Statistics and Its Interface Volume 2 (2009) 133–144

# Missing data methods for linkage analysis of IBS and incomplete IBD from affected sib-pairs

Dennis W. Buckman† and Zhaohai Li*

We derive linkage statistics for situations where the number of marker alleles shared identical-by-descent (IBD) is incomplete, and the number of marker alleles shared identical-by-state (IBS) is known. The linkage statistics are based on the assumption that the parental genotypes are missing at random (MAR). We first assume the marker IBD is unambiguous for a sib-pair if the parental genotypes are available. Then we relax this assumption to assess the impact of marker ambiguity. The derivation of each statistic involves a Taylor series expansion of a log likelihood that is a function of the recombination fraction and nuisance parameters and incorporates the missing data situation. The first derivative of the log likelihood is zero under the null hypothesis of no recombination, so the Taylor series expansion of the log likelihood reveals a linkage statistic that is proportional to the second derivative of the log likelihood evaluated at the null hypothesis value of the recombination fraction. We prove that the standardized linkage statistics have asymptotic normal distributions, and we provide required sample sizes and simulation results and consider the impact of parent availability and marker ambiguity.

Keywords and phrases: Incomplete data, Missing at random, Identity-by-descent, Identity-by-state, Complex traits, Required sample size, Power, Type I error.

## 1. INTRODUCTION

Development of statistical methods for missing data problems is an ongoing area of statistical research. Nicolae et al. (2008) discuss the importance of determining the impact of missing data on the performance of hypothesis tests and propose methods for quantifying the amount of available information relative to the complete data. They discuss linkage analysis and other applications of their methods and consider further methodological work. In our work, we develop missing data methods for linkage analysis using affected sib-pairs, and we consider the impact of parent availability and marker ambiguity. The statistics that we derive involve consideration of missing data methods for tabulated data. Chen and Fienberg (1974) considered the likelihood function for incomplete contingency table data, and Nordheim (1984) applied the missing at random assumption of Rubin (1976) to contingency tables. Before discussing our methods, we provide a brief review of the relevant statistical genetics literature.

In the statistical genetics literature, linkage analysis methods have been developed for the study of complex diseases, that is, diseases that do not follow usual Mendelian inheritance (Bishop and Williamson, 1990; Risch and Zhang, 1995; Olson et al., 1999). Some examples of complex diseases include: forms of cancer, Alzheimer's disease, multiple sclerosis, and insulin-dependent diabetes. To study complex traits, authors have proposed linkage analysis methods that utilize the identity-by-descent (IBD) distribution or utilize the identity-by-state (IBS) distribution for various types of relative-pairs (Haseman, 1970; Haseman and Elston, 1972; Suarez et al., 1978; Lange, 1986; Risch, 1990a-c; Bishop and Williamson, 1990; Risch, 1992; Risch and Zhang, 1995, 1996; Zhang and Risch, 1996; Feingold, 2001; Li and Gastwirth, 2001; Feingold, 2002). Kong and Cox (1997) proposed a likelihood approach that allows for incomplete identity by descent information, and these methods have been incorporated into the nonparametric linkage (NPL) test in commonly used software such as MERLIN (Abecasis et al., 2002).

Haseman and Elston's (1972) linkage analysis method calculates the squared difference of the trait values and the proportion of marker alleles shared IBD for each sib-pair and regresses the squared differences on the IBD proportions. Suarez et al. (1978) provided a linkage analysis method that attempts to detect an increased number of alleles shared IBD among affected sib-pairs. Risch (1990b) extended the affected sib-pair approach by parameterizing using the recurrence risk ratio for specific types of relatives and proposed a likelihood ratio linkage test. He also considered the use of additional relatives to provide more information in situations where the marker is not 100 percent polymorphic (Risch, 1990c, 1992). Dudoit and Speed (1999, 2000) used a Taylor series expansion of their log likelihood function to develop linkage statistics that utilize marker IBD data. They considered the following hypotheses. Under the null hypothesis, the marker locus is not linked to the trait locus, and the recombination fraction, $\theta$, equals $\frac{1}{2}$. Under the alternative hypothesis, the marker and trait loci are linked. Dudoit and Speed's linkage test is based on the second derivative of their log likelihood function evaluated at $\theta = \frac{1}{2}$.

*The work of the second author was supported in part by NIH grant EY014478.

†Corresponding author.

Lange (1986) discussed the use of IBS for situations where IBD is unavailable, such as a disease with late onset. For example, if elderly sib-pairs are sampled, their parents may be unavailable. Weeks and Lange (1988) considered affected members of pedigrees and used an IBS approach with weighting that allows incorporation of the allele frequency. Bishop and Williamson (1990) considered the use of the IBS distribution for affected relative-pairs. Thomson and Motro (1994) compared four methods that use IBS data from affected sib-pairs. They considered each of these IBS methods with and without the presence of linkage disequilibrium. Holmans (1993) used a restriction criteria for his likelihood ratio test and made comparisons with a restriction used by Risch (1990c, 1992), and Holmans used his asymptotic results to discuss situations where typing parents could be considered inefficient compared with typing additional siblings. However, he also mentioned that "other considerations may apply, such as the supply of affected pairs being limited, which might make typing the parents preferable."

We present linkage statistics for situations where the number of marker alleles shared identical-by-descent is incomplete, and the number of marker alleles shared identical-by-state is known (Buckman, 2005). In Section 2.1, we present a likelihood function and linkage statistic based on the assumption that the parental genotypes are missing at random (MAR). Also, in Section 2.1, we assume that the marker IBD is known unambiguously if the parents are available. In Section 2.2, we relax the assumption of Section 2.1 to allow for marker ambiguity and present a likelihood function and linkage statistic that allows the marker IBD to be estimated from sib-pair and parental genotypes and allows the parental genotypes to be MAR. The statistics defined in Sections 2.1 and 2.2 are linear combinations of multinomial random vectors, and we present asymptotic results. In Section 3.1, required sample sizes are presented for various genetic models and recombination fraction values, and comparisons are made for various levels of missing data. Simulation results in Section 3.2 consider finite sample sizes and evaluate the properties of the statistics for various genetic models, recombination fraction values, and levels of missing data, and we compare some of our results to results from MERLIN linkage analysis software. Our asymptotic results, required sample size results, and simulation results allow for evaluation of the impact of parent availability and marker ambiguity.

## 2. METHODS

In Sections 2.1 and 2.2, we consider linkage analysis for affected sib-pairs where parental genotypes are missing at random. The methods in Section 2.1 consider a contingency table of the IBS, IBD, and missing status variables. The methods in Section 2.2 allow for ambiguous marker IBD and consider a contingency table of the IBS, IBD estimate, and missing status variables.

### 2.1 Method for tabulated IBS, IBD, and missing status

#### 2.1.1 Likelihood

The likelihood function is derived for marker IBS and marker IBD data from affected sib-pairs where the parental genotypes are missing at random. Also, in this section, we assume that the marker IBD is unambiguous whenever the parental genotypes are available. Before considering the tabulated data, we must consider that an allele shared IBD is always shared IBS; therefore, if marker IBS is zero, then the marker IBD must be zero. Therefore, even though the parents are MAR, the marker IBD is only MAR when the marker IBS is 1 or 2, and we utilize the MAR and likelihood results for contingency tables as discussed by Chen and Fienberg (1974) and Nordheim (1984).

To derive the likelihood function, define indicator random variables, $Q_k$, such that $Q_k = 1$ if the marker IBD is observed for the $k^{th}$ affected sib-pair and $Q_k = 0$ otherwise. For convenience, the index $k$ is dropped. The conditional probability of observing the marker IBD, given the marker IBS, is

$$\tau_i = P(Q = 1 | IBS_M = i),$$

for $i = 0, 1, 2$. The multinomial random vector of interest is

$$M = (M_{00}, M_{10}, M_{20}, M_{11}, M_{21}, M_{22}, R_1, R_2)'$$

where $M_{ij}$ represents the random number of affected sib-pairs with $Q = 1$, $IBS_M = i$, and $IBD_M = j$, and $R_i$ represents the random number of affected sib-pairs with $Q = 0$, $IBS_M = i$, and $IBD_M$ missing.

Before specifying the distribution of $M$, relevant results from Risch (1990b) and Bishop and Williamson (1990) are reviewed. Let $X$ represent the number of affected siblings in a sib-pair, and since we consider affected sib-pairs, $X$ equals 2. Risch (1990b) gives the following marker IBD probabilities for affected sib-pairs,

$$\begin{aligned}
Z_{S_0}(\theta) &= P(IBD_M = 0 | X = 2) \\
&= \frac{1}{4} - \frac{1}{4\lambda_S}(2\psi - 1)[(\lambda_S - 1) + 2(1 - \psi)(\lambda_S - \lambda_O)], \\
Z_{S_1}(\theta) &= P(IBD_M = 1 | X = 2) \\
&= \frac{1}{2} - \frac{1}{2}(2\psi - 1)^2 \frac{1}{\lambda_S}(\lambda_S - \lambda_O), \\
Z_{S_2}(\theta) &= P(IBD_M = 2 | X = 2) \\
&= \frac{1}{4} + \frac{1}{4\lambda_S}(2\psi - 1)[(\lambda_S - 1) + 2\psi(\lambda_S - \lambda_O)]
\end{aligned}$$

where $\psi = \theta^2 + (1 - \theta)^2$, $\theta$ is the recombination fraction between marker and trait loci, and $\lambda_S$ and $\lambda_O$ are the sibling and parent-offspring recurrence risk ratios, respectively. For an arbitrary locus, labeled L, in Hardy-Weinberg equilibrium, Bishop and Williamson (1990) give the following

*Table 1. Multinomial probabilities for affected sib-pairs with $IBS_M$ known and $IBD_M$ missing at random*

| | | $Q=1$ $IBD_M$ | | | $Q=0$ $IBD_M$ |
|---|---|---|---|---|---|
| | | 0 | 1 | 2 | unobserved |
| $IBS_M$ | 0 | $T_{00}Z_{S_0}(\theta)$ | 0 | 0 | 0 |
| | 1 | $\tau_1 T_{10} Z_{S_0}(\theta)$ | $\tau_1 T_{11} Z_{S_1}(\theta)$ | 0 | $(1-\tau_1)\sum_{j=0}^{1} T_{1j} Z_{S_j}(\theta)$ |
| | 2 | $\tau_2 T_{20} Z_{S_0}(\theta)$ | $\tau_2 T_{21} Z_{S_1}(\theta)$ | $\tau_2 Z_{S_2}(\theta)$ | $(1-\tau_2)\sum_{j=0}^{2} T_{2j} Z_{S_j}(\theta)$ |

conditional probabilities of IBS given IBD,

$$
\begin{aligned}
T_{00} &= \sum_{s\neq t} p_s p_t (1-p_s-p_t)^2 + \sum_s p_s^2 (1-p_s)^2,\\
T_{10} &= 4\sum_{s\neq t} p_s p_t^2 (1-p_s-p_t) + 4\sum_s p_s^3 (1-p_s),\\
T_{20} &= 2\sum_{s\neq t} p_s^2 p_t^2 + \sum_s p_s^4,\\
T_{11} &= \sum_s p_s (1-p_s),\\
T_{21} &= \sum_s p_s^2,\\
T_{22} &= 1
\end{aligned}
$$

where $T_{ij} = P(IBS_L = i | IBD_L = j)$, and $p_s$ and $p_t$ represent the allele frequencies, and $s, t = 1, 2, \ldots, J$ index the alleles.

The multinomial probabilities for $M$ are given by the interior cells of Table 1 which is derived in Appendix A. An allele shared IBD is also shared IBS; therefore, $\tau_0 = 1$, and Table 1 has four structural zeros. Considering the factors of the multinomial probability mass function that involve $\theta$, we define the likelihood function of $\theta$ as

$$
\begin{aligned}
L(\theta; M) = &\left[\prod_{j=0}^{2} (Z_{S_j}(\theta))^{M_{\cdot j}}\right]\left[\sum_{u=0}^{1} T_{1u} Z_{S_u}(\theta)\right]^{R_1}\\
&\cdot \left[\sum_{v=0}^{2} T_{2v} Z_{S_v}(\theta)\right]^{R_2}.
\end{aligned} \tag{1}
$$

To obtain the linkage statistics presented in the next two sections, the log of this likelihood function and its Taylor series expansion are considered.

### 2.1.2 Linkage test with nuisance parameters known

For the linkage statistic presented in this section, it is assumed that the nuisance parameters, $\tau_1$, $\tau_2$, and $(T_{00}, T_{10}, T_{20}, T_{11}, T_{21})$, are known, and in the next section, these parameters are estimated using maximum likelihood estimators. The parameter of interest is the recombination fraction, $\theta$. Under the null hypothesis, $H_0 : \theta = \frac{1}{2}$, the marker and trait loci are not linked, and under the alternative hypothesis, $H_1 : 0 < \theta < \frac{1}{2}$, the marker and trait loci are linked.

Consider the likelihood given by Equation 1, the null value $\theta_0 = \frac{1}{2}$, and a particular alternative value $\theta_1 = \theta_0 - h = \frac{1}{2} - h$ where $h$ is a sufficiently small positive value. Since the first derivative of the log likelihood is zero at the null value of $\theta$, the Taylor series expansion of the log likelihood is

$$
\begin{aligned}
&logL\left(\tfrac{1}{2} - h; M\right)\\
&\quad = logL\left(\tfrac{1}{2}; M\right) + \frac{1}{2}h^2 \frac{d^2}{d\theta^2} logL(\theta; M)\,|_{\theta=\frac{1}{2}} + R(\xi)
\end{aligned}
$$

where $R(\xi)$ is the remainder and $\xi\ \epsilon\ (\theta_1, \theta_0) = \left(\frac{1}{2} - h, \frac{1}{2}\right)$.

Under the null hypothesis of no linkage, the $Z_{S_j}(\theta)$ parameters simplify to

$$
Z_{S_0}\left(\tfrac{1}{2}\right) = \frac{1}{4}, \qquad Z_{S_1}\left(\tfrac{1}{2}\right) = \frac{1}{2}, \qquad \text{and} \quad Z_{S_2}\left(\tfrac{1}{2}\right) = \frac{1}{4}.
$$

Therefore, the second derivative of the log likelihood, evaluated at $\theta = \frac{1}{2}$, is

$$
\frac{d^2}{d\theta^2} logL(\theta; M)\,|_{\theta=\frac{1}{2}} = \frac{8(2\lambda_S - \lambda_O - 1)}{\lambda_S} T
$$

where

$$
T = M_{\cdot 2} - M_{\cdot 0} + \frac{1 - T_{20}}{T_{20} + 2T_{21} + 1} R_2 - \frac{T_{10}}{T_{10} + 2T_{11}} R_1.
$$

Given the vector,

$$
C = \left(-1, -1, -1, 0, 0, 1, \frac{-T_{10}}{T_{10} + 2T_{11}}, \frac{1 - T_{20}}{T_{20} + 2T_{21} + 1}\right)',
$$

the linkage statistic is $T = C'M$.

Appendix B presents the derivation of the mean and variance of $T$ under the null and alternative hypotheses. Under the null hypothesis, $H_0 : \theta = \frac{1}{2}$, the mean of $T$ is $E_{H_0}(T) = 0$, and the variance is

$$
\begin{aligned}
Var_{H_0}(T) = \frac{N}{4}\Bigg(&T_{00} + \tau_1 T_{10} + \tau_2 T_{20} + \tau_2\\
&+ (1-\tau_1)\frac{T_{10}^2}{T_{10} + 2T_{11}} + (1-\tau_2)\frac{(1-T_{20})^2}{T_{20} + 2T_{21} + 1}\Bigg).
\end{aligned}
$$

Since the nuisance parameters are known, the variance of $T$, under the null hypothesis, is also known, and the standardized linkage statistic is

$$
\frac{T}{\sqrt{Var_{H_0}(T)}}.
$$

In Appendix C, it is established that this standardized statistic is asymptotically distributed as $N(0,1)$ under the null hypothesis. Therefore, for the one-sided $\alpha$ level test, the null hypothesis of no linkage is rejected when this standardized statistic is at least as great as $z_{1-\alpha}$ where $\Phi(z_{1-\alpha}) = 1-\alpha$, and $\Phi$ is the standard normal distribution function.

In order to determine the number of affected sib-pairs required to detect linkage at prespecified power and type I error levels, the distribution of the test statistic, $T$, under the alternative hypothesis is also considered. Appendix B presents the mean and variance of $T$ under the alternative hypothesis of linkage, $H_1$. To specify an equation for required sample sizes, consider the following notation: $NV_0 = Var_{H_0}(T)$, $N\mu_1 = E_{H_1}(T)$, and $NV_1 = Var_{H_1}(T)$. The derivation in Appendix C shows that $N^{-1}T$ has an asymptotic normal distribution under the null or alternative hypothesis; therefore, the number of affected sib-pairs required to achieve approximate power of $1-\beta$ is

$$N = \frac{\left(z_{1-\alpha}\sqrt{V_0} + z_{1-\beta}\sqrt{V_1}\right)^2}{\mu_1^2}.$$

#### 2.1.3 Linkage test with nuisance parameters estimated

When the nuisance parameters are unknown, the standardized linkage statistic is defined using maximum likelihood estimators (MLEs) for these unknown parameters. These MLEs are obtained by considering the MLEs for the parameters of the multinomial distribution and applying the invariance property.

The MLEs for the nuisance parameters, $\tau_1$ and $\tau_2$, are

$$\hat{\tau}_1 = \frac{M_{10}+M_{11}}{M_{10}+M_{11}+R_1} \quad \text{and} \quad \hat{\tau}_2 = \frac{M_{20}+M_{21}+M_{22}}{M_{20}+M_{21}+M_{22}+R_2},$$

and the maximum likelihood estimator for the nuisance parameter vector, $(T_{00}, T_{10}, T_{20}, T_{11}, T_{21})$, is

$$(\hat{T}_{00}, \hat{T}_{10}, \hat{T}_{20}, \hat{T}_{11}, \hat{T}_{21}) = \frac{1}{N}\left(\frac{M_{00}}{\hat{Z}_{S_0}(\theta)}, \frac{M_{10}}{\hat{\tau}_1\hat{Z}_{S_0}(\theta)}, \frac{M_{20}}{\hat{\tau}_2\hat{Z}_{S_0}(\theta)}, \frac{M_{11}}{\hat{\tau}_1\hat{Z}_{S_1}(\theta)}, \frac{M_{21}}{\hat{\tau}_2\hat{Z}_{S_1}(\theta)}\right),$$

where

$$\hat{Z}_{S_0}(\theta) = \frac{1}{N}\left[M_{00} + \frac{M_{10}}{\hat{\tau}_1} + \frac{M_{20}}{\hat{\tau}_2}\right] \quad \text{and}$$
$$\hat{Z}_{S_1}(\theta) = \frac{1}{N}\left[\frac{M_{11}}{\hat{\tau}_1} + \frac{M_{21}}{\hat{\tau}_2}\right].$$

Therefore, the variance of $T$ under the null hypothesis, $\sigma_0^2 = Var_{H_0}(T)$, can be estimated by

$$\hat{\sigma}_0^2 = \frac{N}{4}\left(\hat{T}_{00} + \hat{\tau}_1\hat{T}_{10} + \hat{\tau}_2\hat{T}_{20} + \hat{\tau}_2 + (1-\hat{\tau}_1)\frac{\hat{T}_{10}^2}{\hat{T}_{10}+2\hat{T}_{11}} + (1-\hat{\tau}_2)\frac{(1-\hat{T}_{20})^2}{\hat{T}_{20}+2\hat{T}_{21}+1}\right).$$

Since we now assume the nuisance parameters must be estimated, $T$ involves unknown parameters, so the linkage statistic is defined as

$$\tilde{T} = M_{.2} - M_{.0} + \frac{1-\hat{T}_{20}}{\hat{T}_{20}+2\hat{T}_{21}+1}R_2 - \frac{\hat{T}_{10}}{\hat{T}_{10}+2\hat{T}_{11}}R_1,$$

and dividing $\tilde{T}$ by $\hat{\sigma}_0$ yields a standardized linkage statistic that is asymptotically distributed as $N(0,1)$ under the null hypothesis (Appendix C) as in Section 2.1.2.

## 2.2 Method for tabulated IBS, IBD estimate, and missing status

#### 2.2.1 Likelihood

In this section, a linkage test is derived for affected sib-pairs with marker IBS known and marker IBD estimated from sib-pair and parental genotypes, and the parental genotypes are missing at random. Haseman and Elston (1970, 1972) provide an estimator, $\hat{\pi}_{Ms} = \frac{1}{2}f_{s1} + f_{s2}$, of the proportion of marker alleles that the $s^{th}$ sib-pair share IBD, $\pi_{Ms}$, where $f_{sj}$ is the conditional probability that the sib-pair share $j$ marker alleles IBD given their genotypes and their parents' genotypes. Haseman and Elston provide the $f_{sj}$ conditional probabilities for a multiallelic marker locus in Hardy-Weinberg equilibrium. Multiplying Haseman and Elston's estimator by 2 yields $\widehat{IBD}_{Ms} = 2\hat{\pi}_{Ms} = f_{s1} + 2f_{s2}$ which is an estimator for the number of marker alleles that the $s^{th}$ sib-pair share IBD.

Dropping the index $s$ for notational convenience, the estimator, $\widehat{IBD}_M$, is considered along with the $IBS_M$ and $IBD_M$ random variables. Consider the joint probability

$$P\left(\widehat{IBD}_M = \tfrac{k}{2}, IBS_M = i, IBD_M = j\right)$$

where $k = 0,1,2,3,4$ and $i,j = 0,1,2$ and $i \geq j$. Then summing this joint probability over all possible values of $k$ yields the marginal probability $P(IBS_M = i, IBD_M = j)$. Define $V_{kij}$ as the conditional probability

$$V_{kij} = P\left(\widehat{IBD}_M = \tfrac{k}{2} | IBS_M = i, IBD_M = j\right) = \frac{P\left(\widehat{IBD}_M = \frac{k}{2}, IBS_M = i, IBD_M = j\right)}{P\left(IBS_M = i, IBD_M = j\right)}$$

where $k = 0,1,2,3,4$ and $i,j = 0,1,2$ and $i \geq j$. Examining the mating types and sib-pair types and the $f_{sj}$ conditional probabilities given by Haseman and Elston (1970, 1972) reveals that 19 of the conditional probabilities, $V_{kij}$, are always zero. Therefore, only 11 of these conditional probabilities need to be considered in more detail. Given the condition that $IBS_M = 0$ and $IBD_M = 0$, the possible mating type ($MT$) and sib-pair type ($ST$) combinations determine that the conditional probability, $V_{000}$, equals 1. Similarly, evaluating the possible $MT$ and $ST$ combinations given that $IBS_M = 2$ and $IBD_M = 0$ reveals that the conditional

Table 2. *Multinomial random variables for affected sib-pairs with $IBS_M$ known and $\widehat{IBD}_M$ missing at random*

| | | $\tilde{Q}=1$ $\widehat{IBD}_M$ | | | | | $\tilde{Q}=0$ $\widehat{IBD}_M$ |
|---|---|---|---|---|---|---|---|
| | | 0.0 ($k=0$) | 0.5 ($k=1$) | 1.0 ($k=2$) | 1.5 ($k=3$) | 2.0 ($k=4$) | unobserved |
| $IBS_M$ | 0 | $N_{00}$ | 0 | 0 | 0 | 0 | 0 |
| | 1 | $N_{10}$ | $N_{11}$ | $N_{12}$ | 0 | 0 | $R_1$ |
| | 2 | 0 | 0 | $N_{22}$ | $N_{23}$ | $N_{24}$ | $R_2$ |

probability, $V_{220}$, equals 1. To illustrate the calculation of the conditional probabilities, $V_{kij}$, consider

$$V_{221} = \frac{\frac{1}{2}P(MT=I, ST=I) + \frac{1}{2}P(MT=II, ST=V)}{P(IBS_M=2, IBD_M=1)} = \frac{P(MT=I, ST=I) + P(MT=II, ST=V)}{T_{21}}$$

where $MT=I$ and $MT=II$ refer to mating types $I$ and $II$ and $ST=I$ and $ST=V$ refer to sib-pair types $I$ and $V$, respectively, as defined by Haseman and Elston (1970, 1972). Also, considering the possible $MT$ and $ST$ combinations given that $IBS_M=2$ and $IBD_M=1$, it is clear that $V_{321} = 1 - V_{221}$.

The above results will be used in the derivation of the linkage statistic for affected sib-pairs. The data for the affected sib-pair linkage test are described as follows. If the affected sib-pair's genotype and their parents' genotype are available, then $\tilde{Q}=1$, and the affected sib-pair will have an $\widehat{IBD}_M$ value of 0.0, 0.5, 1.0, 1.5, or 2.0; otherwise, $\widehat{IBD}_M$ is missing, and $\tilde{Q}=0$. The marker IBS is available for all of the affected sib-pairs. Suppose $\tilde{\tau}_i = P(\tilde{Q}=1|IBS_M=i)$ for $i=0,1,2$, and $\tilde{\tau}_0=1$ as in similar results from Section 2.1.1. The number of affected sib-pairs with $IBS_M=i$ and missing $\widehat{IBD}_M$ values are represented by $R_i$ where $i=1,2$. Table 2 presents the counts, $N_{ik}$, for affected sib-pairs with observed $\widehat{IBD}_M$ and $IBS_M$ values and the counts, $R_i$, for affected sib-pairs with unobserved $\widehat{IBD}_M$ values.

Considering the entries of Table 2, the random vector

$$\tilde{M} = (N_{00}, N_{10}, N_{11}, N_{12}, N_{22}, N_{23}, N_{24}, R_1, R_2)'$$

has a multinomial distribution with parameter vector

$$\Gamma = (\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5, \gamma_6, \gamma_7, \gamma_8, \gamma_9)'$$

where $\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5, \gamma_6$, and $\gamma_7$ are given by

$$\tilde{\tau}_i P\big(IBS_M=i, \widehat{IBD}_M=\tfrac{k}{2}|X=2\big) = \tilde{\tau}_i \sum_{j=0}^{i} V_{kij} T_{ij} Z_{S_j}(\theta)$$

and $\gamma_8$ and $\gamma_9$ are

$$\gamma_8 = (1-\tilde{\tau}_1)[T_{10}Z_{S_0}(\theta) + T_{11}Z_{S_1}(\theta)],$$
$$\gamma_9 = (1-\tilde{\tau}_2)[T_{20}Z_{S_0}(\theta) + T_{21}Z_{S_1}(\theta) + Z_{S_2}(\theta)].$$

Section 2.1.1 gives expressions for $T_{ij}Z_{S_j}(\theta) = P(IBS_M=i, IBD_M=j|X=2)$.

Considering the factors of the multinomial probability mass function that involve $\theta$, we define the likelihood function of $\theta$ as

$$L(\theta;\tilde{M}) = \left[\prod_{i,k}\left(\sum_{j=0}^{i} V_{kij}T_{ij}Z_{S_j}(\theta)\right)^{N_{ik}}\right] \cdot \left[\sum_{u=0}^{1} T_{1u}Z_{S_u}(\theta)\right]^{R_1} \left[\sum_{v=0}^{2} T_{2v}Z_{S_v}(\theta)\right]^{R_2}. \quad (2)$$

To obtain the linkage statistic presented in the next section, the log of this likelihood function and its Taylor series expansion are considered.

### 2.2.2 Linkage test using IBS and estimated IBD with parental genotypes MAR

Consider the likelihood given by Equation 2, the null value $\theta_0 = \frac{1}{2}$, and a particular alternative value $\theta_1 = \theta_0 - h = \frac{1}{2} - h$ where $h$ is a sufficiently small positive value. Since the first derivative of the log likelihood is zero at the null value of $\theta$, the Taylor series expansion of the log likelihood reveals a linkage statistic proportional to the second derivative of the log likelihood evaluated at $\theta = \frac{1}{2}$. Section 2.1.2 provides justification for this approach. The second derivative of the log likelihood, evaluated at $\theta = \frac{1}{2}$, is

$$\frac{d^2}{d\theta^2} log L\big(\theta;\tilde{M}\big)|_{\theta=\frac{1}{2}} = \frac{8(2\lambda_S - \lambda_O - 1)}{\lambda_S} U$$

where

$$U = -N_{00} - \sum_{k=0}^{2} N_{1k}\frac{V_{k10}T_{10}}{V_{k10}T_{10} + 2V_{k11}T_{11}} + \sum_{k=2}^{4} N_{2k}\frac{V_{k22} - V_{k20}T_{20}}{V_{k20}T_{20} + 2V_{k21}T_{21} + V_{k22}} - R_1\frac{T_{10}}{T_{10}+2T_{11}} + R_2\frac{1-T_{20}}{T_{20}+2T_{21}+1}.$$

After considering the $V_{kij}$ values, the definition of $U$ simplifies. Also, to take advantage of results from Section 2.1.2, we write the statistic $U$ as a linear combination of $\tilde{M}$ using

the vector of constants

$$\tilde{C} = \left(-1, -1, b_1, 0, b_2, b_3, 1, \frac{-T_{10}}{T_{10} + 2T_{11}}, \frac{1 - T_{20}}{T_{20} + 2T_{21} + 1}\right)'$$

where

$$b_1 = \frac{-V_{110}T_{10}}{V_{110}T_{10} + 2V_{111}T_{11}} \quad \text{and}$$
$$b_k = \frac{V_{k22} - V_{k20}T_{20}}{V_{k20}T_{20} + 2V_{k21}T_{21} + V_{k22}} \quad \text{for } k = 2, 3.$$

Therefore, the linkage statistic is $U = \tilde{C}'\tilde{M}$ which is a linear combination of the multinomial vector $\tilde{M}$. Thus, the standardized statistic and the asymptotic distribution are all obtained by substitution, using the multinomial results presented in Section 2.1.2.

## 3. RESULTS

### 3.1 Required sample sizes and impact of incomplete marker information

For various genetic models and parameter values, Tables 3 and 4 provide the number of affected sib-pairs required to detect linkage using the methods from Section 2.1 with 0.80 power for the one-sided $\alpha = 0.0001$ level test that allows the marker IBD to be missing at random when marker IBS is 1 or 2. These results are based on a marker locus with eight equally likely alleles and a trait locus with alleles $D$ and $d_1, d_2, \ldots, d_K$ and allele frequencies $P(D)$ and $P(d_1), P(d_2), \ldots, P(d_K)$, respectively. For notational convenience, let $d$ represent any of the alleles $d_1, d_2, \ldots, d_K$; thus, $P(d) = \sum_{i=1}^{K} P(d_i) = 1 - P(D)$. The genetic models are defined using penetrances $f_0 = 0.1, f_1$, and $f_2$ which correspond to trait genotypes $dd, Dd$, and $DD$, respectively. The genetic models that are considered include a dominant model where $f_2 = \gamma^2 f_0$ and $f_1 = f_2$, a recessive model where $f_2 = \gamma^2 f_0$ and $f_1 = f_0$, a multiplicative model where $f_2 = \gamma^2 f_0$ and $f_1 = \gamma f_0$, and an additive model where $f_2 = \gamma^2 f_0$ and $f_1 = \frac{1}{2}(f_0 + f_2)$.

Considering the dominant, multiplicative, and additive models and fixed $\theta$ and $P(D)$ values, Table 3 shows the required sample size is 6.0% to 6.5% larger when $\tau_1 = \tau_2 = 0.8$ compared to the sample size required when the marker IBD is not missing, i.e. when $\tau_1 = \tau_2 = 1.0$. For example, 196 affected sib-pairs are required for the dominant model with $P(D) = 0.05$ and $\theta = 0.001$ and $\tau_1 = \tau_2 = 0.8$; however, when $\tau_1 = \tau_2 = 1.0$, only 184 affected sib-pairs are required to achieve the same power. The recessive models in Table 3 reveal that the required sample size is 3.9% to 4.9% larger when $\tau_1 = \tau_2 = 0.8$ compared to the sample size required when $\tau_1 = \tau_2 = 1.0$. Table 3 also illustrates that a low trait allele frequency of $P(D) = 0.05$ leads to extremely large required sample sizes for recessive models and fairly large required sample sizes for multiplicative models. The smallest sample sizes in Table 3 are for recessive models with $P(D) = 0.20$ and dominant models with $P(D) = 0.05$.

For various genetic models and parameter values, Table 4 presents the required sample size for other values of $\tau_i$ including 0.0, 0.2, and 0.4, and the required sample size is presented for the mean IBS test which is based on the average number of alleles that an affected sib-pair shares identical-by-state. Considering the affected sib-pairs linkage statistic,

*Table 3. $IBD_M$ known (i.e. $\tau_1 = \tau_2 = 1.0$) vs. $IBD_M$ incomplete: The number of affected sib-pairs required to detect linkage with 0.80 power for a one-sided $\alpha = 0.0001$ level test that uses missing at random IBD and known IBS*

| | $P(D) = 0.05$ | | | $P(D) = 0.20$ | | |
|---|---|---|---|---|---|---|
| | $\tau_i = 0.8^{\ddagger}$ | $\tau_i = 0.9$ | $\tau_i = 1.0$ | $\tau_i = 0.8$ | $\tau_i = 0.9$ | $\tau_i = 1.0$ |
| Dominant$^{\dagger}$ | | | | | | |
| $\theta = 0.001$ | 196 | 190 | 184 | 395 | 383 | 372 |
| $\theta = 0.050$ | 298 | 288 | 280 | 598 | 580 | 563 |
| Recessive$^{\dagger}$ | | | | | | |
| $\theta = 0.001$ | 8089 | 7933 | 7783 | 193 | 188 | 184 |
| $\theta = 0.050$ | 12262 | 11998 | 11745 | 291 | 284 | 278 |
| Multiplicative$^{\dagger}$ | | | | | | |
| $\theta = 0.001$ | 2100 | 2037 | 1977 | 563 | 546 | 531 |
| $\theta = 0.050$ | 3178 | 3081 | 2990 | 853 | 827 | 803 |
| Additive$^{\dagger}$ | | | | | | |
| $\theta = 0.001$ | 568 | 550 | 534 | 588 | 570 | 552 |
| $\theta = 0.050$ | 861 | 834 | 809 | 891 | 863 | 837 |

$^{\ddagger}\tau_i = P(Q = 1 | IBS_M = i)$ for $i = 1, 2$.
$^{\dagger}$Dominant: $f_2 = \gamma^2 f_0, f_1 = f_2$; Recessive: $f_2 = \gamma^2 f_0, f_1 = f_0$; Multiplicative: $f_2 = \gamma^2 f_0, f_1 = \gamma f_0$; Additive: $f_2 = \gamma^2 f_0, f_1 = \frac{1}{2}(f_0 + f_2)$ where penetrances $f_0 = 0.1$, $f_1$, and $f_2$ correspond to trait genotypes $dd$, $Dd$, and $DD$ and $\gamma = 3$.

*Table 4. $IBD_M$ unavailable (i.e. $\tau_1 = \tau_2 = 0.0$) vs. $IBD_M$ incomplete: The number of affected sib-pairs required to detect linkage with 0.80 power for a one-sided $\alpha = 0.0001$ level test that uses either the statistic for missing at random IBD and known IBS or the mean IBS statistic*

| | | IBD MAR and IBS Known | | |
|---|---|---|---|---|
| | Mean IBS | $\tau_i = 0.0^{\ddagger}$ | $\tau_i = 0.2$ | $\tau_i = 0.4$ |
| Dominant$^{\dagger}$ | | | | |
| $\theta = 0.001$ | 526 | 523 | 484 | 450 |
| $\theta = 0.050$ | 799 | 796 | 735 | 683 |
| Recessive$^{\dagger}$ | | | | |
| $\theta = 0.001$ | 242 | 234 | 222 | 211 |
| $\theta = 0.050$ | 371 | 361 | 341 | 322 |
| Multiplicative$^{\dagger}$ | | | | |
| $\theta = 0.001$ | 748 | 743 | 688 | 641 |
| $\theta = 0.050$ | 1136 | 1129 | 1045 | 972 |
| Additive$^{\dagger}$ | | | | |
| $\theta = 0.001$ | 793 | 793 | 730 | 676 |
| $\theta = 0.050$ | 1201 | 1202 | 1105 | 1023 |

$^{\ddagger}\tau_i = P(Q = 1|IBS_M = i)$ for $i = 1, 2$.
$^{\dagger}$Dominant: $f_2 = \gamma^2 f_0$, $f_1 = f_2$; Recessive: $f_2 = \gamma^2 f_0$, $f_1 = f_0$; Multiplicative: $f_2 = \gamma^2 f_0$, $f_1 = \gamma f_0$; Additive: $f_2 = \gamma^2 f_0$, $f_1 = \frac{1}{2}(f_0 + f_2)$ where penetrances $f_0 = 0.1$, $f_1$, and $f_2$ correspond to trait genotypes $dd$, $Dd$, and $DD$, $\gamma = 3$, and $P(D) = 0.20$.

$T$, and a dominant model with $\theta = 0.001$, Table 4 shows that the required numbers of affected sib-pairs are 523, 484, and 450, for $\tau_1 = \tau_2 = 0.0$, $\tau_1 = \tau_2 = 0.2$, and $\tau_1 = \tau_2 = 0.4$, respectively, as compared to 526 for the mean IBS statistic. According to the results presented in Table 4, the power of the linkage statistic, $T$, increases as the value of $\tau_i$ increases, and $T$ is more powerful than the mean IBS statistic for all of the nonzero values of $\tau_i$ that are considered.

Our asymptotic results also allow computation of required sample sizes for our $U$ statistic, so comparisons can be made for various levels of parent availability and with and without the assumption of unambiguous marker IBD for sib-pairs with available parents.

## 3.2 Simulations

In Section 3.2.1, we compare simulation results with asymptotic results from our methods and consider the impact of parent availability. In Section 3.2.2, we compare simulation results from our methods with results from current linkage software that implements Kong and Cox's (1997) methods for incomplete marker data.

Our simulation approach imitates the random biological process of gamete formation and transmission and uses penetrance values to determine which siblings are affected. From this simulated data, we only select affected sib-pairs; therefore, for each simulated data set, the number of families simulated to obtain N affected sib-pairs follows a negative binomial distribution. In Sections 3.2.1 and 3.2.2, we provide a detailed description of our simulation approach and the parameter values that we used.

### 3.2.1 Evaluation of simulation results, asymptotic results, and impact of incomplete marker information

The first simulation study evaluates the performance of the affected sib-pairs linkage statistics of Section 2.1 using finite sample sizes. For various genetic models and nuisance parameter values, the distribution of the statistic is evaluated under the null hypothesis, $H_0 : \theta = 0.5$, and under the alternative hypothesis, $H_1 : \theta = \theta_1$ where $\theta_1 = 0.001$ or 0.05.

Parental genotypes were simulated assuming that marker alleles $A_1, A_2, \ldots, A_J$ have equal allele frequencies, and trait alleles $D$ and $d_1, d_2, \ldots, d_K$ have allele frequencies $P(D)$ and $P(d_1), P(d_2), \ldots, P(d_K)$, respectively. For notational convenience, let $d$ represent any of the alleles $d_1, d_2, \ldots, d_K$; thus, $P(d) = \sum_{i=1}^{K} P(d_i) = 1 - P(D)$. Considering the genotypes of randomly mating parents and the specified recombination fraction, the formation and transmission of parental gametes was simulated to obtain the genotype for each of the siblings in a sib-pair. The marker IBS was obtained by comparing the marker genotypes in each sib-pair. Since the sib-pair genotypes were simulated from the parental genotypes, the true marker IBD was also available for each sib-pair. The observed marker IBD was obtained by considering the true IBD and assigning missing values with probability $1 - \tau_i$ where $i = 1, 2$.

To obtain affected sib-pairs, affection status was assigned using penetrances $f_0 = 0.1, f_1$, and $f_2$ where $f_1$ and $f_2$ are defined according to each genetic model considered including: dominant, recessive, multiplicative, and additive. These models correspond to the models considered in the previous section; therefore, the required sample sizes from Table 3 also appear in Tables 5 and 6.

*Table 5. Estimated (and asymptotic) mean and variances of linkage statistic $\frac{\tilde{T}}{\sqrt{N}}$ where $\tau_i^{\ddagger} = 0.8$*

| | N | $Var_{H_0}\left(\frac{\tilde{T}}{\sqrt{N}}\right)^{\sharp}$ | $E_{H_1}\left(\frac{\tilde{T}}{\sqrt{N}}\right)^{\flat}$ | $Var_{H_1}\left(\frac{\tilde{T}}{\sqrt{N}}\right)^{\flat}$ |
|---|---|---|---|---|
| Dominant$^{\dagger}$, $P(D) = 0.05$ | | | | |
| $\theta = 0.001$ | 196 | 0.4704 (0.4697) | 3.1180 (3.1013) | 0.4269 (0.4223) |
| $\theta = 0.050$ | 298 | 0.4700 (0.4697) | 3.1147 (3.1097) | 0.4481 (0.4382) |
| Recessive$^{\dagger}$, $P(D) = 0.20$ | | | | |
| $\theta = 0.001$ | 193 | 0.4702 (0.4697) | 3.1457 (3.1448) | 0.5001 (0.4887) |
| $\theta = 0.050$ | 291 | 0.4699 (0.4697) | 3.1386 (3.1353) | 0.4843 (0.4822) |
| Multiplicative$^{\dagger}$, $P(D) = 0.20$ | | | | |
| $\theta = 0.001$ | 563 | 0.4698 (0.4697) | 3.1293 (3.1214) | 0.4611 (0.4606) |
| $\theta = 0.050$ | 853 | 0.4698 (0.4697) | 3.1270 (3.1235) | 0.4646 (0.4636) |
| Additive$^{\dagger}$, $P(D) = 0.05$ | | | | |
| $\theta = 0.001$ | 568 | 0.4698 (0.4697) | 3.1121 (3.1150) | 0.4656 (0.4520) |
| $\theta = 0.050$ | 861 | 0.4698 (0.4697) | 3.1230 (3.1190) | 0.4609 (0.4579) |

$^{\sharp}$Estimated by $\frac{\sum_{k=1}^{B} \hat{V}_{0k}}{B}$ where $\hat{V}_{0k}$ is the estimate of $Var_{H_0}\left(\frac{\tilde{T}}{\sqrt{N}}\right)$ and $B = 10{,}000$.

$^{\flat}$Estimated by empirical mean and variance given by $\frac{\bar{\tilde{T}}}{\sqrt{N}} = \frac{\sum_{k=1}^{B} \tilde{T}_k}{B\sqrt{N}}$ and $\frac{\sum_{k=1}^{B} (\tilde{T}_k - \bar{\tilde{T}})^2}{(B-1)N}$, respectively.

$^{\ddagger}\tau_i = P(Q = 1 | IBS_M = i)$ for $i = 1, 2$.

$^{\dagger}$Dominant: $f_2 = \gamma^2 f_0, f_1 = f_2$; Recessive: $f_2 = \gamma^2 f_0, f_1 = f_0$; Multiplicative: $f_2 = \gamma^2 f_0, f_1 = \gamma f_0$; Additive: $f_2 = \gamma^2 f_0, f_1 = \frac{1}{2}(f_0 + f_2)$ where penetrances $f_0 = 0.1$, $f_1$, and $f_2$ correspond to trait genotypes $dd$, $Dd$, and $DD$ and $\gamma = 3$.

*Table 6. $IBD_M$ known (i.e. $\tau_1 = \tau_2 = 1.0$) vs. $IBD_M$ incomplete: The estimated power (and expected power) for a one-sided $\alpha = 0.0001$ level test that uses missing at random IBD and known IBS*

| | N | $T$ Statistic with $\tau_i = 1.0^{\ddagger}$ | $T$ Statistic with $\tau_i = 0.8$ | $\tilde{T}$ Statistic with $\tau_i = 0.8$ |
|---|---|---|---|---|
| Dominant$^{\dagger}$, $P(D) = 0.05$ | | | | |
| $\theta = 0.001$ | 196 | 0.8558 (0.8420) | 0.8085 (0.8024) | 0.8081 (0.8024) |
| $\theta = 0.050$ | 298 | 0.8398 (0.8406) | 0.8047 (0.8016) | 0.8049 (0.8016) |
| Recessive$^{\dagger}$, $P(D) = 0.20$ | | | | |
| $\theta = 0.001$ | 193 | 0.8296 (0.8298) | 0.8008 (0.8031) | 0.8001 (0.8031) |
| $\theta = 0.050$ | 291 | 0.8386 (0.8300) | 0.8064 (0.8009) | 0.8031 (0.8009) |
| Multiplicative$^{\dagger}$, $P(D) = 0.20$ | | | | |
| $\theta = 0.001$ | 563 | 0.8369 (0.8368) | 0.8027 (0.8006) | 0.8038 (0.8006) |
| $\theta = 0.050$ | 853 | 0.8462 (0.8372) | 0.8064 (0.8007) | 0.8059 (0.8007) |
| Additive$^{\dagger}$, $P(D) = 0.05$ | | | | |
| $\theta = 0.001$ | 568 | 0.8387 (0.8390) | 0.7940 (0.8002) | 0.7961 (0.8002) |
| $\theta = 0.050$ | 861 | 0.8404 (0.8388) | 0.8060 (0.8003) | 0.8045 (0.8003) |

$^{\ddagger}\tau_i = P(Q = 1 | IBS_M = i)$ for $i = 1, 2$.

$^{\dagger}$Dominant: $f_2 = \gamma^2 f_0, f_1 = f_2$; Recessive: $f_2 = \gamma^2 f_0, f_1 = f_0$; Multiplicative: $f_2 = \gamma^2 f_0, f_1 = \gamma f_0$; Additive: $f_2 = \gamma^2 f_0, f_1 = \frac{1}{2}(f_0 + f_2)$ where penetrances $f_0 = 0.1$, $f_1$, and $f_2$ correspond to trait genotypes $dd$, $Dd$, and $DD$ and $\gamma = 3$.

Considering a nominal $\alpha$ level of 0.0001, the size of the test is evaluated using 100,000 data sets simulated under the null hypothesis with a sample size of $N = 300$. Assuming 8 equally likely marker alleles, the estimated size of the test, based on the $\tilde{T}$ statistic with $\tau_1 = \tau_2 = 0.8$, is $0.9 \times 10^{-4}$ as compared to $1.3 \times 10^{-4}$ for the ideal situation where $\tau_1 = \tau_2 = 1.0$. Assuming only 4 equally likely marker alleles, the estimated size of the test, based on the $\tilde{T}$ statistic with $\tau_1 = \tau_2 = 0.8$, is $0.7 \times 10^{-4}$ as compared to $1.3 \times 10^{-4}$ for the situation where $\tau_1 = \tau_2 = 1.0$. Considering various values of $\tau_i$ and various numbers of marker alleles and sample sizes, data sets simulated under the null hypothesis demonstrated that the estimates of $Var_{H_0}\left(\frac{\tilde{T}}{\sqrt{N}}\right)$ and the estimates of the nuisance parameters are very close to the true parameter

values. For example, considering 10,000 data sets of $N = 100$ affected sib-pairs with $\tau_1 = \tau_2 = 0.8$ and 8 marker alleles with allele frequencies $\frac{1}{8}$, the simulated null variance estimate is 0.4702, and when $N = 500$ the simulated null variance estimate is 0.4698, as compared to the asymptotic null variance value of 0.4697.

Table 5 provides the estimated and asymptotic mean and variances of the linkage statistic $\frac{\tilde{T}}{\sqrt{N}}$ for the genetic models considered in Table 3. Considering the specified genetic model with $\tau_1 = \tau_2 = 0.8$ and 8 equally likely marker alleles, 10,000 simulated data sets were generated to obtain estimates for each row of Table 5. Estimates for the null variance were calculated as $\frac{\sum_{k=1}^{B} \hat{V}_{0k}}{B}$ where $\hat{V}_{0k}$ is the estimate of $Var_{H_0}\left(\frac{\tilde{T}}{\sqrt{N}}\right)$ and $B = 10{,}000$. When rounded to 3 decimal places, these null variance estimates are identical to the asymptotic value of 0.470. Considering the alternative hypothesis, each row of mean and variance estimates was calculated from the empirical distribution that corresponds to that row. The empirical mean and variance are $\frac{\bar{\tilde{T}}}{\sqrt{N}} = \frac{\sum_{k=1}^{B} \tilde{T}_k}{B\sqrt{N}}$ and $\frac{\sum_{k=1}^{B} \left(\tilde{T}_k - \bar{\tilde{T}}\right)^2}{(B-1)N}$, respectively. For the dominant model with $\theta = 0.001$, the empirical mean and variance are 3.1180 and 0.4269 as compared to the asymptotic values of 3.1013 and 0.4223, respectively.

Table 6 provides the estimated power and expected power for the one-sided $\alpha = 0.0001$ level test that allows the marker IBD to be missing at random when marker IBS is 1 or 2. Each row of estimates was calculated from 10,000 simulated data sets, the same data sets used for the corresponding row of Table 5. Table 6 shows that the estimated power values are in close agreement with the expected power values. The estimated power for the $T$ and $\tilde{T}$ statistics, with $\tau_1 = \tau_2 = 0.8$, are nearly identical, and the power increases as the proportion of available IBD values increases to $\tau_1 = \tau_2 = 1.0$. For example, considering a multiplicative model with $\theta = 0.001$ and $P(D) = 0.20$, the estimated power for the $\tilde{T}$ statistic with $\tau_1 = \tau_2 = 0.8$ is 0.8038, and the estimated power when $\tau_1 = \tau_2 = 1.0$ is 0.8369.

#### 3.2.2 Simulation results comparing $U$ statistic with linkage analysis software

We performed additional simulations to compare the estimated power for our $U$ statistic with the estimated power from MERLIN linkage analysis software. To perform each comparison, we generated 500 simulated data sets using the approach introduced in Section 3.2.1. The MERLIN results were obtained using the simulated data sets as input data and using the "npl" and "fe" options and the p-values calculated from the Kong and Cox (1997) LOD scores.

For our first comparison, we considered a marker locus with 8 equally likely marker alleles, a recombination fraction of $\theta = 0.05$, and an additive genetic model with $f_2 = 9f_0,\ f_1 = \frac{1}{2}(f_0 + f_2),\ f_0 = 0.1$, and $P(D) = 0.05$, and we assumed that 80% of the parents were available. The expected power was 0.93 for the one-sided $\alpha = 0.001$ level test, and our simulation yielded estimated power results of 0.94 from our $U$ statistic and 0.94 from the MERLIN software.

For our second comparison, we considered a marker locus with 8 equally likely marker alleles, a recombination fraction of $\theta = 0.001$, and a dominant genetic model with $f_2 = 9f_0,\ f_1 = f_2,\ f_0 = 0.1$, and $P(D) = 0.2$, and we assumed that 80% of the parents were available. The expected power was 0.93 for the one-sided $\alpha = 0.001$ level test, and our simulation yielded estimated power results of 0.93 from our $U$ statistic and 0.93 from the MERLIN software.

## 4. DISCUSSION

Buckman (2005) derived the $U$, $T$, and $\tilde{T}$ linkage statistics and provided additional details and numeric results and derivations of similar linkage statistics for affected relative-pairs and extreme discordant sib-pairs.

In our derivations, we allow the parental genotypes to be missing at random, and we consider the implications of this MAR assumption for the tabulated genetic data where $IBS_M = 0$ implies that $IBD_M$ also equals zero. Considering missing data methods for contingency tables allows us to specify the appropriate multinomial distributions for the tabulated data presented in Sections 2.1 and 2.2. In Section 2.1, we assume the marker IBD is unambiguous for an affected sib-pair if the parental genotypes are available. Then, we drop this assumption in Section 2.2 and derive the $U$ statistic. Our $U$ statistic provides a linkage test for affected sib-pairs with marker IBS known and marker IBD estimated from sib-pair and parental genotypes where the parental genotypes may be MAR. In Section 3.2.2, we provide some simulation results that demonstrate that our $U$ statistic had the same estimated power as obtained from MERLIN linkage analysis software using the "npl" and "fe" options and using p-values that we calculated directly from the Kong and Cox (1997) LOD scores in order to gain better precision for p-values that were very close to the $\alpha$ level.

Asymptotic results for the $U$ statistic are derived using the same approach as presented in Section 2.1 because the statistics, $U$ and $T$, are each defined as a linear combination of a random multinomial vector. Since we considered the missing data situation in the derivation of our statistics and asymptotic results, the impact of missing parents can be ascertained directly by using our asymptotic results to calculate expected power values or required sample sizes. In addition to asymptotic results, we have also presented simulation results to evaluate our methods with finite sample sizes and compare the performance of our methods with various levels of missing data. Since our $U$ statistic allows for marker ambiguity, and therefore, relaxes the assumption made for the $T$ and $\tilde{T}$ statistics, these statistics can be used together to consider the impact of marker ambiguity.

## APPENDIX A. DERIVATION OF PROBABILITIES IN TABLE 1

According to the missing at random assumption,

$$P(Q=1|IBS_M=i)=P(Q=1|IBS_M=i,IBD_M=j)$$

for $i,j=0,1,2$ and $i\geq j$. Also, assume $Q$ is conditionally independent of the affection status of the sib-pair, given the marker IBS and IBD. Therefore,

$$\begin{aligned}(3)\qquad &P(Q=1|IBS_M=i)\\ &\quad=P(Q=1|IBS_M=i,IBD_M=j,X=2)\end{aligned}$$

for $i,j=0,1,2$ and $i\geq j$.

Also assume that the marker IBS is conditionally independent of the affection status, given the marker IBD. Then applying Equation 3 and combining results from Risch (1990b) and Bishop and Williamson (1990), yields the following probabilities for the $Q=1$ cells of Table 1,

$$\begin{aligned}&P(Q=1,IBS_M=i,IBD_M=j|X=2)\\ &\quad=P(Q=1|IBS_M=i,IBD_M=j,X=2)\\ &\qquad\cdot P(IBS_M=i,IBD_M=j|X=2)\\ &\quad=\tau_i T_{ij}Z_{S_j}(\theta)\end{aligned}$$

for $i,j=0,1,2$ and $i\geq j$. Similarly, the probabilities for the $Q=0$ cells of Table 1 are

$$\begin{aligned}&P(Q=0,IBS_M=i|X=2)\\ &\quad=\sum_{j=0}^{i}[1-P(Q=1|IBS_M=i,IBD_M=j,X=2)]\\ &\qquad\cdot P(IBS_M=i,IBD_M=j|X=2)\\ &\quad=(1-\tau_i)\sum_{j=0}^{i}T_{ij}Z_{S_j}(\theta)\quad\text{for}\quad i=1,2.\end{aligned}$$

## APPENDIX B. DERIVATION OF MEAN AND VARIANCE OF T

The multinomial random vector $M$ has parameter vector

$$\begin{aligned}\Pi=\big(&T_{00}Z_{S_0}(\theta),\tau_1T_{10}Z_{S_0}(\theta),\tau_2T_{20}Z_{S_0}(\theta),\\ &\tau_1T_{11}Z_{S_1}(\theta),\tau_2T_{21}Z_{S_1}(\theta),\tau_2Z_{S_2}(\theta),\\ &(1-\tau_1)[T_{10}Z_{S_0}(\theta)+T_{11}Z_{S_1}(\theta)],\\ &(1-\tau_2)[T_{20}Z_{S_0}(\theta)+T_{21}Z_{S_1}(\theta)+Z_{S_2}(\theta)]\big)'\end{aligned}$$

where $0<\theta\leq\frac{1}{2}$. Since $T=C'M$, the mean and variance of $T$ are $C'E(M)$ and $C'Var(M)C$, respectively. Let $N\Pi_k$ and $N\Sigma_k$ represent the mean vector and covariance matrix of $M$ under $H_k$ where the alternative hypothesis is $H_1:0<\theta<\frac{1}{2}$ and the null hypothesis is $H_0:\theta=\frac{1}{2}$. Therefore, under $H_1$, the mean and variance of $T$ are

$$\begin{aligned}E_{H_1}(T)&=NC'\Pi_1\\ &=N\Big(\tau_2Z_{S_2}(\theta)-T_{00}Z_{S_0}(\theta)-\tau_1T_{10}Z_{S_0}(\theta)\\ &\quad-\tau_2T_{20}Z_{S_0}(\theta)+(1-\tau_2)\\ &\quad\cdot[T_{20}Z_{S_0}(\theta)+T_{21}Z_{S_1}(\theta)+Z_{S_2}(\theta)]\frac{1-T_{20}}{T_{20}+2T_{21}+1}\\ &\quad-(1-\tau_1)[T_{10}Z_{S_0}(\theta)+T_{11}Z_{S_1}(\theta)]\frac{T_{10}}{T_{10}+2T_{11}}\Big)\end{aligned}$$

and

$$\begin{aligned}Var_{H_1}(T)&=NC'\Sigma_1C\\ &=N\Big([T_{00}Z_{S_0}(\theta)+\tau_1T_{10}Z_{S_0}(\theta)+\tau_2T_{20}Z_{S_0}(\theta)]\\ &\quad\cdot(1-[T_{00}Z_{S_0}(\theta)+\tau_1T_{10}Z_{S_0}(\theta)+\tau_2T_{20}Z_{S_0}(\theta)])\\ &\quad+\tau_2Z_{S_2}(\theta)[1-\tau_2Z_{S_2}(\theta)]\\ &\quad+\left[\frac{T_{10}}{T_{10}+2T_{11}}\right]^2(1-\tau_1)[T_{10}Z_{S_0}(\theta)+T_{11}Z_{S_1}(\theta)]\\ &\quad\cdot(1-(1-\tau_1)[T_{10}Z_{S_0}(\theta)+T_{11}Z_{S_1}(\theta)])\\ &\quad+\left[\frac{1-T_{20}}{T_{20}+2T_{21}+1}\right]^2(1-\tau_2)\\ &\quad\cdot[T_{20}Z_{S_0}(\theta)+T_{21}Z_{S_1}(\theta)+Z_{S_2}(\theta)]\\ &\quad\cdot(1-(1-\tau_2)[T_{20}Z_{S_0}(\theta)+T_{21}Z_{S_1}(\theta)+Z_{S_2}(\theta)])\Big)\\ &\quad+2N\Big([T_{00}Z_{S_0}(\theta)+\tau_1T_{10}Z_{S_0}(\theta)+\tau_2T_{20}Z_{S_0}(\theta)]\tau_2Z_{S_2}(\theta)\\ &\quad+(\tau_2Z_{S_2}(\theta)-[T_{00}Z_{S_0}(\theta)+\tau_1T_{10}Z_{S_0}(\theta)+\tau_2T_{20}Z_{S_0}(\theta)])\\ &\quad\cdot\left[\frac{T_{10}}{T_{10}+2T_{11}}(1-\tau_1)[T_{10}Z_{S_0}(\theta)+T_{11}Z_{S_1}(\theta)]\right.\\ &\quad-\frac{1-T_{20}}{T_{20}+2T_{21}+1}(1-\tau_2)\\ &\quad\left.\cdot[T_{20}Z_{S_0}(\theta)+T_{21}Z_{S_1}(\theta)+Z_{S_2}(\theta)]\right]+\left[\frac{T_{10}}{T_{10}+2T_{11}}\right]\\ &\quad\cdot\left[\frac{1-T_{20}}{T_{20}+2T_{21}+1}\right](1-\tau_1)[T_{10}Z_{S_0}(\theta)+T_{11}Z_{S_1}(\theta)]\\ &\quad\cdot(1-\tau_2)[T_{20}Z_{S_0}(\theta)+T_{21}Z_{S_1}(\theta)+Z_{S_2}(\theta)]\Big),\end{aligned}$$

respectively. Similarly, under $H_0$ the mean and variance of $T$ are $E_{H_0}(T)=NC'\Pi_0=0$ and

$$\begin{aligned}Var_{H_0}(T)&=N\left[C'Diag(\Pi_0)C-(C'\Pi_0)^2\right]\\ &=NC'Diag(\Pi_0)C\end{aligned}$$

where $Diag(\Pi_0)$ is a diagonal matrix with $m^{th}$ element of $\Pi_0$ as $m^{th}$ diagonal entry.

## APPENDIX C. DERIVATION OF ASYMPTOTIC DISTRIBUTION OF $N^{-1}T$ AND $N^{-1}\tilde{T}$

The asymptotic distribution of $N^{-1}T$ is derived. Using convergence of moment generating functions, Bishop, Fienberg, and Holland (1975, pp. 469–470) demonstrate the asymptotic normality of the multinomial distribution. Thus, as $N \to \infty$, $N^{-1}M$ converges in distribution to $Z$ where $Z = (Z_1, Z_2, \ldots, Z_8)$ has a multivariate normal distribution with the same mean vector and covariance matrix as $N^{-1}M$. Considering $g\left(N^{-1}M\right) = N^{-1}C'M = N^{-1}T$ and applying Rao's (1973, p. 124) result xii, yields

$$g\left(N^{-1}M\right) \xrightarrow{D} g\left(Z\right) \qquad \text{as} \qquad N \to \infty.$$

The mean and variance of $g(Z) = C'Z$ are obtained by applying Rao's (1973, p. 107) result concerning the mean and variance of a linear combination, and many authors, such as Rohatgi (1976, p. 234), have demonstrated that a linear combination of a multivariate normal vector yields a normal random variable. Therefore, it is established that the asymptotic distribution of $g\left(N^{-1}M\right) = N^{-1}T$ is

$$N\left(C'\Pi,\ N^{-1}C'\Sigma C\right).$$

Next, the asymptotic distribution of $N^{-1}\tilde{T}$ is derived. Since maximum likelihood estimators are consistent, $\hat{\tau}_i$ converges in probability to the nuisance parameter, $\tau_i$, for $i = 1, 2$, and $(\hat{T}_{00}, \hat{T}_{10}, \hat{T}_{20}, \hat{T}_{11}, \hat{T}_{21})$ converges in probability to the nuisance parameter vector, $(T_{00}, T_{10}, T_{20}, T_{11}, T_{21})$. Similarly, the variance estimator, $\hat{\sigma}_0^2$, converges in probability to $\sigma_0^2$. Therefore, applying Slutsky's theorem (Lachin, 2000), the linkage statistic $N^{-1}\tilde{T}$ converges in distribution to $N^{-1}T$ as $N \to \infty$, and under $H_0$, the standardized statistic, $\tilde{T}\hat{\sigma}_0^{-1}$, is asymptotically distributed as $N(0, 1)$.

*Received 9 September 2008*

## REFERENCES


Abecasis, G. R., Cherny, S. S., Cookson, W. O. and Cardon, L. R. (2002). MERLIN – rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.* **30** 97–101.

Bishop, D. T. and Williamson, J. A. (1990). The power of identity-by-state methods for linkage analysis. *Am. J. Hum. Genet.* **46** 254–265.

Bishop, Y. M. M., Fienberg, S. E., and Holland P. W. (1975). *Discrete multivariate analysis – theory and practice.* The MIT Press, Cambridge. MR0381130

Buckman, D. W. (2005). *Linkage tests for relative-pairs with incomplete IBD and known IBS.* Ph.D. Dissertation. Department of Statistics, George Washington University, Washington, DC.

Chen, T. and Fienberg, S. E. (1974). Two-dimensional contingency tables with both completely and partially cross-classified data. *Biometrics* **30** 629–642. MR0403086

Dudoit, S. (1999). *Linkage analysis of complex human traits using identity by descent data.* Ph.D. Dissertation. Department of Statistics, University of California, Berkeley.

Dudoit, S. and Speed, T. P. (2000). A score test for the linkage analysis of qualitative and quantitative traits based on identity by descent data from sib-pairs. *Biostatistics* **1** 1–26.

Feingold, E. (2001). Methods for linkage analysis of quantitative trait loci in humans. *Theor. Popul. Biol.* **60** 167–180.

Feingold, E. (2002). Regression-based quantitative-trait-locus mapping in the 21st Century. *Am. J. Hum. Genet.* **71** 217–222.

Haseman, J. K. (1970). *The genetic analysis of quantitative traits using twin and sib data.* Ph.D. Dissertation. University of North Carolina, Chapel Hill.

Haseman, J. K. and Elston, R. C. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.* **2** 3–19.

Holmans, P. (1993). Asymptotic properties of affected sib-pair linkage analysis. *Am. J. Hum. Genet.* **52** 362–374.

Kong, A. and Cox, N. J. (1997). Allele-sharing models: LOD scores and accurate linkage tests. *Am. J. Hum. Genet.* **61** 1179–1188.

Lachin, J. M. (2000). *Biostatistical methods – the assessment of relative risks.* John Wiley & Sons, New York. MR1783165

Lange, K. (1986). The affected sib-pair method using identity by state relations. *Am. J. Hum. Genet.* **39** 148–150.

Li, Z. and Gastwirth, J. L. (2001). A weighted test using both extreme discordant and concordant sib-pairs for detecting linkage. *Genet. Epidemiol.* **20** 34–43.

Nicolae, D. L., Meng, X. and Kong, A. (2008). Quantifying the fraction of missing information for hypothesis testing in statistical and genetic studies. *Statist. Sci.* **23** 287–312.

Nordheim, E. V. (1984). Inference from nonrandomly missing categorical data – an example from a genetic study on Turner's syndrome. *J. Amer. Statist. Assoc.* **79** 772–780.

Olson, J. M., Witte, J. S. and Elston, R. C. (1999). Tutorial in biostatistics – genetic mapping of complex traits. *Stat. Med.* **18** 2961–2981.

Rao, C. R. (1973). *Linear statistical inference and its applications.* John Wiley & Sons, New York. MR0346957

Risch, N. (1990a). Linkage strategies for genetically complex traits. I. multilocus models. *Am. J. Hum. Genet.* **46** 222–228.

Risch, N. (1990b). Linkage strategies for genetically complex traits. II. the power of affected relative pairs. *Am. J. Hum. Genet.* **46** 229–241.

Risch, N. (1990c). Linkage strategies for genetically complex traits. III. the effect of marker polymorphism on analysis of affected relative pairs. *Am. J. Hum. Genet.* **46** 242–253.

Risch, N. (1992). Corrections to "Linkage strategies for genetically complex traits. III. the effect of marker polymorphism on analysis of affected relative pairs" (*Am. J. Hum. Genet.* **46** 242–253, 1990). *Am. J. Hum. Genet.* **51** 673–675.

Risch, N. and Zhang, H. (1995). Extreme discordant sib-pairs for mapping quantitative trait loci in humans. *Science* **268** 1584–1589.

Risch, N. and Zhang, H. (1996). Mapping quantitative trait loci with extreme discordant sib-pairs – sampling considerations. *Am. J. Hum. Genet.* **58** 836–843.

Rohatgi, V. K. (1976). *An introduction to probability theory and mathematical statistics.* John Wiley & Sons, New York. MR0407916

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. MR0455196

Suarez, B. K., Rice, J. and Reich, T. (1978). The generalized sib-pair IBD distribution – its use in the detection of linkage. *Ann. Hum. Genet.* **42** 87–94.

Thomson, G. and Motro, U. (1994). Affected sib-pair identity by state analyses. *Genetic Epidemiology* **11** 353–364.

Weeks, D.E. and Lange, K. (1988). The affected-pedigree-member method of linkage analysis. *Am. J. Hum. Genet.* **42** 315–326.

Zhang, H. and Risch, N. (1996). Mapping quantitative-trait loci in humans by use of extreme concordant sib-pairs – selected sampling by parental phenotypes. *Am. J. Hum. Genet.* **59** 951–957.

Dennis W. Buckman
Information Management Services, Inc.
2501 Prosperity Dr., Suite 200
Silver Spring, MD 20904
and
Department of Statistics
George Washington University
2140 Pennsylvania Ave., N.W.
Washington, DC 20052
E-mail address: buckmand@imsweb.com

Zhaohai Li
Department of Statistics
George Washington University
2140 Pennsylvania Ave., N.W.
Washington, DC 20052
and
Biostatistics Branch
Division of Cancer Epidemiology and Genetics
National Cancer Institute
Rockville, MD 20892
E-mail address: zli@gwu.edu