Statistica Sinica **12**(2002), 297-309

A UNIFIED MULTIPOINT LINKAGE ANALYSIS OF QUALITATIVE AND QUANTITATIVE TRAITS FOR SIB-PAIRS

I-Shou Chang[1,2], Miao-Ying Chen[2], Chin-Fu Hsiao[2] and Chao A. Hsiung[2]

[1]*National Central University and* [2]*National Health Research Institutes*

Abstract: By introducing functions of the phenotypes of a sib-pair as weight functions in the study of IBD processes, we present a unified non-parametric approach to linkage analysis of qualitative and quantitative traits in sib-pairs based on IBD data obtained from a set of polymorphic markers. With the introduction of weight functions and an appropriate conditional expectation of IBD processes, these statistical methods should be more efficient in the detection of genetic factors for complex diseases. These methods will be also useful in planning genetic studies. Large sample properties of these methods are demonstrated.

Key words and phrases: Genome-wide scan, IBD process, linkage analysis, quantitative traits, sib-pairs.

1. Introduction

Most traditional methods for the analysis of quantitative trait loci (QTL) are based on examining phenotypes conditional on IBD (identity by descent) sharing, see for example Haseman and Elston (1972), Amos (1994) and Kruglyak and Lander (1995). Noticing the low power of these standard QTL methods to detect linkage to loci influencing complex traits, Risch and Zhang (1995, 1996) recommended ascertaining sib-pairs with extremely discordant or highly concordant phenotypes, and making analysis conditional on phenotypes so as to obtain the greatest power to detect linkage.

Recently, Dudoit and Speed (2000) considered the likelihood of IBD data from sib-pairs conditional on the phenotypes of the pairs, and studied the problem of testing the null hypothesis of no linkage between a marker locus and a gene influencing the trait using a score test in the recombination fraction between the two loci. This approach unifies the linkage analysis of qualitative and quantitative traits into a single inferential framework and allows the selection of sib-pairs based on their trait values and the analysis of only those pairs having the most informative phenotypes. An extension of this score-test statistic can be used in the context of a genome scan to test for linkage at loci which are not necessarily typed marker loci.

Another interesting development is due to Liang, Huang, and Beaty (2000), who proposed a unified sampling approach for multipoint mapping of genes for both qualitative and quantitative traits in sib-pairs. Their method builds on a parametric representation for the expected IBD statistic at an arbitrary locus, conditional on an event reflecting the sampling scheme, such as affected sib-pairs for qualitative traits, or extremely discordant sib-pairs for quantitative traits.

In this paper, we present a unified non-parametric approach to linkage analysis of qualitative and quantitative traits in sib-pairs based on IBD data obtained from a set of polymorphic markers. Our approach can be regarded as an extension of the non-parametric allele-sharing methods for qualitative traits in sib-pairs. The extension is made possible by considering functions of the phenotypes of a sib-pair as weight functions in the study of the IBD process.

We study two non-parametric statistical methods. The first consists of a class of tests for the null hypothesis that the phenotypes of a sib-pair are independent of the IBD process. If the null hypothesis is rejected, we provide an estimate to indicate the region of enriched or diluted identity of descent on a chromosome. By choosing appropriate weight functions, these methods can be used for linkage analysis.

We note that statistical methods for genome-wide search for a trait locus using relative pairs were studied by Feingold (1993), Feingold, Brown and Siegmund (1993), Dupuis, Brown and Siegmund (1995), Feingold and Siegmund (1997), and Tu and Siegmund (1999) for qualitative traits, and by Goldgar (1990), Fulker and Cardon (1994), and Guo (1994) for quantitative traits. The methods for quantitative traits are different from ours in that they are extensions of Haseman and Elston (1972) and based on examining phenotypes conditional on IBD sharing, while our work is more in line with the methods for qualitative traits. Even for qualitative traits, our approach is different in the assumptions on the IBD process. In particular, we make no assumption on the map functions.

We need some notation to facilitate the discussion. For $k = 1, 2, \ldots$, let (I_k, M_k, ϕ_k) be i.i.d. random elements with $I_k(t), t \in [0, 1]$, being the IBD process on a particular chromosome region, M_k being the set of genotypes taken at certain markers in this region, and ϕ_k being the phenotype values for the kth sib-pair, respectively. Here, for simplicity, we assume the chromosome region is parametrized by $[0, 1]$.

Since the IBD process $I_k(t)$ is rarely observable, it is natural to consider the conditional expectation of $I_k(t)$ given M_k, $E(I_k(t)|M_k)$, as a substitute. For the calculation of $E(I_k(t)|M_k)$, one still needs to know the distribution of the IBD process $I_k(\cdot)$. For example, Kruglyak and Lander (1995) assume $I_k(\cdot)$ is a stationary Markov chain and calculate $E(I_k(t)|M_k)$ by means of computational techniques developed for hidden Markov models. With the recent availability of

linkage databases (cf. Bishop (1999), pp. 47-48), we can get an estimate of the distribution of the IBD process. In this paper we propose to use the distribution of $I_k(\cdot)$, as estimated by Chang, Wang, Chen and Hsiung (2001) using one such linkage database, as the prior distribution and to calculate $E(I_k(t)|M_k)$ using Bayes Theorem. Further elaboration on this approach is given in Section 4, where a remark on the computational feasibility is included. In view of the chiasma interference reported by Broman and Weber (2000), the crossover process on a meiosis product, and hence the IBD process, may not be stationary Markov. We think our approach to $E(I_k(t)|M_k)$ is a useful alternative.

Let

$$G_K(t) \equiv \frac{1}{\sqrt{K}} \sum_{k=1}^{K} \sum_{j=1}^{J} a_j W_j(\phi_k) E(I_k(t) - 1|M_k). \tag{1.1}$$

Here W_j is a weight function and $\sum_{j=1}^{J} a_j^2 = 1$. We show that G_K converges weakly to a mean zero Gaussian process, indexed by $a = (a_1, \cdots, a_J)$ and t, under the null hypothesis. This suggests that we reject the null hypothesis if $\sup_{a,t} |G_K|$ is large. In case $\sum_j a_j W_j \geq 0$ gives more weight to highly concordant sib-pairs, we would reject the hypothesis that there is no trait locus in the region if $\sup_t G_K(t)$ is large. In case $\sum_j a_j W_j \geq 0$ gives more weight to extremely discordant sib-pairs, we would reject this hypothesis if $\inf_t G_K(t)$ is a large negative number.

To estimate the location of a gene influencing the trait when there is one on the chromosome, we consider the M-estimator τ_K maximizing the empirical criterion function

$$\mathcal{P}_K m(t) \equiv \frac{1}{K} \sum_{k=1}^{K} \sum_{j=1}^{J} (a_j W_j(\phi_k))^2 E((I_k(t) - 1)^2|M_k). \tag{1.2}$$

Weight functions are useful in both the testing and estimation problems. We can increase the power of the test in certain situations, and decrease the variance of the estimate, by putting more emphasis on sib-pairs having more informative phenotypes. In fact, for a given set of data, we can even use our statistics to indicate the sib-pairs having more informative phenotypes. See Section 3.

In view of the concerns over the low power of linkage analysis to detect genetic factors for complex diseases (cf. Risch and Merikangas (1996)), we hope the introduction of weight functions and the more appropriate conditional expectation of IBD processes will result in more efficient linkage analyses.

The paper is organized as follows. Section 2 studies the linkage analysis under the idealistic assumption that the IBD process is observable. This helps us to clarify certain arguments and to simplify the presentation. The only assumptions we make on the IBD process are the existence of an upper bound on the total number of discontinuities of the process, and the existence of bounded densities

for the joint distributions of these discontinuities. For each procedure, we present an asymptotic distribution result that is proved by empirical process theory (cf. van der Vaart and Wellner (1996)).

Section 3 treats linkage analysis under the more realistic assumption that only the phenotype values and the genotypes at each of a set of polymorphic markers are available. The asymptotic distributions of the test statistic and the estimator are established by refining the arguments in Section 2. For the test statistic, we note that a bootstrap method may be used to obtain an approximate p-value. Section 4 introduces two methods to calculate $E(I(t)|M)$. Section 5 is a brief discussion, indicating possible extensions in terms of the informativeness of the phenotypes. Readers are referred to McPeek (1996), Speed (1996), Hauser and Boehnke (1998), Sham (1998), and Ott (1999) for an introduction to recombination and linkage analysis.

2. Inference Based on IBD Process

This section proposes statistical methods for linkage analysis under the (idealistic) assumption that the IBD process is itself observable. We produce a test statistic for the null hypothesis that the genes influencing the traits are not in the chromosome region under study, and an estimator of the map position of an unobserved susceptibility gene under the assumption of some preliminary evidence of linkage in the region.

In order to describe the IBD process precisely, we introduce the following notation. Let $0 = T_0 \le T_1 \le \cdots \le T_{n_0}$ be a sequence of random variables, and suppose $(T_1, \ldots, T_{n_0})$ has a bounded density on $T^{n_0} \equiv \{(t_1, \ldots, t_{n_0}) | 0 \le t_1 \le \cdots \le t_{n_0}\}$ relative to Lebesgue measure on $\mathcal{R}^{n_0}$.

Let $X_1, \ldots, X_{n_0+1}$ be a sequence of $\{0,1,2\}-$valued random variables with $P(X_{i+1} = 1|X_i = 2) = P(X_{i+1} = 1|X_i = 0) = 1$, $P(X_{i+1} = 2|X_i = 1) = P(X_{i+1} = 0|X_i = 1) = 1/2$, $i = 1, 2, \ldots, n_0$. Assume $(T_1, \ldots, T_{n_0})$ and $(X_1, \ldots, X_{n_0+1})$ are independent.

Define $I(t) = \sum_{i=1}^{n_0+1} X_i 1_{[T_{i-1}, T_i)}(t)$, where $T_{n_0+1} = \infty$. Our IBD process is $I(t)$, $t \in [0,1]$. Let ϕ be a random vector valued in $[0,1]^2$. We use X_i to represent the number of IBD sharing at the ith discontinuity T_i of the process if $T_i < 1$. Here we denote the phenotype of a sib-pair by a point in $[0,1]^2$. Let (I_k, ϕ_k) be a sequence of independent random elements with distribution identical to that of (I, ϕ).

2.1. A class of test statistics

Consider the stochastic process

$$\bar{G}_K(t, a_1, \ldots, a_J) \equiv \frac{1}{\sqrt{K}} \sum_{k=1}^{K} \sum_{j=1}^{J} a_j W_j(\phi_k)(I_k(t) - 1), \tag{2.1}$$

where $W_j : [0,1]^2 \to \mathcal{R}$ is a bounded function and the a_j satisfy $\sum_{j=1}^J a_j^2 = 1$. Let $S = \{(a_j, \ldots, a_J) | \sum_{j=1}^J a_j^2 = 1\}$. Let $\ell_\infty([0,1] \times S)$ denote the space of real-valued bounded functions on $[0,1] \times S$. In view of the following theorem, we propose to reject the hypothesis H_0 that $I(\cdot)$ and ϕ are independent if $\sup_{t,a} |\bar{G}_K|$ is too large.

Theorem 2.1. *If $E(I(t)|\phi) = 1$ for every $t \in [0,1]$, $\bar{G}_K(\cdot)$ converges weakly to a mean zero tight Gaussian process $\bar{G}(\cdot)$ in $\ell_\infty([0,1] \times S)$.*

Proof. The proof is an application of empirical process theory as presented, for example, in van der Vaart and Wellner (1996) (hereafter V and W). Since $\bar{G}_K$ is an i.i.d. sum of bounded random elements, it suffices to show that the relevant class of functions is Donsker in order to get the desired weak convergence.

Observe that, for $a > 0$,

$$\begin{aligned}
&\Big\{(T_1, \ldots, T_{n_0}, X_1, \ldots, X_{n_0+1}, s) | s < aI(t)\} \\
&= (T^{n_0} \times \{0,1,2\}^{n_0} \times (-\infty, 0)) \\
&\quad \bigcup \{(T_1, \ldots, T_{n_0}, X_1, \ldots, X_{n_0+1}, s) | s \geq 0, s < aI(t)\} \\
&= (T^{n_0} \times \{0,1,2\}^{n_0} \times (-\infty, 0)) \\
&\quad \bigcup_{i=1}^{n_0+1} \{(T_1, \ldots, T_{n_0}, X_1, \ldots, X_{n_0+1}, s) | s \geq 0, s < aX_i 1_{[T_{i-1}, T_i)}(t)\} \\
&= (T^{n_0} \times \{0,1,2\}^{n_0} \times (-\infty, 0)) \\
&\quad \bigcup_{i=1}^{n_0+1} \{(T_1, \ldots, T_{n_0}, X_1, \ldots, X_{n_0+1}, s) | s \geq 0, T_{i-1} \leq t < T_i, s < aX_i\}. \qquad (2.2)
\end{aligned}$$

It is straightforward to show that the class of sets $\{(T_1, \ldots, T_{n_0}, X_1, \ldots, X_{n_0+1}, s) | s \geq 0, T_{i-1} \leq t < T_i, s < aX_i\}$, $t \in [0,1]$ and $a \in \mathcal{R}$, is a VC-class (cf. Problem 2.6.14 in V and W). Hence we can conclude from (2.2) and the permanence property, Lemma 2.6.17 (iii) in V and W, that (2.2) indexed by $t \in [0,1]$ and $a \in \mathcal{R}$ is a VC-class. Therefore, the class of functions $(T_1, \ldots, T_{n_0}, X_1, \ldots, X_{n_0+1}, \phi) \to aI(t)$, indexed by $t \in [0,1]$ and $a \in \mathcal{R}$, is a VC-class. This together with the permanence property, Lemma 2.6.18 (vi) in V and W, that the class of functions

$$(T_1, \ldots, T_{n_0}, X_1, \ldots, X_{n_0+1}, \phi) \to a_j W_j(\phi) I(t), \qquad (2.3)$$

indexed by $t \in [0,1]$ and $a_j \in \mathcal{R}$, is a VC-class. We note also that (2.3) is a P-measurable class for every P, because $[0,1] \times \mathcal{R}$ has a countable dense subset (cf. Example 2.3.4 in V and W).

Since a VC-class satisfies the uniform entropy condition, we know (2.3) is Donsker (cf. Theorem 2.6.7 and Theorem 2.5.2 in V and W). It now follows from a permanence property for Donsker classes that the class of functions

$$(T_1, \ldots, T_{n_0}, X_1, \ldots, X_{n_0+1}, \phi) \to \sum_{j=1}^{J} a_j W_j(\phi) I(t), \tag{2.4}$$

indexed by $t \in [0,1]$, $a = (a_1, \ldots, a_J) \in S$, is Donsker (cf. Corollary 2.10.13 in V and W). This completes the proof.

2.2. A class of estimators

Let $\bar{m}_a(t)$ be the function defined by

$$\bar{m}_a(t)(T_1, \ldots, T_{n_0}, X_1, \ldots, X_{n_0+1}, \phi) = \sum_{j=1}^{J} (a_j W_j(\phi)(I(t) - 1))^2 \tag{2.5}$$

for $t \in [0,1]$, $a = (a_1, \ldots, a_J) \in S$. Because $(T_1, \ldots, T_{n_0})$ has a bounded density, we know the maps $t \to E\bar{m}_a(t)$ and $(t_1, t_2) \to E\bar{m}_a(t_1)\bar{m}_a(t_2)$ are twice continuously differentiable. Let

$$\mathcal{P}_K \bar{m}_a(t) \equiv \frac{1}{K} \sum_{k=1}^{K} \sum_{j=1}^{J} (a_j W_j(\phi_k)(I_k(t) - 1))^2 \tag{2.6}$$

and let $\tilde{\tau}_K$ be a maximizer of $\mathcal{P}_K \bar{m}_a(t)$. (A more illuminative notation for $\tilde{\tau}_K$ is $\tilde{\tau}_K(a)$.) Let $\bar{D} = \{(\tau, a) | a \in S, 0 < \tau < 1, E\bar{m}_a(t) \le E\bar{m}_a(\tau)$ for every $t \in [0,1]$ and $t \to E\bar{m}_a(t)$ has non-zero second derivative at $\tau\}$.

Theorem 2.2. *Assume $\bar{D}$ is non-empty and let $(\tau, a) \in \bar{D}$ be given. Then $\tilde{\tau}_K \to \tau$ in outer probability, and $\sqrt{K}(\tilde{\tau}_K - \tau)$ is asymptotically normal with mean 0 and variance $-(\frac{\partial^2}{\partial t^2} E\bar{m}_a(t)|_{t=\tau})^{-1} (\frac{\partial^2}{\partial t_1 \partial t_2} E\bar{m}_a(t_1)\bar{m}_a(t_2)|_{(t_1,t_2)=(\tau,\tau)})$.*

Proof. Using (2.4) and the arguments in the proof of Theorem 2.1, we know the class of functions $(T_1, \ldots, T_{n_0}, X_1, \ldots, X_{n_0+1}, \phi) \to \sum_{j=1}^{J} (a_j W_j(\phi)(I(t) - 1))^2$, indexed by $t \in [0,1]$, is Donsker. This implies that $\sup_{t \in [0,1]} |\mathcal{P}_K \bar{m}_a(t) - E\bar{m}_a(t)|$ converges to 0 in probability. This together with the Argmax Theorem shows that $\tilde{\tau}_K \to \tau$ in outer probability (cf. Corollary 3.2.3 in V and W).

We now establish the asymptotic normality by checking the conditions in Problem 3.2.1 of V and W, which concerns the asymptotic normality of an M-estimator.

Let $\mathcal{M}_\delta = \{\bar{m}_a(t) - \bar{m}_a(\tau) | |t - \tau| < \delta\}$ for $\delta > 0$. Since the arguments leading to (2.3) imply that the class of functions $(T_1, \ldots, T_{n_0}, X_1, \ldots, X_{n_0+1}, \phi) \to (a_j W_j(\phi)(I(t) - 1))^2$, indexed by $t \in [0,1]$, is a VC-class, we can use Theorem

2.6.7 in V and W to show that it satisfies the uniform entropy integral bound (3.2.8) of V and W. Using this and Theorem 2.10.20 in V and W, we know $\mathcal{M}_\delta$ also satisfies (3.2.8). Since all the other conditions in Problem 3.2.1 can be easily verified, the conclusion of the theorem follows.

3. Inference Based on Marker Data

This section presents linkage analysis under a more realistic situation. Let $M(t)$ be a $\{1, 2, \ldots, L\}-$valued stochastic process indexed by $T = \{t_1, \ldots, t_J\} \subset (0, 1)$. We use $M(t)$ to denote the genotype values of a sib-pair at locus t, and use M to denote the set of these values at each locus t in T. Let (I_k, M_k, ϕ_k) be a sequence of independent random elements distributed as (I, M, ϕ). We assume the conditional expectation of $I_k(t)$ given M_k, $E(I_k(t)|M_k)$, is available. The problem of calculating $E(I_k(t)|M_k)$ is postponed to Section 4.

Consider

$$G_K(t, a_1, \ldots, a_J) \equiv \frac{1}{\sqrt{K}} \sum_{k=1}^{K} \sum_{j=1}^{J} a_j W_j(\phi_k) E(I_k(t) - 1|M_k). \tag{3.1}$$

In view of the following, (3.1) can be used to test the null hypothesis that $I(\cdot)$ and ϕ are independent.

Theorem 3.1. *If $E(I(t)|\phi) = 1$ for every t, $G_K(\cdot)$ converges weakly to a mean zero tight Gaussian process $G(\cdot)$ in $\ell_\infty([0, 1] \times S)$.*

Proof. Observe that

$$\begin{aligned} E^{1/2}(E(I(t_1)|M) - E(I(t_2)|M))^2 &= E^{1/2}E^2(I(t_1) - I(t_2)|M) \\ &\le E^{1/2}E((I(t_1) - I(t_2))^2|M) = E^{1/2}(I(t_1) - I(t_2))^2. \end{aligned}$$

Then the covering number, defined by the L_2 norm, for the class of functions $(T_1, \ldots, T_{n_0}, X_1, \ldots, X_{n_0+1}, \phi) \to a_j W_j(\phi) E(I(t)|M)$, indexed by $t \in [0, 1]$ and $a \in S$, is no bigger than that for (2.3). The definition of covering number in this context can be found in V and W. Therefore, using the same arguments in the proof of Theorem 2.1, we conclude that $G_K(\cdot)$ in (3.1) converges weakly as desired.

For Theorem 3.1 to be useful it is desirable to know the distribution of G, and the following corollary suggests a bootstrap method to approximate it. We note that the validity of the bootstrap method in this context is guaranteed by Theorem 3.6.3 in V and W.

Let $b : \{1, 2, \ldots\} \to \{1, 2, \ldots\}$ be a map and let

$$G_K^{(b)}(t, a_1, \ldots, a_J) \equiv \frac{1}{K} \sum_{k=1}^{K} \sum_{j=1}^{J} a_j W_j(\phi_{b(k)}) E(I_k(t) - 1|M_k). \tag{3.2}$$

A corollary to Theorem 3.1 is the following.

Corollary 3.2. *If $I_k(\cdot)$ and $\phi_{b(k)}$ are independent for every $k = 1, 2, \ldots$, then $G_K^{(b)}(\cdot)$ converges weakly to the mean zero tight Gaussian process $G(\cdot)$ in $\ell_\infty([0,1]\times S)$.*

Let $Y_k(t, a_1, \ldots, a_J) = \sum_{j=1}^J a_j W_j(\phi_{b(k)}) E(I_k(t) - 1|M_k)$. Let $\bar{S} \subset [0,1] \times S$ be a finite subset. Consider Y_k as a function defined on $\bar{S}$. By Corollary 3.2, we can estimate the distribution of G under null hypothesis by resampling from $\{Y_k|\ k = 1, \ldots, K\}$ and forming (3.2) with, for example, $b(k) = k + 1$ for $k = 1, 2, \ldots, K - 1$ and $b(K) = 1$.

The following is a corresponding extension for the estimation problem. Let $m_a(t)\ (T_1, \ldots, T_{n_0}, X_1, \ldots, X_{n_0}, \phi) = \sum_{j=1}^J (a_j W_j(\phi))^2 E((I(t) - 1)^2|M)$, and set

$$\mathcal{P}_K m_a(t) \equiv \frac{1}{K} \sum_{k=1}^K \sum_{j=1}^J (a_j W_j(\phi_k))^2 E((I_k(t) - 1)^2|M_k).$$

Let $\hat{\tau}_K$ be a maximizer of $\mathcal{P}_K m_a(t)$. Let $D = \{(\tau, a)|a \in S, 0 < \tau < 1, Em_a(t) \leq Em_a(\tau)$ for every $t \in [0,1]$ and $t \to Em_a(t)$ has non-zero second derivative at $\tau\}$.

Theorem 3.3. *Assume D is non-empty and let $(\tau, a) \in D$ be given. Then $\hat{\tau}_K \to \tau$ in outer probability, and $\sqrt{K}(\hat{\tau}_K - \tau)$ is asymptotically normal with mean 0 and variance*

$$-(\frac{\partial^2}{\partial t^2} Em_a(t)|_{t=\tau})^{-1} (\frac{\partial^2}{\partial t_1 \partial t_2} Em_a(t_1) m_a(t_2)|_{(t_1,t_2)=(\tau,\tau)}). \tag{3.3}$$

Proof. Using the argument in the proof of Theorem 3.1, we know the covering number for the class of functions $(a_j W_j(\phi))^2 E((I(t) - 1)^2|M)$, indexed by $t \in [0,1]$, is no bigger than that for $(a_j W_j(\phi))^2 (I(t) - 1)^2$, which is a VC-class. With this observation, we can follow the proof of Theorem 2.2 to obtain the desired result.

We note that

$$\frac{\partial^2}{\partial t^2} Em_a(t) = \sum_{j=1}^J a_j^2 \frac{\partial^2}{\partial t^2} \lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^K W_j^2(\phi_k) E((I_k(t) - 1)^2|M_k), \tag{3.4}$$

$$\begin{aligned} \frac{\partial^2}{\partial t_1 \partial t_2} Em_a(t_1) m_a(t_2) = \sum_{j,l=1}^J a_j^2 a_l^2 \frac{\partial^2}{\partial t_1 \partial t_2} [\lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^K W_j^2(\phi_k) W_l^2(\phi_k) \\ \times E((I_k(t_1) - 1)^2|M_k) E((I_k(t_2) - 1)^2|M_k)], \end{aligned} \tag{3.5}$$

which provide a method to approximate (3.3).

Remarks on Weight Functions

We now indicate a scenario in which the results in Section 3 may be used. Suppose we are interested in mapping genes based on the phenotypic values of certain sib-pairs and their genotype data taken at a set of markers. We may decompose the space of the phenotypic values into, for example, four disjoint regions Φ_1, Φ_2, Φ_3, and Φ_4, where Φ_1 is a region for concordant sib-pairs with high phenotypic values, Φ_2 is a region for concordant low values, Φ_3 is a region for extreme discordant sib-pairs, and Φ_4 is the complement of $\Phi_1 \bigcup \Phi_2 \bigcup \Phi_3$. At this moment, we try not to make these terms rigorous and we do not consider the problem of how to decide the decomposition. Let $W_j(\phi) = 1_{\Phi_j}(\phi)$ for $j = 1, 2, 3, 4$. With this decomposition, we may reject the hypothesis of no linkage if $\sup G_K(t, a_1, a_2, a_3, a_4)$ is too large, where the supremum is taken over $t \in [0, 1]$, $(a_1, a_2, a_3, a_4) \in S$ and $a_1 \geq 0$, $a_2 \geq 0$, $a_3 < 0$. For this test, Theorem 3.1 and Corollary 3.2 can be used to provide an approximate p-value.

We can also use Theorem 3.3 to obtain an estimate of τ, the disease locus. In addition to the point estimate, we can choose an $a \in S$ with $a_1 \geq 0$, $a_2 \geq 0$, and $a_3 < 0$ so that the variance (3.3) is small. Here we note that both (3.4) and (3.5) are needed. The relative values of a_1, a_2, and $-a_3$ can be used to plan future gentic studies. For example, if $-a_3 > a_1$ and $-a_3 > a_2$, we would tend to select more extreme discordant sib-pairs for further genotyping and then carry out finer mapping of the disease gene.

We know from Risch and Zhang (1995) and Allison, Heo, Schork, Wong and Elston (1998) that, without knowledge of the mode of inheritance, it is hard to improve efficiency in terms of cost savings through reducing required number of sib-pairs to be genotyped for the same statistical power. We think the approach outlined above may offer some opportunities to improve the efficiency in this sense.

4. Conditional IBD Distribution

This section studies the problem of calculating $E(I(t)|M)$ under the assumption that the distribution of the IBD process is available. In fact, we discuss methods to calculate $P(I(t) = i|M = \mu)$, the conditional probability of sharing i alleles identical by descent at the point t given the genotype data. Here $i = 0, 1, 2$. These conditional probabilities can be used to obtain $E(I(t)|M)$ and the variance of the conditional distribution $P(I(t) \leq i|M = \mu)$. These can be used in turn to calculate the information-content mapping, introduced by Kruglyak and Lander (1995), to measure the extent to which all inheritance information has been extracted at the locus t. The information-content mapping helps to decide if additional genotype data at other nearby markers are needed for linkage analysis.

We now elaborate on the distribution of an IBD process. We recall that the spatial stochastic point process on $[0,1]$ describing the locations of the crossover events in a single strand product of meiosis, is called a crossover process. (cf. Speed(1996)). Since the distribution of an IBD process can be derived from that of a crossover process under the assumption of random mating, and the distribution of a crossover process can be estimated by using pedigree genotype data, it seems reasonable to assume that the distribution of an IBD process, instead of the process itself, is known. We note that, assuming a generalized count-location model, Chang et al. (2001) provides a non-parametric estimate of the distribution of a crossover process on the basis of genotype data obtained from a dense set of markers in seven three-generation CEPH references families. See Broman, Murray, Sheffield, White and Weber (1998) and Broman and Weber (2000) for more information about this data set.

Considering that $E(I(t)|M=\mu) = \{2P(M=\mu|I(t)=2)/P(M=\mu)\}$ $P(I(t)=2) + \{P(M=\mu|I(t)=1)/P(M=\mu)\}P(I(t)=1)$, $P(M=\mu|I(t)=i)$ $= E(E(1_{[M=\mu]}|I(\cdot))|I(t)=i)$, and the availability of the distribution of the IBD process, it suffices to know one of

$$P(M=\mu|I(t_1)=i_1,\ldots,I(t_J)=i_J,I(t)=i), \tag{4.1}$$

$$P(M=\mu,I(t_1)=i_1,\ldots,I(t_J)=i_J,I(t)=i), \tag{4.2}$$

to obtain $E(I(t)=i|M=\mu)$. We recall that $M=\mu$ is the genotypes of the sib-pair taken at the loci $t_1,\ldots,t_J$.

We now introduce two approaches to (4.1) or (4.2). The first assumes that the allele frequencies of the markers at $t_1,\ldots,t_J$ are known and the haplotypes at these markers are the result of a random combination of alleles. This is realistic unless some of the loci are extremely close to each other. Let $v_1(s) = (v_{11}(s), v_{12}(s))$ and $v_2(s) = (v_{21}(s), v_{22}(s))$ be the inheritance vectors, respectively, for the first and second sibs. Here $v_{11}(\cdot)$ and $v_{21}(\cdot)$ refer to the paternally derived meiosis, $v_{12}(\cdot)$ and $v_{22}(\cdot)$ refer to the maternally derived meiosis and, for every $i=1,2$ and $j=1,2$, $v_{ij}(s)$ equals 0 if the allele at s comes from one of the grandmothers and equals 1 otherwise. We note that $I(s) = \sum_{i=1}^{2} 1_{[v_{1i}(s)=v_{2i}(s)]}$.

Now (4.2) can be decomposed according to the parental haplotypes and the inheritance vectors. Using the fact that the haplotypes of the parents are independent of the inheritance vectors, each of the summands is a product of the haplotype frequencies and probabilities involving distributions of the crossover process. This gives the value for (4.2). We note that many summands in the above calculation are zero, because $v_{ij}(\cdot)$ changes its value only occasionally.

The second approach assumes

$$\begin{aligned} &P(M(t_1)=\mu_1,\ldots,M(t_J)=\mu_J|I(t_1)=i_1,\ldots,I(t_J)=i_J)\\ &=\prod_{j=1}^{J}P(M(t_j)=\mu_j|I(t_j)=i_j), \end{aligned} \tag{4.3}$$

and only makes use of $E(I(t)=i|M=\mu)$ for $t\in\{t_1,\ldots,t_J\}$. In this case, $P(M(t_j)=\mu_j|I(t_j)=i_j)$ can be estimated from the data in view of the likelihood derived from

$$\begin{aligned} &P(M(t_1)=\mu_1,I(t_1)=i_1)\\ &=P(M(t_1)=\mu_1|I(t_1)=i_1)P(I(t_1)=i_1) \end{aligned} \tag{4.4}$$

and hence we do not need to know the allele frequencies in order to calculate (4.1) or (4.2).

Remark The feasibility of the above computation depends largely on J, the number of markers where the genotype data are used in calculating $E(I(t)|M)$. When J is large, we can use $E(I(t)|\tilde{M}(t))$ instead, where $\tilde{M}(t)$ is the set of genotype data taken at the markers that are in a neighborhood of the locus t. Our initial experience using a desktop PC indicates that it is computationally feasible if there are four or six markers involved in $\tilde{M}(t)$.

5. Discussion

We have proposed two statistical methods for genetic linkage analysis of qualitative and quantitative traits in sib-pairs based on genotype data obtained from a set of polymorphic markers. With the introduction of weight functions and the more appropriate conditional expectation of IBD processes, we hope our statistical methods will be useful for the detection of genetic factors for complex diseases. In order to illustrate the numerical performance of these methods, we are conducting some simulation studies and examining some real data sets. These will be reported on elsewhere. Preliminary studies indicate that our methods are computationally feasible. In addition to computational issues, we pay particular attention to the choice of weight functions so as to indicate sib-pairs having more informative phenotypes in various situations. We also use the information-content mapping to decide if the set of markers is dense enough or if additional genotype data at other markers are needed.

There are two important components in deriving these procedures. One involves the conditional distribution of the IBD process given the genotype data, and the other involves weight functions of the phenotypes. Examining these closely, we think the present theory can be extended to other pedigrees and

functionals of the IBD process. In fact, our methods are genome-wide versions of the classical means test, which is only one of several tests useful for linkage analysis of sib-pair genotype data. Readers are referred to Whittemore and Tu (1998) and references therein for these methods. We will report extensions to the present theory in simulation studies. It would be interesting if our approach sheds light on the question of informativeness in other pedigree designs (cf. Teng and Siegmund (1997), and Feingold and Siegmund (1997)). Such results would be useful in planning genetic studies.

Acknowledgements

We are grateful to an anonymous referee for valuable comments and suggestions which led to the improvement of this paper.

References

Allison, D. B., Heo, M., Schork, N. J., Wong, S. L. and Elston, R. C. (1998). Extreme selection strategies in gene mapping studies of oligogenic quantitative traits do not always increase power. *Hum. Hered.* **48**, 97-107.

Amos, C. I. (1994). Robust variance-components approach for assessing genetic linkage in pedigrees. *Amer. J. Hum. Genet.* **54**, 535-543.

Billingsley, P. (1995). *Probability and Measure*. John Wiley, New York.

Bishop, M. (1999). *Genetics Databases*. Academic Press, London and San Diego.

Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L. and Weber, J. L. (1998). Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Amer. J. Hum. Genet.* **63**, 861-869.

Broman, K. W. and Weber, J. L. (2000). Characterization of human crossover interference. *Amer. J. Hum. Genet.* **66**, 1911-1926.

Chang, I. S., Wang, W. C., Chen, W. C. and Hsiung, C. A. (2001). An estimate of the distribution of crossover points in human meiosis. Submitted.

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.

Dupuis, J., Brown, P. O. and Siegmund, D. (1995). Statistical methods for linkage analysis of complex traits from high-resolution maps of identity by descent. *Genetics* **140**, 843-856.

Feingold, E. (1993). Markov processes for modeling and analyzing a new genetic mapping method. *J. Appl. Probab.* **30**, 766-779.

Feingold, E., Brown, P. O. and Siegmund, D. (1993). Gaussian models for genetic linkage analysis using complete high resolution maps of identity-by-descent. *Amer. J. Hum. Genet.* **53**, 234-251.

Feingold, E. and Siegmund, D. (1997). Strategies for mapping heterogeneous recessive traits by allele-sharing methods. *Amer. J. Hum. Genet.* **60**, 965-978.

Fulker, D. W. and Cardon, L. R. (1994). A sib-pair approach to interval mapping of quantitative trait loci. *Amer. J. Hum. Genet.* **54**, 1092-1103.

Goldgar, D. E. (1990). Multipoint analysis of human quantitative genetic variation. *Amer. J. Hum. Genet.* **47**, 957-967.

Guo, S.-W. (1994). Computation of identity-by-descent proportions shared by two siblings. *Amer. J. Hum. Genet.* **54**, 1104-1109.

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

Hauser, E. R. and Boehnke, M. (1998). Genetic linkage analysis of complex genetic traits by using affected sibling pairs. *Biometrics* **54**, 1238-1246.

Kruglyak, L. and Lander, E. S. (1995). Complete multipoint sib-pair analysis of qualitative and quantitative traits. *Amer. J. Hum. Genet.* **57**, 439-454.

Liang, K. Y., Huang, C. Y. and Beaty, T. H. (2000). A unified sampling approach for multipoint analysis of qualitative and quantitative traits in sib pairs. *Amer. J. Hum. Genet.* **66**, 1631-1641.

McPeek, M. S. (1996). An introduction to recombination and linkage analysis. In *Genetic Mapping and DNA Sequencing* (Edited by T. P. Speed and M. S. Waterman). Volume 81 of *IMA Volumes in Mathematics and its Applications.* Springer-Verlag, New York.

Ott, J. (1999). *Analysis of Human Genetic Linkage.* The Johns Hopkins University Press, Baltimore and London.

Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science* **273**, 1516-1517.

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. *Amer. J. Hum. Genet.* **58**, 836-843.

Sham, P. (1998). *Statistics in Human Genetics.* John Wiley, New York.

Speed, T. P. (1996). What is a genetic map function? In *Genetic Mapping and DNA Sequencing* (Edited by T. P. Speed and M. S. Waterman). Volume 81 of *IMA Volumes in Mathematics and its Applications.* Springer-Verlag, New York.

Teng, J. and Siegmund, D. (1997). Combining information within and between pedigrees for mapping complex traits. *Amer. J. Hum. Genet.* **60**, 979-992.

Tu, I. P. and Siegmund, D. (1999). The maximum of a function of a Markov chain and application to linkage analysis. *Advances in Applied Probability* **31**, 510-531.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes.* Springer, New York.

Whittemore, A. S. and Tu, I. P. (1998). Simple, robust linkage tests for affected sibs. *Amer. J. Hum. Genet.* **62**, 1228-1242.

Department of Mathematics, National Central University, Chungli, Taiwan.

E-mail: chang@math.ncu.edu.tw

Division of Biostatistics and Bioinformatics, National Health Research Institutes, Nankang, Taipei, Taiwan.

Division of Biostatistics and Bioinformatics, National Health Research Institutes, Nankang, Taipei, Taiwan.

E-mail: chinfu@nhri.org.tw

Division of Biostatistics and Bioinformatics, National Health Research Institutes, Nankang, Taipei, Taiwan.

E-mail: hsiung@nhri.org.tw

(Received February 2001; accepted October 2001)