{T\in(FS,HS,FC,U,PO)} z_i^T \, P(T|\text{Aff}, A_j) \tag{1.19}$$

and we maximize log-likelihood ratio for all affected pairs, $\sum_j \log LR_j$ over z_i^T to get z-RULS as in equation (1.12).

In order to compute the likelihood ratio, we proceed as follows. In equation (1.30), we can not actually estimate $P(T|\text{Aff}, A)$, but can only estimate $P(T|A)$, as shown below:

$$\begin{aligned} P(T|\text{Aff}, A) &= P(A|T, \text{Aff}) \frac{P(T)}{P(A)} \frac{P(\text{Aff}|T)}{P(\text{Aff}|A)} \\ &= P(A|T) \frac{P(T)}{P(A)} \frac{P(\text{Aff}|T)}{P(\text{Aff}|A)} \\ &= P(T|A) \frac{P(\text{Aff}|T)}{P(\text{Aff}|A)} \end{aligned} \tag{1.20}$$

As we do not know the true disease model, the value of $P(\text{Aff}|T)$ and $P(\text{Aff}|A)$ are unknown to us. Hence, we cannot estimate $P(T|\text{Aff}, A)$. Next we show that the term $P(T|\text{Aff}, A)$ in equation (1.30) can be approximated by $P(T|A)$. Let us assume $P(T_l|A) \approx 1$ and $P(T_k|A) \approx 0$ for $k \neq l$. Also we have

$$\begin{aligned} P(\text{Aff}|A) &= \sum_k P(\text{Aff}|T_k, A)P(T_k|A) \\ &= \sum_k P(\text{Aff}|T_k)P(T_k|A) \end{aligned} \tag{1.21}$$

Under the above assumption, we can write equation (1.21) as

$$P(\text{Aff}|A) \approx P(\text{Aff}|T_l) \tag{1.22}$$

From equations 1.20 and 1.22 we get,

$$\Rightarrow P(T|\text{Aff}, A) \approx P(T|A)$$

under the assumption that $P(T|A)$ consists of one 1 and the rest zeros, so we can approximate $P(T|\text{Aff}, A)$ by an estimate of $P(T|A)$, denoted here as $r_j^{T|A}$. In both the SP and LP datasets we have $r_j^{T|A}$ satisfying the condition that $P(T|A)$ consists of one 1 and the rest zeros, showing that approximating $P(T|\text{Aff}, A)$ by $P(T|A)$ works well for these two datasets. One has to though remember that $r_j^{T|A}$ may consist exactly of one 1 and rest zeros only when there is no genotyping error. In the presence of genotyping error, $r_j^{T|A}$ might not consist exactly one 1 and rest 0's, but might be close to that, and $P(T|\text{Aff}, A)$ might still be estimated by $P(T|A)$.

In order to estimate $r_j^{T|A}$, we consider the genome-wide marker data of the pair j under H_0:

$$\begin{aligned} P(G_j|A_j) &= \sum_T P(G_j|T)P(T|A_j) \\ &= \sum_T P(G_j|T)r_j^{T|A} \end{aligned} \tag{1.23}$$

We maximize $P(G_j|A_j)$ over $r^{T|A}$ to get the estimates for pair j. The $P(G_j|T)$ in equation (1.23) are obtained (under the assumption of no intereference) by modifying the PREST program [31]. We consider FS, HS, FC, U and PO as the possible true relationships as they cover the space for outbred relationships in the relationship triangle (except the MZ twins corner and the impossible region) [47]. We use quasi-Newton optimization as implemented in the SEARCH program [26], to optimize equation (1.23) over $r_j^{T|A}$, subject to the constraint $\sum_T r_j^{T|A} \leq 1$ for each affected pair j. In the flowchart (Figure 5) to compute RULS, these steps are shown by denoting $r_j^{T|A}$ as $r^{T|A}$ for an affected pair.

We then calculate the posterior probability of sharing alleles IBD as,

$$F_{ij}^A \approx \sum_{T \in (FS,HS,FC,U,PO)} f_i^T r_j^{T|A} \tag{1.24}$$

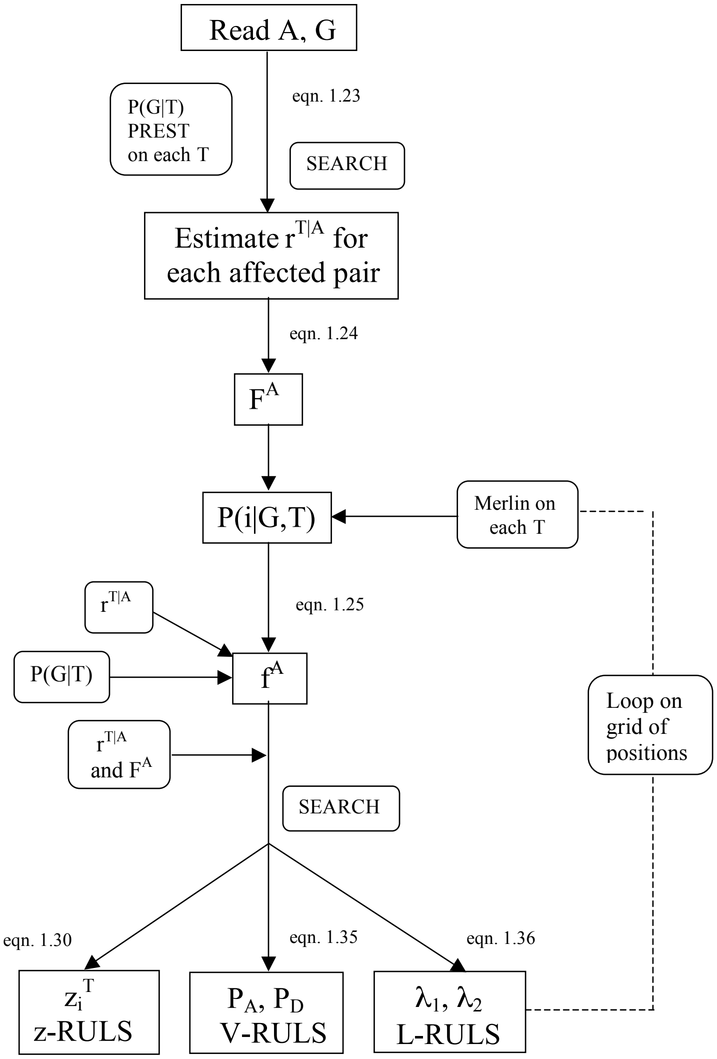


Figure 5: Flowchart showing the steps to calculate the RULS. See section 1.5 for the notation used in this flowchart.

where f_i^T is prior probability of sharing i alleles IBD for true relationship T. Once $r_j^{T|A}$'s are obtained, we calculate F_{ij}^A from equation (1.24) as illustrated in the flowchart Figure 5, and F_{ij}^A is denoted as F^A.

The term f_{ij}^A in equation (1.12) is obtained as

$$\begin{aligned} f_{ij}^A &= P(i|A_j, G_j) \\ &\approx \sum_{T \in (FS,HS,FC,U,PO)} P(i|A_j, G_j, T)P(T|A_j, G_j) \\ &= \sum_T P(i|G_j, T)\frac{P(G_j|T)P(T|A_j)}{P(G_j|A_j)} \\ &= \sum_{T \in (FS,HS,FC,U,PO)} P(i|G_j, T)\frac{P(G_j|T)r_j^{T|A}}{\sum_T P(G_j|T)r_j^{T|A}}. \end{aligned} \tag{1.25}$$

where $P(i|G_j, T)$ is computed at a grid of positions on the genome.

Now, it might at first seem that for T=PO, $P(i|G_j, T)$ may be undefined if one observes, as for a genotype configuration not consistent with 1 IBD sharing. For example, we can discuss about such a scenario at a marker, as assuming non interference between markers, $P(i|G_j, T)$ depends on genotype data at the marker. Consider the genotype configuration G_j at a marker for a PO pair is (1/1, 2/2). In this case, at this marker G_j, T=PO is not consistent and also, $i = 2$ is not consistent with either G_j or T=PO. Thus, for this genotype configuration at the marker, $P(i = 2|G_j, T = PO)$ is undefined.

But, we can show as follows, that, taking genotype error model, $P(i|G_j, T)$ exists and becomes 0, 1, 0 for i = 0, 1, 2 if the error is non-zero. To start with, we consider an error model (see model below) for an individual with observed genotype G_o at a marker and true genotype G_t at the marker, and this error model is similar to that given by Sobel et al. [43].

$$P(G_o|G_t) = \begin{cases} 1-\epsilon & \text{if } G_o = G_t \\ \frac{\epsilon}{m-1} & \text{otherwise} \end{cases}$$

where ϵ is the error rate per genotype, both genotypes are unordered, and there are m genotypes in all common for the pair. We note that $P(i|G_o, T)$ is given by:

$$P(i|G_o, T) = \frac{P(i, G_o, T)}{P(G_o, T)} \tag{1.26}$$

where $P(i, G_o, T)$ can be written in terms of G_t as:

$$\begin{aligned} P(i, G_o, T) &= \sum_{G_t} P(i, G_o, G_t, T) \\ &= \sum_{G \subset G_t} P(i, G_o, G, T) \qquad (1.27) \end{aligned}$$

In step (1.27), G is a subset of the true genotype configurations G_t, those that are consistent with i and $T = PO$, and equality at this step holds as $P(i, G_o, G_t, T) = 0$ for other G_t's where i, G_t, G_o, T are not consistent with each other. So, in either case of $\epsilon > 0$ or $= 0$, $i = 0, 2$ is not consistent with $T = PO$, implying that

$$P(i, G_o, G, T = PO) = \begin{cases} 0 & \text{if i=0,2} \\ > 0 & \text{if i=1} \end{cases}$$

Now, $P(G_o, T = PO)$ of equation (1.26) is given by,

$$\begin{aligned} P(G_o, T = PO) &= \sum_{j} P(i, G_o, T = PO) \\ &= \sum_{j} \sum_{G} P(i, G_o, G, T) \\ &= \sum_{G} \sum_{j} P(i, G_o, G, T) \\ &= \sum_{G} P(i = 1, G_o, G, T = PO) \qquad (1.28) \end{aligned}$$

where step (1.28) holds true by the previous argument.

So, to prove that $P(i|G_o, T)$ in equation (1.26) exists for any i, we have to essentially show that $\sum_G P(i = 1, G_o, G, T) > 0$.

$$\begin{aligned} P(i = 1, G_o, G, T) &= \sum_{G} P(T|i = 1, G_o, G) P(i = 1|G_o, G) P(G_o|G) P(G) \\ &= \sum_{G} P(T|G, i = 1) P(i = 1|G) P(G_o|G) P(G), \qquad (1.29) \end{aligned}$$

where, G being consistent with $i = 1$ and $T = PO$, the probabilities $P(T = PO|G, i)$ and $P(i = 1|G)$ are positive. Also, as for $\epsilon > 0$, $P(G_o|G)$ is also positive, and thus from equation (1.29), $\sum_G P(i = 1, G_o, G, T) > 0$, *i.e.* $P(G_o, T) > 0$. Hence, considering error model, we have showed that for $\epsilon > 0$, the probability $P(i|G_o, T)$ exists.

Now, as we have already showed that assuming the genotype error model, $P(i, G_o, G, T) = 0$ for $i = 0, 2$, and $P(G_o, T) > 0$, the conditional probabilities $P(i|G_o, T) = 0$ for $i = 0, 2$, and this implies that $P(i = 1|G_o, T) = 1$. Hence we can conclude that for $T = PO$, if we allow for a small error, $P(i|G_j, T = PO) = (0, 1, 0)$.

The term $P(i|G_j, T)$ in equation (1.25) for pair j is obtained by Merlin [1] at a grid of positions (1 cM apart on genome) on the true structure T. See flowchart Figure 5 for steps to calculate $f^A = f^A_{ij}$ at grid of positions. To compute the z-RULS, we use the SEARCH optimization method [26] on equation (1.12) under equality constraints $\sum_{i=0}^{2} z_i^{FS} = 1$, $\sum_{i=0}^{1} z_i^{HS} = 1$ with $z_2^{HS} = 0$ and $\sum_{i=0}^{1} z_i^{FC} = 1$ with $z_2^{FC} = 0$.

Here we note that the z-RULS as given by

$$\sum_j log_{10}[\sum_i \frac{f^A_{ij}}{F^A_{ij}} \sum_{T \in (FS,HS,FC,U,PO)} z_i^T \, P(T|\text{Aff}, A_j)] \tag{1.30}$$

is analogous to Cordell et al. MLS (see equation (1.13)) as expressed in Table 2 where each of the three terms (z_i,$\hat{f}_{ij}$,f_{ij}) in equation (1.13) is replaced by the appropriately weighted average of their respective values for each of true relationships.

Table 2: Comparison between Cordell et al's MLS and z-RULS.

MLS	z_i	$\hat{f}_{ij}$	f_{ij}
z-RULS	$\sum_i z_i^T r_j^{T\mid A}$	$f^A_{ij} \sim \sum_T P(i\mid G_j, T) \frac{P(G_j\mid T) r_j^{T\mid A}}{\sum_T P(G_j\mid T) r_j^{T\mid A}}$	$F^A_{ij} \sim \sum_T f_i^T r_j^{T\mid A}$

1.5.2 V-RULS

To derive our V-RULS, we begin with the likelihood ratio for affected pair j:

$$LR_j = \sum_i \frac{f^A_{ij}}{F^A_{ij}} \sum_{T \in \{FS,HS,FC,U,PO\}} z_i^T P(T|\text{Aff}, A_j)$$

Now, z_i^T can be expressed, as given by Cordell et al. [9], in terms of two parameters λ_i and λ_T and one known term that depends only on T, f_i^T:

$$z_i^T = \frac{\lambda_i f_i^T}{\lambda_T}. \tag{1.31}$$

Here, f_i^T is the prior probability of sharing i alleles IBD for relationship T , λ_i are the relative risks to an individual who shares i alleles IBD with an affected individual, and λ_T is the risk ratio of an individual related as T with an affected individual. The above likelihood for pair j can then be written as,

$$LR_j = \sum_i \frac{f_{ij}^A \lambda_i}{F_{ij}^A} \sum_{T \in \{FS,HS,FC,U,PO\}} f_i^T r_j^{T|A} / \lambda_T. \tag{1.32}$$

As derived by James [20], we can express λ_T in terms of the covariance of affection status of the affected pair and the population prevalence K:

$$\lambda_T = 1 + \frac{Cov(X_1, X_2)}{K^2}, \tag{1.33}$$

where X_i is the phenotype of person i defined to be 0 or 1, according to whether the person is affected or unaffected, and Cov denotes the covariance. James [20] showed that $Cov(X_1, X_2)$ can be expressed in terms of the additive and dominance variance (V_A, V_D) caused by disease locus, and kinship coefficient (r_A) and prior probability of sharing 2 alleles IBD (f_2^T). The covariance is given by:

$$Cov = 2r_A V_A + u_R V_D \tag{1.34}$$

where $r_A = .5f_2^T + .25f_1^T$ and $u_R = f_2^T$. This expression and equation (1.33) show that both λ_T and λ_i are functions of V_A and V_D, and hence are functions of $P_A = V_A/K^2$ and $P_D = V_D/K^2$. Thus V-RULS can be written as:

$$V\text{-}RULS = \sum_j \log_{10} \sum_i \frac{f_{ij}^A \lambda_{i(P_A,P_D)}}{F_{ij}^A} \sum_{T \in \{FS,HS,FC,U,PO\}} f_i^T P(T|\text{Aff}, A_j) / \lambda_{T(P_A,P_D)} \tag{1.35}$$

We use SEARCH [26] to optimize equation (1.35) with respect to $P_A \geq 0$, $P_D \geq 0$ to compute the V-RULS. The steps to compute V-RULS are also shown in Figure 5.

1.5.3 L-RULS

To derive the L-RULS, based on the approach of Olson [33], we write the likelihood ratio for affected pair j is

$$\begin{aligned}
LR_j = \frac{P(G_j|\text{Aff}, A_j)}{P(G_j|A_j)} &= \frac{P(\text{Aff}|G_j, A_j)}{P(\text{Aff}|A_j)} \\
&= \frac{P(\text{Aff}_1)P(\text{Aff}_2|G_j, A_j, \text{Aff}_1)}{P(\text{Aff}_1)P(\text{Aff}_2|A_j, \text{Aff}_1)} \frac{P(\text{Aff}_2)}{P(\text{Aff}_2)} \\
&= \frac{P(\text{Aff}_2|G_j, A_j\text{Aff}_1)/P(\text{Aff}_2)}{P(\text{Aff}_2|A_j, \text{Aff}_1)/P(\text{Aff}_2)} \\
&= \frac{\sum_i [P(\text{Aff}_2|i, \text{Aff}_1)/P(\text{Aff}_2)]P(i|G_j, A_j)}{\sum_i [P(\text{Aff}_2|i, \text{Aff}_1)/P(\text{Aff}_2)]P(i|A_j)} \\
&= \frac{\sum_i \lambda_i f_{ij}^A}{\sum_i \lambda_i F_{ij}^A}
\end{aligned}$$

where Aff_1 and Aff_2 denote individual 1 and 2 is affected respectively, λ_i are the relative risk to an individual who shares i alleles IBD with an affected relative, $i = 0, 1, 2$, $\lambda_0 = 1$. The terms f_{ij}^A and F_{ij}^A are obtained as in z-RULS. Thus, the log likelihood ratio taking all affected pairs can be written as:

$$L\text{-}RULS = \sum_j \log_{10}[\frac{\sum_i \lambda_i f_{ij}^A}{\sum_i \lambda_i F_{ij}^A}]. \tag{1.36}$$

To compute L-RULS, we use SEARCH [26] to optimize equation (1.38) under the condition $\lambda_0 = 1$, $\lambda_1 \geq 1$, $\lambda_2 \geq 2\lambda_1 - 1$ so that the genetic constraints hold. The steps to compute L-RULS is also shown in flowchart Figure 5.

Here we will show that V-RULS and L-RULS become equal when $r^{T|A}$ contains one 1 and others zeros. The likelihood ratio for pair j for V-RULS can be written as:

$$\begin{aligned}
&\sum_i \frac{f_{ij}^A}{F_{ij}^A} \sum_T \frac{\lambda_i f_i^T}{\lambda_T} r_j^{T|A} \\
&\quad = \sum_i \frac{\lambda_i f_{ij}^A}{\sum_T f_i^T r_j^{T|A}} \sum_T \frac{f_i^T r_j^{T|A}}{\lambda_T}
\end{aligned} \tag{1.37}$$

and that for L-RULS is,

$$
\begin{aligned}
\frac{\lambda_i f_{ij}^A}{\sum_i F_{ij}^A \lambda_i} &= \frac{\sum_i f_{ij}^A \lambda_i}{\sum_T r_j^{T|A} \sum_i f_i^T \lambda_i} \\
&= \frac{\sum_i f_{ij}^A \lambda_i}{\sum_T r_j^{T|A} \lambda_T}. \qquad (1.38)
\end{aligned}
$$

Now, for $r_j^{T|A}$ containing one 1 and others zeros, the term $f_i^T r_j^{T|A} = f_i^T$ for that T where $r_j^{T|A} = 1$. So, equation (1.37) becomes equal to equation (1.38), implying that V-RULS is same as L-RULS when $r_j^{T|A}$ has one 1 and others zeros.

All our RULS computations are done at a pairwise level for an ARP, and changing the structure to a true structure is also done for the pair, as shown in flowchart Figure 5. For both MLS and S_{all}, the probability of genotype data is computed from the entire family data unlike considering only ARPs in RULS.

We implemented the RULS (as discussed in this section and in flowchart Figure 5) in a software program. This software will be available at http://watson.hgen.pitt.edu/register/. Please see Appendix A for software documentation, and Appendix B for code developed to compute RULS, MLS and S_{all} on SP and LP dataset.

1.6 SIMULATION

We performed a simulation study to evaluate our new RULS, and to compare them to the MLS [9] and the exponential S_{all} nonparametric LOD score from Merlin [1] using true and discarded structures. We simulated two datasets, one having small pedigrees (SP) and another having large pedigrees (LP). As affected relative pairs in the same pedigree may not be mutually independent, we have empirically computed the significance thresholds for our RULS by simulation [9].

1.6.1 Data structure

The SP dataset consists of 300 pedigrees with 660 affected pairs having several underlying true relationship types between the affected individuals (see Figure 6A). We first simulate data using the true pedigree structures and then change the true structures to the apparent structures, where all affecteds are apparently related to each other as full sibs (FS). So the number of pairs in SP having relationship errors is 40 (true HS stated as apparent FS, *i.e.* HS→ FS), 20 (HS→ FS), 40 (U→ FS) and 80 (FC→ FS). This implies there are 180 out of 660 ARPs, *i.e.* 27% of ARPs have erroneous relationships. To get the discarded data for SP dataset, we removed those individuals who are not truly full sibs.

The LP dataset consists of 60 large pedigrees (structures I and II) with several underlying true relationships, e.g., full sibs, half sibs, second cousins (see Figure 6B). To create the apparent structure, we randomly moved an individual to another sibship in the terminal generation based on an assumed error rate. In structure I (Figure 6B), we randomly moved an individual from one sibship to become an apparent member of another sibship with probabilities as given in Table 3, and for structure II, we randomly moved an individual from sibship d to e and vice versa with probability 0.2. Unlike SP, LP dataset contains a more realistic relationship error proportion ranging between 10% to 16%. To create the discarded structure, we removed those individuals who are known to be erroneous *i.e.* those individuals who are moved.

1.6.2 Marker data and disease models

The simulated marker data consists of 367 autosomal markers with an average 10 cM spacing throughout the genome. We used realistic microsatellite marker allele frequencies and realistic map distances. Chromosome 10 contains the disease locus (at 52.53 cM) which is simulated using the underlying genetic models in Table 4. For given values of K and the penetrances, the rest of the parameters are calculated.

We simulated marker data on the non-disease chromosomes using Simulate [45] and on the disease chromosome conditional on the assigned disease status using Allegro v1.2c [17]. The marker simulation is done without any genotyping errors and the proportion of linked

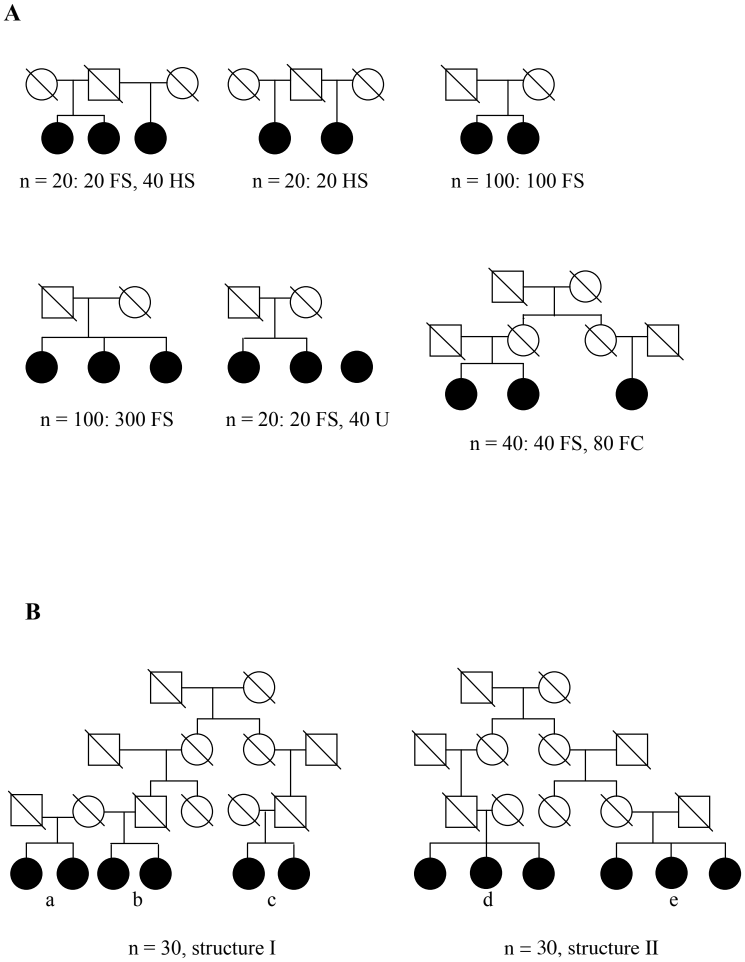


Figure 6: True structures of the small pedigree (SP) and large pedigree (LP) datasets. FS, HS, U and FC represent full sib, half sib, unrelated and first cousin, and n is the number of families with the given structure. Circles and squares denote females and males respectively, the blackened symbols indicate affected individuals, the clear symbols with a slash denote deceased individuals (who are neither phenotyped or genotyped).

Table 3: Probabilities of randomly moving an individual in structure I in LP to create the apparent structure

		To		
	Sibship	a	b	c
	a	.7	.2	.1
From	b	.2	.7	.1
	c	.1	.1	.8

Table 4: Genetic models used in our simulations, where K: prevalence, q: disease allele frequency, pen: penetrance, λ_s: relative risk to sib, and λ_o: relative risk to offspring

Model	q	pen$_1$	pen$_2$	pen$_3$	λ_s	λ_o	Description
K=.13:							
1	.12	.00	.60	.60	2.75	2.80	Dominant, no phenocopies
2	.13	.01	.50	.50	2.27	2.22	Dominant, phenocopies
3	.46	.00	.00	.60	2.48	2.15	Recessive, no phenocopies
4	.49	.01	.01	.50	2.09	1.87	Recessive, phenocopies
5	.13	.00	.50	1.0	2.67	2.67	Additive, no phenocopies
6	.21	.01	.30	0.6	1.83	1.83	Additive, phenocopies

families varied between models and pedigree structures (see legend for Figures 7 & 8). In the SP dataset, 80% of the families are linked for models 1-5 and 100% are linked for model 6. We have taken all families to be linked for all models in LP dataset, so that power of the statistics is not too small to compare between each other. Though genetic models 1-5 have genetic heterogeneity, the values of relative risks λ_s and λ_o (in Table 4) were computed under the assumption of genetic homogeneity.

1.6.3 Data generation

For each replicate, we computed the RULS at a grid of positions of 1 cM throughout the genome (see section 1.5 and flowchart Figure 5 for computational details). We also computed the MLS using code from Cordell et al. [9] and S_{all} using Merlin [1] on both true and discarded structures.

1.6.4 Computation

We simulated 1,000 replicates using all of the unlinked chromosomes to compute the empirical genome-wide threshold for each statistic. The threshold τ is taken to be the value for which $P(\text{maximum of RULS over all unlinked chromosomes} \geq \tau) = \alpha$, where α is the significance level, 0.01 and 0.05. For the power calculation, we simulated 400 replicates. Power is calculated at both levels of significance, as the proportion of replicates with a statistic value greater than the empirical threshold τ anywhere within +/- 10 cM of the true location of the disease locus.

1.7 RESULTS

Here we discuss the results obtained from applying RULS on both SP and LP datasets. First, in both SP and LP datasets, estimate of the weights $r^{T|A}$ consist of one 1 and rest zeros, implying that $P(T|\text{Aff}, A)$ can be approximated by $P(T|A)$. This is due to the fact that both SP and LP have genome-wide markers simulated without any error in the genotype data. Also we have checked from the true structure of SP and LP (see Figures 6A and B) that the T for which $r^{T|A}$ is 1 is the correct one for that specific pair. The relationship types of true structure of ARPs in SP dataset belong to the five T's we have considered. But in LP dataset, there are ARPs related as second cousins and half second cousins, which does not belong to the set of five T's we considered. For such ARPs, $r^{T=U|A} = 1$, as both second cousins and half second cousins have IBD sharing very close to an unrelated affected pair. Second, as it is a simulated study, the true structure is known, and we verified that the estimates of F^A for a pair from apparent pedigree structure matched with that of the pair's true structure.

Thirdly, we have calculated the genome-wide empirical thresholds for all the statistics. For some statistics we compared the empirical thresholds with the analytical thresholds obtained from their asymptotic distributions. As L-RULS, V-RULS and MLS have genetic constraints on their parameter spaces, it is difficult to derive the mixture distributions, and so analytical thresholds are not obtained for L-RULS and V-RULS. For other statistics, empirical thresholds are reasonably close to the expected analytical thresholds derived from their asymptotic distributions (Table 5).

Fourth, for both the SP and LP datasets, the L-RULS and V-RULS perform better than MLS and S_{all} on the discarded structure and their powers are reasonably close to MLS and S_{all} on the true structure (Figures 7 and 8). In both power graphs, a horizontal line is drawn along L-RULS, so that it is easier to compare the power of L-RULS with MLS and S_{all} on true and discarded structures.

Table 5: Genome-wide empirical thresholds and number of parameters estimated for small pedigree (SP) and large pedigree (LP), based on 1,000 replicates at the 0.05 and 0.01 significance levels. Analytical thresholds based on the asymptotic distributions of the statistics also given.

Level:							
(Dataset)	z-RULS	L-RULS	V-RULS	MLS	MLS	S_{all}	S_{all}
				true	discarded	true	discarded
Empirical:							
0.05 (SP)	5.60 (4)	3.50	3.50	3.04	3.08	3.10 (1)	3.06 (1)
0.05 (LP)	6.31(4)	4.21	4.21	5.81	5.72	2.65 (1)	2.61 (1)
Analytical:							
0.05	5.11 (4)					3.29 (1)	3.29 (1)
Empirical:							
0.01 (SP)	6.69 (4)	4.30	4.30	3.70	3.95	3.73 (1)	3.70 (1)
0.01 (LP)	6.70 (4)	4.52	4.51	6.10	5.80	3.50 (1)	4.10 (1)
Analytical:							
0.01	5.67 (4)					3.78 (1)	3.78 (1)

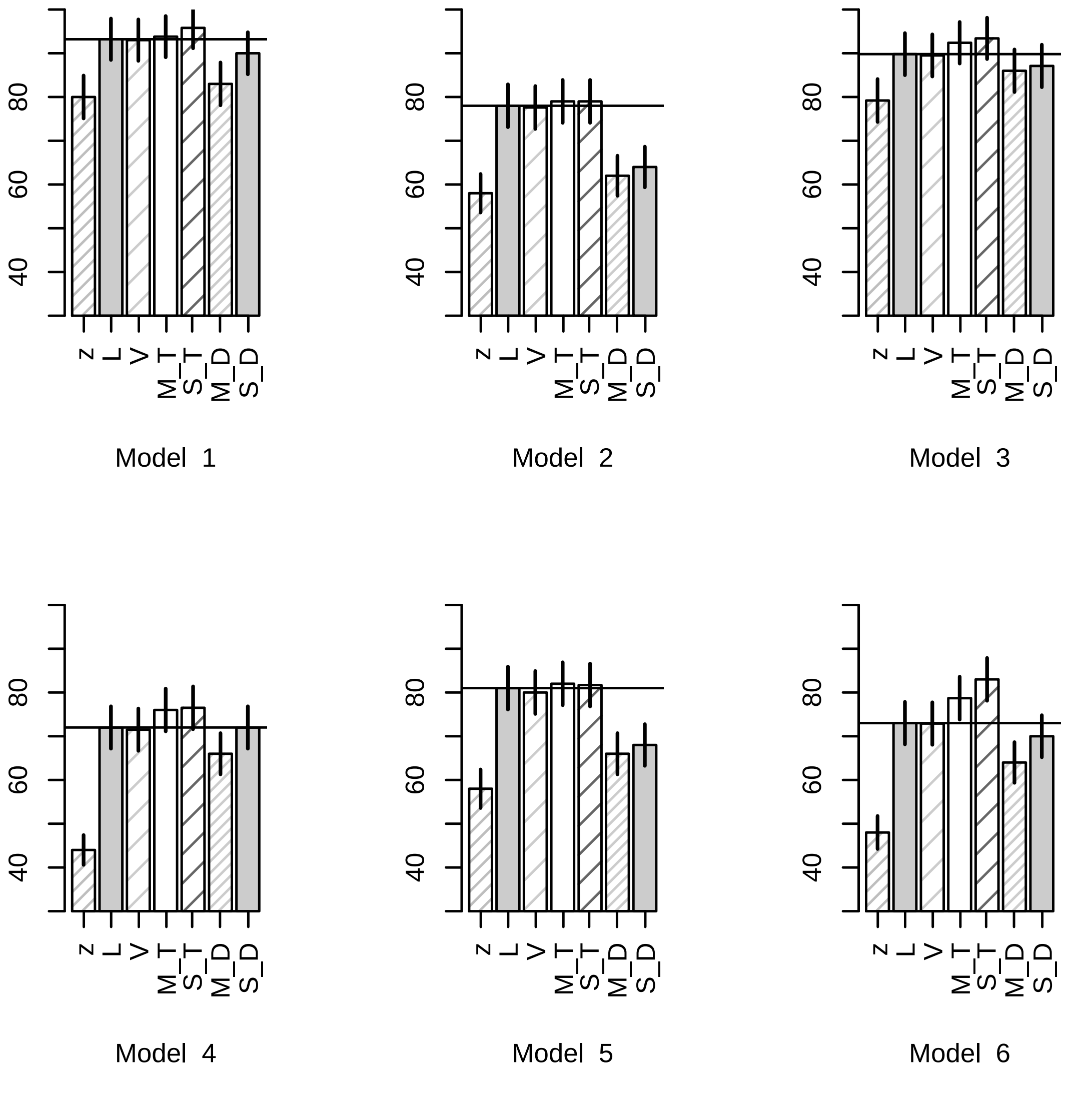


Figure 7: Power (95% CI) in +/-10 cM window from disease locus at the significance level of 0.01, for different disease models and small pedigree (SP), based on 400 replicates. z: z-RULS, L: L-RULS, V: V-RULS, M_T: MLS on true structure, S_T: S_{all} on true structure, M_D: MLS on discarded structure, S_D: S_{all} on discarded structure. Here 80% of the families are linked for models 1-5 and 100% are linked for model 6.

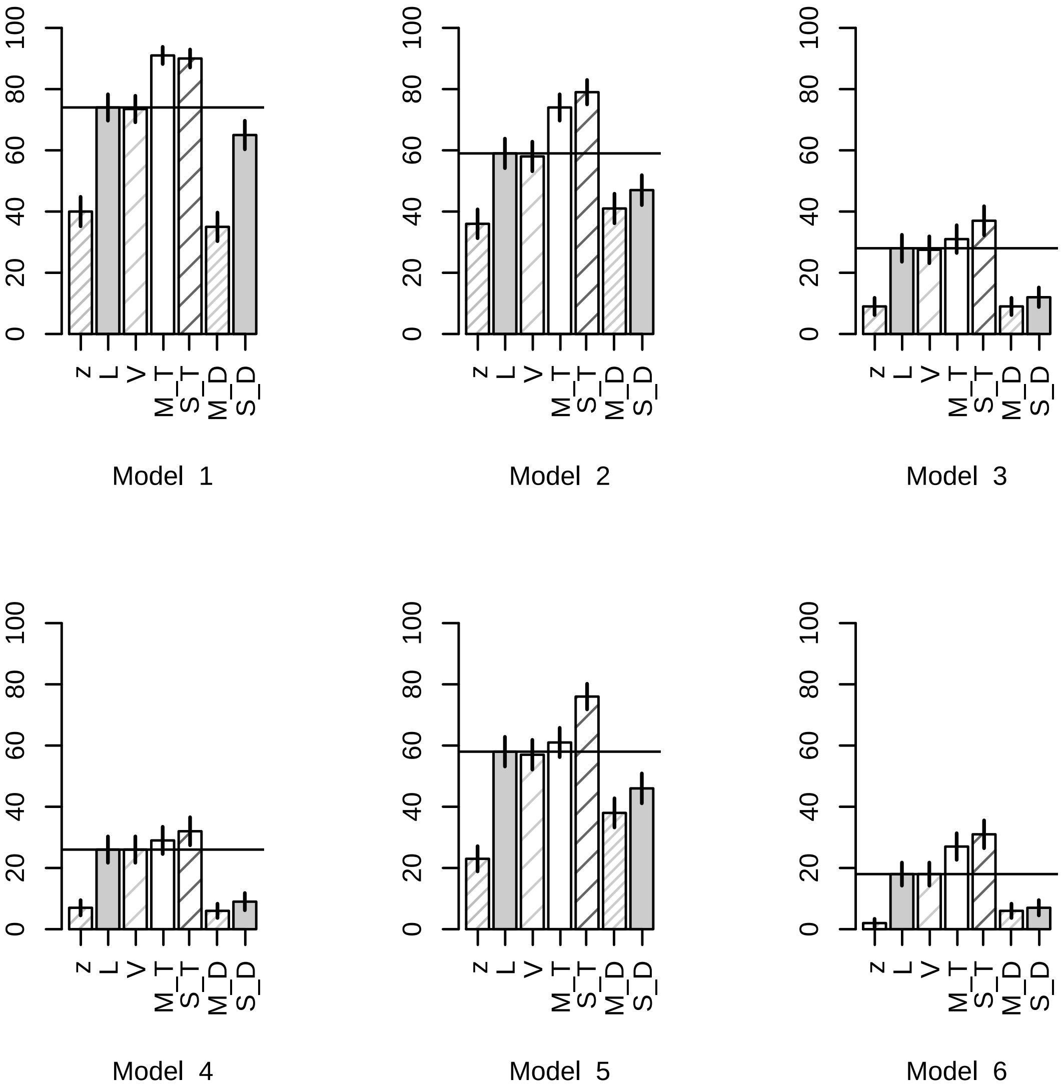


Figure 8: Power (95% CI) in +/-10 cM window from disease locus at the significance level of 0.01, for different disease models and large pedigree (LP), based on 400 replicates. z: z-RULS, L: L-RULS, V: V-RULS, M_T: MLS on true structure, S_T: S_{all} on true structure, M_D: MLS on discarded structure, S_D: S_{all} on discarded structure. Here 100% families are linked to disease for all models.

The power calculations for the LP dataset (Figure 8) indicate a similar power pattern as for the SP dataset (Figure 7), though the power of the RULS is not as close to the power of the MLS and S_{all} on the true structure as is the case for the SP dataset. For example, using dominant Model 1, L-RULS gave 94% power for SP dataset and 74% power for LP dataset, both being higher than MLS (83% for SP and 35% for LP) and S_{all} (90% for SP and 65% for LP) using the discarded structure. The power pattern for both pedigree structures at both levels of significance is:

$$\begin{array}{ccccc} \text{MLS and } S_{all} & \leq & \text{L-RULS and V-RULS} & \leq & \text{MLS and } S_{all} \\ \text{on discarded structure} & & \text{on apparent structure} & & \text{on true structure} \end{array}$$

The power of the z-RULS is dramatically lower than that of the other RULS because z-RULS estimates more parameters than the other RULS. The z-RULS does not perform better than MLS and S_{all} on discarded structure for both SP and LP datasets and hence it is not desirable to use the z-RULS. Also, as it is easier to interpret λ_i's for $i = 0, 1, 2$, L-RULS is preferred over V-RULS. Thus the above results suggest that ideally one should use L-RULS on the apparent structure in presence of relationship uncertainty in the data.

1.8 DISCUSSION

Here we developed several relationship uncertainty linkage statistics (RULS) that statistically model relationship uncertainty by properly weighing over the possible true relationships. We carried out a simulation study to assess the RULS and to compare them to the MLS and S_{all} nonparametric LOD score for both the true pedigree structure and discarded structure where individuals with erroneous relationships are discarded from the pedigree. In our simulation study we considered both small pedigree (SP) and large pedigree (LP) datasets under several disease models. The results showed that both L-RULS and V-RULS have power nearly as high as MLS and S_{all} on the true pedigree structure. Also, both the RULS have significantly higher power than MLS and S_{all} on the discarded structure. So, it is better to compute these RULS on the apparent pedigree structure than to discard erroneous individuals.

The weights $r^{T|A}$ estimated in both SP and LP consist of one 1 and other zeros, where 1 appears at the correct place, for example, an ARP truly related as half sib has estimated weight $r^{T=HS|A=HS} = 1$ and other $r^{T|A=HS} = 0$ for every other T's. This implies that true relationship structure of an ARP could be inferred quite precisely for both the datasets. But it is tougher to infer true structure of the entire family consisting of more than one ARP. In real data scenario, sometimes it might be difficult to construct the true family structure just by knowing true structure of individual ARPs. One might use Mendel v7.0 [32] to obtain true structure of a family, and we will try to apply this approach to construct an alternative structure for a family, as discussed in next section for future work.

The V-RULS, L-RULS and MLS have mixture chi-square distributions, and thus analytical thresholds are not given for these statistics in critical threshold table (see Table 5). The genome-wide empirical thresholds computed for z-RULS and S_{all} for both datasets and for the MLS for the SP dataset are reasonably close to the analytical thresholds obtained using the asymptotic distributions. The MLS for the LP dataset have empirical thresholds much higher than the analytical thresholds at both levels of significance. It is likely that the lack of independence of the affected pairs in the LP dataset might explain the higher empirical threshold for MLS. Also, for both SP and LP datasets, as estimated weights $r^{T|A}$ consist of only one 1 and rest zeros, V-RULS becomes of the same form as MLS [9]. But we have to remember that only an affected pair genotype data is used for the V-RULS computations, whereas genotype data for an entire family is considered for computing MLS. Thus the thresholds of V-RULS and MLS using true structure differ for both SP and LP datasets.

In this study, we find that proper modeling of relationship errors in linkage analysis gives substantially better power than the commonly-used approach of discarding erroneous individuals. Another possible solution for handling relationship errors is to construct the most likely alternative structure for the erroneous ARPs. This approach is similar to the statistically dubious approach of treating the most likely haplotype as the “certain” haplotype for an individual in association studies. Thus, this approach of using most likely alternative structure is less statistically appropriate than our approach, which, instead of arbitrarily choosing to use only the single most likely alternative structure, essentially considers all possible alternative structures through use of proper probabilistic weights.

We can compute the RULS on a real data set by following the steps shown in the flowchart Figure 5, though significance thresholds cannot be computed through simulation as the true relationship structures will not be known. Empirical thresholds corresponding to SP or LP dataset can be used if the data structure is similar to SP or LP respectively.

In this study, we considered as the set of true relationships only outbred relationships [47]. Many real studies have inbred relationships and failure to consider such inbred relationships in the RULS might give lower power for such studies. We plan to determine the performance of an extended RULS, which will include some inbred relationships in the set of the true relationships.

1.9 L-RULS APPLIED ON A REALISTIC SIMULATED DATA

Our L-RULS has been applied to a realistic simulated dataset that has a realistic proportion of families with relationship and genotyping errors. This application gave an idea of how L-RULS performed in comparison to applying S_{all} on the apparent, discarded and alternative structures.

A realistic data (it is a simulated dataset with realistic microsatellite marker allele frequencies, and has some relationship and genotyping error) will be now used as our apparent pedigree structure. In these data, 575 individuals from 218 families were typed for 373 microsatellite markers from autosomal chromosomes. Relationship testing by PREST [31] identified 5 families with error when testing at a 0.01 level of significance, *i.e.* 2.3% families have relationship errors. A conservative structure is constructed, where individuals with erroneous relationships are removed. We also changed, by hand, the structures of each of the 5 erroneous families to an alternative structure. This was done by inferring a possible alternative structure for a pair from the IBD sharing estimated in PREST. Multipoint S_{all} LOD scores were computed for all three structures: apparent, conservative and alternative structures. L-RULS is then computed on the apparent structure. There are 371 affected relative pairs in these data.

The results on two chromosomes 3 and 21 are shown in Figures 9 and 10. Chromosome 3 has significant linkage peaks with $S_{all} \geq 3.36$, marginal p-value≤ 0.00004 for all three structures (Figure 9, parts A, B, and C). Now this realistic dataset has similar apparent structure as that of our SP dataset. Thus, we used the empirical threshold we obtained for L-RULS in SP dataset as the threshold for L-RULS in this realistic dataset. See Table 5, where the empirical thresholds for L-RULS in the SP dataset at 0.05 and 0.01 level of significance are 3.50 and 4.30. This implies L-RULS=4.38 on chromosome 3 (see Figure 9D), is significant. Chromosome 21 has suggestive linkage peaks with $S_{all} \geq 2.15$, marginal p-value ≤ 0.0008 for all three structures (Figure 10A, B, and C). We again compare the L-RULS obtained in this dataset with empirical threshold obtained in SP dataset. This comparison showed that L-RULS=3.24 is suggestive and not significant at a 0.05 level of significance.

Though using empirical thresholds obtained in SP dataset might not be the most accurate one to use, it gives a close approximation to the threshold, as the apparent structures are similar between SP and this realistic dataset. This application shows that: first, removing or correcting relationship errors gave higher LOD scores; second, L-RULS identified linkage peaks at the same location as S_{all}; and third, L-RULS gave similar conclusions about linkage peak as that obtained from S_{all}.

1.10 FUTURE WORK

Here we discuss the possible future work that we would like to explore in applying RULS. Due to poor performance of z-RULS in this study, we won't further explore behavior of z-RULS. L-RULS and V-RULS are same when $r^{T|A}$ contains one 1 and rest zeros, but one might encounter situations where these weights differ from this condition. We will compute only L-RULS for the following future work. Firstly, we will apply RULS on a real data. We have already tried to implement a part of RULS computations on Otitis Media with Effusion data from Caucasian families (see chapter 2, section 2.3.7). Secondly, we will consider computing RULS on the alternative structure. Though it is hard to automate a code that can select a best possible alternative structure, it would be interesting to compare performance of MLS, S_{all} on alternative structure from a simulation study, with those on true and discarded

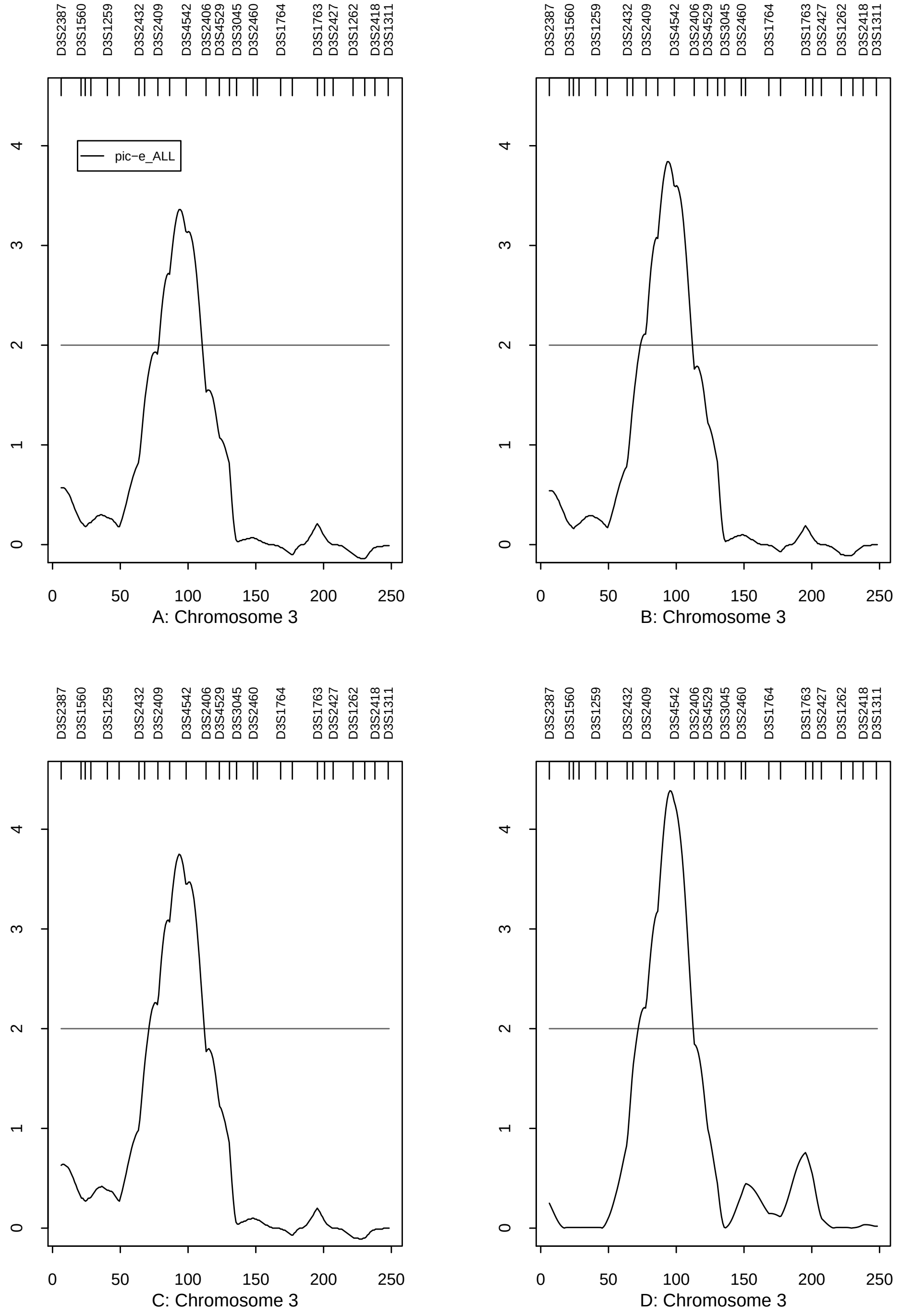


Figure 9: Chromosome 3 linkage peaks for A. Sall=3.36 on apparent structure, B. Sall=3.84 on conservative structure, C. Sall=3.75 on alternative structure, and D. L-RULS=4.38 on apparent structure.

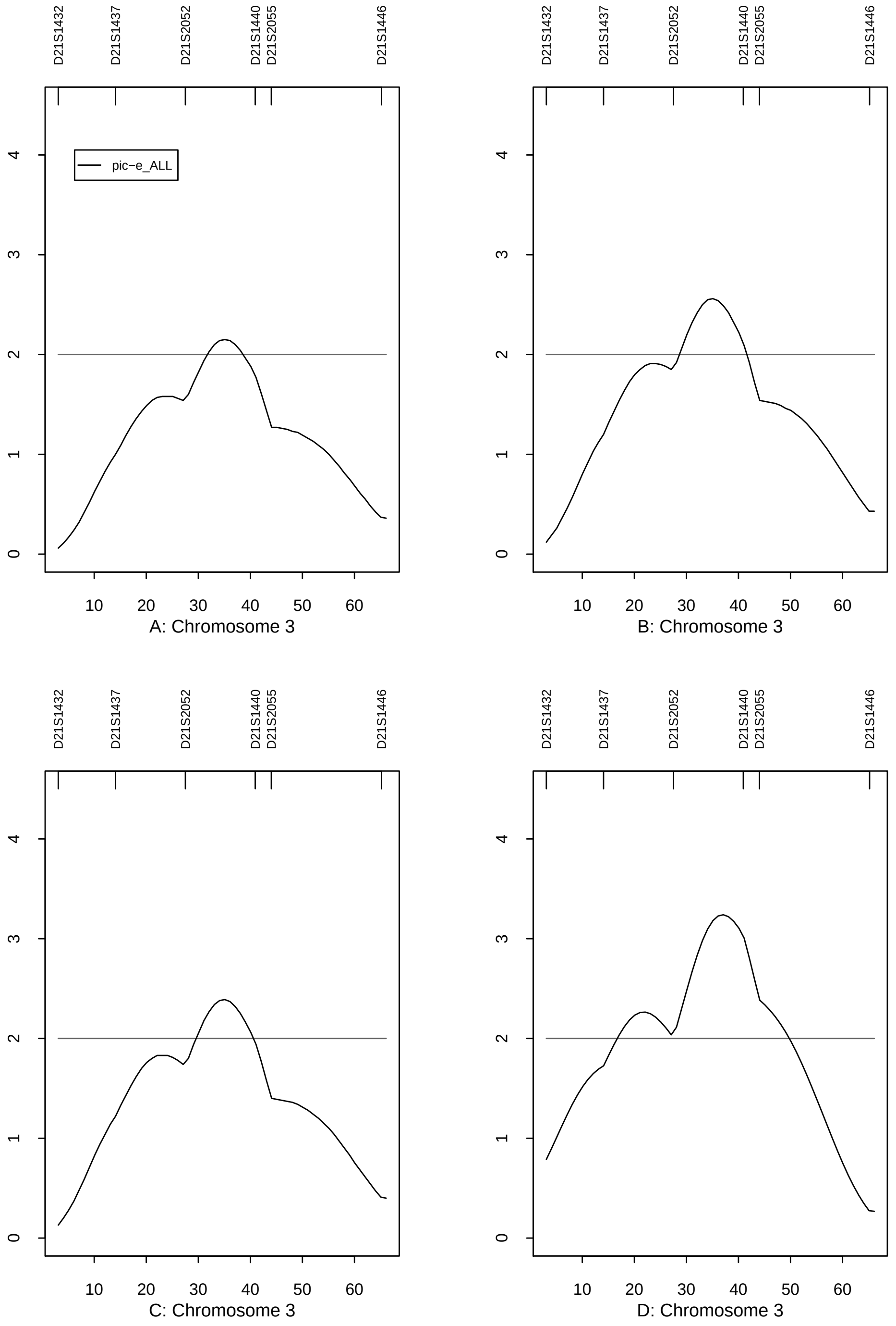


Figure 10: Chromosome 21 linkage peaks for A. Sall=2.15 on apparent structure, B. Sall=2.56 on conservative structure, C. Sall=2.39 on alternative structure, and D. L-RULS=3.24 on apparent structure.

structures and RULS on apparent structure that have already been done. Also, in a real data scenario, as we won't know the true structure, it might be fruitful to use empirical p-values of RULS on the most likely alternative structure. Thirdly, we want to incorporate inbred relationships in extended RULS, as it might reflect the real data scenario better. Fourthly, error in genotype data will be allowed to assess the performance of RULS.

1.10.1 Comparison to Alternative structure

In this study, our RULS using the observed apparent pedigree structures are compared to MLS [9] and S_{all} [21] using both true and discarded structures. As a part of our future work, we will perform similar comparison to MLS and S_{all} using the most likely alternative structure for each pedigree. It seemed that replacing an apparent structure with an alternative structure might not be the best statistically sensible solution to overcome relationship error. But we found out that for an ARP, estimated weight $r^{T|A}$ is close to one 1 and rest zeros, with the 1 appearing at the right place. We have checked that for both SP and LP datasets, the 1 in $r^{T|A}$ appears at the correct place where A matches with the underlying T. Now, the true relationship for which $r^{T|A}$ is close to 1, can be taken as the alternative structure for the ARP. One has to note that having an alternative structure for a pair is not the same as constructing the alternative structure for the whole pedigree. Either we have construct the alternative structure for the family using alternative structure of each ARP manually, or we can apply Mendel v7.0 with certain options to get the alternative structure of the entire family. Now, in reality, sometimes it might be hard to manually reconstruct alternative structure of the family using that of the pairs. Also, it would be tough to automate the process of constructing alternative structure of family by using alternative structure of pairs. Thus to perform power study for L-RULS and MLS on alternative structure, we will use Mendel v7.0. Thus, considering an alternative structure might be a solution to relationship error. Hence, in future, we will compare RULS on apparent structure to MLS and S_{all} on alternative structure.

We will consider the same apparent structure A as we have used to compute RULS. As before, modified PREST code and SEARCH will be applied to obtain the estimated weights for an ARP. Now, once $r^{T|A}$ is computed for an ARP, the relationship for which this estimated weight is close to 1, will be noted. Then the structure of an ARP will be changed to that specific alternative structure, and this will be coded in R software. Once we get the apparent pedigree structure for all ARPs uniquely identified for a family, we will manually construct the alternative structure for the family. MLS and S_{all} will be computed on the apparent structure. We can also use ALTERTEST from PREST package [31] to obtain alternative structure for a relative pair. This process might be tough in certain realistic situations, and also to automate this method. Mendel v7.0 [32] analysis option 9 with model 2 (that tests for all relationship types) can be used to obtain alternative structure for an entire family. A simulation study on SP and LP datasets will be done to evaluate power of MLS and S_{all} on apparent structure.

Also, it would be interesting to construct an alternative structure for a real dataset. As we only have apparent structure, we won't know the true structure for a real data, and hence, we will use the best possible alternative structure as a surrogate for the true structure. ARPs can be part of a very large pedigree, and in that case, using our above mentioned approach of constructing alternative seeing $r^{T|A}$, or using ALTERTEST can give the alternative structure only for the particular ARP, and does not construct the entire alternative pedigree structure. Hence, we will use Mendel v7.0 analysis option 9, model 2 to construct alternative structure. One has to give a prior probability to all possible alternative structures, and a reasonable prior reflecting the relationship types in the population, will be provided to Mendel. We will then simulate marker data for the alternative structure, change to apparent structure, and compute RULS on the apparent. A simulation study will be done implementing the above algorithm to get p-values for RULS on the real data. This way we can use an empirical p-value for RULS.

1.10.2 Incorporating Inbred relationships

As discussed in section 1.4.1, till now we have considered only outbred relationships as possible true underlying relationships to weight the relationship uncertainties in RULS. In a real data scenario, inbreeding is possible, and failure to consider such inbred relationships in

RULS might lead to a false positive linkage signals. Hence we will incorporate an inbred relationship in the set of true relationships considered for RULS. Performance of this extended RULS by incorporating inbred relationship will then be assessed by computing genome-wide threshold and power through a simulation study, and then compared with our L-RULS, and MLS, S_{all}, as before.

To incorporate inbred relationships to RULS, we will include inbred full sibs whose parents are first cousins in the set of possible true relationships to derive our extended RULS. First cousins marriage seems to be the most common consanguinity marriages over various populations, and so, this inbred relationship might be a realistic one to include in our set. As discussed in section 1.3.5, an inbred relationship falls in the impossible region of the relationship triangle [47] given for noninbreds. We will also try to plot this inbred relationship in the triangle using prest.R code (written by Daniel E. Weeks) in PREST, to inspect coverage of our new set of relationships to be considered in extended RULS.

First the pedigrees with true structure as given in Figure 11 will be simulated. 40 large families of which 10 have inbreeding with family (structure II), and rest 30 having non inbred relationships, will be simulated for true structure. We will take the same marker data, and the same location to simulate the disease locus with Model 1 (Table 4), as in SP or LP datasets in our original study. Second, an apparent structure will be obtained by randomly moving an individual to another sibship in the terminal generation based on an assumed error rate, and also removing the inbreeding loop in some families. For structure I, we take the same error rate same as that of structure I in LP dataset, and is given by probabilities as shown in Table 3. For structure II, in 5 families, we take the same error rate as in structure II in LP dataset, *i.e.* an individual will be randomly moved from sibship d to e and vice versa with probability 0.2. In the remaining 5 families we would remove the inbreeding loop by adding a dummy parent. To create the discarded structure, we will remove those individuals who are known to be erroneous, *i.e.* those individuals who are moved.

Thirdly, to incorporate inbreeding in RULS, we will take inbred full sibs whose parents are first cousins as the inbred relationship that is to be included in the set of true relationships. So, for the extended RULS, the set of possible true relationships will now be consist of six relationships, *i.e.* T∈(FS, HS, FC, U, PO, InFS), where InFS stands for the inbred full sibs.

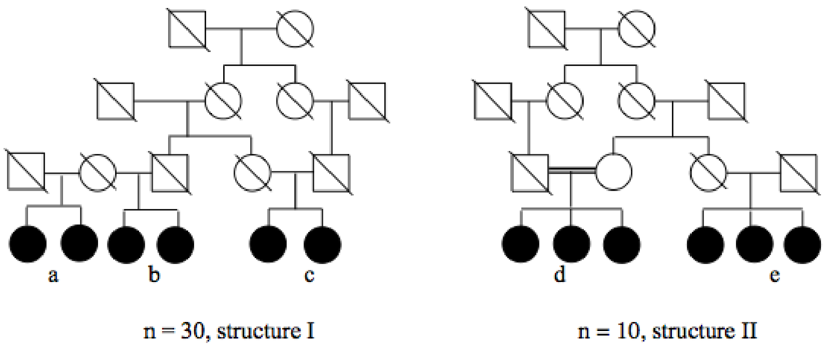


Figure 11: True structure of inbred dataset for extended RULS, where n is the number of families with the given structure. Circles and squares denote females and males respectively, the blackened symbols indicate affected individuals, the clear symbols with a slash denote deceased individuals (who are neither phenotyped or genotyped).

As discussed before, only L-RULS will be extended. Now we will discuss how to obtain the extended L-RULS by following through the steps of flowchart Figure 5. For each ARP, we will compute $r^{T|A}$ by calculating $P(G|T)$ from Merlin [1] software with *- -likelihood* option, as Merlin can handle inbreeding within a family. Then we will compute F^A, and change the structure to those six possible true structures. In order to give an ARP the inbred FS structure, one has to put the same dummy parents for both the parents of the ARP, *i.e.* parents of ARP are full sibs. We will then compute $P(i|G,T)$ from Merlin software with *- -ibd* option at each grid of positions over the genome. These steps will help to calculate f^A as given in section 1.5. Finally, as shown in the flowchart, we will compute the extended L-RULS at each grid of positions. A simulation study will be performed to assess the extended RULS.

1.10.3 Impact of genotyping error

We will explore the impact of genotyping error on RULS as a part of our future work. In this study, both the datasets SP and LP are simulated free of any genotyping error, and thus these datasets might not reflect a real scenario. Also, Douglas et al. [11] showed that even small genotyping errors might substantially lower the power to detect linkage using ASPs. Hence, we will implement a moderate genotype error rate of 1% [14] while simulating the true pedigree structure, compute our RULS and then compare performance of RULS in both scenarios, one without genotyping error and another, genotyping error present in data.

We will consider the same true pedigree structure as in SP dataset (Figure 6A). The unlinked markers will be simulated by Simulate [45]. We will simulate disease locus at the same position as taken for SP dataset, using Allegro [17], and with disease model as Model 1 (Table 4). Microsatellite marker data will be simulated with 1% genotyping error rate with uniform error model, using the genotyping error simulation module in Mega2 [29]. Once we construct the final pedigree file, we will follow the steps given in flowchart Figure 5 to get the RULS. To explore the impact of genotyping error on RULS, we have to compute RULS on the same pedigree structure without genotyping error. Another dataset will be simulated with the same pedigree structure, and same unlinked marker data. This time, we do not generate any genotyping error, *i.e.* *err* option will be kept at 0. RULS will be computed for this dataset without genotyping error. We will then compare RULS in both scenarios, in the presence and absence of genotyping error. Similar to the previous two proposed works, here also, we will perform simulation study for each scenario, and compare RULS to detect the impact of genotyping error on the performance of RULS.

2.0 GENOME-WIDE LINKAGE SCAN FOR OTITIS MEDIA WITH EFFUSION AMONG THE CAUCASIANS

2.1 INTRODUCTION

Otitis Media (OM) is a middle ear infection and is the most common cause of hearing loss among young children. Children suffering from Otitis Media with effusion (OME) have fluid in the middle ear that causes mild hearing loss without any other symptoms. OME is caused when the auditary tube becomes blocked, and as for children, this tube being more horizontal than that of adults, OME is more common among children. Tympanostomy tube placement is recommended if fluid is still present after 4-6 months. In this study, subjects with history of tympanostomy tube insertion, along with their families (affected or unaffected siblings and parents) are recruited at University of Pittsburgh Medical Center, and 1976 from 500 families have been enrolled in this study. As this is an ongoing study, 1317 people are genotyped till date, using Affymetrix 10K SNP chip technology. Though the enrolled families included both Caucasian and African-American or biracial families, since the number of Caucasian families recruited is much higher, we perform linkage analysis using only the Caucasian families. We check for relationship and genotype error prior to linkage analysis, and 1235 samples are then used for linkage analysis. We perform nonparametric multipoint linkage analysis using Merlin software [1], and observe exponential S_{all} LOD score [21].

2.2 BACKGROUND

Otitis media (OM) is the most common among children under 15 years of age, with almost 25 million affected in 1990 [40]. Otitis Media with effusion (OME) is the most common cause of mild hearing loss among children, and OME is caused by fluid in middle ear. As OME is asymptomatic, affection status is determined by history of tympanostomy tube insertion. This study is conducted using subjects from Pittsburgh, and are enrolled by University of Pittsburgh Medical Center. In Pittsburgh, 6.5% of African American and 5% Caucasian infants of two years of age have tympanostomy tube insertion [6]. There is also a study using nationwide representative sample of US children that showed by 3 years of age, 6.8% had tympanostomy tubes inserted [22].

There are previous studies regarding some important risk factors related to Otitis Media with effusion. There are studies showing environmental risk factors that affect the incidence of the disease. Also several studies have shown that there is association between allergy and OME. Some literature also suggests that children with a variant of gene producing high level of gamma interferon might not develop ear infection after getting a cold like other children (Web resources [C]). Thus the risk factors of OME include characteristics of a subject (age, sex, race, allergy, cleft palate/craniofacial abnormalities and genetics), and also environment (day care, school, passive smoking, socio-economic status etc.).

There are also several studies indicating the role of genetics in OME. Firstly, though several studies showed there is difference of OME incidence between population, two recent studies [6],[34] found that incidence of OME did not differ among Caucasian and African-American children. In addition, a positive family history gives an increased risk to OME, and this shows familial aggregation of OME [38]. Thirdly, there are several studies that indicate siblings are at higher risk of getting OME. One study by Rasmussen [35] showed that incidence of OME among children with affected siblings is four times higher than other children. Fourthly, the most convincing evidence of genetic component in susceptibility to OME is obtained from twin study results. A preliminary twin study showed an estimated heritability of .73, indicating that OME has a strong genetic component. The twin and triplet cohort study by Casselbrant et al. [27] showed that there is strong genetic component to the amount of time with middle ear effusion in children. Another longitudinal twin study by Rovers et al. [38] found that heritability is .71 at the age of 4 years with even higher heritability at earlier ages. Studies show that heritability estimates of OME are in the range of common diseases whose susceptibility genes have been mapped using affected sib pair linkage analysis. All these evidence imply that there is significant genetic role in incidence of OME, and though there might be individual risk factors, genetics might have a higher contribution to the susceptibility of OME.

Though there is evidence in literature that OME might be a genetic disease, there has been only one genome screen with affected relative pairs. Daly et al. [10] studied 133 families recruited by the University of Minnesota Otitis Media center, and 591 samples were genotyped at 404 microsatellite markers. Their genome-wide linkage analysis showed evidence of linkage in the chromosomal regions of 10q and 19q with multipoint nonparametric LOD scores of 2.61 and 2.53 respectively.

2.3 MATERIALS AND METHODS

2.3.1 Family structure and affection status

Subjects with a history of tympanostomy tube insertion, along with their affected and unaffected siblings and one or both parent(s) were recruited. An individual is considered affected if they had tympanostomy tubes inserted. The total number of individuals enrolled for this study is 1976 and total number of families is 545, out of which 670 are Caucasian families and 39 African American or biracial families. As the number of African American or biracial families is low, we used only Caucasian families for the statistical analyses. There are families with one parent present, Table 6 shows the number of families with number of typed parents.

Table 6: Pedigree information. First two rows give the number of families with either one parent missing or have both parents typed. The next two rows give the number of families according to the number of siblings in the family.

Number of parent	1	2
Number of families	84	234

Number of siblings	1	2	3	4	5	6
Number of families	26	215	60	11	5	1

2.3.2 Genotyping

Genotyping was done using Affymetrix 10K SNP chip technology. Two versions of the 10K Affymetrix SNP panels were used, the older version had 11,560 SNPs, while the newer version had 10,204 SNPs without the rs numbers. We combined two versions in order to use maximum number of common markers available for genotyped individuals. The total number of SNPs in the combined data is 11,093 with average spacing (over all chromosomes except 3, 13, 16 and 22) of less than 0.5 cM, and with much higher spacing at the end of the mentioned four chromosomes. Notice that we should have obtained genotype data for at least 11,560 SNPs, and not 11,093 SNPs in the combined data. But since we could not obtain physical positions for more than 467 SNPs, we had to remove those SNPs. Out of 1,976 enrolled individuals, 1,317 were genotyped. Population specific differences in the incidence of OM have been reported, and in our study, among the 704 genotyped affected individuals, 670 are from Caucasian families and 34 are from African American or biracial families.

2.3.3 New map

We have developed a more accurate SNP genetic map for linkage analysis. To estimate the genetic position of our SNPs, we used interpolation between the physical position of the SNPs from dbSNP (Build 135) and that from the Rutgers Combined Linkage Map [23] (also Build 135 version). The later map combines genotype data from both the CEPH and deCODE pedigrees, and provides genetic map with more than 28,000 SNPs. In this new map, we have 11046 SNPs with average spacing of .5 cM. As we believe this new combined map is better for linkage analysis as genetic map distances are more accurate, we will perform linkage analysis using this interpolated map.

2.3.4 Relationship and Genotyping error checking

The first approach towards data analysis was to identify any relationship error present in the data. We did two rounds of relationship error checking: each time we performed relationship testing using PREST [31] and then checked the erroneous pairs with the clinical lab. In both rounds of error checking, we inferred from the relationship testing based on genome-wide marker data and the apparent relationships that several pairs have significantly different relationship than their apparent relationship. We checked both the p-values for IBS test and the estimated identity by descent (IBD) probabilities to detect the pairs with erroneous relationships. Due to multiple testing issues, we used a stringent p-value of 0.001.

2.3.4.1 Within pedigree relationship error check We started with 2,001 genotyped relative pairs and relationship testing in the first round detected 28 pairs with erroneous relationships and we inferred on their true relationships: 6 full sib pairs inferred as MZ twins, 3 full sibs as half sibs, 1 full sib as unrelated, 16 parent-offspring pairs as unrelated, and 2 parent-offspring pairs as MZ twins. These erroneous pairs were referred back to the lab, and after checking these pairs, they gave us the updated information on the pedigree data. In the second round of relationship testing with 1,976 pairs, we again found 63 pairs from 22 families with significant relationship errors. Second round of error checking showed more families with errors that first round, because in the first round we checked error in full sibs and parent-offspring relationships, unlike all relationships in second round. As before, from the p-values and estimated IBD probabilities, we inferred the possible true relationships for the erroneous pairs: 20 full sibs inferred as half sibs, 1 full sib as unrelated, 13 parent-offspring pairs as unrelated, etc. So, 10 families out of these 22 families suggested changing the full sib structure to half sib structure, and 9 families showed an apparent paternity problem. This time also, we informed the lab with the list of families that showed relationship errors. They confirmed the errors for majority of those families. As we could not confirm the inferred structure for all the families at this stage, after completing the across family error checks, we constructed a conservative structure by discarding the minimal number of individuals that eliminate the majority of problematic relationships.

2.3.4.2 Between pedigree relationship error check As PREST performs relationship testing for pairs within a family, we also checked for across family relationship errors by using RELPAIR [14] on our data. We have used only 5,007 polymorphic SNPs in RELPAIR to identify whether there is any relationship error across families, *i.e.* the families are connected. The results showed that 2 sets of two families are connected and the lab confirmed that one set has duplicated families and another has sample mix-ups. One of the duplicated families was removed and another with no relationship error within the family was used for the analysis. So after we went through all the steps of relationship error checking, we detected 313 families out of 322 total families, and 295 out of 305 Caucasian families are free of any relationship errors. Individuals with erroneous relationships with the rest of the individuals in their family were removed from the pedigree to prepare the conservative structure. The conservative dataset has 1,307 individuals from 313 families. As the number of Caucasian families genotyped is much more than African-American families, we henceforth will analyze only the Caucasian families. We prepared the conservative dataset with the Caucasian families, and it has 1,235 individuals from 295 families.

2.3.5 Genotype error check

We checked this conservative structure for genotyping error using PedCheck (O'Connell and Weeks 1998), and an entire family is zeroed out at each marker showing Mendelian inconsistency for that family. Out of total 11,093 SNPs, 7,426 had to be set missing for several families, maximum being for SNP rs6509245 in 53 families.

2.3.6 Nonparametric Linkage analysis

Multipoint nonparametric linkage analysis was performed using Merlin software (Abecasis et al. 2002). The conservative structure for Caucasian families was used for this analysis. The S_{all} LOD score is computed at each marker and also at a position in between two markers. The SNPs along with the LOD score and marginal p-value were noted for the regions with a maximum LOD score on a chromosome.

2.3.7 Applying RULS on OME

Our Otitis Media with effusion data for Caucasian families indicate 10 families out of 305 families, *i.e.* 3.3% of the families have relationship errors. We have removed erroneous individuals with relationship errors from Caucasian data to perform linkage analysis on the conservative dataset. As discussed in chapter 1, using RULS on the given structure might be more powerful than using the conservative structure. Hence, we have already tried to implement RULS on our OME data from the original 305 Caucasian families with relationship errors, but only could proceed to get the estimates of the weights of each possible true underlying relationship. Please refer to the notations given in chapter 1 for the following explanation. We have encountered some programming problems in RULS while computing $P(G|T)$ to estimate $P(T|A)$. Individuals are typed for 10K SNPs in this OME dataset, and to get $P(G|T)$ for each affected relative pair, probability terms for each marker are multiplied, leading to an underflow problem (cumulative probabilities over all markers lower than $4.940656e^{-324}$ are set to 0). This issue of a severe underflow problem was not considered previously in RULS, because we validated RULS code using microsatellite markers, and they being much fewer in number, did not cause serious underflow problem like SNP data. We have implemented a check for underflow, and added a positive number to $\ln P(G|T)$ for each ARP, to finally obtain $P(G|T)$. Since this term $P(G|T)$ occurs both in the numerator and denominator of f^A for each ARP, there is no need to adjust for the added constant to overcome the underflow problem. Now, we will continue from here to compute RULS on those 305 Caucasian families.

2.4 RESULTS

Nonparametric linkage analysis identified several regions in the genome having suggestive linkage, out of which regions on six chromosomes have LOD scores ≥ 1.5 (marginal p-value ≤ 0.002). The maximum LOD score of 2.36 (p-value 0.0005) is observed at rs1345938 on chromosome 7. There are other regions on this chromosome that showed LOD scores higher than 1.5. SNP rs2812415 on chromosome 10 is another region showing LOD>2. This marker has LOD 2.06, p-value 0.001. The regions on other four chromosomes with the suggestive evidence of linkage are at rs924266 (LOD=1.67) on chromosome 1, rs725395 (LOD=1.76) on chromosome 2, rs2133507 (LOD=1.79) on chromosome 4, and rs958653 (LOD=1.65) on chromosome 8.

2.5 FUTURE WORK

Here we discuss several directions to be explored in future. Firstly, as in this OME study, we are using SNPs to identify susceptibility loci for linkage, we need to assess the quality of SNPs to be used in the analysis. Secondly, we will validate the interpolated map we have developed, and then use it for linkage analysis. Thirdly, we will perform association analysis to identify risk alleles to OME. The literature of OME suggests that OME is a common disease and hence doing association analysis might be fruitful.

To assess the quality of SNPs, we have to perform several steps of quality control and filter out the low quality SNPs from further analysis. Several recent study performing genome wide association assess quality of SNPs. Also, one study by Hinds et al. [18], examining common genetic variation pattern between three populations, explained their steps of quality assessment done to select a better set of SNPs. We will follow the similar screening to obtain a higher quality of SNPs, and will reject data for SNPs that perform poorly in this quality control steps. Firstly, we will remove SNPs with high rate of missing genotypes. Percentage of SNP with amount of missing genotypes, and overall frequency of successful genotype calls will be recorded, and SNPs with significant missing genotypes will be removed. Secondly,

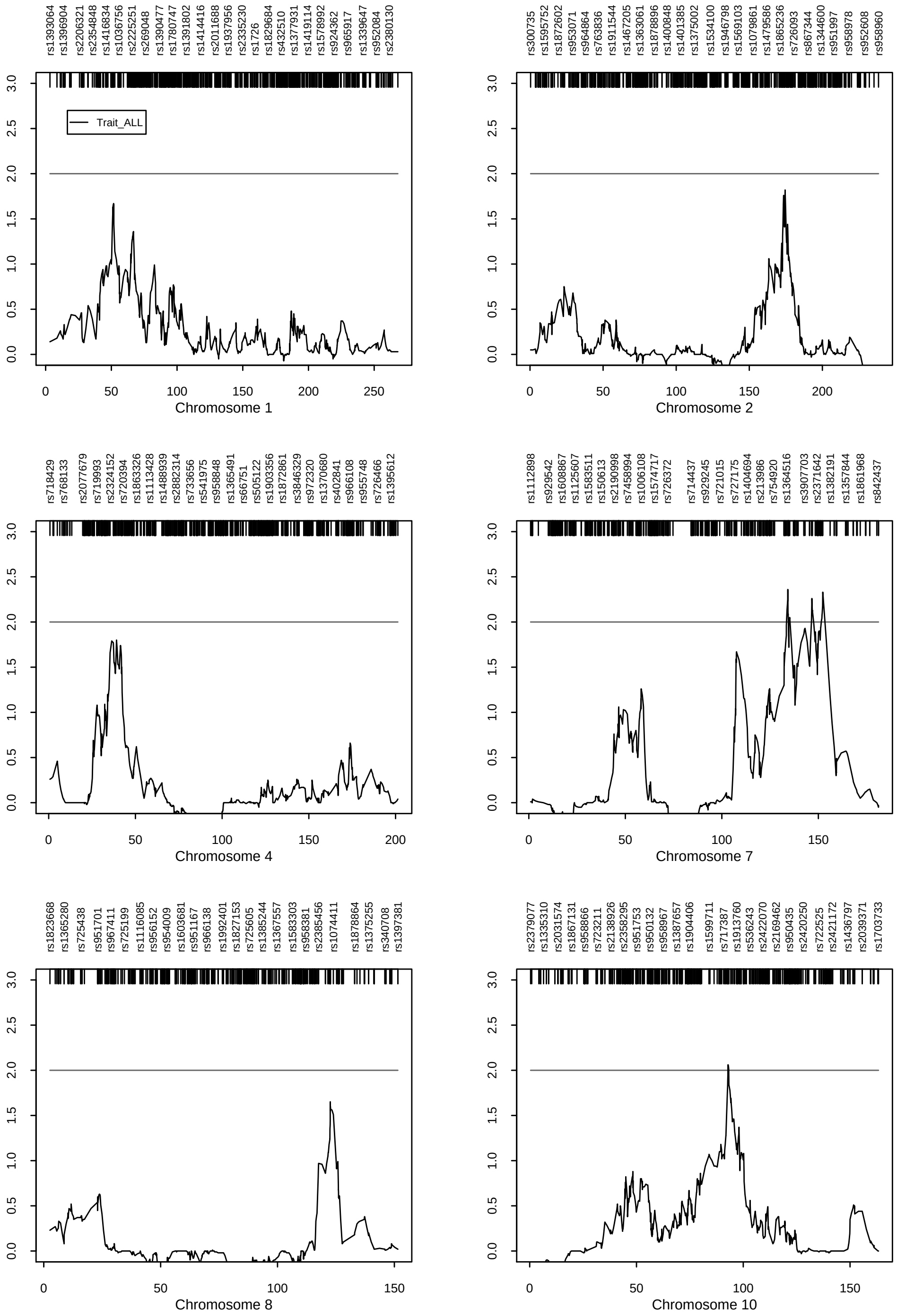


Figure 12: Linkage analysis of OME on conservative structure for Caucasian families. Chromosomes 1, 2, 4, 7, 8 and 10 show multipoint S_{all} LOD score ≥ 1.5.

SNPs deviating from Hardy-Weinberg equilibrium will be discarded. For linkage purpose, removing SNPs significantly out of Hardy-Weinberg might give a better set of error free genotypes that are consistent with the equilibrium. Thirdly, we then want to compare allele frequency of SNPs, as estimated from our data using only Caucasian families, with that of the same set of SNPs from the HapMap project and Affymetrix Caucasian data. Similar to some studies, we can submit the SNP assay details to National Center for Biotechnology Information (NCBI)'s SNP database. From these, we might get an idea how our SNPs frequencies agree with either of the mentioned databases. Once we get the better set of SNPs with higher percentage of non-missing genotypes, we will then take only Caucasian family data for the following steps. Our fourth step is to first select SNPs with minor allele frequency ≥ 0.05 in Caucasian data, and then analyze for linkage disequilibrium between SNPs. As SNPs are densely located, there might be high linkage disequilibrium between them, hence, we will also select tag SNPs with a high correlation coefficient with every SNP in a strong LD block. This approach might give a smaller set of SNPs (tag SNPs) to be used for association analysis, and at the same time this set should have approximately the same power for detecting any disease association as the entire set. SNP-SNP linkage disequilibrium between close proximity SNPs also influences linkage signals, tag SNPs will be used for linkage analysis. We will identify tagging SNPs by Haploview [2]. Genotype data for founders, and physical location of SNPs are the inputs in Haploview, and we will take squared correlation coefficient of $r^2 \geq 0.8$ to identify tag SNPs. Using the Tagger option, we will obtain the tag SNPs in each haplotype block.

We will validate and use for linkage analysis, the interpolated map that we have developed as a part of this study. Initial comparisons with several sources like dbSNP, Map-O-Mat and Devlin et al.'s interpolated map (all Build 135) showed some inconsistencies with our interpolated map. The interpolated map is constructed using physical location of dbSNP markers that are common with our SNPs from OME study, and the genetic map positions from Rutgers Map-O-Mat SNPs (see the interpolated map section). As we have obtained the interpolated map, we want to cross check it with several other maps of same build, so that we can be assured of the map quality we use in linkage analysis. Map quality is important for multipoint linkage analysis as the LOD scores from multipoint analysis is sensitive to intermarker map distances. We will compare the intermarker genetic map distance, and overall map sequence for each chromosome. Then the new interpolated map will be used to perform linkage analysis.

We will perform association analysis to identify risk alleles for OME. There are two approaches of doing that, either we will test several SNPs in close proximity with SNPs showing significant linkage signals, or we will perform genome wide association analysis involving all SNPs, rather ideally the tag SNPs. There are again several approaches of doing association analysis using family data. We will perform TDT type association analysis, as this approach is not influenced by population substructure. In our data, we will check for population substructure within the Caucasian families by STRUCTURE [48]. As in this study we are not taking trios, *i.e.* parents and an affected offspring, the simple TDT statistic [44] won't be applicable. This OME study involves nuclear families with parents (either typed or missing) and their offspring (affected and affected siblings). Thus family based association test [28] (FBAT) or pedigree disequilibrium test [25] (PDT) will be more applicable. Unlike FBAT which uses only nuclear families, PDT is a more general approach of association test which can use general pedigrees and is not restricted to only nuclear families. Now, in this OME study, we consider only nuclear families with typed or missing parents and affected/unaffected siblings, it would be appropriate to use FBAT itself. We will analyze the SNPs by FBAT software (see web resources) with analysis option *fbat [-m] [marker(s)]*, and we will not consider any covariate in our analysis. For the first approach of analyzing SNPs in the region of significant linkage, we will use an added analysis option *[-e]* as this tests for association in presence of linkage.

2.6 ACKNOWLEDGEMENT

This work was supported by the University of Pittsburgh and NIH grants D43TW006180, and R01DC005630.

2.7 WEB RESOURCES

A. http://www.ncbi.nlm.nih.gov/projects/SNP/

B. http://www.affymetrix.com/index.affx

C. http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=hstat6.chapter.23362

D. http://www.sph.umich.edu/csg/abecasis/Merlin/download

E. http://galton.uchicago.edu/ mcpeek/software/prest/download.html

F. http://www.biostat.harvard.edu/ fbat/default.html

G. http://compgen.rutgers.edu/mapomat/

H. http://watson.hgen.pitt.edu/register/

3.0 APPENDIX A: SOFTWARE DOCUMENTATION FOR RULS

The RULS.tgz file contains code to compute the RULS on affected relative pair (ARP) linkage data, as proposed in ”Relationship uncertainty linkage statistics (RULS): Affected relative pair linkage statistics that model relationship uncertainty” by Amrita Ray and Daniel E. Weeks.

3.1 REQUIREMENTS FOR RULS

1. Unix Operating System
2. cc, C compiler
3. g77, Fortran compiler
4. gpc, Pascal compiler
5. R (http://www.r-project.org/)
6. MEGA2 (http://watson.hgen.pitt.edu/register/)
7. SIMULATE (http://www.genemapping.cn/simulate.htm)
8. Allegro v1.2c (http://www.decode.com/software/allegro)
9. SEARCH (http://www.biomath.ucla.edu/faculty/klange/register.html)
10. threelocfull.tar.gz (http://www.staff.ncl.ac.uk/heather.cordell/threeloc.html)
11. MERLIN (http://www.sph.umich.edu/csg/abecasis/Merlin/download)
12. GENEHUNTER (http://www.broad.mit.edu/ftp/distribution/software/genehunter/)

3.2 INSTALL INSTRUCTIONS

Download RULS.tgz from http://watson.hgen.pitt.edu/register/ and untar it by typing tar zxf RULS.tgz, this should create a RULS folder containing a copy of this documentation (README.txt) and two folders, example/ and real/.

The example/ folder contains sim.sh and code.tgz, and the real/ folder contains ruls.sh and code1.tgz.

Download and install items 1-8 and 11-12 of the above list (in Requirements for RULS) to /usr/local/bin. For items 9, and 10, do as follows: After obtaining SEARCH and threelocfull.tar.gz from their respective websites, then

A. Copy SEARCH.FOR into the RULS folder.

B. Copy threelocfull.tar.gz into the RULS folder, untar it by typing

gunzip threelocfull.tar.gz
tar xvf threelocfull.tar,
and go to folder THREELOCFULL/ and change onelocarp.f by inserting a new line after line 707: write(10,36) vadd(1), vdom(1)) so that it prints V_A and V_D.

3.3 INPUT FILES

Three input files required to compute RULS and other statistics, MLS and S_{all}, are: Pedigree file containing pedigree information, Locus file with affection status and marker information, and Map file with marker position information. Details behind each input file is discussed now.

3.3.1 Pedigree file

The pedigree file (pedfile.dat) should be in the pre-Makeped LINKAGE format. The pre-Makeped columns are pedigree ID, person ID, father ID, mother ID, gender (1=Male and 2=Female), affection status (0 = unknown, 1 = normal, 2 = affected) and genotypes (To code a codominant marker locus phenotype, list the two numbered alleles with at least one space or tab between the alleles, the unknown genotype is coded as 0 0). Everyone must have either two parents or no parents in the data file.

As the RULS work well only when one has marker data on all the autosomal chromosomes, pedfile.dat should contain genotype information from genome-wide marker data. Also, Allegro [17] and Genehunter [24] requires that there should be less than 17 non-founders in any given pedigree.

Example of pedigree file for a genotyped affected sib pair (showing marker information for 2 loci):

```
1  1  0  0  1  0  0  0  0  0  ...
1  2  0  0  2  0  0  0  0  0  ...
1  3  1  2  2  2  6  5  3  3  ...
1  4  1  2  2  2  5  5  3  3  ...
```

3.3.2 Locus file

The locus data file (datafile.dat) should be in standard LINKAGE format with the addition of locus names (after the number of alleles in marker locus, one should put a # sign followed by the locus name), which must be specified. It contains information about the markers to be used in the analysis, including marker names, and their population frequencies. The locus file also contains information regarding affection status locus, and the first locus should be the single affection status locus with single liability class. Also, the locus file has to be matched to the pedigree file, with the loci in exactly the same order.

Example of locus file with 367 markers and 1 trait locus with single liability class:

368 0 0 4
0 0.0 0.0 0
1 2 3 4 5 ··· 367 368
1 2 #diabetes ≪ AFFECTION, NO. OF ALLELES
0.8850 0.1150 ≪GENE FREQUENCIES
1 ≪ NO. OF LIABILITY CLASSES

0.000 0.600 0.600 ≪ PENETRANCES

3 9 # D1S468

0.395890 0.002740 0.093150 0.093150 0.005480 0.012330 0.194520 0.200000 0.002740

3 10 # D1S214

0.085990 0.002770 0.006930 0.027740 0.485440 0.289880 0.090150 0.008320 0.001390 0.001390

⋮

3 11 # D22S274

0.001400 0.253500 0.015410 0.002800 0.113450 0.249300 0.138660 0.001400 0.135850 0.043420 0.044820

0 0 ≪ SEX DIFFERENCE, INTERFERENCE (IF 1 OR 2)

0.1 0.1 0.1 0.1 0.1⋯ 0.1

1 0.10000 0.45000 ≪ REC VARIED, INCREMENT, FINISHING VALUE

3.3.3 Map file

The map file contains the (relative) map position of genome-wide markers in Kosambi centiMorgans (cM).

Example of mapfile:

CHROMOSOME	KOSAMBI	NAME
1	0	D1S468
1	9.82	D1S214
1	16.39	D1S450
1	20.46	D1S2667
	...	

3.4 RULS ON REAL DATA

To analyze your own data with RULS, go to the real/ folder, and follow the instructions below.

A. Copy the your 3 input files: pedigree, locus and map files to the real/ folder (the files should be named as pedfile.dat, datafile.dat and mapfile.dat respectively). The required format of the input files is given above.

B. To compute RULS on your data, type

   ./ruls.sh

C. The result file ruls.tgz will be in real/ folder. Untar ruls.tgz by typing

   tar zxf ruls.tgz,

   to get the output files zruls*.txt, lruls*.txt, vruls*.txt (for RULS), where * denotes the chromosome number.

As you will see, computation of RULS can be time-consuming, thus, if you wish to get results on a specific chromosome, change the chromosome loop in ruls.sh, uncomment that specific chromosome and comment others.

For a real data, you can not create a true structure, so there is no comparison with S_{all} or MLS, thus ruls.sh only computes our RULS.

3.5 OUTPUT FILES FROM RULS.SH

From ruls.sh we only get output files containing our RULS. All the RULS compute multipoint non-parametric LOD scores at a grid of 1cM throughout the genome. The output file zruls*.txt contains z-RULS, estimates of z's (z^T as in our paper), and analytical p-value at each grid of position on chromosome (* represents chromosome number). Similarly, lruls*.txt contains L-RULS, estimates of λ_1, λ_2, p-values; and vruls*.txt contains V-RULS, estimates of V_A/K^2, V_D/K^2, p-values.

3.6 RULS ON EXAMPLE DATA

For example files, go to the example/ folder. See example.pdf for the example pedigree structure. These data were simulated to have a disease locus at 52.54 cM on chromosome

10 segregating under a dominant model with penetrance (0,0.6,0.6).

If you want to see the example pedigree, locus and map file, untar code.tgz by typing
tar zxf code.tgz,
and the respective files (ped*.pre, datafile.dat and newmap.dat) are in code/ folder.

To analyze the example data, continue to next steps in example/ folder, and follow as given below.

A. To simulate the example data and to compute RULS, MLS and S_{all}, type
./sim.sh, and to change the random number seed for simulation, see instructions in sim.sh
.

B. Once you run sim.sh, the result file sim.tgz will be in example/ folder. Untar sim.tgz file by typing
tar zxf sim.tgz

to get the output files. The output files from sim.sh are apparentped (apparent pedigree structure), trueped (true pedigree structure), discard.txt (discarded pedigree structure), zruls*txt, lruls*.txt, vruls*.txt (RULS on apparentped), mlstrue*.out, mlsdiscard*.out (MLS on true and discarded pedigree structures) and lodtrue*.txt, loddiscard*.txt (S_{all} on true and discarded pedigree structures), * denotes for all chromosomes. To get results on a specific chromosome, change the chromosome loop in sim.sh, uncomment that specific chromosome and comment others. See section Output files for details.

3.7 OUTPUT FILES FROM SIM.SH

Here we give the details behind all the output files from sim.sh. All the statistics (RULS, MLS and S_{all}) compute multipoint non-parametric LOD scores at a grid of 1cM throughout the genome. The output file zruls*.txt contains z-RULS, estimates of z's (z^T as in our paper), and analytical p-value at each grid of position on chromosome (* represents chromosome number). Similarly, lruls*.txt contains L-RULS, estimates of λ_1, λ_2, p-values; and vruls*.txt contains V-RULS, estimates of V_A/K^2, V_D/K^2, p-values.

The output files containing MLS (Cordell et al. 2000) on true and discarded pedigree structure are mlstrue*.out and mlsdiscard*.out respectively, where * represents chromosome number. The file mlstrue*.out gives MLS on true structure, estimates of z's (see Cordell et al. 2000), and analytical p-value at each grid of position on chromosome *. Similarly, mlsdiscard*.out gives MLS on discarded structure, z's and p-value.

The output files lodtrue*.txt and loddiscard*.txt contain S_{all} LOD score using true and discarded pedigree structure respectively, where * represents chromosome number. The file lodtrue*.txt contains S_{all} on true sturcure and p-value at each grid of position on chromosome *. Similarly, loddiscard*.txt contains S_{all} on discarded structure.

3.8 JOB SUBMIT

We have performed the simulation on our computer cluster with Sun Grid Engine. To compute genome-wide empirical thresholds we simulated 1000 replicates, and to compute power we simulated 400 replicates. These replicates are obtained by submitting the script (sim.sh modified, see later) to different nodes (one replicate per node).

The sim.sh script is set up to let you run one replicate on a specific node. If you want to submit jobs, look for instructions (commented sections) in sim.sh for job submit. You have to uncomment the job submit sections in sim.sh, and provide the node number to run sim.sh (for example, ./sim.sh 1, where 1 denotes the node number).

3.9 ACKNOWLEDGEMENT FOR THE SOFTWARE

This work was supported by the University of Pittsburgh and, NIH grants 5D43TW006180 and 5R01DK055406. We thank Dr. Mary S. McPeek and Dr. Lei Sun to give us permission to distribute our modified version of their PREST.

3.10 LICENSE

We request that use of this software be cited in publications as Ray A, Weeks DE (2007) ”Relationship uncertainty linkage statistics (RULS): Affected relative pair linkage statistics that model relationship uncertainty”

4.0 APPENDIX B: CODE TO COMPUTE RULS

4.1 RULS.SH

This code, ruls.sh computes RULS using apparent structure from a real data (see software documentation chapter on how to use ruls.sh).

```
#!/bin/tcsh


# Local job. RULS on apparent pedigree structure. #


########### for job submit ###########
#$ -cwd
#set node=$1
#echo JOB_ID: $JOB_ID JOB_NAME: $JOB_NAME HOSTNAME: $HOSTNAME
#mkdir /tmp/$$/
#unalias cp
#unalias mv
#unalias rm
#cp code1.tgz /tmp/$$/
#set HomeDir=‘pwd‘
#cd /tmp/$$
#tar zxf code1.tgz
#cd code1/
#####################################
```

```
set node=1
date

unalias cp
unalias mv
unalias rm

tar zxf code1.tgz     # code1.tgz: codes to compute RULS, MLS and S_all #
cd code1/

# Check for R and C, Fortran, Pascal compiler #
set  a = ‘which R‘
if ( "$a" =˜ *Command?not?found* ) then
  echo ERROR: R not found
  exit
endif

set  a = ‘which awk‘
if ( "$a" =˜ *Command?not?found* ) then
  echo ERROR: awk not found
  exit
endif

set  a = ‘which gcc‘
if ( "$a" =˜ *Command?not?found* ) then
  echo ERROR: gcc, C compiler not found
  exit
endif
```

```
set  a = ‘which g77‘
if ( "$a" =˜ *Command?not?found* ) then
  echo ERROR: g77, Fortran compiler not found
  exit
endif

cp ../../SEARCH.FOR ./SEARCH.F

if !(-e SEARCH.F) then
  echo ERROR: SEARCH.F not found
  exit
endif

make all # compiles .c and .F files #

cp ../pedfile.dat ped
cp ../datafile.dat .
cp ../mapfile.dat newmap.dat

if !(-e ped) then
  echo ERROR: pedigree file not found
  exit
endif

if !(-e datafile.dat) then
  echo ERROR: locus file not found
  exit
endif

if !(-e newmap.dat) then
```

```
  echo ERROR: map file not found
  exit
endif

mega2 -nosave MEGA2.BATCH2>& /dev/null          # MEGA2 with PREST option #

rm gprob.dat

echo Computing P(G|T) # T=FS, HS, FC, U, PO #
./hmm1wolog prest_ped.all prest_chrom.all 2 > gprob.dat
# modified PREST to get P(G|T) #
cp prest_out2 prestout2
 sed ’/Prest/,/is/d’ gprob.dat > ng1.dat
  mv ng1.dat gprob.dat

# to get P(G|T) for 5 T’s #
awk ’$5==1||$5==2||$5==5||$5==6||$5==10 {print $0}’ gprob.dat > gprob1.dat

echo Changing structure from A to T

R CMD BATCH step.R   # Computes r^T|A, F^A, and changes A to 5 T’s #
echo F^A computed

awk ’$6==2 {print $1"_"$2}’ trial1.txt >aff.txt

# Compute RULS at grid of positions on chromosome i. #
#Uncomment chromosome loop if results on other chromosomes are required. #
set i=1
while ( $i<= 9 )
echo Computing RULS for chr$i
```

```
set j=1
while ( $j<= 3 )
cp trial$j.txt trial.txt
cp MEGA2.BATCH3 MEGA2.BATCH    # Mega2 with merlin option to get P(i|G,T) #
echo Chromosome_Single=$i >> MEGA2.BATCH
mega2 -nosave MEGA2.BATCH << MENU1
--ibd --grid 1
MENU1
merlin -p merlin_ped.0$i -d merlin_data.0$i -m merlin_map.0$i -f
merlin_freq.0$i --ibd --grid 1 >! merlin_out
awk ’{print $1,$1"_"$2,$1"_"$3,$4,$5,$6,$7}’ merlin.ibd >! ibd.txt
awk ’NR==FNR {s[$1]} NR!=FNR && ($2 in s) && ($3 in s)’ aff.txt
ibd.txt >! ibd1.txt
awk ’$2!=$3 {print $0}’ ibd1.txt >! ibd2.txt
sed ’s/_/ /g’ ibd2.txt >! ibd3.txt
awk ’{print $1,$3,$5,$6,$7,$8,$9}’ ibd3.txt >! merlinaff$j.txt
# These steps to get P(i|G,T) for affected pairs #
rm merlin_out
rm merlin.ibd
rm ibd.txt
rm ibd1.txt
rm ibd2.txt
rm ibd3.txt
@ j++
end
R CMD BATCH ruls.R        # computes f^A and the RULS #
cp zruls.txt zruls0$i.txt     # position, z-RULS, z’s #
cp lruls.txt lruls0$i.txt      # position, L-RULS, lambda_1, lambda_2 #
cp vruls.txt vruls0$i.txt    # position, V-RULS, V_A, V_D #
echo RULS for chr$i done
```

```
@ i++
end

set i=10
while ( $i<= 22 )
echo Computing RULS for chr$i
set j=1
while ( $j<= 3 )
cp trial$j.txt trial.txt
cp MEGA2.BATCH3 MEGA2.BATCH
echo Chromosome_Single=$i >> MEGA2.BATCH
mega2 -nosave MEGA2.BATCH << MENU1
--ibd --grid 1
MENU1
merlin -p merlin_ped.$i -d merlin_data.$i -m merlin_map.$i -f merlin_freq.$i
--ibd --grid 1 >! merlin_out
awk ’{print $1,$1"_"$2,$1"_"$3,$4,$5,$6,$7}’ merlin.ibd >! ibd.txt
awk ’NR==FNR {s[$1]} NR!=FNR && ($2 in s) && ($3 in s)’ aff.txt
ibd.txt >! ibd1.txt
awk ’$2!=$3 {print $0}’ ibd1.txt >! ibd2.txt
sed ’s/_/ /g’ ibd2.txt >! ibd3.txt
awk ’{print $1,$3,$5,$6,$7,$8,$9}’ ibd3.txt >! merlinaff$j.txt
rm merlin_out
rm merlin.ibd
rm ibd.txt
rm ibd1.txt
rm ibd2.txt
rm ibd3.txt
@ j++
end
```

```
R CMD BATCH ruls.R        # computes f^A and the RULS #
cp zruls.txt zruls$i.txt      # position, z-RULS, z's #
cp lruls.txt lruls$i.txt       # position, L-RULS, lambda_1, lambda_2 #
cp vruls.txt vruls$i.txt     # position, V-RULS, V_A, V_D #
echo RULS for chr$i done
@ i++
end

if !(-e zruls.txt) then
echo ERROR: RULS files not created,
echo check if you have the necessary files given in README.txt
exit(1)
endif

echo RULS done

# Compress the files into a single gzipped tar file
tar zcf ruls.tgz *ruls*.txt
cp ruls.tgz ../

########## for job submit ##########
# Compress the files into a single gzipped tar file
tar zcf ruls.$node.tgz *ruls*.txt

## Copy back the compressed archive file
unalias cp
cp -f ruls.$node.tgz $HomeDir

cd $HomeDir
####################################
```

```
##remove all the files on node
if (-e ruls.tgz) then
echo ruls.tgz copied back, so removing the folder code1/
rm -rf code1/
endif

date
exit
```

4.1.1 step.R

This code, step.R computes $r^T|A_j$, F_{ij}^A, and changes structure A to five T's for each affected pair (*i.e.* $r^T|A$, F^A and changing structure steps in the flowchart Figure 5).

```
# calculates r^T|A and F^A #
gprob<-read.table("gprob1.dat")     # output from modified PREST,
columns=ped id, per1, per2 (affected pairs),A=FS,T, P(G|T) #
gprob[ ,c(2,3)]<-gprob[ ,c(2,3)]+gprob[ ,1]*1000
ped1<-read.table("ped")
ped2<-ped1[ped1$V10==2,]                # for premakeped, it is column 6#
ped2[ ,2]<-ped2[ ,2]+ped2[ ,1]*1000
data<-ped2[ ,c(2,11:ncol(ped2))]        # marker data for unique person id #
aff<-data[,1]
m<-match(x=gprob[,2],table=aff)
m1<-gprob[c(which(is.na(m)==FALSE)),]
m2<-match(x=m1[,3],table=aff)
m3<-m1[c(which(is.na(m2)==FALSE)),]
pr1<-m3[ ,5:6]
n1<-nrow(pr1)/5                # n1=number of affected pairs #
f1<-matrix(0,n1,5)             # matrix each row will be P(G|T) #
```

```
d<-NULL
d1<-NULL
d2<-NULL
for(i in 1:n1){
tmp<-pr1[((((i-1)*5)+1):(i*5)),2]
m<-max(abs(tmp))
n<--740+m
d1<-rbind(d1,n)
tmp1<-exp(n+tmp)
d2<-rbind(d2,tmp1)
write.table(tmp1,"tmp.txt",quote=FALSE,row.names=FALSE,col.names=FALSE)
        system("./clust1.o>OUTS2.DAT") # optimize to get r^T|A
          for each affected pair #
        system("awk -f findmax.awk -v maxiter=‘awk -f max.awk OUTS2.DAT‘
          OUTS2.DAT> r1.txt")
        r<-scan("r1.txt")
        d<-rbind(d,r[4:8])
}
write.table(d,"r.txt",eol="\n",row.names=FALSE,col.names=FALSE,quote=FALSE)
I<-matrix(nrow=5,ncol=3,c(.25,.5,.25,.5,.5,0,.75,.25,0,1,0,0,0,1,0),byrow=TRUE)
F<-t(t(I)%*%t(d))
write.table(F,"F.txt",eol="\n",row.names=FALSE,col.names=FALSE,quote=FALSE)
write.table(d1,"underflow.txt",quote=FALSE,row.names=FALSE,col.names=FALSE)
write.table(d2,"pofg.txt",quote=FALSE,row.names=FALSE,col.names=FALSE)

# Changes apaprent structure of affected relative pairs to T #
g<-m3[m3$V5==1,]
n<-nrow(g)                              # number of affected pairs #

# takes affected pair and makes it FS #
```

```
d<-NULL
A<-matrix(nrow=2,ncol=5,c(100,0,0,1,0,200,0,0,2,0),byrow=T)
for(i in 1:n){
   B<-matrix(nrow=2,ncol=5,c(g[i,2],100,200,2,2,g[i,3],100,200,2,2),byrow=T)
        tmp<-rbind(A,B)                 # pedigree with affecteds as FS #
        d<-rbind(d,tmp)
}
a<-seq(1:nrow(d))
d<-cbind(d,a)
m<-merge(x=d,y=data,by.x=c(1),by.y=c(1),all.x=TRUE)
m1<-sort(m[ ,6],index.return=T)
d<-m[m1$ix,c(1:5,7:ncol(m))]
b<-rep(c(1:n),each=4)           # ped id, 4 individuals for T=FS #
d<-cbind(b,d)
d[,2]<-round(1000*(d[,2]/1000-floor(d[,2]/1000)))
write.table(d,"trial1.txt",na="0",eol="\n",quote=FALSE,
row.names=FALSE,col.names=FALSE)

# takes affected pair and makes it HS #
d<-NULL
A<-matrix(nrow=3,ncol=5,c(100,0,0,1,0,200,0,0,2,0,300,0,0,2,0),byrow=T)
for(i in 1:n){
   B<-matrix(nrow=2,ncol=5,c(g[i,2],100,200,2,2,g[i,3],100,300,2,2),byrow=T)
        tmp<-rbind(A,B)                 # pedigree with affecteds as HS #
        d<-rbind(d,tmp)
}
a<-seq(1:nrow(d))
d<-cbind(d,a)
m<-merge(x=d,y=data,by.x=c(1),by.y=c(1),all.x=TRUE)
m1<-sort(m[ ,6],index.return=T)
```

```
d<-m[m1$ix,c(1:5,7:ncol(m))]
b<-rep(c(1:n),each=5)           # ped id, 5 individuals for T=HS #
d<-cbind(b,d)
d[,2]<-round(1000*(d[,2]/1000-floor(d[,2]/1000)))
write.table(d,"trial2.txt",na="0",eol="\n",quote=FALSE,
row.names=FALSE,col.names=FALSE)

# takes affected pair and makes it FC #
d<-NULL
A<-matrix(nrow=6,ncol=5,c(100,0,0,1,0,200,0,0,2,0,300,100,200,2,0,
400,0,0,1,0,500,100,200,2,0,600,0,0,1,0),byrow=T)
for(i in 1:n){
   B<-matrix(nrow=2,ncol=5,c(g[i,2],400,300,2,2,g[i,3],600,500,2,2),byrow=T)
        tmp<-rbind(A,B)                 # pedigree with affecteds as FC #
        d<-rbind(d,tmp)
}
a<-seq(1:nrow(d))
d<-cbind(d,a)
m<-merge(x=d,y=data,by.x=c(1),by.y=c(1),all.x=TRUE)
m1<-sort(m[ ,6],index.return=T)
d<-m[m1$ix,c(1:5,7:ncol(m))]
b<-rep(c(1:n),each=8)                   # ped id, 8 individuals for T=FC #
d<-cbind(b,d)
d[,2]<-round(1000*(d[,2]/1000-floor(d[,2]/1000)))
write.table(d,"trial3.txt",na="0",eol="\n",quote=FALSE,
row.names=FALSE,col.names=FALSE)

#change to U#
d<-NULL
for(i in 1:n){
        A<-matrix(nrow=1,ncol=5,c(100,g[i,2],g[i,3],1,0),byrow=T)
        B<-matrix(nrow=2,ncol=5,c(g[i,2],0,0,1,2,g[i,3],0,0,2,2),byrow=T)
        tmp<-rbind(A,B)
        d<-rbind(d,tmp)
}
a<-seq(1:nrow(d))
d<-cbind(d,a)
m<-merge(x=d,y=data,by.x=c(1),by.y=c(1),all.x=TRUE)
m1<-sort(m[ ,6],index.return=T)
d<-m[m1$ix,c(1:5,7:ncol(m))]
b<-rep(c(1:n),each=3)          # ped id, 3 individuals for T=U  #
d<-cbind(b,d)
d[,2]<-round(1000*(d[,2]/1000-floor(d[,2]/1000)))
d[,3]<-round(1000*(d[,3]/1000-floor(d[,3]/1000)))
d[,4]<-round(1000*(d[,4]/1000-floor(d[,4]/1000)))
write.table(d,"trial4.txt",na="0",eol="\n",quote=FALSE,
row.names=FALSE,col.names=FALSE)

# change to PO#
d<-NULL
A<-matrix(nrow=1,ncol=5,c(100,0,0,2,0),byrow=T)
for(i in 1:n){
   B<-matrix(nrow=2,ncol=5,c(g[i,2],0,0,1,2,g[i,3],g[i,2],100,2,2),byrow=T)
        tmp<-rbind(A,B)
        d<-rbind(d,tmp)
}
a<-seq(1:nrow(d))
d<-cbind(d,a)
m<-merge(x=d,y=data,by.x=c(1),by.y=c(1),all.x=TRUE)
m1<-sort(m[ ,6],index.return=T)
```

```
d<-m[m1$ix,c(1:5,7:ncol(m))]
b<-rep(c(1:n),each=3)                   # ped id, 3 individuals for T=PO #
d<-cbind(b,d)
d[,2]<-round(1000*(d[,2]/1000-floor(d[,2]/1000)))
d[,3]<-round(1000*(d[,3]/1000-floor(d[,3]/1000)))
write.table(d,"trial5.txt",na="0",eol="\n",quote=FALSE,
row.names=FALSE,col.names=FALSE)

q("no")
```

4.1.2 ruls.R

This code, ruls.R computes f_{ij}^A and the RULS at grid of positions on a chromosomes (*i.e.* f^A and RULS steps in the flowchart Figure 5).

```
# computes RULS #
ng<-read.table("pofg.txt")
ng<-as.matrix(ng)
aff1<-read.table("merlinaff1.txt")    #ibd of affected pairs, P(i|G,T=FS) #
aff2<-read.table("merlinaff2.txt")    #ibd of affected pairs, P(i|G,T=HS)#
aff3<-read.table("merlinaff3.txt")    #ibd of affected pairs, P(i|G,T=FC)#
tmp<-unique(aff1[ ,4])
q<-read.table("F.txt")          #reads F^A#
r<-read.table("r.txt")          #reads r^T|A, here A=FS#
write.table(nrow(r),"pair.txt",eol="\n",quote=FALSE,
row.names=FALSE,col.names=FALSE)
# pair.txt gives number of affected pairs #
d2<-NULL
d3<-NULL
d4<-NULL
```

```
d<-vector("numeric",length=nrow(r))
for(k in 1:length(tmp)){        #loop over positions#
        tmp1<-aff1[aff1$V4==tmp[k],5:7]  #P(i|G,T) at each position#
        tmp2<-aff2[aff2$V4==tmp[k],5:7]
        tmp3<-aff3[aff3$V4==tmp[k],5:7]
        d<-diag(ng%*%t(r))
tmp10<-(tmp1*ng[,1]*r[,1]+tmp2*ng[,2]*r[,2]+tmp3*ng[,3]*r[,3]+(matrix
(nrow=nrow(r),ncol=3,c(1,0,0),byrow=T))*ng[,4]*r[,4]+(matrix
(nrow=nrow(r),ncol=3,c(0,1,0),byrow=T))*ng[,5]*r[,5])*(1/d)
        f3<-cbind(tmp10,q) #puts f^A and F^A together to calculate L-RULS#
write.table(f3,"file3.txt",eol="\n",quote=FALSE,
row.names=FALSE,col.names=FALSE)
        tmp10[ ,1]<-tmp10[ ,1]/q[ ,1]
        tmp10[ ,2]<-tmp10[ ,2]/q[ ,2]
        tmp10[ ,3]<-tmp10[ ,3]/q[ ,3]
f2<-cbind(tmp10,r)
#puts vector f^A/F^A and r together to calculate z-RULS and V-RULS#
write.table(f2,"file.txt",eol="\n",quote=FALSE,
row.names=FALSE,col.names=FALSE)
system("./clust2.o > OUTS3.DAT")   #SEARCH to compute z-RULS#
system("./clust3.o")  #SEARCH to compute L-RULS#
system("./clust4.o > OUTS6.DAT")  #SEARCH to compute V-RULS#
system("awk -f findmax.awk -v maxiter=‘awk -f max.awk OUTS3.DAT‘
 OUTS3.DAT>amr.txt")
# tmp.awk and max.awk finds the iteration at which the function attains #
# minimum and pulls out the row of parameter estimates #
system("awk -f findmax.awk -v maxiter=‘awk -f max.awk OUTS4.DAT‘
 OUTS4.DAT>amr1.txt")
system("awk -f findmax.awk -v maxiter=‘awk -f max.awk OUTS6.DAT‘
 OUTS6.DAT>amr2.txt")
```

```
        amr<-scan("amr.txt")
        amr1<-scan("amr1.txt")
        amr2<-scan("amr2.txt")
        d2<-rbind(d2,amr)
        d3<-rbind(d3,amr1)
        d4<-rbind(d4,amr2)
}
d30<-d2[ ,3:10]    #  values of function and z #
d30<-abs(d30)
#takes the positive value as minimum obtained for negative of the function #
d40<-cbind(tmp,d30)               #adds position vector#
write.table(d40,"zruls.txt",eol="\n",quote=FALSE,row.names=FALSE,
col.names=FALSE)
#file with pos, z-RULS, z’s#
d31<-d3[ ,3:5]      #  values of function and lambda_1 and lambda_2 #
d31<-abs(d31)
d41<-cbind(tmp,d31)
write.table(d41,"lruls.txt",eol="\n",quote=FALSE,row.names=FALSE,
col.names=FALSE)
#file with pos, L-RULS, lambda_1 and lambda_2#
d32<-d4[ ,3:5]       # values of function and V_A and V_D #
d32<-abs(d32)
d42<-cbind(tmp,d32)
write.table(d42,"vruls.txt",eol="\n",quote=FALSE,row.names=FALSE,
col.names=FALSE)
#file with pos, V-RULS, V_A, V_D#
q("no")
```

4.2 SIM.SH

This code, sim.sh computes RULS on the example data files (see software documentation on how to use sim.sh).

```
#!/bin/tcsh


# Local job. RULS on apparent pedigree structure, as obtained #
# from simulated true structure. #


########### for job submit ###########
#$ -cwd
#set node=$1
#echo JOB_ID: $JOB_ID JOB_NAME: $JOB_NAME HOSTNAME: $HOSTNAME
#mkdir /tmp/$$/
#unalias cp
#unalias mv
#unalias rm
#cp code.tgz /tmp/$$/
#set HomeDir=‘pwd‘
#cd /tmp/$$
#tar zxf code.tgz
#cd code/
#####################################


set node=1
date


unalias cp
unalias mv
unalias rm
```

```
tar zxf code.tgz  # code.tgz has the necessary codes to
compute RULS, MLS and S_all #
cd code/

set  a = ‘which R‘
if ( "$a" =˜ *Command?not?found* ) then
  echo ERROR: R not found
  exit
endif

set  a = ‘which awk‘
if ( "$a" =˜ *Command?not?found* ) then
  echo ERROR: awk not found
  exit
endif

set  a = ‘which gcc‘
if ( "$a" =˜ *Command?not?found* ) then
  echo ERROR: gcc, C compiler not found
  exit
endif

set  a = ‘which g77‘
if ( "$a" =˜ *Command?not?found* ) then
  echo ERROR: g77, Fortran compiler not found
  exit
endif

cp ../../SEARCH.FOR ./SEARCH.F
cp ../../maxfun.f .
cp ../../onelocarp.f .

if !(-e SEARCH.F) then
  echo ERROR: SEARCH.F not found
  exit
endif

if !(-e onelocarp.f) then
  echo ERROR: onelocarp.f not found
  exit
endif
if !(-e maxfun.f) then
  echo ERROR: maxfun.f not found
  exit
endif

make all

touch seed$node.txt
echo $node >! node.txt

echo Beginning simulation replicate 1

### For pedigree simulation ###
R CMD BATCH rand.R  # Creates different seed files for each pedigree type #
#To change the random number seed: as rand.R generates one random number #
# seed at a time, to get a new number, you have to run rand.R #
#for more replicates or change the node number. #
```

```
set a  = ( 2 2 10 10 2 4 ) # Number of pedigrees for each type #


set i=1
while ( $i<= 6 )
R CMD BATCH ped$i.R  # ped*.pre has the true pedigree structure, ped*.R #
#adds to ped*.pre: 1 1 for typed individuals and 0 0 for untyped ones, #
#required for SIMULATE #
@ i++
end
# You have to change ped*.pre files to your input true pedigree structure. #
#ped*.pre has 6 columns: pedigree ID, person ID, parent ID's, #
# sex and affection status #


set j=1
while ( $j<= 6 )
cp ped$j.pre ped.pre
mega2 MEGA2.BATCH1>& /dev/null << MENU
# MEGA2 with SIMULATE option, to simulate null chromosomes #
1
$a[$j]
0
MENU
set k=1              # Loop over chromosomes #
while ( $k<= 9 )
cp nproblem$j.$node problem.dat   # nproblem* is the seed file from rand.R #
echo -n $j 0$k >> seed$node.txt
cat problem.dat >> seed$node.txt
cp simdata.0$k simdata.dat
cp simped.0$k simped.dat
simulate >& /dev/null
```

```
cp problem.dat nproblem$j.$node  # seed changed by SIMULATE#
rm simdata.dat
rm simped.dat
mv pedfile.dat pedfile.0$k   # Pedigree file simulated for chromosome #
@ k++
end
set k=11   # Disease chromosome is 10, to be simulated by Allegro #
while ( $k<= 22 )
cp nproblem$j.$node problem.dat
echo -n $j $k >> seed$node.txt
cat problem.dat >> seed$node.txt
cp simdata.$k simdata.dat
cp simped.$k simped.dat
simulate >& /dev/null
cp problem.dat nproblem$j.$node
rm simdata.dat
rm simped.dat
mv pedfile.dat pedfile.$k
@ k++
end

R CMD BATCH try.R  # Binds the pedfiles over the null chromosomes #
cp ped ped$j

R CMD BATCH trial$j.R
# Changes the pedigree structure from T to A=FS. #
#If you also require A=FS, you need to only change person ID in trial*.R #
@ j++
end
```

```
R CMD BATCH totalped.R
# Gives apparent and true pedigree without chr10 simulated.#
#For your input pedigree, you have to change the pedigree ID in totalped.R #

cp simdata.10 ex1.dat
# ex1.dat: locus data file for disease chr. 10, required for Allegro #

set i=1                                 # Loop on pedigree type #
while ($i<= 6)
cp ex1.opt$i ex1.opt
cp ex1.pre$i ex1.pre
# ex1.pre* is the true pedigree structure with 8 columns: first 6 same
as ped*.pre and last two, 0 0 for untyped individuals and 1 1 for
typed individuals. This format is reuired for Allegro. #
# For your input files, change ex1.pre* #

allegro ex1.opt >& /dev/null   # ALLEGRO to simulate disease chromosome #
R CMD BATCH newped$i.R  # True structure changed to apparent, A=FS. #
#If you also require A=FS, you need to only change the person ID
 in newped*.R #
@ i++
end
R CMD BATCH ped.R
# Merge T and A structures over disease and null chromosomes. #
#ped.R outputs T (trueped) and A (apparentped) pedigree structures #

### RULS computation ###
mega2 -nosave MEGA2.BATCH2>& /dev/null    # MEGA2 with PREST option #

rm gprob.dat
```

```
echo Computing P(G|T) # T= FS, HS, FC, U, PO #
./hmm1wolog prest_ped.all prest_chrom.all 2 > gprob.dat
# modified PREST to get P(G|T) #
cp prest_out2 prestout2
 sed ’/Prest/,/is/d’ gprob.dat > ng1.dat
  mv ng1.dat gprob.dat

# to get P(G|T) for 5 T’s #
awk ’$5==1||$5==2||$5==5||$5==6||$5==10 {print $0}’ gprob.dat > gprob1.dat

echo Changing structure from A to T

R CMD BATCH step.R   # Computes r^T|A, F^A, and changes A to 5 T’s #
echo F^A computed

awk ’$6==2 {print $1"_"$2}’ trial1.txt >aff.txt

# Compute RULS at grid of positions on chromosome i. #
#Uncomment the chromosome loop if results on other
chromosomes are required. #
set i=1
while ( $i<= 9 )
echo Computing RULS for chr$i
set j=1
while ( $j<= 3 )
cp trial$j.txt trial.txt
cp MEGA2.BATCH3 MEGA2.BATCH    # Mega2 with merlin option to get P(i|G,T) #
echo Chromosome_Single=$i >> MEGA2.BATCH
mega2 -nosave MEGA2.BATCH << MENU1
```

```
--ibd --grid 1
MENU1
merlin -p merlin_ped.0$i -d merlin_data.0$i -m merlin_map.0$i
-f merlin_freq.0$i
--ibd --grid 1 >! merlin_out
awk ’{print $1,$1"_"$2,$1"_"$3,$4,$5,$6,$7}’ merlin.ibd >! ibd.txt
awk ’NR==FNR {s[$1]} NR!=FNR && ($2 in s) && ($3 in s)’ aff.txt
ibd.txt >! ibd1.txt
awk ’$2!=$3 {print $0}’ ibd1.txt >! ibd2.txt
sed ’s/_/ /g’ ibd2.txt >! ibd3.txt
awk ’{print $1,$3,$5,$6,$7,$8,$9}’ ibd3.txt >! merlinaff$j.txt
# These steps to get P(i|G,T) for affected pairs #
rm merlin_out
rm merlin.ibd
rm ibd.txt
rm ibd1.txt
rm ibd2.txt
rm ibd3.txt
@ j++
end
R CMD BATCH ruls.R        # computes f^A and the RULS #
cp zruls.txt zruls0$i.txt     # position, z-RULS, z’s #
cp lruls.txt lruls0$i.txt      # position, L-RULS, lambda_1, lambda_2 #
cp vruls.txt vruls0$i.txt    # position, V-RULS, V_A, V_D #
echo RULS for chr$i done
@ i++
end

set i=10
while ( $i<= 22 )
```

```
echo Computing RULS for chr$i
set j=1
while ( $j<= 3 )
cp trial$j.txt trial.txt
cp MEGA2.BATCH3 MEGA2.BATCH
echo Chromosome_Single=$i >> MEGA2.BATCH
mega2 -nosave MEGA2.BATCH << MENU1
--ibd --grid 1
MENU1
merlin -p merlin_ped.$i -d merlin_data.$i -m merlin_map.$i -f merlin_freq.$i
--ibd --grid 1 >! merlin_out
awk ’{print $1,$1"_"$2,$1"_"$3,$4,$5,$6,$7}’ merlin.ibd >! ibd.txt
awk ’NR==FNR {s[$1]} NR!=FNR && ($2 in s) && ($3 in s)’ aff.txt
ibd.txt >! ibd1.txt
awk ’$2!=$3 {print $0}’ ibd1.txt >! ibd2.txt
sed ’s/_/ /g’ ibd2.txt >! ibd3.txt
awk ’{print $1,$3,$5,$6,$7,$8,$9}’ ibd3.txt >! merlinaff$j.txt
rm merlin_out
rm merlin.ibd
rm ibd.txt
rm ibd1.txt
rm ibd2.txt
rm ibd3.txt
@ j++
end
R CMD BATCH ruls.R        # computes f^A and the RULS #
cp zruls.txt zruls$i.txt      # position, z-RULS, z’s #
cp lruls.txt lruls$i.txt       # position, L-RULS, lambda_1, lambda_2 #
cp vruls.txt vruls$i.txt     # position, V-RULS, V_A, V_D #
echo RULS for chr$i done
```

```
@ i++

end


if !(-e zruls.txt) then

echo ERROR: RULS files not created,

echo check if you have the necessary files given in README.txt

exit(1)

endif


echo RULS done


### MLS computation ###

cp trueped ped


# MEGA2 with gh option on trueped #

mega2 MEGA2.BATCH4>& /dev/null<< MENU2

0

0

MENU2

cp gh_ped.10 gh_ped


set i=1

while ( $i<= 9 )  # Loop on chromosome #

printf "load gh_dat.0$i\n scan gh_ped.0$i\n dump ibd\n ibd0$i\n

quit\n" >! input2.txt

gh < input2.txt >& /dev/null

sed ’s/,/ /g’ ibd0$i >! tmp

sed ’1d’ tmp >! tmp1

rm tmp

R CMD BATCH cor.R
```

```
# Makes the prior and posterior files, required for Cordell’s
code onelocarp to run #
rm tmp1
cat prevalence.txt d.txt a1.txt >! tmp0$i
# prevalence.txt has value of K, d.txt has # of affected pairs, #
# and a1.txt has # of marker positions, #
#these three lines required at the beginning of posterior file #
# for onelocarp to compute MLS #

cat oneposterior.dat >> tmp0$i
mv tmp0$i oneposterior.dat
./onelocarp >! z.out   # Computes MLS and also prints V_A,V_D #
R CMD BATCH result.R   # Prints position, MLS, z_0, z_1, z_2 #
mv onemls.out mlstrue0$i.out
@ i++
end

set i=10
while ( $i<= 22 )
printf "load gh_dat.$i\n scan gh_ped.$i\n dump ibd\n ibd$i\n
quit\n" >! input2.txt
gh < input2.txt >& /dev/null
sed ’s/,/ /g’ ibd$i >! tmp
sed ’1d’ tmp >! tmp1
rm tmp
R CMD BATCH cor.R
rm tmp1
cat prevalence.txt d.txt a1.txt >! tmp$i
cat oneposterior.dat >> tmp$i
mv tmp$i oneposterior.dat
```

```
./onelocarp >! z.out
R CMD BATCH result.R
mv onemls.out mlstrue$i.out
@ i++
end
rm gh_ped*
rm gh_dat*
rm ibd*

# mega2 with gh option on discard.txt #
cp discard.txt ped
mega2 MEGA2.BATCH4>& /dev/null<< MENU2
0
0
MENU2
cp gh_ped.10 gh_ped

set i=1
while ( $i<= 9 )                # Loop on chromosome #
printf "load gh_dat.0$i\n scan gh_ped.0$i\n dump ibd\n ibd0$i\n
quit\n" >! input4.txt
gh < input4.txt >& /dev/null
sed ’s/,/ /g’ ibd0$i >! tmp
sed ’1d’ tmp >! tmp1
rm tmp
R CMD BATCH cor1.R
# Makes the prior and posterior files, required for Cordell’s
code onelocarp to run #
rm tmp1
cat prevalence.txt d.txt a1.txt >! tmp0$i
cat oneposterior.dat >> tmp0$i
mv tmp0$i oneposterior.dat
./onelocarp >! z.out
R CMD BATCH result.R
mv onemls.out mlsdiscard0$i.out
@ i++
end

set i=10
while ( $i<= 22 )
printf "load gh_dat.$i\n scan gh_ped.$i\n dump ibd\n ibd$i\n
quit\n" >! input4.txt
gh < input4.txt >& /dev/null
sed ’s/,/ /g’ ibd$i >! tmp
sed ’1d’ tmp >! tmp1
rm tmp
R CMD BATCH cor1.R
rm tmp1
cat prevalence.txt d.txt a1.txt >! tmp$i
cat oneposterior.dat >> tmp$i
mv tmp$i oneposterior.dat
./onelocarp >! z.out
R CMD BATCH result.R
mv onemls.out mlsdiscard$i.out
@ i++
end

if !(-e onemls.out) then
echo ERROR: MLS files not created
exit(1)
```

```
endif

echo MLS done

# Merlin #
# [4] mega2 and merlin on trueped #

cp trueped ped
mega2 MEGA2.BATCH5>& /dev/null<< MENU2
--npl --deviates --grid 1
0
MENU2
set i=1
while ( $i<= 9 )        # Loop on chromosome #
merlin -p merlin_ped.0$i -d merlin_data.0$i -m merlin_map.0$i
-f merlin_freq.0$i --npl --deviates --grid 1 >! merlin_out
./merlinmax.awk merlin_out >! out
sed '1,2d' out>!out1
R CMD BATCH sgn.r  # multiplies the sign of delta to the LOD scores, S_all #
mv lod.txt lodtrue0$i.txt
@ i++
end

set i=10
while ( $i<= 22 )
merlin -p merlin_ped.$i -d merlin_data.$i -m merlin_map.$i
-f merlin_freq.$i --npl --deviates --grid 1 >! merlin_out
./merlinmax.awk merlin_out >! out
sed '1,2d' out>!out1
R CMD BATCH sgn.r  # multiplies the sign of delta to the LOD scores, S_all #
```

```
mv lod.txt lodtrue$i.txt
@ i++
end

# [5] mega2 and merlin on discard.txt (ped structure discarding
 the individual with erroneous relationship), datafile.dat#
# mega2 and prest on discard.txt, datafile.dat and newmap.dat #

cp discard.txt ped
mega2 MEGA2.BATCH5>& /dev/null<< MENU2
--npl --deviates --grid 1
0
MENU2

set i=1
while ( $i<= 9 )
merlin -p merlin_ped.0$i -d merlin_data.0$i -m merlin_map.0$i
-f merlin_freq.0$i --npl --deviates --grid 1 >! merlin_out
./merlinmax.awk merlin_out >! out
sed ’1,2d’ out>!out1
R CMD BATCH sgn.r       # multiplies the sign of delta to the LOD scores #
mv lod.txt loddiscard0$i.txt
@ i++
end

set i=10
while ( $i<= 22 )
merlin -p merlin_ped.$i -d merlin_data.$i -m merlin_map.$i
-f merlin_freq.$i --npl --deviates --grid 1 >! merlin_out
./merlinmax.awk merlin_out >! out
```

```
sed '1,2d' out>!out1
R CMD BATCH sgn.r       # multiplies the sign of delta to the LOD scores #
while ( $i<= 9 )
merlin -p merlin_ped.0$i -d merlin_data.0$i -m merlin_map.0$i
-f merlin_freq.0$i --npl --deviates --grid 1 >! merlin_out
./merlinmax.awk merlin_out >! out
sed '1,2d' out>!out1
R CMD BATCH sgn.r   # multiplies the sign of delta to the LOD scores #
mv lod.txt loddiscard0$i.txt
@ i++
end

set i=10
while ( $i<= 22 )
merlin -p merlin_ped.$i -d merlin_data.$i -m merlin_map.$i
-f merlin_freq.$i --npl --deviates --grid 1 >! merlin_out
./merlinmax.awk merlin_out >! out
sed '1,2d' out>!out1
R CMD BATCH sgn.r       # multiplies the sign of delta to the LOD scores #
mv lod.txt loddiscard$i.txt
@ i++
end

echo MERLINDONE

# Compress the files into a single gzipped tar file
tar zcf sim.tgz apparentped trueped *ruls*.txt mls*.out lod*.txt
cp sim.tgz ../          # comment this for cluster job submission #

## Copy back the compressed archive file
```

```
unalias cp
cp -f sim.$node.tgz $HomeDir

cd $HomeDir
####################################

##remove all the files on node
if (-e sim.tgz) then
echo sim.tgz copied back, so removing the folder code1/
rm -rf code1/
endif

date
exit
```

BIBLIOGRAPHY

[1] Abecasis GR, Cherny SS, Cookson WO, Cardon LR *Merlin-rapid analysis of dense genetic maps using sparse gene flow trees*, Am J Hum Genet 30:97-101, 2002.

[2] Barrett Fry, Maller, Daly *Haploview: analysis and visualization of LD and haplotype maps*, Bioinformatics, 21(2):263-265, 2005.

[3] Baum LE *An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes*, Inequalities 3:1-8, 1972.

[4] Blackwelder WC, Elston RC *A comparison of sib-pair linkage tests for disease susceptibility loci*, Genet Epidemiol 2:85-97, 1985.

[5] Boehnke M, Cox NJ *Accurate inference of relationships in sib-pair linkage studies*, Am J Hum Genet 61:423-429, 1997.

[6] Casselbrant ML, Mandel EM, Kurs-Lasky M, Rockette HE, Bluestone CD *Otitis media in a population of black American and white American infants, 0-2 years of age*, Int J Ped Otorhinolaryngol 33:1-16, 1995.

[7] O'Connell JR, Weeks DE *PedCheck: A program for identifying genotype incompatibilities in linkage analysis*, Am J Hum Genet 63:259-266, 1998.

[8] Cordell HJ, Todd JA, Bennett ST, Kawaguchi Y, Farrall M *Two-locus maximum LOD score analysis of a multifactorial trait: joint consideration of IDDM2 and IDDM4 with IDDM1 in type 1 diabetes*, Am J Hum Genet 57:920-934, 1995.

[9] Cordell HJ, Wedig GC, Jacobs KB, Elston RC *Multilocus linkage tests based on affected relative pairs*, Am J Hum Genet 66:1273-1286, 2000.

[10] Daly KA, Brown WM, Segade F, Bowden DW, Keats BJ, Lindgren BR, Levine SC, et al. *Chronic and recurrent otitis media: a genome scan for susceptibility loci*, Am J Hum Genet 75:988-997, 2004.

[11] Douglas JA, Boehnke M, Lange K *A multipoint method for detecting genotyping errors and mutations in sibling-pair linkage data*, Am J Hum Genet. 66:1287-97, 2000.

[12] Ehm M, Wagner M *A test statistic to detect errors in sib-pair relationships*, Am J Hum Genet. 62(1):181-188, 1998.

[13] Ehm MG, Karnoub MC, Sakul H, Gottschalk K, Holt DC, Weber JL, Vaske D, et al *Genomewide search for type 2 diabetes susceptibility genes in four American populations*, Am J Hum Genet 66:1871-1881, 2000.

[14] Epstein MP, Duren WL, Boehnke M *Improved inference of relationships for pairs of individuals*, Am J Hum Genet. 67:1219-1231, 2000.

[15] Faraway JJ *Improved sib-pair linkage test for disease susceptibility loci*, Genet Epidemiol 10:225-233, 1993.

[16] Göring HH, Ott J *Relationship estimation in affected sib pair analysis of late-onset diseases*, Eur J Hum Genet 5:69-77, 1997.

[17] Gudbjartsson DF, Thorvaldsson T, Kong A, Gunnarsson G, Ingolfsdottir A *Allegro version 2*, Nat Genet 37:1015-1016, 2005.

[18] Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR *Whole-Genome Patterns of Common DNA Variation in Three Human Populations*, Science, 307:1072-1079, 2005.

[19] Holmans P *Asymptotic properties of affected sib-pair linkage analysis*, Am J Hum Genet 52:362-374, 1993.

[20] James JW *Frequency in relatives for an all-or-none trait*, Ann Hum Genet 35:47-49, 1971.

[21] Kong A, Cox NJ *Allele-sharing models: LOD scores and accurate linkage tests*, Am J Hum Genet. 61:1179-1188, 1997.

[22] Kogan MD, Overpeck MD, Hoffman HJ, Casselbrant ML. Factors associated with tympanostomy tube insertion among preschool-aged children in the United States. Am J Public Health 90(2):245-50, 2000.

[23] Kong X, Murphy K, Raj T, He C, White PS, Matise TC *A Combined Linkage-Physical Map of the Human Genome*, Am J Hum Genet, 75(6):1143-8, 2004.

[24] Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES *Parametric and nonparametric linkage analysis: a unified multipoint approach*, Am J Hum Genet 58:1347-1363, 1996.

[25] Laird, Horvath, Xu *Implementing a unified approach to family based tests of association*, Genetic Epi, supp 1, 19:36-42, 2000.

[26] Lange K *SEARCH , Version 3.0*, 1985-1991.

[27] Margaretha L, Casselbrant ML, Mandel EM, Fall PA, Rockette HE, Lasky MK, Bluestone CD, Ferrell RE *The Heritability of Otitis Media, A Twin and Triplet Study*, JAMA, 282:2125-2130, 1999.

[28] Martin E.R., Monks SA, Warren LL, Kaplan NL *A test for linkage and association in general pedigrees: The pedigree disequilibrium test*, Am. J. Hum. Genet 67:146-154, 2000.

[29] Mukhopadhyay N, Almasy L, Schroeder M, Mulvihill WP, Weeks DE *Mega2: data-handling for facilitating genetic linkage and association analyses*, Bioinformatics May 15;21(10):2556-7, 2005

[30] McPeedk *Optimal allele-sharing statistics for genetic mapping using affected relatives*, Genetic Epidemiology 16:225-49, 1999.

[31] McPeek MS, Sun L *Statistical tests for detection of misspecified relationships by use of genome-screen data*, Am J Hum Genet 66:1076-1094, 2000.

[32] Lange K, Cantor R, Horvath S, Perola M, Sabatti C, Sinsheimer J, Sobel E *Mendel version 4.0: A complete package for the exact genetic analysis of discrete traits in pedigree and population data sets* Amer J Hum Genetics 69(supplement):A1886, 2001.

[33] Olson JM *A general conditional-logistic model for affected-relative-pair linkage studies*, Am J Hum Genet 65:1760-1769, 1999.

[34] Paradise JL, Rockette HE, Colburn DK, Bernard BS, Smith CG, Kurs-Lasky M, Janosky JE *Otitis media in 2253 Pittsburgh-area infants: prevalence and risk factors during the first two years of life*, Pediatrics 99:318-333, 1997.

[35] Rasmussen F *Protracted secretory otitis media: The impact of familial factors and day-care center attendance*, Int J Pediatr Otorhinolaryngol 26:29-37, 1993.

[36] Risch N *Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs*, Am J Hum Genet 46:242-53, 1990.

[37] Rohde K, Fuerst R *Haplotyping and estimation of haplotype frequencies for closely linked biallelic multilocus genetic phenotypes including nuclear family information*, Hum Mutat 17:289-295, 2001.

[38] Rovers MM, Zielhuis GA, Straatman H, Ingels K, van der Wilt GJ, van den Broek P *Prognostic factors for persistent otitis media with effusion in infants*, Arch Otolaryngol Head Neck Surg 125:1203-1207, 1999.

[39] Rovers M, Haggard M, Gannon M, Schomerus GK, Plomin R *Heritability of Symptom Domains in Otitis Media: A Longitudinal Study of 1,373 Twin Pairs*, Am J Epidem, 10:958-964, 2002.

[40] Schappert SM *Office visits for otitis media: United States, 1975-90.* Adv Data 222:1-19, 1992.

[41] Segade F, Daly KA, Allred D, Hicks PJ, Cox M, Brown M, Hughes RE *Association of the FBXO11 gene with chronic otitis media with effusion and recurrent otitis media: the Minnesota COME/ROM Family Study*, Arch Otolaryngol Head Neck Surg, Jul;132(7):729-33, 2006.

[42] Shmulewitz D, Heath SC, Blundell ML, Han Z, Sharma R, Salit J, Auerbach SB, et al *Linkage analysis of quantitative traits for obesity, diabetes, hypertension, and dyslipidemia on the island of Kosrae, Federated States of Micronesia*, Proc Natl Acad Sci USA 103:3502-3509, 2006.

[43] Sobel E, Papp JC, Lange K *Detection and Integration of Genotyping Errors in Statistical Genetics*, Am J Hum Genet. 70(2):496-508, 2002.

[44] Spielman RS, McGinnis RE, Ewens WJ *Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM)*, Am J Hum Genet. 52(3):506-16, 1993.

[45] Terwilliger JD, Speer M, Ott J *Chromosome-based method for rapid computer simulation in human genetic linkage analysis*, Genet Epidemiol 10:217-224, 1993.

[46] Thompson EA *The estimation of pairwise relationships*, Ann Hum Genet 39:173-188, 1975.

[47] Thompson EA *Pedigree Analysis in Human Genetics*, The Johns Hopkins University Press, Baltimore, 1986.

[48] Thompson CL, Rybicki BA, Iannuzzi MC, Elston RC, Iyengar SK, Gray-McGuire C *Reduction of sample heterogeneity through use of population substructure: an example from a population of African American families with sarcoidosis*, Am J Hum Genet. Oct;79(4):606-13, 2006.

[49] Whittemore AS, Halpern J *A class of tests for linkage using affected pedigree members*, Biometrics 50:118-127.

[50] Shih MC, Whittemore AS *Allele-sharing among affected relatives: non parametric methods for identifying genes*, Statistical Methods in Medical Research, 10:27-55, 2001.