Linkage analysis of ordinal traits for pedigree data

Rui Feng[†], James F. Leckman[‡], and Heping Zhang[†‡§]

[†]Department of Epidemiology and Public Health and [‡]Yale Child Study Center, Yale University of School of Medicine, New Haven, CT 06250-8034

Edited by David O. Siegmund, Stanford University, Stanford, CA, and approved September 29, 2004 (received for review June 29, 2004)

Linkage analysis is used routinely to map genes for human diseases and conditions. However, the existing linkage-analysis methods require that the diseases or conditions either be dichotomized or measured by a quantitative trait, such as blood pressure for hypertension. In the latter case, normality is generally assumed for the trait. However, many diseases and conditions, such as cancer and mental and behavioral conditions, are rated on ordinal scales. The objective of this study was to establish a framework to conduct linkage analysis for ordinal traits. We propose a latent-variable, proportional-odds logistic model that relates inheritance patterns to the distribution of the ordinal trait. We use the likelihood-ratio test for testing evidence of linkage. By means of simulation studies, we find that the power of our proposed model is substantially higher than that of the binary-trait-based linkage analysis and that our test statistic is robust with regard to certain parameter misspecifications. By using our proposed method, we performed a genome scan of the hoarding phenotype in a data set with 53 nuclear families, which were collected by the Tourette Syndrome Association International Consortium for Genetics (TSAICG). Standard linkage scans using hoarding as a dichotomous trait were also performed by using GENEHUNTER and ALLEGRO. Both GENEHUNTER and ALLEGRO failed to reveal any marker significantly linked to the binary hoarding phenotypes. However, our method identified three markers at 4q34-35 ($P = 0.0009$), 5q35.2-35.3 ($P = 0.0001$), and 17q25 ($P = 0.0005$) that manifest significant allele sharing.

Linkage analysis has been very useful for mapping disease genes, such as the breast cancer gene BRCA1 (1, 2). Statistical methods for linkage analysis are well established for both quantitative and binary traits (3–9). However, methods for linkage analysis have not been established for ordinal traits, although many human conditions (e.g., cancer) are rated on discrete, ordinal scales. Studies have suggested genetic heritabilities for some ordinal traits (10–13). Ordinal traits are typically first dichotomized into binary traits, and then analyzed by standard linkage-analysis programs such as GENEHUNTER (8) and ALLEGRO (14).

We propose a statistical method for linkage analysis of general pedigree data with ordinal traits and demonstrate the gain of power when the ordinal, rather than dichotomized, traits are directly used. Other researchers (15) have also observed loss of power for linkage analysis due to the dichotomization of trichotomous phenotypes.

It is widely recognized that demographic and environmental factors, together with genetic mechanisms, are important determinants in most complex diseases. Thus, our model was developed to accommodate selected demographic and environmental factors.

Methods

Methods for linkage analysis include two main steps (8). The first step is to infer information about the inheritance pattern of a pedigree by means of the so-called inheritance vector. This is a unified step, regardless of the properties of a trait; hence, we adopt the same method as used in ref. 8. In the second step, the linkage of a marker to a disease locus is established if the inheritance pattern of the marker is associated with the trait because, in the absence of linkage, the inheritance pattern is expected to be independent of the trait. This second step obviously depends on the distribution of the trait, and it is the main focus of this article.

Inheritance Vector and Its Distribution. Briefly, we present the necessary notation and steps for deriving the distribution of inheritance vector (16). Given a pedigree, a founder refers to a subject whose parents are not included in the pedigree. In a nuclear family with two parents (founders) and n siblings (nonfounders), the inheritance pattern at a marker location t is described completely by an inheritance vector $v(t) = (v_1, v_2, v_3, v_4, \ldots, v_{2n-1}, v_{2n})'$, whose elements describe the outcomes of the paternal and maternal meioses transmitted to the n siblings. Specifically, $v_{2j-1} = 1$ or 2, according to whether the grandpaternal or grandmaternal allele is transmitted in the paternal meiosis to the jth sibling. v_{2j} carries the similar information for the corresponding maternal meiosis, namely, $v_{2j} = 3$ or 4, according to whether the grandpaternal or grandmaternal allele was transmitted in the maternal meiosis to the jth sibling. For example, if $v(t) = (1, 4, 1, 3)'$, the first child received the father's paternally derived allele and the mother's maternally derived allele, and the second child received the father's paternally derived allele and the mother's paternally derived allele.

For a more complex pedigree with f founders and n nonfounders, we can index the alleles of the f founders as (1, 2), (3, 4), (5, 6), . . . , $(2f - 1, 2f)$. Then, we can define the inheritance vector for the n nonfounders similarly. The inheritance vector completely specifies which of the $2f$ distinct founders' alleles are inherited by every nonfounder. There are a total of 2^{2n} possible inheritance vectors, which can be grouped into 2^{2n-f} distinct configurations.

For clarity, we will use nuclear families with two founders below. The inheritance distribution is the conditional probability distribution over the possible inheritance vectors that conform the alleles observed at the marker locus t, which we denote by $p\{v(t) = w\}$ for all inheritance vectors $w \in V$; here, V is the set of all possible inheritance vectors. In the absence of any genotypic information, all inheritance vectors are equally likely according to Mendel's first law, and the probability distribution is uniform (denoted as p_{unif}). As genotypic information is enhanced, the probability distribution becomes concentrated in certain inheritance vectors. In any case, available standard software can be used to derive the inheritance distribution for genotyped pedigrees. We used GENEHUNTER in our computation.

A Latent-Variable Model (LVM) in Linkage Analysis. To infer whether the inheritance pattern at a given marker is associated with an ordinal trait, we propose a latent-variable, proportional-odds logistic model, which is a further development of the LVM for segregation analysis (13, 17).

We consider a trait Y taking an ordinal value from 0, 1, . . . , $K(K \geq 1)$. Let $\mathbf{x}$ be a p-vector of covariates that is also available for every study subject. For the ith family, we assume there exist two

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: GTS, Gilles de Tourette syndrome; lod, logarithm of odds; LVM, latent-variable model; NPL, nonparametric linkage; OCD, obsessive–compulsive disorder; plod, penalized lod; TSAICG, Tourette Syndrome Association International Consortium for Genetics.

[§]To whom correspondence should be addressed. E-mail: heping.zhang@yale.edu.

www.pnas.org/cgi/doi/10.1073/pnas.0404623101

types of latent random variables, U_1^i and U_2^i, which represent (*i*) the common genetic or environmental factors in a family that are not observed through the covariates and (*ii*) the genetic susceptibility of the family founders and nonfounders, respectively. The genetic susceptibility due to a particular gene is accommodated through the inheritance vector in the pedigree as follows.

Let $U_{2,1}^i, \ldots, U_{2,2f}^i$ be the genetic susceptibility of the f founders. In a simple pedigree with two founders, we have four latent variables in which $U_{2,1}^i$ and $U_{2,2}^i$ represent the genetic susceptibility associated with the two parental alleles at marker location t. Likewise, $U_{2,3}^i$ and $U_{2,4}^i$ represent the genetic susceptibility associated with the two maternal alleles at t. The genetic susceptibility is reflected through the inheritance vector v. That is, for the jth nonfounder in the ith family, his/her latent variables are $U_{2,v_{2j-1}}^i$ and $U_{2,v_{2j}}^i$. If we consider the additive susceptibility due to the gene linked to locus t, then for the jth sibling, $U_{2,v_{2j-1}}^i + U_{2,v_{2j}}^i$ for $j = 1, 2, \ldots, n$, reflects the genetic susceptibility due to part of the genomes on the two chromosomes transmitted from the parents. For example, in a five-member pedigree with two parents and three offspring, if the inheritance vector of the three offspring is (1, 3, 1, 4, 2, 3), then their latent variables U_2 values are ($U_{2,1}$, $U_{2,3}$), ($U_{2,1}$, $U_{2,4}$), and ($U_{2,2}$, $U_{2,3}$).

For the latent variables introduced above, we assume that $P(U_1^i = 1) = 1 - P(U_1^i = 0) = \theta_1$, where θ_1 is an unknown parameter. Also, for a founder, $P(U_2^i = 1) = 1 - P(U_2^i = 0) = \theta_2$, where θ_2 is an unknown parameter. This Bernoulli possibility basically assumes that there exists a single major susceptibility locus with alleles, A and a, at marker location t and that the frequency of allele A is θ_2. Furthermore, all U_1^i values and the U_2^i values of the founders are assumed to be distributed independently and identically across families.

Conditional on all of the latent variables, denoted by U^i, and inheritance vectors v^i, within the ith family, the traits of all nonfounders are independent and follow the following distribution:

$$\operatorname{logit}(P\{Y_j^i \leq k | U^i, v^i\}) = x_j^i\beta + \alpha_k + U_1^i\gamma_1 + U_{2(j)}^i\gamma_2, \qquad k = 0,1, \ldots, K-1, \quad [1]$$

where $U_{2(j)}^i = U_{2,v_{2j-1}}^i + U_{2,v_{2j}}^i$, and β is a p vector of parameters, reflecting the covariate effects on the trait. The α_k variables are the trait-level-dependent intercepts, and they reflect the differences between cumulative probabilities $P(Y_j^i \leq k)$. We must have $\alpha_0 \leq \alpha_1 \leq \cdots \leq \alpha_K$, so that the category probabilities $P(Y_j^i = k)$ are nonnegative (18). $\gamma = (\gamma_1, \gamma_2)'$ indicates the familial and genetic contributions to the trait. Without the latent variables (or $\gamma = 0$), model **1** is the commonly used proportional-odds logistic model for an ordinal response (18). In this article, individuals in category 0 are "most severely affected," individuals in category K are "unaffected," and individuals in other categories are between unaffected and most severely affected. Thus, we expect $\gamma_2 > 0$.

Parameter Estimates. Suppose that there are n families and n_i siblings in the ith family. If both the trait and the latent variables were observable, given a particular inheritance vector v, the complete log-likelihood function, $l^i(\alpha, \beta, \gamma, \theta \mid U, v)$, would be equal to the following:

$$U_1^i \log(\theta_1) + (1 - U_1^i)\log(1 - \theta_1) + \sum_{j=1}^{4}[U_{2j}^i \log(\theta_2) + (1 - U_{2j}^i)\log(1 - \theta_2)] + \sum_{j=1}^{n_i}\sum_{k=0}^{K}[I(Y_j^i = k)\log(\pi_{kj}^i - \pi_{k-1,j}^i)], \quad [2]$$

where $I(\cdot)$ is the indicator function, and

$$\pi_{kj}^i = \frac{\exp(x_j^i\beta + \alpha_k + U_1^i\gamma_1 + U_{2(j)}^i\gamma_2)}{1 + \exp(x_j^i\beta + \alpha_k + U_1^i\gamma_1 + U_{2(j)}^i\gamma_2)} \quad \text{for } k = 0,1, \ldots, K-1, \ \pi_{-1,j}^i = 0, \ \pi_{Kj}^i = 1.$$

The EM algorithm (19) is used to find the maximum-likelihood estimation (MLE) in a similar way to Zhang *et al.* (13). In general, the inheritance vector v involves uncertainty at a marker location. As usual (8), we can take the expectation of the likelihood L^i over the inheritance distribution inferred from the genotyped data, namely, $\Sigma_{v \in V} L^i(\alpha, \beta, \gamma, \theta \mid v)p_{\text{comp}}(v)$, where $p_{\text{comp}}(v)$ is the inheritance distribution consistent with the marker information and V is the set of 2^{2n-f} equivalent classes of inheritance vectors. Then, after the integration over the distributions of the latent variables, the likelihood becomes the following:

$$L_*^i(\alpha, \beta, \gamma, \theta) = E_{\theta_1\theta_2}\left[\sum_{v \in V} p_{\text{comp}}(v)L^i\right] = \sum_{v \in V}[p_{\text{comp}}(v)E_{\theta_1\theta_2}L^i]. \quad [3]$$

After obtaining the MLEs of the parameters, we can estimate the covariance matrix of the MLEs by using an existing method (20, 21).

Logarithm of Odds (lod) Score Calculation. The likelihood-ratio statistic is used commonly in linkage analysis. The null hypothesis is that a disease gene is not in linkage with locus t. The alternative hypothesis is that locus t is linked to the disease gene. The likelihood ratio of $LR = p(Y | v(t))/\Sigma_{w \in V}\, p(y \mid w)p_{\text{unif}}(w)$, where $p(Y | v(t))$ is the probability of the observed ordinal trait Y conditional on $v(t)$, V is the set of 2^{2n-f} equivalent classes of inheritance vectors, and $p_{\text{unif}}(w)$ is the uniform distribution over V (i.e., the distribution in the absence of linkage). The lod score, defined as the decimal log-likelihood ratio, is commonly used (22).

In the LVM, $p(Y \mid v(t))$ is a function of parameters $\omega = (\alpha, \beta, \gamma_1, \theta_1)$ and (γ_2, θ_2). In the absence of linkage, $\gamma_2 = 0$ in model **1**. Therefore, for a given θ_2,

$$LR(\theta_2) = \frac{\max\limits_{\omega,\ \gamma_2} \prod\limits_i L_*^i(\omega, \ \gamma_2, \ \theta_2)}{\max\limits_{\omega} \prod\limits_i L_*^i(\omega, \ \gamma_2 \mid \gamma_2 = 0)}, \quad [4]$$

and the lod score at locus t for a given θ_2 becomes $\text{lod}(\theta_2) = \log_{10}LR(\theta_2)$.

Based on asymptotic theory involving nonstandard conditions, the distribution of the standard likelihood-ratio statistic in our model may be complicated (13). One reason for the complication is that θ_2 is not identifiable when $\gamma_2 = 0$. In fact, this is the case for many statistical models in genetic studies. To test the presence of linkage, we follow the discussions in ref. 13 and consider the penalized log likelihood $PLR(\theta_2) = \log LR(\theta_2) + \lambda \log[4\theta_2(1 - \theta_2)]$. Under standard regularity conditions (13, 23), the maximum penalized log-likelihood-ratio statistic $2 \max_{\theta_2} PLR(\theta_2)$ follows $(1/2)\chi_0^2 + (1/2)\chi_1^2$ $(\chi_0^2 \equiv 0)$. Thus, when θ_2 is not specified, $2 \max_{\theta_2} PLR(\theta_2)$ is used to determine the significance level, and the penalized lod (plod) score, defined as $plod = \max_{\theta_2} PLR(\theta_2)/\log(10)$, will be reported for model **1**. However, for a given θ_2, we use $\text{lod}(\theta_2)$, which is a function of θ_2.

Table 1. Power comparison

		Marker						
Score	Method	1–3	4–6	7–9	10–11	12–14	15–17	18–20
plod >2*	LVM	54	62	74	85	80	61	57
lod >2*	GH	0	0	0	0	0	0	0
lod >2*	AL	1	0	4	5	6	2	0
plod >3*	LVM	26	31	52	66	57	34	25
lod >3*	GH	0	0	0	0	0	0	0
lod >3*	AL	0	0	1	1	1	0	0
lod >2†	LVM	49	51	66	80	74	50	47
lod >2†	GH	0	0	5	19	12	0	2
lod >3†	LVM	16	25	41	56	47	25	20
lod >3†	GH	0	0	1	0	0	0	0

Frequencies at which linkage is detected in the 100 data sets by the LVM, GENEHUNTER (GH), and ALLEGRO (AL) are shown.

*Nonparametric lod scores converted from NPL in GENEHUNTER and Z_{lr} from ALLEGRO.

†Parametric lod scores calculated by using $\theta_2 = 0.3$ for LVM and GH.

Simulation Studies

To demonstrate the promise of our model in mapping genes for ordinal traits, we report a series of simulations that are designed to examine three specific aims. First, in addition to our theoretical understanding of the asymptotic distribution of the plod score, we demonstrate numerically how the distribution appears under the null hypothesis. Second, as a critical incentive for the use of ordinal traits, we present evidence for gain of power of detecting linkage when there is a linked disease gene. The comparison is made with the use of GENEHUNTER and ALLEGRO, in which an individual's disease status is represented by a binary trait. Last, our model depends on a number of assumptions. It is important to scrutinize the robustness of our estimates when the model is misspecified.

Distribution of Likelihood-Ratio Statistic Under the Null Hypothesis. For each data set, we generated 200 five-member pedigrees with two founders and three offspring in each pedigree. One covariate, x_1, was generated from independent uniform (0, 1) distribution among all family members. Without loss of generality, we suppose that the response, Y, takes values from three ordinal levels 0, 1, and 2. In the absence of latent variable U_1, Y was simulated from an ordinal logistic model with the two logit links being $-1\ -0.9x_1$ (for $Y = 0$) and $-0.9x_1$ (for $Y \le 1$).

For each founder in a pedigree, 20 highly polymorphic markers with 10 equally likely alleles (24), spaced 5 cM apart, were generated on one chromosome. Recombination fractions were converted to map distances without interference, and there was no linkage disequilibrium among markers (25). After the genotypes were generated for the founders, the genotypes of nonfounders were generated subsequently based on the recombination fractions.

The simulation was replicated 30,000 times (30,000 data sets). At a fixed marker locus, the empirical P values from plod scores are 0.5018, 0.1971, 0.1027, 0.0536, 0.0129, and 0.0016 when the asymptotic values are 0.5, 0.2, 0.1, 0.05, 0.01, and 0.001, respectively. Given the size of our simulation replications, the nominal and empirical P values are numerically within reasonable ranges of each other.

We also examined the number of false-positive errors when the significant linkage was declared at plod scores of >2 and >3 by using our model. When we focused on a specific locus, there were 60 of 30,000 data sets ($P = 0.002$) with plod scores of >2, and four with plod scores of >3 ($P = 0.0001$). We used Bonferroni correction to project locus-specific P values to a chromosome or the genome. For example, a plod score of 3 corresponds to ≈0.05 genomewide significance level for a typical genotypic data set of 370 microsatellite markers, as used in *Application on Obsessive–Compulsive Disorder* (*OCD*). The method given in ref. 26 could be used to explore a better approach.

Power Comparison. For the power comparison, the data sets were simulated similarly to those in the previous experiment. However, a locus between the 10th and 11th markers was set to be the disease locus, and the disease-causing allele was set at frequency 0.3. For the comparison purpose with GENEHUNTER and ALLEGRO, we do not consider any covariates in this simulation. In addition, the parameters involved in the logit link functions were set at $\gamma_1 = 0$, $\gamma_2 = 2$, $\alpha_0 = -2$, and $\alpha_1 = -1$. These parameters were unknown in the linkage analysis and were estimated from the data.

We conducted linkage analysis with 100 replications by using our method, GENEHUNTER, and ALLEGRO. For GENEHUNTER and ALLEGRO, the trait was dichotomized as $Y = 0$ or $Y \ge 1$. We also considered $Y \le 1$ versus $Y = 2$, and the results were consistent and, hence, not shown. Table 1 contrasts the results from the LVM and the nonparametric linkage (NPL) score method in GENEHUNTER and the nonparametric method (Z_{lr}) in ALLEGRO. Under the null hypothesis, the NPL score (8) follows a normal distribution asymptotically and, thus, can be transformed into lod score units according to the formula: $NPL^2/(2^*\log(10)) = lod$. To provide numerical support to this asymptotic property, Ulgen *et al.* (27) carried out a simulation study of 2,400 replications of 100 nuclear families and verified that the observed distribution of (2ln10)lod fit well with $(1/2)\chi_0^2 + (1/2)\chi_1^2$ under the null hypothesis. Thus, plod from our model for ordinal traits and the transformed lod score from NPL from the other software for binary traits have the same asymptotic distribution. Consequently, our plod and the NPL from GENEHUNTER could be viewed as counterparts of each other under the null hypothesis. This observation is useful to ensure the validity of the power comparison, because the type I error rates for different methods must be similar.

We also compared the lod scores from the LVM with the parametric method in GENEHUNTER by specifying the disease allele frequency to be the "true" frequency of 0.3. We used the simulated model to specify the penetrances as (0.12, 0.50, 0.88), which were based on the true parameters and, hence, expected to be in favor of GENEHUNTER.

In Table 2, we used allele frequencies of 0.2 and 0.5 to examine the impact of misspecifying the allele frequency. It is evident from this table that these misspecifications have little effect on power. Therefore, the LVM is very robust in terms of power with

Table 2. Power comparison under misspecified allele frequencies

		Marker						
Score	Specified allele frequency	1–3	4–6	7–9	10–11	12–14	15–17	18–20
lod >2	0.2	40	42	57	74	71	43	40
lod >2	0.5	52	58	73	83	78	61	55
lod >3	0.2	13	21	31	44	37	20	14
lod >3	0.5	23	30	50	65	57	31	23

Frequencies at which linkage is detected in the 100 data sets are shown.

respect to misspecification of the allele frequency, which is the only parameter to be specified after model **1** is assumed.

Application on Obsessive–Compulsive Disorder (OCD)

Background. OCD is a potentially disabling condition affecting nearly 5 million people in the United States (28). Patients with OCD become obsessed with unwanted worries or unpleasant images, such as persistent fears that harm may come to a loved one, an unreasonable concern with becoming contaminated, or an excessive need to do things correctly. OCD occurs in a spectrum from mild to severe, and if severe and untreated, it can destroy a person's capacity to function at work, at school, or in the home (29).

The causes of OCD have yet to be established. However, growing evidence shows that biological factors are likely to contribute to the disorder. Evidence for a genetic component in OCD comes from twin studies, family genetics studies, and segregation analyses (30).

Hoarding, as a major factor in OCD, affects 2.4 million Americans. In adults, hoarding has been defined as the gathering of articles without clear, conscious motivation or control (31). Many studies have been done on hoarding, but its etiology is not understood. Only a few studies have focused specifically on the role that genetic factors play in the transmission and expression of hoarding symptoms. One such study was undertaken by Leckman *et al.* (32) as part of the TSAICG. They found evidence in support of a recessive mode of transmission for the hoarding symptom dimension in families with two affected siblings with Gilles de la Tourette syndrome (GTS) (32). Zhang *et al.* (33) performed linkage study of hoarding in the sibling pairs in the same study sample that we used here. They found strong evidence of linkage to three different regions by using a numerical factor scale from two hoarding items as a quantitative trait. The goal of the present study is to conduct and compare genome scans by treating hoarding symptoms as either an ordinal or dichotomous trait.

The Study Sample. The study sample was obtained through GTS patients. Since Gilles de la Tourette noted in 1885 that OCD symptoms were present in GTS patients, studies have shown prevalences of obsessive–compulsive symptoms of 11–80% among individuals with GTS (34). In contrast to the prevalence of OCD of 1–3% in the general population (35), the elevated prevalence of obsessive–compulsive symptoms are found in both clinical samples composed of GTS patients and nonreferred individuals with tics and their relatives in community samples (36).

All families include at least two siblings with GTS. In the original ascertainment, families were excluded if both parents were affected with GTS or if one parent had GTS, chronic tics (CT), OCD, and/or subclinical OCD and the other parent also had CT, OCD, and/or subclinical OCD. All diagnoses were made by use of Diagnostic and Statistical Manual-III-R criteria (37). The final sample included in the genome scan consisted of 53 families, with a total of 223 individuals. No information has been collected to perform a justifiable ascertainment adjustment for hoarding, which was a limitation of this study, as it was for ref. 33.

The significant hoarding symptoms were recorded as "present" when one or both of the hoarding items on the Yale–Brown Obsessive–Compulsive Scale symptom checklist were rated as present by clinicians and as "absent" otherwise. In addition to treating hoarding as a dichotomous outcome, we also considered it as an ordinal trait; that is, we recorded 2 if both

Table 3. Comparison of results for hoarding from LVM, GENEHUNTER, and ALLEGRO

	P values				
	Parametric, $\theta_2 = 0.3$		Nonparametric		
Marker (location in cM)	LMV	GH	LMV	GH*	AL†
4q34-35					
DS42431 (163.26)	0.006	0.101	0.001	0.120	0.156
D4S2417 (169.00)	0.005	0.072	0.0009	0.154	0.192
D4S408 (182.13)	0.012	0.063	0.006	0.068	0.091
D4S1652 (195.14)	0.003	0.040	0.004	0.126	0.136
5q35.2-35.3					
D5S1471 (172.13)	0.003	0.122	0.001	0.560	0.563
D5S1456 (174.80)	0.002	0.139	0.0003	0.628	0.640
D5SMfd154 (182.89)	0.0006	0.095	0.00006	0.299	0.299
D5S408 (195.49)	0.0002	0.030	0.00001	0.133	0.100
17q25					
D17S1301 (99.39)	0.005	0.066	0.0005	0.052	0.024
D17S784 (116.23)	0.002	0.034	0.0006	0.019	0.007

AL, ALLEGRO; GH, GENEHUNTER.

*Based on NPL.

†Based on Z_{lr}.

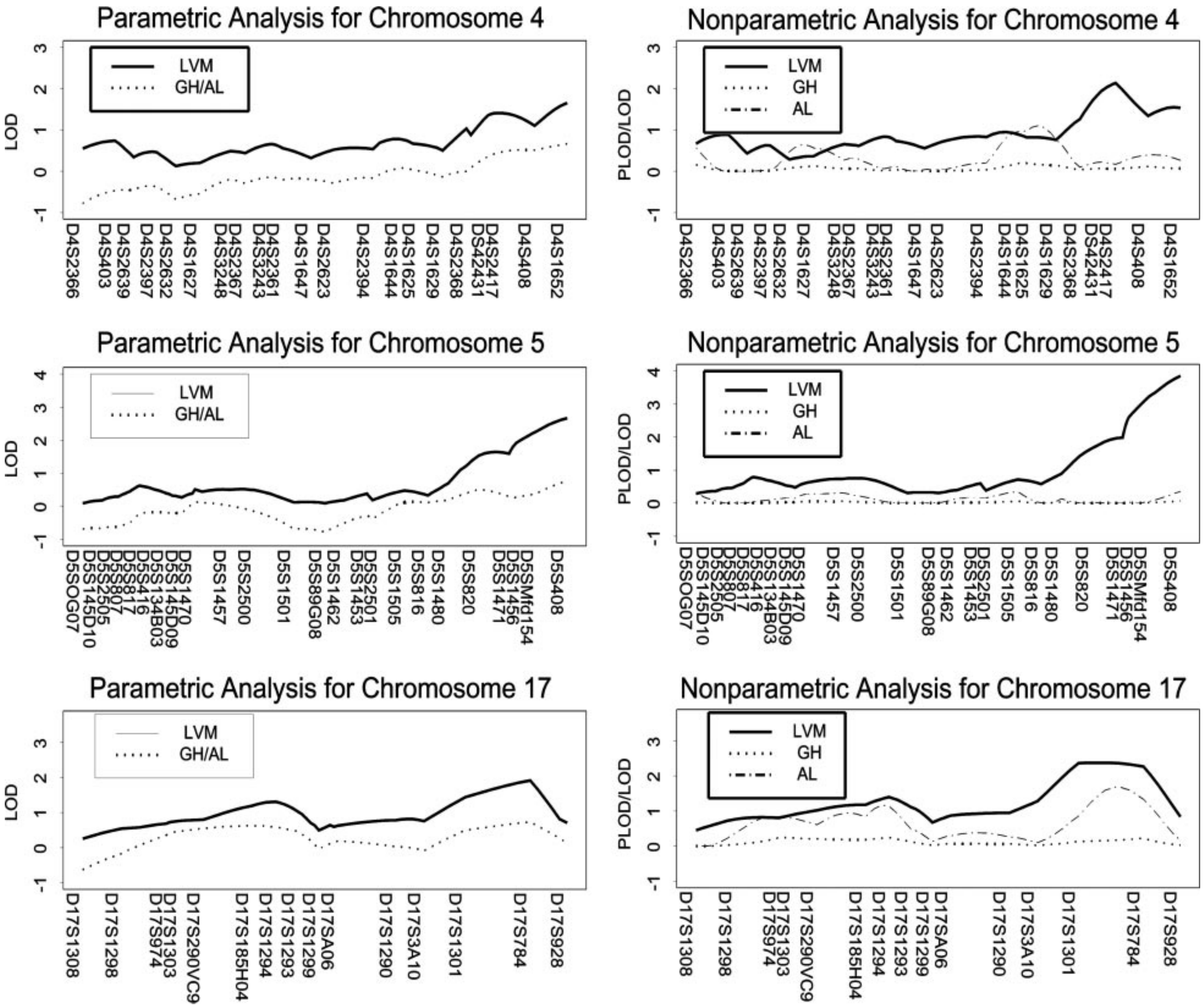

Fig. 1. LOD/PLODs produced by LMV, GENEHUNTER, and ALLEGRO on chromosomes 4, 5, and 17.

items were absent, 1 if only one item was present, and 0 if both items were present.

The panel of genotyped markers included 370 DNA markers with an average spacing of 9.1 cM in the male meiotic map on 22 autosomal chromosomes. A detailed description of the markers and map is given by TSAICG. Zhang *et al.* (33) presented a detailed description of the study design and data collection.

Data Analysis. Allele frequencies for the genetic markers were established by gene counting in genotyped parents. For each family, the inheritance distribution was estimated by multipoint analyses in GENEHUNTER (8) and ALLEGRO (14). In the multipoint analysis, maximum-likelihood scores (38) were computed for 3,000 different locations relative to the markers (average step size, ≈1 cM). As a standard procedure, genotype data were checked against errors before the linkage analysis.

The status of hoarding was first examined as a binary trait. Parametric and nonparametric analyses were done with GENEHUNTER and ALLEGRO. For parametric analysis, disease allele frequency was set at 0.3, and the penetrances were set at 0.125, 0.575, and 0.75 (33).

The main results reported here are from the linkage analyses of the ordinal hoarding trait. The analyses were completed by using the LVM for the nuclear families. For comparison, we obtained the likelihood-ratio statistics with an unknown θ_2 as well as a fixed value at $\theta_2 = 0.3$.

Analyses with GENEHUNTER and ALLEGRO did not reveal any evidence of linkage for this binary trait. However, the results of the analyses using the ordinal trait revealed linkage to three regions on three chromosomes (4q, 5q, and 17q). Thus, the power of detecting linkage was increased by the use of this ordinal trait in chromosomes 4 and 5. The significances of plod/lod scores are shown in Table 3. Specifically, in the region of 4q34–35, the best significance levels are .0009. In the region of 5q35.2–35.3, the best significance levels are .00001; and in the region of 17q25, the best significance level is .0005.

Fig. 1 provides graphical comparisons of the parametric and nonparametric lod scores from GENEHUNTER and ALLEGRO, and plod/lod scores from the LVM. These values are comparable with those obtained by Zhang *et al.* (33) when hoarding was treated as a continuous variable derived from factor analyses. It is important to note that although Zhang *et al.* (33) assumed a continuous trait, the trait has only three distinct values.

Discussion

We proposed a framework to map candidate genes for an ordinal trait by using a latent-variable, proportional-odds model. Our simulation studies clearly support the substantial gain of power by using an ordinal, as opposed to a dichotomized, trait. One may wonder whether this gain of power is at the cost of an increased type I error. Our simulations demonstrate that the locus-specific type I errors are close to the nominal level for the LVM. Using a specified allele frequency, our simplified parametric approach proves to be highly robust in detecting linkage when the specified frequency is very different from the true one. Because the severity of many health conditions are recorded on ordinal scales, our model can be employed successfully to study the genetic basis of complex traits.

Our model is very basic, and it is still in its infancy. Within our framework, it is important to investigate issues including ascertainment bias, genetic heterogeneity, inbreeding, and imprinting. Although our simulations demonstrate the great promise of our model, our experiments are still limited. Further simulation and theoretical studies of our model are warranted to understand the properties of our model.

We have used a special set of latent variables, namely, Bernoulli random variables. Preliminary evidence suggests that these latent variables work reasonably well even if the underlying latent variables are continuous (13, 17).

We applied our model to conduct a genome scan of hoarding. Significant linkage to specific loci on 4q, 5q, and 17q were found. The 4q site is in proximity to D4S1625, which was identified by the TSAICG as a region linked to the GTS phenotype. The other two regions, 5q and 17q, show the strongest evidence for linkage. Duggirala *et al.* (39) found strong evidence for linkage of smoking behavior to a genetic location near D5S1456. Increased allele sharing at this marker was also in pedigrees with bipolar subjects in a National Institute of Mental Health collaborative initiative for genetics of bipolar disorders (40). The region on 17q has been linked to many other diseases, including autoimmune diseases (41) and schizophrenia (42).

Overall, our results are consistent with the analysis given in ref. 33. They performed a genomewide scan of the same family data by treating the hoarding score as a quantitative trait. Normality is generally assumed for analyzing quantitative traits. In this data set, the hoarding score used in Zhang *et al.* (33) came from two yes–no questions, and the normality is clearly a concern. It is important that we confirmed the earlier results by the new model and a more rigorous approach.

We thank Dr. David Pauls and the TSAICG for the reuse of the hoarding data. This work was supported in part by National Institute on Drug Abuse Grants DA012468, DA017713, and DA016750.

1. Hall, J. M., Lee, M. K., Newman, B., Morrow, J. E., Anderson, L. A., Huey, B. & King, M. C. (1990) *Science* **250,** 1684–1689.
2. Claus, E. B., Risch, N. J. & Thompson, W. D. (1990) *Am. J. Epidemiol.* **131,** 961–972.
3. Ott, J. (1999) *Analysis of Human Genetic Linkage* (John Hopkins Univ. Press, Baltimore), 3rd Ed.
4. Blackwelder, W. C. & Elston, R. C. (1985) *Genet. Epidemiol.* **2,** 85–97.
5. Goldgar, D. E. (1990) *Am. J. Hum. Genet.* **47,** 957–967.
6. Schork, N. J. (1993) *Am. J. Hum. Genet.* **53**, 1306–1319.
7. Amos, C. I. (1994) *Am. J. Hum. Genet.* **54,** 535–543.
8. Kruglyak, L., Daly, M. J., Reeve-Daly, M. P. & Lander, E. S. (1996) *Am. J. Hum. Genet.* **58,** 1347–1363.
9. Blangero, J. & Almasy, L. (1997) *Genet. Epidemiol.* **14,** 959–964.
10. Heath, A. C. & Nelson, E. C. (2002) *Alcohol Res. Health* **26,** 193–201.
11. Steinke, J. W., Borish, L. & Rosenwasser, L. J. (2003) *J. Allergy Clin. Immunol.* **111,** S495–S501.
12. Vergne, L., Bourgeois, A., Mpoudi-Ngole, E., Mougnutou, R., Mbuagbaw, J., Liegeois, F., Laurent, C., Butel, C., Zekeng, L., Delaporte, E. & Peeters, M. (2003) *Virology* **310,** 254–266.
13. Zhang, H. P., Feng, R. & Zhu, H. (2003) *J. Am. Stat. Assoc.* **98,** 1023–1034.
14. Gudbjartsson, D. F., Jonasson, K., Frigge, M. L. & Kong, A. (2000) *Nat. Genet.* **25,** 12–13.
15. Corbett, J., Gu, C. C., Rice, J. P., Reich, T., Province, M. A. & Rao, D. C. (2004) *Hum. Heredity* **57,** 21–27.
16. Lander, E. S. & Green, P. (1987) *Proc. Natl. Acad. Sci. USA* **81,** 3443–3446.
17. Zhang, H. P. & Merikangas, K. (2000) *Biometrics* **56,** 815–823.
18. McCullagh, P. & Nelder, J. A. (1989) *Generalized Linear Models* (Chapman & Hall, London), 2nd Ed.
19. Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977) *J. Royal. Stat. Soc. B* **39,** 1–38.
20. Louis, T. A. (1982) *J. R. Stat. Soc. B* **44,** 226–233.
21. Meilijson, I. (1989) *J. R. Stat. Soc. B* **51,** 127–138.
22. Morton, N. E. (1995) *Genetics* **140,** 7–12.
23. Self, S. G. & Liang, K. Y. (1987) *J. Am. Stat. Assoc.* **82,** 605–610.
24. Speer, M. C., Terwilliger, J. D. & Ott, J. (1995) *Genet. Epidemiol.* **12,** 561–564.
25. Hodge, S. E. (1995) *Genet. Epidemiol.* **12,** 555–560.
26. Feingold, E., Brown, P. O. & Siegmund, D. S. (1993) *Am. J. Hum. Genet.* **53,** 234–251.
27. Ulgen, A., Yoo, J. Y., Gordon, D., Finch, J. S. & Mendell, R. N. (2004) *Hum. Heredity* **57,** 39–48.
28. Karno, M., Golding, J. M., Sorenson, S. B. & Burnam, M. A. (1988) *Arch. Gen. Psychiatry* **45,** 1094–1099.
29. Christensen, D. D. & Greist, J. H. (2001) *Prim. Psychiatry* **8,** 79–86.
30. Alsobrook, J. P., II, Zohar, A. H., Leboyer, M., Chabane, N., Ebstein, R. P. & Pauls, D. L. (2002) *Am. J. Med. Genet.* **114,** 116–120.
31. Greenberg, D., Witztum, E. & Levy, A. (1990) *J. Clin. Psychiatry* **51,** 417–421.
32. Leckman, J. F., Pauls, D. L., Zhang, H. P., Rosario-Campos, M. C., Katsovich, L., Kidd, K. K., Pakstis, A. J., Alsobrook, J. P., Robertson, M. M., Walkup, J. T., *et al.* (2003) *Am. J. Med. Genet.* **116B,** 60–68.
33. Zhang, H. P., Leckman, J. F., Pauls, D. L., Tsai, C.-P., Kidd, K. K. & Campos, M. R. (2002) *Am. J. Hum. Genet.* **70,** 896–904.
34. King, R. A., Leckman, J. F., Scahill, L. D. & Cohen, D. J. (1998) in *Tourette's Syndrome Tics, Obsessions, Compulsions: Developmental Psychopathology and Clinical Care*, eds. Leckman, J. F. & Cohen, D. J. (Wiley, New York), pp. 43–62.
35. Horwath, E. & Weissman, M. M. (2000) *Psychiatr. Clin. N. Am.* **23,** 493–507.
36. Pauls, D. L., Raymond, C. L., Stevenson, J. M. & Leckman, J. F. (1991) *Am. J. Hum. Genet.* **48,** 154–163.
37. American Psychiatric Association (1994) *Diagnostic and Statistical Manual of Mental Disorders* (Washington, DC), 4th Ed.
38. Risch, N. (1990) *Am. J. Hum. Genet.* **46,** 229–241.
39. Duggirala, R., Almasy, L. & Blangero, J. (1999) *Genet. Epidemiol.* **17S1,** S139–S144.
40. Edenberg, H. J., Foroud, T., Conneally, P. M., Sorbel, J. J., Carr, K., Crose, C., Willig, C., Zhao, J., Miller, M., Bowman, E., *et al.* (1997) *Am. J. Med. Genet.* **74,** 238–246.
41. Jawaheer, D., Seldin, M. F., Amos, C. I., Chen, W. V., Shigeta, R., Monteiro, J., Kern, M., Criswell, L. A., Albani, S., Nelson, J. L., *et al.* (2001) *Am. J. Hum. Genet.* **68,** 927–936.
42. Riley, B. P., Tahir, E., Rajagopalan, S., Mogudi-Carter, M., Faure, S., Weissenbach, J., Jenkins, T. & Williamson, R. (1997) *Psychiatr. Genet.* **7,** 57–74.