Interval mapping for loci affecting unordered categorical traits

T Hayashi and T Awata

Laboratory of Animal Genome, National Institute of Agrobiological Sciences, 2 Ikenodai, Tsukuba-shi, Ibaraki 305-0901, Japan

Many traits including shapes and colors of flowers, fruits and seeds in plants, as well as coat colors and some behavioral properties in animals, are recorded in discrete categories. If categories are ordered, genetic analyses of the categorical traits are often performed using the threshold model, which considers a latent continuous variable, called the liability, underlying a trait and assumes the monotonic relationship between the phenotype and the liability. In some categorical traits, however, descriptions of phenotypes are purely nominal and the phenotypic scores cannot be ordered. The threshold model is unreasonable for the analyses of such unordered categorical traits. In this study, we developed a method for interval mapping of loci affecting unordered categorical traits with more than two categories. The probability of the phenotype of an individual falling in each

of the categories was expressed by a polychotomous logistic model, in which the log-odds for each category relative to the reference category were assumed to follow a linear model including genotype at a locus affecting a trait as covariate. Based on the model, the interval mapping using a maximum likelihood method was devised for the analysis of complex categorical traits described with unordered categories. We confined ourselves to the case of F₂ populations derived from a cross between two inbred lines, although this approach can easily be extended to the analyses for other populations of general structures. As results of analyses of simulated data show, the method showed high efficiency in detecting the loci affecting unordered categorical traits.

Heredity (2006) **96**, 185–194. doi:10.1038/sj.hdy.6800783; published online 4 January 2006

Keywords: categorical trait; unordered category; interval mapping; maximum likelihood; linkage analysis

Introduction

Many traits including shapes and colors of flowers, fruits and seeds in plants, as well as coat colors and some behavioral properties in animals, are recorded in discrete categories. The scores of sensory tests evaluating eating-quality traits such as flavors or tastes of fruits and meats are also described with categories. If categorical phenotypes are ordered, the discrete distribution of phenotypes is often treated using the threshold model, where a latent continuous variable called the liability underlying a categorical trait is considered and the categorical phenotype and the liability are linked through the fixed thresholds, assuming monotonic relationship between the phenotype and the liability. The threshold model was mainly developed for the genetic analysis of binary categorical traits (Dempster and Lerner, 1950) and extended to the case of polychotomous traits with more than two categories (Gianola, 1979; Sorensen *et al.*, 1995). For locating loci affecting categorical traits on the linkage map, Xu and Atchley (1996) developed a method of interval mapping for the analysis of binary traits based on the threshold model, where liability was expressed as a linear model including the genotypes of loci involved as covariates. Yi and Xu (2000) devised a statistical procedure of Bayesian estimation for mapping binary trait loci using

a similar model. The methods of interval mapping based on the threshold model were extended to the ordered categorical traits with more than two categories (Rao and Xu, 1998; Rao and Li, 2001; Xu *et al.*, 2005).

For some categorical traits, however, descriptions of phenotypes are purely nominal and phenotypes are scored with unordered categories. When unordered categorical traits are controlled by multiple loci, the decomposition of the phenotypic score into the contribution of each individual locus using the genetic markers is required for mapping each of the loci to clarify the number of loci involved and inheritance mode of the traits as in the case of QTL analyses. The threshold model assuming a liability monotonically related with the phenotypic value is not applicable to the analyses of unordered categorical traits.

The variations in flower colors (Clegg and Durbin, 2000), coat patterns and colors of seeds (McClean *et al.*, 2002) and coat colors of animals (Klungland and Vage, 2000) are typical examples for unordered categorical traits. Moreover, scores obtained by the sensory test for eating qualities of fruits or meats, which are composite traits affected by many constituent biochemical attributes, cannot be described by a simple quantitative genetic model, thus, should be regarded as unordered categorical traits, although the scores may be ordered according to the evaluation of flavors, textures and so on, and have been analyzed with a method based on the conventional linear model (King *et al.*, 2000; Causse *et al.*, 2001, 2002).

There are few methods suitable for linkage analyses of unordered categorical traits affected by multiple loci. In this study, we developed a method for interval mapping

Correspondence: T Hayashi, Laboratory of Animal Genome, National Institute of Agrobiological Sciences, 2 Ikenodai, Tsukuba-shi, Ibaraki 305-0901, Japan. E-mail: hayatk@affrc.go.jp

Received 6 April 2005; accepted 23 November 2005; published online 4 January 2006

of loci affecting categorical traits with more than two categories, phenotypic scores of which are regarded as unordered and purely nominal. We considered that the probability of phenotype of an individual being classified into each of the categories depended on covariates including the genotypes of putative loci affecting a trait. This probability was described using a polychotomous logistic regression model, in which the log-odds for each category relative to a reference category were expressed using a linear regression model. Based on the model, the interval mapping using a maximum likelihood method was devised to analyze a categorical trait for mapping loci affecting the trait.

In the study of Ayyadevara *et al* (2003), a method utilizing logistic regression for the probability of category was applied for dichotomous phenotype and referred to as categorical trait interval mapping (CTIM). Categorical trait interval mapping was based on a latent class model in which unobservable genotypes at a locus affecting a trait were regarded as latent classes and performed a maximum likelihood estimation with iterative reweighted least squares (IRLS) algorithm (Galecki *et al*, 2001). However, a statistical model underlying CTIM has never been explicitly described. In our method, although logistic regression was utilized for modeling the probability of category as well, mixture multinomial distribution was considered as a likelihood function and EM algorithm was used for maximum likelihood estimation instead of a latent class model and IRLS algorithm, respectively. Our method is more flexible for inclusion of additional covariates in the model and extension to mapping of multiple loci than CTIM.

In the threshold model for ordered categorical traits, the probability of an individual being classified into each category could be expressed as the difference between two cumulative normal probabilities corresponding to two thresholds defining borders of a category assuming normality for a distribution of the liability. A logistic distribution has been used as the approximation of the normal probability in the analyses of binary traits by Xu and Atchley (1996) and ordered categorical traits by several authors for the purpose of computational feasibility (Hackett and Weller, 1995; Rao and Xu, 1998; Rao and Li, 2001). For treating multinomial probabilities of unordered categories, polychotomous logistic regression model is also a suitable tool.

We describe a new method for the case of analysis of an unordered categorical trait affected by multiple loci segregating in an F2 population derived from a cross between two inbred lines, although the method can easily be applied to the analyses of other populations with different structures. The power of detecting loci affecting a categorical trait was evaluated for the developed method with simulated F2 data. As a result of analyses of simulated data, the method showed high efficiency in detecting the true association between genomic regions and phenotypes.

Materials and methods

Statistical model

Let us consider an F2 population consisting of n individuals derived from a cross between two inbred lines, P1 and P2, fixed for alternative allele at loci

affecting an unordered categorical trait of interest and markers. We assume that the marker genotypes are obtained for P1, P2 and F2 individuals and the phenotypes of a categorical trait are observed for F2 individuals with the phenotypic score of each individual being classified into one of $m+1$ unordered categories, denoted by $C_0, C_1, \dots, C_m$. In the procedure of interval mapping, a putative locus, L , affecting the trait is located at the tested position of a linkage map and the probabilities of three possible genotypes at L for the i th F2 individual ($i = 1, 2, \dots, n$), denoted as $q_{i1} = \text{Prob}(QQ)$, $q_{i2} = \text{Prob}(Qq)$ and $q_{i3} = \text{Prob}(qq)$, where Q and q indicate two alleles at L transmitted from P1 and P2 to the F2 individual, respectively, are inferred based on the genotypes of markers flanking L .

The probability of the phenotypic score of the i th individual, y_i , being classified into C_j is denoted by p_{ijk} ($i = 1, 2, \dots, n$; $j = 0, 1, \dots, m$; $k = 1, 2, 3$), when the individual has genotype k , where the three genotypes QQ , Qq and qq at a putative locus L located on a tested position are indicated by 1, 2 and 3, respectively. When there really exists a locus affecting a trait near a tested position, there should be the dependence of probabilities p_{ijk} on the genotypes k ($k = 1, 2, 3$). We adopted a polychotomous logistic function, $h_{ijk} = h(p_{ijk})$, as link function of a linear model for p_{ijk} , with C_0 used as a reference category (McCullagh and Nelder, 1989), that is, $h_{ijk} = \log(p_{ijk}/p_{i0k})$ ($j = 1, 2, \dots, m$), and used a logistic regression for modeling the dependence of p_{ijk} on the genotype of a locus,

$$h_{ijk} = \sum_{l=1}^f \alpha_{jl} c_{il} + \beta_j u_{ik} + \gamma_j v_{ik}, \quad (1)$$

where (u_{ik}, v_{ik}) are the indicator variables for the genotypes of a putative locus located in the tested position taking values (1, 0), (0, 1) and $(-1, 0)$ for $k = 1, 2$ and 3 corresponding to genotypes QQ , Qq and qq , respectively; β_j and γ_j are regression coefficients corresponding to explanatory variables u_{ik} and v_{ik} , respectively; α_{jl} indicates the l th fixed effect ($l = 1, 2, \dots, f$) that may include the overall mean, sex effect, systematic environmental effect and some genetic contributions from other loci than the tested locus; and c_{il} is a component of the design matrix relating α_{jl} to h_{ijk} . Each of the probabilities, p_{ijk} ($j = 0, 1, \dots, m$), is expressed in terms of h_{ijk} as

$$p_{i0k} = \frac{1}{1 + \sum_{j=1}^m \exp(h_{ijk})}$$

and

$$p_{ijk} = \frac{\exp(h_{ijk})}{1 + \sum_{j=1}^m \exp(h_{ijk})}, \quad (j = 1, 2, \dots, m) \quad (2)$$

Maximum likelihood procedure for mapping loci

The method of maximum likelihood is used to estimate regression coefficients. Significant values of regression coefficients indicate significant evidence for the existence of a locus affecting a trait around the tested position. Given the phenotypic observations for F2 individuals, the likelihood function L_1 under model (1) is expressed as

$$L_1 = \prod_{i=1}^n \left(\sum_{k=1}^3 q_{ik} \prod_{j=0}^m p_{ijk}^{z_{ij}} \right)$$

where z_{ij} ($j=0,1,\dots,m$) is a variable indicating the category to which the phenotype of the i th individual belongs, that is, for y_i being classified into C_j , we obtain $z_{ij}=1$ and $z_{ik}=0$ ($k\neq j$). To obtain maximum likelihood estimates for regression coefficients, numerical calculation is required. For a binary trait ($m=1$), Xu and Atchley (1996) applied an EM algorithm to maximize the likelihood function. We modified the algorithm for the case of polychotomous categorical traits to maximize L_1 (see Appendix A).

Under the null model, in which we assume that there are no loci affecting the trait, linked to the tested position, both regression coefficients of explanatory variables for genotype of L , β_j and γ_j , are set at 0. The likelihood function under the null model is denoted by L_0 . The log-likelihood ratio test statistic, $LRT=2\log(\max L_1/\max L_0)$, is calculated at each tested position and used for detecting the association between the tested position and loci involved in a categorical trait. The significance level for LRT can be obtained by numerical methods such as permutation test (Churchill and Doerge, 1994).

Simulation studies

Designs of simulation experiments

To evaluate the power of detecting loci affecting an unordered categorical trait using the proposed method, we carried out simulation experiments. A genome consisting of two chromosomes (chromosome 1 and chromosome 2), each with length 100 cM, on each of which 11 markers were located at every 10 cM, were considered for an F2 population derived from crossing two inbred lines, P1 and P2. Two loci, L1 and L2, affecting a categorical trait were located at 55 cM on chromosome 1 and at 35 cM on chromosome 2, respectively. The two alleles at L1 derived from P1 and P2 were denoted by Q1 and q1, respectively, and those at L2 were denoted by Q2 and q2, respectively. The phenotypic score of a categorical trait was determined for each F2 individual depending on the underlying continuous variable determined by genetic effects at L1 and L2 and the environmental effects, which we called the liability as in the threshold model for ordered categorical traits. However, we assumed that the relation between the phenotypic score and the liability was not monotonic to make categories unordered. The genetic effects at L1 were denoted by a_1 , d_1 and $-a_1$ for genotypes Q1Q1, Q1q1 and q1q1, respectively, and those at L2 were denoted by a_2 , d_2 and $-a_2$ for genotypes Q2Q2, Q2q2 and q2q2, respectively. In the simulations, the values of a_1 , d_1 , a_2 and d_2 were fixed at 0.4, 0.0, 0.0 and 0.6, respectively.

Two cases corresponding to different genetic scenarios were considered to simulate categorical phenotype of each F2 individual based on the liability. In the first case (Case I), we assumed that the liability, x , for each individual in the F2 population was given as a sum of genetic effects at two loci and environmental effect, e , sampled from normal distribution with mean 0 and variance 1. There are assumed to be three categories, C_0 , C_1 and C_2 , and phenotype of each F2 individual was scored as one of C_0 , C_1 and C_2 depending on the value of the liability and the two thresholds, which were set at 0.0 and 0.6. We assigned the phenotype of an individual with liability x to C_0 , C_1 and C_2 corresponding to

$0.0\leq x<0.6$, $x\geq 0.6$ and $x<0.0$, respectively. Thus, three phenotypes, C_0 , C_1 and C_2 , which were not monotonically related with the liability, could be regarded as unordered categories.

In the second case (Case II), we assumed that the gene expressions at L1 and L2 were independently controlled in a binary mode and two states of gene expression, that is, 'expressed' and 'not expressed', were denoted by 1 and 0, respectively. For example, we indicated a pattern of gene expression, in which gene at L1 is expressed and gene at L2 is not expressed, with '(1,0)'. Moreover, we assumed that the gene expression at each of L1 and L2 was determined by each of the underlying liabilities, x_1 and x_2 , which were obtained as the sums of the genetic effects and the environmental effects, denoted as e_1 and e_2 , separately for L1 and L2. The environmental effects were assumed to independently influence the gene expression of L1 and L2, respectively, and were sampled from normal distribution with mean 0 and variance 1. We supposed that gene at each of L1 and L2 was expressed when each of the liabilities, x_1 and x_2 , exceeded each of the predetermined threshold values, which were set at 0. The phenotype of a categorical trait was classified into C_0 , C_1 , C_2 and C_3 corresponding to the four combinations of gene expressions at a pair of loci, (1,1), (0,0), (0,1) and (1,0), respectively. Consider, for example, an F2 individual with genotypes Q1Q1Q2Q2 at L1 and L2 and the non-genetic noises $e_1=0.1$ and $e_2=-0.1$. For the individual, the liabilities are obtained as $x_1=a_1+e_1=0.5$ and $x_2=a_2+e_2=-0.1$; thus, the pattern of gene expression at L1 and L2 is denoted as (1,0) meaning that the gene is expressed at L1 and not expressed at L2. Accordingly, the phenotypic score of the F2 individual is classified into C_3 . The phenotypic scores thus obtained were not ordered and the categories were regarded as purely nominal.

F2 populations consisting of 400 individuals were simulated and analyzed with the method developed here. Simulation was repeated 1000 times to evaluate the power in detecting L1 and L2 and the precision in the estimates of their locations and the coefficients in the logistic regression model for each of Case I and Case II. The genome-wide threshold for LRT was determined from empirical null distribution obtained by the analyses of two chromosomes for 5000 data sets of F2 populations with 400 individuals generated under the assumption that all parameters of genetic effects, a_1 , d_1 , a_2 and d_2 , were set at 0 for each case. When a peak of LRT exceeded a threshold of genome-wide 5% significance level on each of two chromosomes; detection of each of L1 and L2 was regarded as successful. The number of simulations with successful detection was counted to obtain the power of detection for each of L1 and L2. At the same time, the estimated locations of L1 and L2 are obtained as the tested positions corresponding to the peak of LRT and the estimated values of logistic regression coefficients included in model (1) at the positions were recorded on each of two chromosomes to obtain the statistical properties of the estimates. For simplicity, only overall mean was included as non-genetic fixed effect in the model used for analyses. For comparison, we applied non-parametric interval mapping (NPIM) method proposed by Kruglyak and Lander (1995), which requires no assumptions about trait distribution, to the same simulated data, treating the j th category, C_j , as a regular quantitative phenotypic value, j ($j=0,1,2$ for Case I and

$j=0,1,2,3$ for Case II). We used the statistic $X_W^2 = Z_A^2 + Z_D^2$ for mapping loci in simulated F2 data sets with NPIM, where Z_A and Z_D were test statistics for detecting additive and dominance effects, respectively (see Kruglyak and Lander, 1995). NPIM was successfully applied for the analysis of a dichotomous categorical trait as well as CTIM in the study of Ayyadevara *et al* (2003).

In each of two cases, the probability of an F2 individual being classified into each of these categories can be theoretically obtained corresponding to the genotype of the individual at L1 and L2. Consider, for example, the probability of an individual with genotype Q1Q1Q2Q2 at L1 and L2 being classified into each of three categories in Case I. The mean of liabilities of individuals with genotype Q1Q1Q2Q2 at L1 and L2 was written as $x = a_1 + a_2 + e = 0.4 + e$. Accordingly, $x < 0.0$, $0.0 \leq x < 0.6$ and $x \geq 0.6$ corresponded to $e < -0.4$, $-0.4 \leq e < 0.2$ and $e \geq 0.2$, respectively; thus, the probabilities of the individual being classified into C_0 , C_1 , and C_2 were calculated as $\Phi(0.2) - \Phi(-0.4) = 0.235$, $1 - \Phi(0.2) = 0.421$, $\Phi(-0.4) = 0.345$, respectively, where $\Phi(x)$ denotes the cumulative distribution function for a normal variable with mean 0 and variance 1. Similarly the probabilities of categories for individuals with other genotypes at L1 and L2 were obtained. Table 1 lists the probability of an individual being classified into each of categories given the genotype at L1 and L2. From Table 1 and model (1), the theoretical expected values of regression coefficients in the simulated data of Case I are obtained as follows: $(\alpha_{11}, \alpha_{21}, \beta_1, \beta_2, \gamma_1, \gamma_2) = (0.559, 0.559, 0.380, -0.380, -0.020, -0.020)$ for L1 and $(\alpha_{11}, \alpha_{21}, \beta_1, \beta_2, \gamma_1, \gamma_2) = (0.257, 0.830, 0.0, 0.0, 0.573, -0.573)$ for L2.

Likewise, we can calculate the probabilities of categories for individuals in Case II. Consider, for example, the probability of an individual with genotype Q1Q1Q2Q2 being classified into each of four categories. For the individual, we obtain $x_1 = 0.4 + e_1$ and $x_2 = 0.0 + e_2$. Thus, the probabilities of $x_1 > 0$ and $x_2 > 0$ are $\Phi(0.4)$ and $\Phi(0.0)$, respectively. Accordingly, the probability that the gene expression pattern is (1,1) for a pair of L1 and L2, which classifies the individual into

C_0 , is $\Phi(0.4)\Phi(0.0) = 0.328$. Similarly, the probabilities that the individual is classified into other categories can be calculated. We list the probability of the individual being classified into each of categories given the genotype at L1 and L2 in Table 2 for Case II. From Table 2 and model (1), the theoretical expected values of regression coefficients in the simulated data are obtained as follows: $(\alpha_{11}, \alpha_{21}, \alpha_{31}, \beta_1, \beta_2, \beta_3, \gamma_1, \gamma_2, \gamma_3) = (-0.460, 0.0, -0.460, -0.646, -0.645, 0.0, 0.045, 0.0, 0.045)$ for L1 and $(\alpha_{11}, \alpha_{21}, \alpha_{31}, \beta_1, \beta_2, \beta_3, \gamma_1, \gamma_2, \gamma_3) = (0.0, 0.0, 0.0, 0.0, 0.0, 0.0, -0.974, 0.0, -0.974)$ for L2.

Results of simulation

Figures 1 and 2 show the plots of *LRT* at each tested position on two chromosomes in the analysis for one of data sets generated in simulations using the proposed method in Case I and Case II, respectively. For comparison, the result of an analysis of the same data set using NPIM method regarding the categories as regular quantitative phenotypic values is also shown in Figures 1 and 2, where threshold values of 5% significance level, obtained based on 5000 data sets simulated from the null model with all of a_1 , d_1 , a_2 and d_2 being set at 0, were also indicated for both methods. In the analyses of the data set, both L1 and L2 could be detected around the true positions by the proposed method, while NPIM method could detect neither of L1 and L2 in Case I and Case II.

The power of detection for L1 and L2 using the proposed method is listed in Tables 3 and 4 for Case I and Case II, respectively, along with the statistical properties of estimates for the positions of detected loci and regression coefficients in the logistic model obtained at the positions. It was shown in Table 3 that L1 and L2 were successfully detected in 846 data sets and 811 data sets, respectively, of 1000 simulated data sets with the proposed method for Case I, while L1 and L2 were only detected in 366 data sets and 363 data sets with NPIM method. Similarly, the proposed method could detect L1 and L2 with higher frequencies than NPIM method for Case II as shown in Table 4, where L1 and L2 were

Table 1 Probability of categories given genotypes at L1 and L2 in Case I

L1	L2			Probability of categories of the given genotypes at L1
	Q2Q2	Q2q2	q2q2	
Q1Q1	$s_{011}^a = 0.235$	$s_{012} = 0.186$	$s_{013} = 0.235$	$p_{01}^b = 0.210$
	$s_{111} = 0.421$	$s_{112} = 0.655$	$s_{113} = 0.421$	$p_{11} = 0.538$
	$s_{211} = 0.345$	$s_{212} = 0.159$	$s_{213} = 0.345$	$p_{21} = 0.252$
Q1q1	$s_{021} = 0.226$	$s_{022} = 0.226$	$s_{023} = 0.226$	$p_{02} = 0.226$
	$s_{121} = 0.274$	$s_{122} = 0.500$	$s_{123} = 0.274$	$p_{12} = 0.387$
	$s_{221} = 0.500$	$s_{222} = 0.274$	$s_{223} = 0.500$	$p_{22} = 0.387$
q1q1	$s_{031} = 0.186$	$s_{032} = 0.235$	$s_{033} = 0.186$	$p_{03} = 0.210$
	$s_{131} = 0.159$	$s_{132} = 0.345$	$s_{133} = 0.159$	$p_{13} = 0.252$
	$s_{231} = 0.655$	$s_{232} = 0.421$	$s_{233} = 0.655$	$p_{23} = 0.538$
Probability of categories given genotypes at L2	$p_{0.1}^c = 0.218$	$p_{0.2} = 0.218$	$p_{0.3} = 0.218$	
	$p_{1.1} = 0.282$	$p_{1.2} = 0.500$	$p_{1.3} = 0.282$	
	$p_{2.1} = 0.500$	$p_{2.2} = 0.282$	$p_{2.3} = 0.500$	

^a s_{ijk} stands for the probability of an individual with genotype j and k at two loci being classified into the category C_i ($i=0,1,2$; $j=1,2,3$; $k=1,2,3$).

^b p_{ij} stands for the probability of category C_i for the genotype j at L1, which is obtained as $p_{ij} = s_{ij1}/4 + s_{ij2}/2 + s_{ij3}/4$.

^c $p_{i.k}$ stands for the probability of category C_i for the genotype k at L2, which is obtained as $p_{i.k} = s_{i1k}/4 + s_{i2k}/2 + s_{i3k}/4$.

Table 2 Probability of categories given genotypes at L1 and L2 in Case II

L1	L2			Probability of categories of the given genotypes at L1
	Q2Q2	Q2q2	q2q2	
Q1Q1	$s_{011} = 0.328$	$s_{012} = 0.476$	$s_{013} = 0.328$	$p_{01.} = 0.402$
	$s_{111} = 0.172$	$s_{112} = 0.095$	$s_{113} = 0.172$	$p_{11.} = 0.133$
	$s_{211} = 0.172$	$s_{212} = 0.250$	$s_{213} = 0.172$	$p_{21.} = 0.211$
	$s_{311} = 0.328$	$s_{312} = 0.180$	$s_{313} = 0.328$	$p_{31.} = 0.254$
Q1q1	$s_{021} = 0.250$	$s_{022} = 0.363$	$s_{023} = 0.250$	$p_{02.} = 0.306$
	$s_{121} = 0.250$	$s_{122} = 0.137$	$s_{123} = 0.250$	$p_{12.} = 0.194$
	$s_{221} = 0.250$	$s_{222} = 0.363$	$s_{223} = 0.250$	$p_{22.} = 0.306$
	$s_{321} = 0.250$	$s_{322} = 0.137$	$s_{323} = 0.250$	$p_{32.} = 0.194$
q1q1	$s_{031} = 0.172$	$s_{032} = 0.250$	$s_{033} = 0.172$	$p_{03.} = 0.211$
	$s_{131} = 0.328$	$s_{132} = 0.180$	$s_{133} = 0.328$	$p_{13.} = 0.254$
	$s_{231} = 0.328$	$s_{232} = 0.476$	$s_{233} = 0.328$	$p_{23.} = 0.402$
	$s_{331} = 0.172$	$s_{332} = 0.095$	$s_{333} = 0.172$	$p_{33.} = 0.133$
Probability of categories given genotypes at L2	$p_{0.1} = 0.250$	$p_{0.2} = 0.363$	$p_{0.3} = 0.250$	
	$p_{1.1} = 0.250$	$p_{1.2} = 0.137$	$p_{1.3} = 0.250$	
	$p_{2.1} = 0.250$	$p_{2.2} = 0.363$	$p_{2.3} = 0.250$	
	$p_{3.1} = 0.250$	$p_{3.2} = 0.137$	$p_{3.3} = 0.250$	

For the notations, s_{ijk} , p_{ij} , and $p_{i.k}$ ($i = 0, 1, 2, 3$; $j = 1, 2, 3$; $k = 1, 2, 3$), see the legends in Table 1.

successfully detected in 683 data sets and 658 data sets, respectively, of 1000 simulated data sets with the proposed method, while the times of successful detection with NPIM method are only 37 and 128 for L1 and L2, respectively, out of 1000 repetitions.

The estimated positions of two loci were almost unbiased, although regression coefficients were estimated exaggeratedly, that is, the absolute values of the means of estimates tended to be larger than those of the theoretical values for both cases. Considering β_1 and β_2 for L1 in Case I, the means of the estimates were 0.415 and -0.407 as shown in Table 3, while the corresponding theoretical expected values were 0.380 and -0.380 , respectively. Larger biases were found for γ_1 and γ_3 at L2 in Case II, where means of estimates for γ_1 and γ_3 were both -1.113 as shown in Table 4 against the theoretical expected values of -0.974 . As the estimates for regression coefficients at each of L1 and L2 were averaged over the repetitions in which each of L1 and L2 were successfully detected, we can consider Beavis effect (Beavis, 1998; Xu, 2003) to explain the large biases of the estimates, as discussed later.

Discussion

So far, for the analyses of ordered categorical traits including binary and polychotomous discrete traits, threshold models have been mainly used (Dempster and Lerner, 1950; Gianola, 1979; Sorensen *et al.*, 1995; Xu and Atchley, 1996; Yi and Xu, 2000; Rao and Li, 2001; Xu *et al.*, 2005). In threshold models, phenotypic values are obtained by discretizing the liabilities with thresholds, where it is assumed that phenotypic values are monotonically related with liabilities. Accordingly, liability underlying an ordered categorical trait can be treated in the same way as the conventional phenotypic values of a quantitative trait and analyzed with a simple linear model including genetic effects at loci involved and environmental effect. However, the threshold model is not applicable for some categorical traits with nominal and unordered categorical phenotypes. Although we can

consider the liability underlying a trait even for such unordered categorical trait, it is difficult to describe the relation between the phenotype and the liability with a simple model. Therefore, in the paper, for the interval mapping of loci affecting unordered categorical traits, the probabilities of categories were modeled with polychotomous logistic regression using marker genotypes as covariates, where no detailed description about the genetic system underlying the phenotype was required for model construction.

We evaluated the usefulness of the proposed method for the analyses of unordered categorical traits with simulations. In simulations, we considered two cases to obtain categorical phenotypes, where phenotypic scores of individuals were generated from the underlying continuous variables affected by genetic effects and environmental effects, which we called as the liabilities as in the ordered categorical traits. In Case I, the liability x , was obtained as a sum of genetic effects contributed from two loci, L1 and L2, and an environmental effect and the categorical phenotypic score was determined corresponding to the three ranges into which the value of the liability fell, that is, the categorical phenotypic score was classified into C_0 , C_1 or C_2 depending on the ranges, $0.0 \leq x < 0.6$, $x \geq 0.6$ or $x < 0.0$, respectively. Therefore, the liability was rendered discrete values, 0, 1 and 2, via threshold values, 0.0 and 0.6, in Case I as in the usual threshold model; however, we assumed the non-monotonic relation between the discrete value and the liability to generate unordered categorical phenotypes.

In Case II, an unordered categorical trait with four categories was considered, where the phenotypes of individuals were determined according to the patterns of gene expression at two loci. Gene expression at each locus was determined as a sum of the genetic effect at the locus and an environmental effect influencing the locus. The phenotype of each individual was scored as C_0 , C_1 , C_2 or C_3 depending on the states of gene expression at L1 and L2, that is, (1,1), (0,0), (0,1) or (1,0), respectively. The genetic scenario used in Case II could be interpreted as the following practical situation. Consider, for example,

Figure 1 Plot of LRT and X^2_W against the position of linkage map obtained by the analyses of one simulated data set in Case I for chromosome 1 (a) and chromosome 2 (b). Solid curves and dotted curves indicate the results obtained by the proposed method and NPIM method, respectively. We used LRT and X^2_W as test statistics for detecting loci in the proposed method and NPIM method, respectively. NPIM method treats the category, C_j , as phenotypic value j ($j=0,1,2$) and requires no assumptions about trait distribution. Horizontal lines are threshold values of genome-wide 5% level obtained by the analyses with the proposed method (solid lines) and NPIM method (dotted line) based on 5000 simulated data sets generated under the null model, where there are no loci affecting the trait. The true positions of loci affecting the trait are indicated by arrows.

aroma of a fruit influenced by two volatiles, the presence of each of which is controlled by each of two different loci, L1 and L2. The aroma quality is assessed as the best (C_0) in the presence of both of the two volatiles, while in the absence of both volatiles the aroma quality is scored as inferior (C_1) and as the worse (C_2 and C_3) under the presence of either one of the two volatiles.

In the analyses of 1000 simulated data sets consisting of 400 F2 individuals, using the proposed method, each of two loci was successfully detected in more than 80% for Case I and in about 70% for Case II of all data sets and the estimates of locations of two loci were unbiased (Table 3 and Table 4). However, the biases were observed in the estimates of regression coefficients for two loci for both cases, where the absolute values of regression coefficients were overestimated. These biases seemed to be caused by Beavis effect (Beavis, 1998; Xu, 2003) as the

estimates of regression coefficients were averaged over only the repetitions with successful detection of each of loci. When the averages were taken over all of 1000 repetitions, irrespective of detecting loci or not, for the estimates of regression coefficients for each of loci obtained at a position showing a peak of LRT on each of two chromosomes, the biases were much reduced. For β_1 and β_2 at L1 in Case I, for example, the averages of estimates were changed from 0.415 to 0.384 and from -0.407 to -0.386 , respectively, and for γ_1 and γ_3 at L2 in Case II, the averages were changed from -1.113 to -0.998 and from -1.113 to -0.958 , respectively. The averages thus obtained showed good agreement with the true values (results not shown).

Using the proposed method for analyzing categorical traits with $m+1$ categories, the number of parameters to be estimated is more or equal to $3m$, and the number of

Figure 2 Plot of LRT and X^2_W against the position of linkage map obtained by the analyses of one simulated data set in Case II for chromosome 1 (a) and chromosome 2 (b). Solid curves and dotted curves indicate the results obtained by the proposed method and NPIM method, respectively. We used LRT and X^2_W as test statistics for detecting loci in the proposed method and NPIM method, respectively. NPIM method treats the category, C_j , as phenotypic value j ($j = 0, 1, 2, 3$) and requires no assumptions about trait distribution. Horizontal lines are threshold values of genome-wide 5% level obtained by the analyses with the proposed method (solid lines) and NPIM method (dotted line) based on 5000 simulated data sets generated under the null model, where there are no loci affecting the trait. The true positions of loci affecting the trait are indicated by arrows.

Table 3 Power of detecting loci and estimates of parameters obtained with 1000 repetitions of simulation in Case I

	Power	Estimate of position	Estimates of α	Estimates of β	Estimates of γ
L1	846 (366) ^a	54.70 ± 10.93 (53.14 ± 17.15)	$\alpha_{11} = 0.575 \pm 0.244$ $\alpha_{21} = 0.575 \pm 0.245$	$\beta_1 = 0.415 \pm 0.241$ $\beta_2 = -0.407 \pm 0.244$	$\gamma_1 = -0.028 \pm 0.374$ $\gamma_2 = -0.020 \pm 0.378$
L2	811 (363)	34.73 ± 5.92 (36.39 ± 11.34)	$\alpha_{11} = 0.254 \pm 0.227$ $\alpha_{21} = 0.878 \pm 0.203$	$\beta_1 = -0.006 \pm 0.238$ $\beta_2 = -0.009 \pm 0.210$	$\gamma_1 = 0.610 \pm 0.328$ $\gamma_2 = -0.627 \pm 0.332$

^aPower and estimate of position are listed in parentheses for NPIM method.

Theoretical expected values of parameters are $(\alpha_{11}, \alpha_{21}, \beta_1, \beta_2, \gamma_1, \gamma_2) = (0.559, 0.559, 0.380, -0.380, -0.020, -0.020)$ and $(0.257, 0.830, 0.0, 0.0, 0.573, -0.573)$ for L1 and L2, respectively.

parameters to be estimated is increased with the number of categories. Accordingly, the large sample size may be required for obtaining high power of detecting loci and reliable estimates for locations and regression coefficients in the case of many categories. Decreasing the sample size of F2 individuals to 300 in the simulations, the power of detecting loci was reduced to about 70% and

50% in Case I and Case II, respectively, (results not shown). When no individuals with specific genotypes are included in some categories for the case of analyzing a population of small size, the parameters corresponding to the categories and genotypes cannot be estimated. In such a case, the modification of model by omitting the relevant parameters would be necessary. The power of

Table 4 Power of detecting loci and estimates of parameters obtained with 1000 repetitions of simulation in Case II

	Power	Estimate of position	Estimates of α	Estimates of β	Estimates of γ
L1	683 (37) ^a	54.91 ± 12.10 (59.59 ± 24.29)	$\alpha_{11} = -0.480 \pm 0.429$ $\alpha_{21} = -0.009 \pm 0.238$ $\alpha_{31} = -0.479 \pm 0.284$	$\beta_1 = -0.757 \pm 0.392$ $\beta_2 = -0.724 \pm 0.208$ $\beta_3 = 0.002 \pm 0.288$	$\gamma_1 = 0.005 \pm 0.429$ $\gamma_2 = 0.016 \pm 0.352$ $\gamma_3 = 0.011 \pm 0.414$
L2	658 (128)	35.40 ± 6.76 (36.55 ± 15.60)	$\alpha_{11} = 0.061 \pm 0.219$ $\alpha_{21} = -0.010 \pm 0.248$ $\alpha_{31} = 0.055 \pm 0.224$	$\beta_1 = 0.004 \pm 0.240$ $\beta_2 = 0.009 \pm 0.249$ $\beta_3 = 0.010 \pm 0.236$	$\gamma_1 = -1.113 \pm 0.345$ $\gamma_2 = 0.014 \pm 0.333$ $\gamma_3 = -1.113 \pm 0.360$

^aPower and estimate of position are listed in parentheses for NPIM method.

Theoretical expected values parameters are $(\alpha_{11}, \alpha_{21}, \alpha_{31}, \beta_1, \beta_2, \beta_3, \gamma_1, \gamma_2, \gamma_3) = (-0.460, 0.0, -0.460, -0.646, -0.645, 0.0, 0.045, 0.0, 0.045)$ and $(0.0, 0.0, 0.0, 0.0, 0.0, 0.0, -0.974, 0.0, -0.974)$ for L1 and L2, respectively.

the method in detecting a locus is much influenced by the difference in probability distributions over categories among genotypes at a locus. The power is increased as the degree of difference in the distributions over categories among genotypes is more. In the extreme case of strict correspondence between a category and a genotype at a locus, where category of an individual is unequivocally determined by its genotype and the different genotypes result in the different categories, this locus would be easily detected. In this case, only one of $m + 1$ categorical probabilities is 1 with other probabilities being 0; thus, the logistic model might cause the problem of convergence in the estimation procedure because the probability cannot take value of 0 or 1 in the model as seen in Equation (2). This problem, however, would be practically circumvented by stopping the cycle in the numerical iterations as the likelihood exceeds the predetermined upper bound. On the other hand, the efficiency of the method is decreased as the difference in the categorical distribution among genotypes at a locus is small, which means that the affection of this locus on a trait is subtle. From this viewpoint, it seems that the difference in categorical probability among genotypes at L1 and L2 in the simulations, shown in Tables 1 and 2 (the parts of p_{ij} and p_{ik}), is not so large, thus, rather unfavorable case for detecting loci was considered. Nevertheless, the efficiencies of the proposed method were high in detecting loci and estimating the positions and parameters. Therefore, it is concluded that the proposed method works well for the analyses of unordered categorical traits without investigating other variable settings in simulations.

The efficiencies in the analyses of unordered categorical traits with NPIM method treating category C_j as phenotypic value j were much reduced as shown in Tables 3 and 4. Regarding a category as a phenotypic value, we can calculate total phenotypic variances and additive and dominance effects contributed from L1 and L2 by applying the conventional method of quantitative genetics to Tables 1 and 2. The total phenotypic variance σ_p^2 was calculated as 0.579, additive effect A_1 and dominance effect D_1 at L1 were as -0.143 and -0.024 , respectively, and additive effect A_2 and dominance effect D_2 at L2 were as 0.0 and -0.218 , respectively, from Table 1 in Case I. Likewise, we obtained the values of σ_p^2 , A_1 , D_1 , A_2 and D_2 as 1.237 , 0.070 , 0.0 , 0.0 and -0.226 , respectively, from Table 2 in Case II. Therefore, the fractions of variances explained by L1 and L2 in total phenotypic variance were obtained for F2 population as $(0.5A_1^2 + 0.25D_1^2)/\sigma_p^2 = 0.018$ and $(0.5A_2^2 + 0.25D_2^2)/\sigma_p^2 = 0.021$, respectively, in Case I, and 0.002 and 0.010 ,

respectively, in Case II. The low fractions explained by the loci in the total phenotypic variance caused the low power in detecting loci with NPIM method, that is, the times of successful detection of L1 and L2 were 366 and 363 in Case I and 37 and 128 in Case II, respectively, in total of 1000 repetitions as shown in Tables 3 and 4. Exchanging categories such that the monotonic relation holds between the liability and the categorical score may enhance the power with NPIM method. However, it would usually be difficult to obtain such suitable categorization for the practical data. Any methods based on the conventional linear models presuming ordered phenotypes are considered to be unsuitable for the analysis of unordered categorical traits.

In CTIM method, which is also applicable to unordered categorical traits, a latent class model was assumed about the categorical frequencies for unobserved genotypes at a putative locus (Ayyadevara *et al*, 2003). In latent class model, the categorical frequencies are considered separately for each combination of covariates in the model (Galecki *et al*, 2001); thus, the inclusion of additional covariates such as sex and seasonal effects in the model may increase the complexity in computation of maximum likelihood estimation. On the other hand, it is easy to implement additional covariates in model (1) and to extend our method to the mapping of multiple loci by inclusion of effects of multiple loci in model (1). A method for mapping of multiple loci affecting unordered categorical traits based on a logistic regression model including genotypes at multiple loci as covariates are now being developed in the framework of Bayesian method.

Colors of flowers and seeds in plants and coat colors in animals are typical examples of unordered categorical traits. Generally, the variation in colors is controlled by many loci involved in the biosynthetic pathway of pigments, including flavonoid, anthocyanin and melanin (Clegg and Durbin, 2000; Klungland and Vage, 2000; McClean *et al*, 2002). The underlying genetics of pigmentation is so complex that understanding the genetic factors causing phenotypic variation is still incomplete, although molecular studies are now advancing for some species. Moreover, some sensory scores regarding the eating quality of fruits and meats are regarded as unordered categorical phenotypes as the absolute values of scores have no biological meanings. Sensory scores are influenced by many biochemical attributes, thus, the genetic background determining the scores is much complicated. The method proposed in this paper enables the mapping of loci affecting a unordered categorical trait with a simple logistic regres-

sion model without assuming the underlying complex genetic model, and thus can provide useful information for subsequent detailed molecular analyses of genes relevant to an unordered categorical trait.

References

- Ayyadevara S, Ayyadevara R, Hou W, Thaden JJ, Shmookler Reis RJ (2003). Genetic loci modulating fitness and life span in *Caenorhabditis elegans*: categorical trait interval mapping in CL2a × Bergerac-BO recombinant-inbred worms. *Genetics* **163**: 557–570.
- Beavis WD (1998). QTL analyses: power, precision, and accuracy. In: Paterson AH (eds) *Molecular Dissection of Complex Trait*. CRC Press: New York. pp 145–162.
- Causse M, Saliba-Colombani V, Lecomte L, Rousselle P, Buret M (2002). QTL analysis of fruit quality in fresh market tomato: a few chromosome regions control the variation of sensory and instrumental traits. *J Exp Bot* **53**: 2089–2098.
- Causse M, Saliba-Colombani V, Lesschaevé I (2001). Genetic analysis of organoleptic quality in fresh market tomato. 2. Mapping QTLs for sensory attributes. *Theor Appl Genet* **102**: 273–283.
- Churchill GA, Doerge RW (1994). Empirical threshold values for quantitative trait mapping. *Genetics* **138**: 963–971.
- Clegg T, Durbin ML (2000). Flower color variation: a model for the experimental study of evolution. *Proc Natl Acad Sci USA* **97**: 7016–7023.
- Dempster ER, Lerner IM (1950). Heritability of threshold characters. *Genetics* **35**: 212–236.
- Galecki AT, Ten Have TR, Molenberghs G (2001). A simple and fast alternative to the EM algorithm for incomplete categorical data and latent class models. *Comput Stat Data Anal* **35**: 265–281.
- Gianola D (1979). Heritability of polychotomous characters. *Genetics* **93**: 1051–1055.
- Hackett CA, Weller JL (1995). Genetic mapping of quantitative trait loci for traits with ordinal distributions. *Biometrics* **51**: 1252–1263.
- King GJ, Maliepaard C, Lynn JR, Alston FH, Durel CE, Evans KM *et al* (2000). Quantitative genetic analysis and comparison of physical and sensory descriptors relating to fruits flesh firmness in apple (*Malus pumila* Mill.). *Theor Appl Genet* **100**: 1074–1084.
- Klungland H, Vage DI (2000). Molecular genetics of pigmentation in domestic animals. *Curr Genom* **1**: 223–242.
- Kruglyak L, Lander ES (1995). A nonparametric approach for mapping quantitative trait loci. *Genetics* **139**: 1421–1428.
- McClellan PE, Lee RK, Otto C, Gepts P, Bassett MJ (2002). Molecular and phenotypic mapping of genes controlling seed coat pattern and color in common bean (*Phaseolus vulgaris* L.). *J Hered* **93**: 148–151.
- McCullagh P, Nelder JA (1989). *Generalized Linear Model*, 2nd edn. Chapman and Hall: London.
- Rao S, Li X (2001). Strategies for genetic mapping of categorical traits. *Genetica* **109**: 183–197.
- Rao S, Xu S (1998). Mapping quantitative trait loci for ordered categorical traits in four-way crosses. *Heredity* **81**: 214–224.
- Sorensen DA, Andersen S, Gianola D, Korsgaard I (1995). Bayesian inference in threshold model using Gibbs sampling. *Genet Sel Evol* **27**: 229–249.
- Xu C, Zhang YM, Xu S (2005). An EM algorithm for mapping quantitative resistance loci. *Heredity* **94**: 119–128.
- Xu S (2003). Theoretical basis of the Beavis effect. *Genetics* **165**: 2259–2268.
- Xu S, Atchley WR (1996). Mapping quantitative trait loci for complex binary disease using line crosses. *Genetics* **143**: 1417–1424.
- Yi N, Xu S (2000). Bayesian mapping of quantitative trait loci for complex binary traits. *Genetics* **155**: 1391–1403.

Appendix A

We maximized the log-likelihood function,

$$l_1 = \log L_1 = \sum_{i=1}^n \log \left(\sum_{k=1}^3 q_{ik} \prod_{j=0}^m p_{ijk}^{z_{ij}} \right)$$

via EM algorithm incorporating the Newton–Raphson iteration. For the Newton–Raphson iteration, a vector of the first partial derivatives and a matrix of the second partial derivatives, denoted by **d** and **J**, respectively, were required. The first partial derivatives of l_1 with respect to the parameters are obtained from (1) and (2) as

$$\frac{\partial l_1}{\partial \alpha_{jl}} = \sum_{i=1}^n c_{il} \left[\sum_{k=1}^3 w_{ik} (z_{ij} - p_{ijk}) \right],$$

$$j = 1, 2, \dots, m; l = 1, 2, \dots, f$$

$$\frac{\partial l_1}{\partial \beta_j} = \sum_{i=1}^n [w_{i1}(z_{ij} - p_{ij1}) - w_{i3}(z_{ij} - p_{ij3})],$$

$$j = 1, 2, \dots, m$$

and

$$\frac{\partial l_1}{\partial \gamma_j} = \sum_{i=1}^n w_{i2}(z_{ij} - p_{ij2}), \quad j = 1, 2, \dots, m$$

where

$$w_{ik} = \frac{q_{ik} \prod_{j=0}^m p_{ijk}^{z_{ij}}}{\sum_{k=1}^3 q_{ik} \prod_{j=0}^m p_{ijk}^{z_{ij}}}$$

is regarded as a posterior probability that the genotype of the i th individual at the tested locus is k .

The second partial derivatives of l_1 , which are the components of **J**, are calculated, regarding the posterior probabilities, w_{ik} ($i = 1, 2, \dots, n$; $k = 1, 2, 3$), as constants, and are given as follows:

$$\frac{\partial^2 l_1}{\partial \alpha_{jl} \partial \alpha_{jh}} = - \sum_{i=1}^n c_{il} c_{ih} \left[\sum_{k=1}^3 w_{ik} p_{ijk} (1 - p_{ijk}) \right]$$

$$\frac{\partial^2 l_1}{\partial \alpha_{j1} \partial \alpha_{j2}} = \sum_{i=1}^n c_{il} c_{ih} \left[\sum_{k=1}^3 w_{ik} p_{ij1k} p_{ij2k} \right]$$

$$\frac{\partial^2 l_1}{\partial \alpha_{jl} \partial \beta_j} = - \sum_{i=1}^n c_{il} [w_{i1} p_{ij1} (1 - p_{ij1}) - w_{i3} p_{ij3} (1 - p_{ij3})]$$

$$\frac{\partial^2 l_1}{\partial \alpha_{j1} \partial \beta_j} = \sum_{i=1}^n c_{il} [w_{i1} p_{ij1} p_{ij2} - w_{i3} p_{ij3} p_{ij2}]$$

$$\frac{\partial^2 l_1}{\partial \alpha_{jl} \partial \gamma_j} = - \sum_{i=1}^n c_{il} w_{i2} p_{ij2} (1 - p_{ij2})$$

$$\frac{\partial^2 l_1}{\partial \alpha_{j1} \partial \gamma_{j2}} = \sum_{i=1}^n c_{il} w_{i2} p_{ij1} p_{ij2}$$

$$\frac{\partial^2 l_1}{\partial \beta_j^2} = - \sum_{i=1}^n [w_{i1} p_{ij1} (1 - p_{ij1}) + w_{i3} p_{ij3} (1 - p_{ij3})]$$

$$\frac{\partial^2 l_1}{\partial \beta_{j1} \partial \beta_{j2}} = \sum_{i=1}^n [w_{i1} p_{ij1} p_{ij2} + w_{i3} p_{ij3} p_{ij2}]$$

$$\frac{\partial^2 l_1}{\partial \beta_j \partial \gamma_j} = \frac{\partial^2 l_1}{\partial \beta_{j1} \partial \gamma_{j2}} = 0$$

$$\frac{\partial^2 l_1}{\partial \gamma_j^2} = - \sum_{i=1}^n w_{i2} p_{ij2} (1 - p_{ij2})$$

and

$$\frac{\partial^2 l_1}{\partial \gamma_{j1} \partial \gamma_{j2}} = \sum_{i=1}^n w_{i2} p_{ij1} p_{ij2},$$

$$l, h = 1, 2, \dots, f, \quad j, j_1, j_2 = 1, 2, \dots, m, \quad j_1 \neq j_2$$

Denoting the parameters by θ , the values of θ maximizing l_1 can be obtained using the EM algorithm and the Newton–Raphson iteration, where, given the current values of θ , the posterior probabilities w_{ik} ($i = 1, 2, \dots, n; k = 1, 2, 3$) are calculated as E-step and the new values of θ , θ^* , are obtained as M-step using the formula, $\theta^* = \theta - J^{-1}d$, in the Newton–Raphson iteration with J and d being evaluated at the current value of θ . The values of θ are updated repeatedly until the convergence is attained. The converged values of θ are the maximum likelihood estimates of the parameters and the corresponding value of l_1 is the maximum log-likelihood.