MINING THE QUANTITATIVE TRAIT LOCI ASSOCIATED WITH ORAL GLUCOSE TOLERANCE IN THE OLETF RAT

A. NAKAYA,[1] H. HISHIGAKI,[2] and S. MORISHITA[1]

[1] *Institute of Medical Science, The University of Tokyo, 4-6-1, Shirokanedai Minato-ku, Tokyo 108-8639, Japan*

[2] *Otsuka GEN Research Institute, Otsuka Pharmaceutical Co., Ltd., 463-10, Kagasuno, Kawauchi-cho, Tokushima 771-0192, Japan*

Although the synergetic effects of multiple marker loci regarding quantitative traits such as blood glucose level have attracted interest, previous conclusions have been based on assumptions that each marker locus behaves independently of the other, leading to approximation. To cope with this problem, this paper focuses on the effects of multiple genetic factors and tries to find significant marker combinations by using conjunctive rules regarding genotypes at multiple marker loci. Application of the proposed method on the OLETF model rat of non-insulin dependent diabetes mellitus (NIDDM) has found significant combinations of marker loci with respect to oral glucose tolerance (OGT).

1 Introduction

Quantitative Trait Loci Analysis. Oral glucose tolerance (OGT), as well as factors such as body weight, fat weight, and insulin resistance, is an important quantitative trait significant to non-insulin dependent diabetes mellitus (NIDDM). Oral glucose tolerance (measured as the postprandial blood glucose level) is considered to be regulated by multiple *quantitative trait loci* (QTLs). In the investigation of these trait-causing loci, model rat strains of NIDDM have been developed, and some OGT-related loci have been mapped on the genome.[2,3] In QTL analysis, the genotypes at marker loci and observation of quantitative trait value in the individuals are given as input data. Marker genotypes take categorical values while quantitative traits take numerical values.

Linear Regression and LOD Score. As shown in previous studies,[1,3] the *interval mapping* method[6] based on a simple regression model has been widely used to map QTLs and find the OGT-related loci. The existence of a QTL within an interval flanked by a pair of neighboring marker loci is estimated along the genome. The logarithm of the likelihood ratio of the linkage between a marker locus and the quantitative trait against no linkage is called the *LOD score* (discussed later) and is calculated at each marker locus (or a putative locus in an interval with the genotype estimated from the flanking marker loci).

The OLETF Model Rat. A previous study employing the LOD score has identified OGT-causing QTLs on chromosomes in an F_2 intercross progeny of the Otsuka Long-Evans Tokushima Fatty (OLETF) rat.[4] The OLETF rat strain is an animal model of non-insulin dependent diabetes mellitus (NIDDM). These rats exhibit hyperglycemia, hyperinsulinemia, insulin resistance, and obesity, as well as showing glucose intolerance.[4] In our study we used F344 rats as a non-diabetic control strain. A cohort of male (female OLETF $\times$ male F344)F_2 intercross progeny including 157 rats was studied.

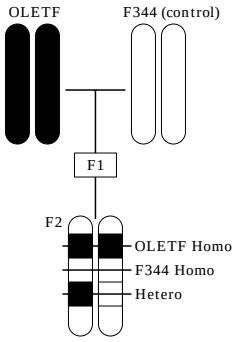

Figure 1: OLETF F_2.

We used 279 microsatellite markers to determine the genotype of each individual. As shown in Fig. 1, in the F_2 intercross progeny, marker loci on the autosomes indicate the genotypes of the OLETF homozygote, the F344 homozygote, and the heterozygote (O/O, F/F, and O/F, respectively). For the marker loci on the X-chromosome, we also use the notations of O/O and F/F for the hemizygote genotypes of O and F. Recently, more than five thousand microsatellite markers have been identified and shown to be densely spaced throughout the entire genome.[10] Thanks to the heigh density of the map, in this study we assume that QTLs are linked to marker loci.

Multiple Factors. However, pointwise estimation of the evidence for a QTL assumes that the markers are not correlated with each other. Therefore, the next task is to clarify the interactions between the trait-causing marker loci. In order to reflect the effects of the multiple marker loci on the explained trait value, a multiple linear regression model provides a theoretical expansion of the simple regression model.

However, as mentioned by Zeng,[11] a multiple linear regression model still

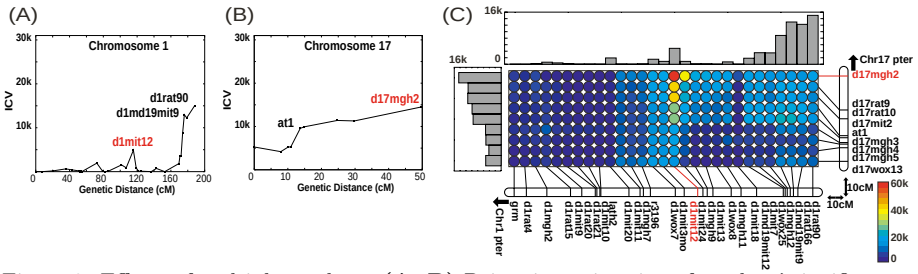

Figure 2: Effects of multiple markers. **(A, B)** Pointwise estimation of markers' significance along chromosomes 1 and 17. **(C)** Significance of marker pairs on chromosomes 1 $\times$ 17 (the red spot indicates a significant pair). Attached bar charts correspond to (A) and (B). We can observe the correlated effects between these chromosomes.

assumes additivity of the QTL effects between loci. Even if one can construct a multiple linear regression model which adequately explains a quantitative trait, it is difficult to interpret the model (i.e., its partial coefficiencies) except when the markers behave independently of each other.

Actually, selectivity and correlation between marker loci have been found in this study. Fig. 2A and Fig. 2B show the pointwise estimation of the significance of O/O genotype at marker loci along chromosome 1 and 17. The significance is expressed by inter-class variance (as discussed later, this quantity is equivalent to the LOD score). On chromosome 1, we have a major peak around marker *D1Rat90*, and a minor peak around marker *D1Mit12* (Fig. 2A). On the other hand, when we focus on the co-existence of O/O genotype at two marker loci on chromosomes 1 and 17, we have a peak only around the marker pair *D1Mit12* $\times$ *D17Mgh2* (the red spot in Fig. 2C). Thus, consideration of the synergetic effects between markers is indispensable for analysis of multiple factors.

2 Method

Association Studies. In a given dataset, association study tries to extract latent rules, for instance; *"If marker A is homozygous and marker B is also homozygous, then the trait value is high."*

The dataset consists of genotype information at marker loci and the quantitative trait values of interest in each individual. If we let $m_{j,i}$ denote the genotype at jth marker locus in the ith individual, and Φ_i denote the trait value in the ith individual, the total data can be summarized in a table as Fig. 3 (M and N are the numbers of markers and individuals, respectively). In this study, we call the judgment whether or not the jth marker locus has the genotype v in the ith individual a *primitive test* on the jth marker and denote it as $(m_{j,i} = v)$.

	Φ_i	$m_{1,i}$	$m_{2,i}$	$\cdots$	$m_{M,i}$
1	Φ_1	$m_{1,1}$	$m_{2,1}$	$\cdots$	$m_{M,1}$
2	Φ_2	$m_{1,2}$	$m_{2,2}$	$\cdots$	$m_{M,2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	Φ_N	$m_{1,N}$	$m_{2,N}$	$\cdots$	$m_{M,N}$

Figure 3: Dataset.

To investigate the relations between particular genotypes at multiple marker loci and the quantitative trait value, we use a conjunctive rule as follows:

$$G_i = (m_{j_1,i} = v_1) \text{ and} \cdots \text{and } (m_{j_k,i} = v_k), \tag{1}$$

where k is a given constant number and G_i returns true if all the primitive tests hold, otherwise returns false. According to whether or not each individual satisfies the rule G_i, we divide the set of individuals S into two subsets S_0 and S_1, letting S_0 and S_1 respectively consist of the individuals that do not satisfy G_i and those that satisfy it (we call this operation *division by* G_i). Here, if the rule can sort out a subset S_1 which contains most of the individuals with high

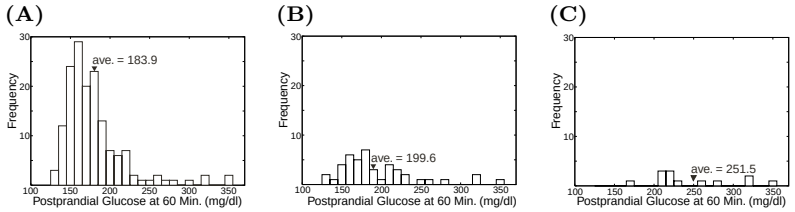

Figure 4: Data division in terms of genotype information. (**A**) The original distribution of the blood glucose level in the population of 157 OLETF rats. (**B**) The distribution in the population of 44 rats which satisfy the rule (*D1Rat90* = O/O). The peak (130-220 mg/dl) is blunted and the average shifted to the right (183.9 to 199.6 mg/dl). (**C**) The distribution in the population of 13 rats which satisfy the rule (*D1Rat90* = O/O) and (*D14Rat13* = O/O). The peak in (A) disappeared and the rats with poor glucose tolerance remained (average is 251.5 mg/dl).

trait values, then the marker loci which constitute the rule G_i are considered to be related to the trait. Fig. 4 shows an example of the division of the population of 157 OLETF rats in terms of genotype information.

The reason why we employed a conjunctive rule is that it can determine the correlated effects of the marker loci on the quantitative trait value. The traditional LOD score, on the other hand, misses the correlated effects since it essentially focuses on only the one-to-one relationships between a marker and the trait value.

Significance of a Rule. To evaluate the significance of the division by a rule, we use inter-class variance (ICV) defined as follows:

$$\mathrm{ICV} = |S_0|(\mu_0 - \mu)^2 + |S_1|(\mu_1 - \mu)^2, \tag{2}$$

where $|S_0|$ and $|S_1|$ are the numbers of individuals in S_0 and S_1 respectively, μ, μ_0, and μ_1 are respectively the average values of the quantitative trait in S, S_0, and S_1. The ICV indicates the degree of shift of the average values in the subsets (μ_0 and μ_1) from the average in the total dataset (μ) with reflecting the sizes of the subsets ($|S_0|$ and $|S_1|$). Greater inter-class variance means the division is more significant.

Note that the definition of the ICV (Eq. 2) is a simple function. On the other hand, as discussed in the next section, calculation of the LOD score requires maximization of the likelihood Eq. 5. Especially in the interval mapping method,[6] the maximization of the likelihood requires a method such as the EM algorithm and calculation steps are iterated until the likelihood converges to the maximum (it is not solved analytically). The evaluation of the putative marker loci between neighboring two loci has introduced the calculation above. However, if we assume that the density of today's markers is

sufficient for pointwise estimation of QTLs along the genome, ICV is suitable for evaluating the combinations of markers due to its simple definition.

3 Statistical Background of ICV

In this section we discuss the relation between the ICV and the LOD score. We introduce the definition and show that finding peaks of the LOD score along the genome is equivalent to finding those of the ICV.

Lod Score. On the assumption that the summation of the effects of multiple marker loci defines the quantitative trait, the LOD score analysis uses a simple regression model. In a population of N individuals, we use the genotypes of M marker loci and observational data of the quantitative trait. Using the above assumption, this method estimates the effect of each marker independently. For a given marker locus, let g_i be an indicator variable which shows the genotype in the ith individual, and when the marker genotype in the ith individual has a particular genotype (e.g., OLETF homozygous), g_i takes 1, otherwise it takes 0, and let Φ_i be the observation of the quantitative trait value in the ith individual. Using these variables we construct a regression model as follows (a and b are regression coefficiencies):

$$\Phi_i = a + bg_i + \varepsilon. \tag{3}$$

In this model, b corresponds to the effects of the genotype at the marker locus. ε is a normal variable with mean 0 and variance v. Using the assumption that the error term ε follows the normal distribution, the probability that ε takes a value x is defined as follows:

$$z(x, v) = \frac{1}{\sqrt{2\pi v}} \exp\left(\frac{-x^2}{2v}\right). \tag{4}$$

Thus, the probability that the observation of the quantitative trait values would have occurred under this parameterized model (likelihood) is

$$L(a, b, v) = \prod_{i=1}^{n} z(\Phi_i - (a + bg_i), v). \tag{5}$$

We determine the unknown parameters a, b, and v so that this likelihood is maximized and let the solutions (called maximum likelihood estimators) be $\hat{a}$, $\hat{b}$, and $\hat{v}$, respectively. The obtained likelihood $L(\hat{a}, \hat{b}, \hat{v})$ is compared to the likelihood that the marker locus has no effect on the quantitative trait (i.e., $b = 0$). Let $L(\hat{\mu}, 0, \hat{v}_0)$ denote the latter likelihood ($\hat{\mu}$ and $\hat{v}_0$ are the average

and the variance of the quantitative trait in the N individuals, respectively). The *LOD score* is the logarithm of the ratio:

$$\mathrm{LOD} = \log_{10}\left(\frac{L(\hat{a},\hat{b},\hat{v})}{L(\hat{\mu},0,\hat{v}_0)}\right) = \alpha(\ln L(\hat{a},\hat{b},\hat{v}) - \ln L(\hat{\mu},0,\hat{v}_0)), \tag{6}$$

where $\alpha = 1/ln10 > 0$ is a constant number. Note that the term $\ln L(\hat{\mu},0,\hat{v}_0)$ is constant. A LOD score greater than a statistical threshold constitutes evidence for a QTL. In this article, *maximization of the LOD score* means finding the marker locus which maximizes the LOD score.

Relation between LOD Score and ICV. Thus far, we introduced two statistics: the ICV and the LOD score. At a glance, they appear to be different. However, we can prove that calculation of the ICV of the data division in terms of the genotype at a single marker locus is essentially equivalent to that of the LOD score.

Let S_0 and S_1 be the populations of individuals whose g_i is 0 and 1, respectively ($S_0 = \{i|g_i = 0\}$ and $S_1 = \{i|g_i = 1\}$). Let μ_0 and μ_1 denote the averages of Φ_i in S_0 and S_1.

Lemma 1 *Maximization of the LOD score is equivalent to that of*

$$\ln L(\hat{a},\hat{b},\hat{v}) \quad = \quad -n\ln\sqrt{2\pi} + -\frac{n}{2}\ln\hat{v} - \frac{n}{2}, \tag{7}$$

where

$$\hat{v} \quad = \quad \frac{1}{n}\sum_{i=1}^{n}(\Phi_i - (\hat{a}+\hat{b}g_i))^2, \tag{8}$$

$$\hat{a} \quad = \quad \sum_{i\in S_0}\Phi_i/|S_0| = \mu_0, \quad \hat{b} = \sum_{i\in S_1}\Phi_i/|S_1| - \sum_{i\in S_0}\Phi_i/|S_0| = \mu_1 - \mu_0. \tag{9}$$

Proof. Likelihood is given by

$$L(a,b,v) \quad = \quad \prod_{i=1}^{n} z(\Phi_i - (a+bg_i), v) \tag{10}$$

$$\quad = \quad \left(\frac{1}{\sqrt{2\pi}}\right)^n \left(\frac{1}{v}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2v}\sum_{i=1}^{n}(\Phi_i - (a+bg_i))^2\right). \tag{11}$$

Log-likelihood is given by

$$\ln L(a,b,v) \quad = \quad -n\ln\sqrt{2\pi} - \frac{n}{2}\ln v - \frac{1}{2v}\sum_{i=1}^{n}(\Phi_i - (a+bg_i))^2. \tag{12}$$

By differentiating the log-likelihood with respect to v, a, and b, and setting the derivatives equal to zero, we have the maximum likelihood estimators of

v, a, and b as follows:

$$\frac{\partial \ln L}{\partial v} = -\frac{1}{2v}\left(n - \frac{1}{v}\sum_{i=1}^{n}(\Phi_i - (a + bg_i))^2\right) = 0, \quad (13)$$

$$\frac{\partial \ln L}{\partial a} = \frac{1}{\sigma^2}\left(\sum_{i\in S_0}(\Phi_i - a) + \sum_{i\in S_1}(\Phi_i - (a+b))\right) = 0, \quad (14)$$

$$\frac{\partial \ln L}{\partial b} = \frac{1}{\sigma^2}\sum_{i\in S_1}(\Phi_i - (a+b)) = 0. \quad (15)$$

$$\hat{v} = \frac{1}{n}\sum_{i=1}^{n}(\Phi_i - (\hat{a} + \hat{b}g_i))^2, \quad (16)$$

$$\hat{a} = \sum_{i\in S_0}\Phi_i/|S_0| = \mu_0, \quad \hat{b} = \sum_{i\in S_1}\Phi_i/|S_1| - \sum_{i\in S_0}\Phi_i/|S_0| = \mu_1 - \mu_0 \quad (17)$$

From Eq. 12 and Eq. 16 we have

$$\ln L(\hat{a}, \hat{b}, \hat{v}) = -n\ln\sqrt{2\pi} - \frac{n}{2}\ln\hat{v} - \frac{n}{2} \quad (18)$$

Lemma 2 *Maximization of the LOD score is equivalent to minimization of $\hat{v}$.*
Proof. $LOD = \alpha(\ln L(\hat{a}, \hat{b}, \hat{v}) - \ln L(\hat{\mu}, 0, \hat{v}_0))$ where $\alpha > 0$ is a constant number. Since $\hat{v} \geq 0$, $\frac{dLOD}{d\hat{v}} = -\frac{\alpha n}{2\hat{v}} \leq 0$. Therefore, maximization of the LOD score is equivalent to minimization of $\hat{v}$ ∎

Lemma 3 *Minimization of $\hat{v}$ is equivalent to maximization of the ICV.*
Proof. Let S_0 and S_1 be the sets of individuals whose g_i is 0 and 1, respectively, and let μ_0 and μ_1 denote the averages of Φ_i in S_0 and S_1. Let S be $S_0 \cup S_1$. We can rewrite $\hat{v} = \frac{1}{n}\left(\sum_{i\in S_0}(\Phi_i - \mu_0)^2 + \sum_{i\in S_1}(\Phi_i - \mu_1)^2\right) = \frac{1}{n}(\sum_{i\in S}\Phi_i^2 - (|S_0|\mu_0^2 + |S_1|\mu_1^2))$. Since $ICV = -n\mu^2 + (|S_0|\mu_0^2 + |S_1|\mu_1^2)$, minimization of $\hat{v}$ is equivalent to maximization of ICV. Thus, maximization of the LOD score is equivalent to that of the ICV∎

Theorem 1 *Maximization of the LOD score is equivalent to that of the ICV.*
Proof. From Lemma 2 and Lemma 3 ∎

Thus, calculation of the LOD score at a marker locus is equivalent to that of the ICV of the data division in terms of the genotype at the marker.

Handling Multiple Markers. Based on the properties above, we can define the LOD score that can evaluate the effects of multiple markers. Calculation of the ICV of the data division by a rule $G_i = (m_{j_1,i} = v_1)$ and $\cdots$ and $(m_{j_k,i} = v_k)$ is equivalent to regression of the data on the model as follows:

$$\Phi_i = a + bG_i + \varepsilon. \quad (19)$$

Except for G_i, definitions of the variables are the same as those in Eq. 3. In this regression model, coefficiency b is the phenotypic effect of the co-existence of particular genotypes $(v_1, \cdots, v_k)$ at multiple marker loci $(m_{j_1,i}, \cdots, m_{j_k,i})$. Here, we can use the same definitions of the likelihood function and the LOD score as those in Eq. 5 and Eq. 6, respectively, since the definitions do not depend on the definition of g_i in Eq. 5. The significant combination of marker loci associated with the quantitative trait value is found by selecting a set of marker loci which maximizes the ICV of the data division according to their genotype information. This provides a multi-dimensional expansion of the LOD score.

4 Search Program

Graph Search for Conjunctions. Consider all the conjunctions of the form $(m_{j_1,i} = v_1) \times \cdots \times (m_{j_k,i} = v_k)$, where $v_n = 0$ or 1. We first remark that it is NP-hard to compute the optimal conjunction that maximizes the ICV.[7] One common approach to such optimization problems is an iterative improvement graph search algorithm that initially selects a candidate conjunction by using a greedy algorithm and then tries to improve the ensemble of candidate conjunctions by local search heuristics. To avoid the repetition of visiting the same node, conventional graph search algorithms maintain the list of visited nodes,[5] which however could be a severe bottleneck of parallel execution. We instead proposed a rule of rewriting a conjunction to others.[8] We first apply the rewriting rule to the initial conjunction to obtain child conjunctions, and then we repeat application of the rule to descendant conjunctions so that we can visit every conjunction just once without maintaining the list of visited conjunctions. Moreover, each application of the rewriting rule can be well parallelized.

For a dataset with a Boolean target attribute (e.g., indicating whether diseased or not), we have developed a branch-and-bound heuristics appropriate for the significance of correlation between a conjunction and the target attribute (expressed by χ^2 value).[8] For a dataset with a numeric target attribute such as the glucose level, a similar heuristics based on the convexity of the ICV is also available to prune the search space.

Implementation. We wrote a search program in the C++ language and parallelized it with the POSIX thread library on two commercially available parallel computers: the Sun Microsystems Enterprise 10000 (64 UltraSPARCII [250MHz] processors) and the SGI Origin2000 (128 R10000 [195MHz] processors). For a dataset of an intercross population, it can use primitive tests of the form $(m_{j,i} = \mathrm{O/O})$, $(m_{j,i} = \mathrm{F/F})$, $(m_{j,i} = \mathrm{O/F})$, and $(m_{j,i} = \mathrm{O/O}$ or $\mathrm{O/F})$ to reveal the dominant and the recessive effects of the marker loci.

In this work, we mainly focused on the effects of the combinations of two markers. Therefore, we first executed the program under the restriction $k = 2$, and then compared the results with those under the condition $k = 1$, 3, and 4. We calculated the ICV of the data division by all the combinations of k markers without the branch-and-bound heuristics. During parallel execution, we used calculation of the ICV of the data division by each marker combination as the unit of computing, since they can be carried out simultaneously. To distribute the computing among multiple P processors we statically divided the set of unit of computing into disjoint P subsets evenly and assigned them to the processors. Each processor iterates the calculation of the ICV indicated in the assigned subset. When all the processors complete the calculation the program terminates. The required computation time for calculation of the ICVs for all the combinations of two markers using the dataset with 157 individuals and 279 markers is 16 seconds (Origin2000) and 37 seconds (Enterprise 10000). These results correspond to an 85-fold and a 50-fold calculation speedup, respectively. Calculation speedup scaled almost linearly with respect to the number of the processors used.

5 Results

To find significant conjunctive rules with respect to oral glucose tolerance (measured as the postprandial blood glucose level at 60 minutes after oral administration) we calculated ICVs of the data division by all the combinations of the k markers out of 279 markers (k = 1, 2, 3, and 4). In this section we focus on the rules which use the primitive tests of the form $(m_{j,i} = O/O)$. For simplicity, we denote a rule $G_i = (m_{j_1,i} = O/O)$ and $\cdots$ and $(m_{j_k,i} = O/O)$ as $m_{j_1} \times \cdots \times m_{j_k}$.

Single Marker. Fig. 5 shows the ICV of the data division in terms of a single marker rule at each marker locus along chromosomes 1, 5, 7, 14, 17, and X. For example, we can observe ICV peaks in the region around markers *D1Rat90*, *D7Wox6*, and *D14Mit5*. These three marker loci on chromosomes 1, 7, and 14 correspond respectively to the significant loci designated *Dmo1*, *Dmo2*, and *Dmo3* which have been found by traditional LOD score analysis.[4] We also have a peak on chromosome 14 near marker *Cckar* 34.9cM apart from *D14Mit5*. This marker also has been known to be related to the *Cc-kar* gene.[4,9] The weak peak around *D1Mit12* on chromosome 1 and *DxMgh2* on chromosome X are also known to be linked to the oral glucose tolerance (OGT).[4,9]

Two Markers. To assess the effects of the co-existence of O/O genotypes at pairs of marker loci, we calculated the ICV of the data division by all the possible conjunctive rules which consist of two markers ($m_{j_1} \times m_{j_2}$, $1 \leq j_1 < j_2 \leq 279$). According to the calculated ICV we sorted the rules and picked up the pairs of chromosomes on which exist marker pairs with high ICVs. We

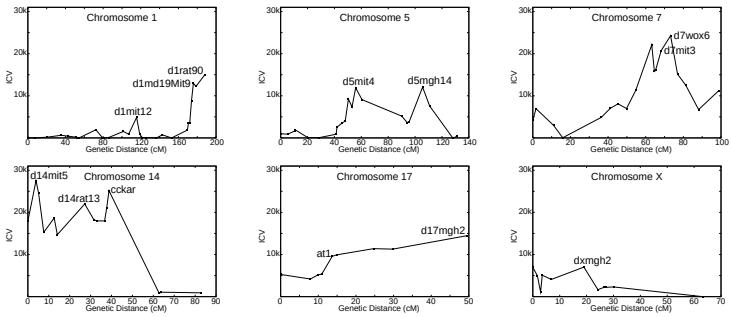

Figure 5: ICV of the data division in terms of the genotype at each marker locus along chromosomes (1, 5, 7, 14, 17, and X). Small rectangular marks correspond to markers. The horizontal axis indicates the genetic distance (cM) between markers.

Table 1: The pairs of markers significant to oral glucose tolerance (OGT).

Chr×Chr	$m_{j_1} \times m_{j_2}$	ICV	$\lvert S_1\rvert$	μ_1	$\lvert S_0\rvert$	μ_0	F
1×14	*D1Rat90*×*D14Rat13*	64762	13	251.5	144	177.8	58.5
1×14	*D1Rat166*×*Cckar*	58426	12	251.0	145	178.4	50.9
7×14	*D7Wox6*×*D14Mit5*	62715	13	250.4	144	177.9	56.0
17×14	*At1*×*D14Mit5*	62428	8	270	149	179.3	55.7
1×5	*D1Md19Mit9*×*D5Mgh14*	57047	13	247.4	144	178.2	49.3
7×X	*D7Wox6*×*DxMgh2*	59115	11	254.6	146	178.6	51.7
1×17	*D1Mit12*×*D17Mgh2*	62670	7	276.4	150	179.6	56.0

have significant pairs of markers on the pairs of chromosomes: 1×14, 7×14, 17×14, 1×17, 7×X, and 1×5. Table 1 lists the pairs of markers with a high ICV on those chromosome pairs.

At first glance, all the markers in the rules have an ICV peak even with a single marker alone (cf. Fig. 5), and each marker works in an additive manner. However, we can observe selectivity among the marker pairs. For example, the first four marker pairs in Table 1 use markers on chromosomes 1, 7, and 17 as the counterparts of those on chromosome 14. The three markers on chromosome 14 used in the pairs *D14Mit5*, *D14Rat13*, and *Cckar* exist in this order on chromosome 14, and their relative distances from *D14Mit5* are respectively 0, 23.3, and 34.9 cM. Other marker pairs on the chromosome pairs do not make the ICV high. For example, the ICV of *D1Rat90* × *D14Mit5* is 34831. This shows that a pair of markers does not make the ICV high even if each of the two markers does alone.

Plotting the ICV on a two-dimensional plane spanned by two chromosomes makes this clearer. Fig. 6 shows all the combinations of markers on the pairs of chromosomes (1×14, 17×14, 1×17, 7×14, 1×5, and 7×X). A spot corresponds

to a pair of markers and its color indicates the ICV (red indicates a high ICV). Bar charts attached to the vertical and the horizontal axes show the the ICV of the data division by a single marker on each axis (they correspond to Fig. 5). Observe that the markers in the region *D14Rat13–Cckar* on chromosome 14 make the ICV high when they are paired with the those around *D1Rat90* on chromosome 1 (Fig. 6A). However, *D14Mit5* does not exert this influence under the same condition. On the other hand, *D14Mit5* makes the ICV high when it is paired with the markers around *At1* on chromosome 17 but those in the region *D14Rat13–Cckar* do not (Fig. 6B). When the markers on chromosome 14 are paired with those on chromosome 7, pairs of markers work almost in an additive manner and we can observe two peaks in the region corresponding to the *D14Rat13–Cckar* and around *D14Mit5* (Fig. 6D). A similar situation can be found on chromosome 1. When the markers on chromosome 1 are paired with those on chromosome 17, the red region around *D1Rat90* in Fig. 6A disappears and another peak appears around *D1Mit12* × *D17Mgh2* instead (Fig. 6C).

More Than Two Markers. When the rules consist of three markers, we found high ICV values in the three-dimensional space, for example, around (*D7Mit16* × *D14Mit5* × *DxMgh2*), (*D7Wox6* × *D14Mit5* × *At1*), and (*D1Md19-Mit9* × *D5Mgh14* × *Cckar*). The ICV of these rules are 86821, 83712, and 80651, respectively. For the rules consisting of four markers, we observed ICV peaks around (*D7Mit16* × *D14Mit5* × *DxMgh2* × *D7Wox6*) and (*D7Wox6* × *D14Mit5* × *At1* × *D17Mgh2*). Note that two markers on the same chromosome appear in the rules. In the former rule, *D7Wox6* and *D7Mit16* are 3.7 cM distant from each other on chromosome 7, and in the latter, *At1* and *D17Mgh2* are 35.8 cM distant on chromosome 17. In these rules, the fourth markers are not effective compared to the above rules which use three markers (the ICV and the number of rats which make the rule true have not changed by the additional markers).

6 Conclusion

We have focused on the relation between the multiple marker loci and the quantitative trait. To investigate the effects of multiple marker loci, we divided the set of the individuals into two subsets according to a judgement whether or not each individual has particular genotypes at multiple marker loci. We formalized the judgement regarding genotypes as a conjunctive rule and estimated the significance of the rule in terms of inter-class variance. The proposed method can determine the significant combinations of marker loci by finding the rule accompanied by a high ICV. We also showed that finding the significant marker loci based on inter-class variance is equivalent to that based on the traditional LOD score.

The application of the above method on the OLETF model rat of non-

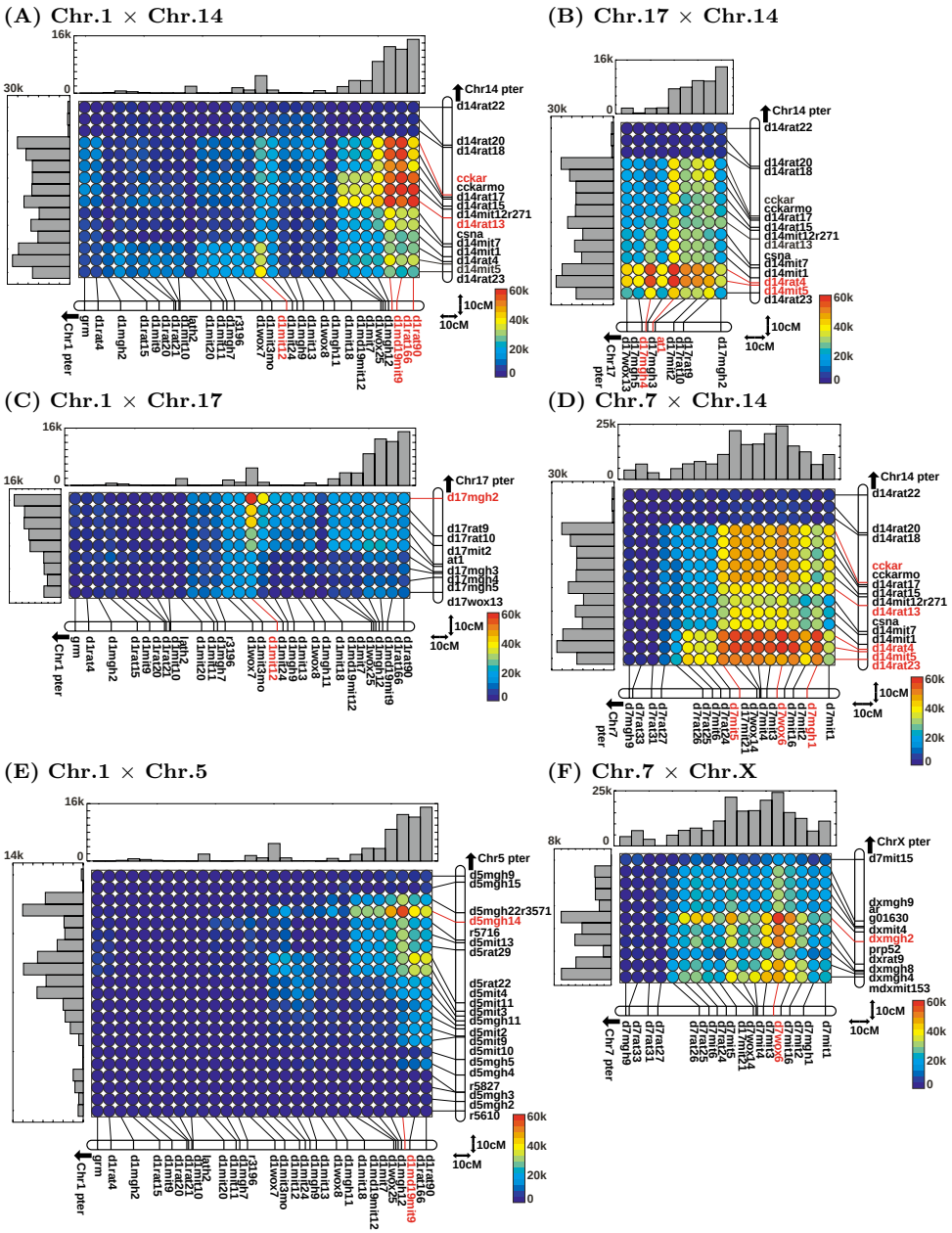

Figure 6: The ICV of the data division by two markers. Attached bar charts show the ICV of the data division by a single marker on each axis.

insulin dependent diabetes mellitus (NIDDM) has found the combinations of marker loci significant to oral glucose tolerance (OGT). Plotting the ICV on a two-dimensional plane spanned by two chromosomes presents clearly the relation among the marker loci. We can observe selectivity in the effects of the marker combinations. This property of the marker loci cannot be discovered solely by analysis of one-to-one relationships between a marker locus and the quantitative trait, as seen in the calculation of the LOD score along chromosomes. Thus, we have proposed a new method of QTLs analysis and showed its usefulness using experimental results in conjunction with real data.

Acknowledgements
This research is partly supported by Grant-in-Aid for Scientific Research on Priority Areas "Genome Science" from the Ministry of Education, Science and Culture, Japan.

References

1. J. Galli *et al.*, *Nature Genet.* **12**:31-37 (1996).
2. D. Gauguiter *et al.*, *Nature Genet.* **12**:38-43 (1996).
3. T. Hirashima *et al.*, *Biochem.Biophys.Res.Commun.* **224**:420-425 (1996).
4. N. Kanemoto *et al.*, *Mamm. Genome* **9**:419-425 (1998).
5. V. Kumar, A. Grama, and G. Karypis, Introduction to Parallel Computing: Design and Analysis of algorithms, Benjamin Cummings, (1993).
6. E. S. Lander and D. Botstein, *Genetics* **121**:185-199 (1989).
7. S. Morishita, On Classificaton and Regression, In *Proc. of Discovery Science, DS'98, Lecture Notes in Artificial Inteligence* **1532**:40-57 (1998).
8. S. Morishita and A. Nakaya, In *Proc. of Workshop on Large-Scale Parallel KDD Systems in conj. with the 5th ACM SIGKDD* :25-34 (1999).
9. S. Takiguchi *et al.*, *Gene* **197**:169-175 (1997).
10. T. K. Watanabe *et al.*, *Nature Genet.* **22**:27-36 (1999).
11. Z.-B. Zeng, *Genetics* **136**:1457-1468 (1994).