Shigeo Kamitsuji · Naoyuki Kamatani

Estimation of haplotype associated with several quantitative phenotypes based on maximization of area under a receiver operating characteristic (ROC) curve

Received: 24 October 2005 / Accepted: 15 December 2005 / Published online: 15 February 2006
© The Japan Society of Human Genetics and Springer-Verlag 2006

Abstract An algorithm for estimating haplotypes associated with several quantitative phenotypes is proposed. The concept of a receiver operating characteristic (ROC) curve was introduced, and a linear combination of the quantitative phenotypic values was considered. This set of values was divided into two parts: values for subjects with and without a particular haplotype. The goodness of its partition was evaluated by the area under the ROC curve (AUC). The AUC value varied from 0 to 1; this value was close to 1 when the partition had high accuracy. Therefore, the strength of association between phenotypes and haplotypes was considered to be proportional to the AUC value. In our algorithm, the parameters representing a degree of association between the haplotypes and phenotypes were estimated so as to maximize the AUC value; further, the haplotype with the maximum AUC value was considered to be the best haplotype associated with the phenotypes. This algorithm was implemented by using R language. The effectiveness of our algorithm was evaluated by applying it to real genotype data of the Calpine-10 gene obtained from diabetics. The results showed that our algorithm was more reasonable and advantageous for use with several quantitative phenotypes than the generalized linear model or the neural network model.

Keywords Area under the ROC curve (AUC) · Generalized linear model · Haplotype estimation ·

Neural network · Quantitative trait locus · Receiver operating characteristic (ROC) curve

Introduction

Quantitative trait locus (QTL) analysis is a method used to find DNA markers associated with a quantitative phenotype. QTL analysis has a long history in statistical genetics, and work in many theoretical and practical studies has involved QTL analysis.

Shibata et al. (2004) recently developed the algorithm *QTLhaplo* for estimating haplotypes associated with an observed quantitative phenotypic value based on the maximum likelihood method. However, in many cases, multiple quantitative phenotypic values are available for a subject. *QTLhaplo* requires that one quantitative phenotype be selected by clinical experience or that all phenotypes be selected in a sequential manner.

Machine learning models are also used for QTL analysis. For example, a Bayesian network model (Sebastiani et al. 2005) and a graphical model (Thomas et al. 2004) have been used to express the association between several genes and phenotypes. An advantage of using machine learning models for QTL analysis is that several phenotypes are available as variables. However, most machine learning models do not consider genetic laws, and it is assumed that the random component of genetic laws can be approximated by the models. Moreover, since the structure of such a machine learning model is very complicated, the characteristics of the data from the results obtained are usually difficult to understand. On the other hand, statistical models have been also used. Yi et al. (2003) applied a linear model to explain the relationship between quantitative phenotypes and markers. In order to consider genetic laws, many parameters are introduced into the model and are estimated by using the Markov Chain Monte Carlo method. However, it is often difficult to understand the characteristics of data obtained from models with many complex parameters. Moreover, the disadvantage of models of this type is that

S. Kamitsuji (✉)
Genome Diversity Team, Integrated Database Group,
Japan Biological Information Research Center (JBIRC),
Japan Biological Informatics Consortium (JBIC),
AIST Bio-IT Research Building, 2-42 Aomi,
Koto-ku, Tokyo 135-0064, Japan
E-mail: skamitsu@jbirc.aist.go.jp
Tel.: +81-3-35998311
Fax: +81-3-35998144

N. Kamatani
Institute of Rheumatology, Tokyo Women's Medical University,
Kawada-cho 10-22, Shinjuku-ku, Tokyo 162-0054, Japan

their results depend on the method used for coding the genotype data, which are considered as categorical data, because the values of the estimated coefficients vary according to the selected coding method (Chambers and Hastie 1992).

In addition, in QTL analysis it is noteworthy that the quantitative phenotypic value is the result of genetic factors, environmental factors, and the interaction of environmental factors with genetic factors (Fisher 1918). Usually, a quantitative phenotypic value is based on the fact that the sum of these factors contributes to a variation in a population. In order to identify such a locus or haplotype, it is important to extract only the genetic variation from the observed quantitative phenotypic value.

In this article, we propose a new method for estimating a haplotype associated with several quantitative phenotypic values from a genotype and several quantitative phenotypic data. The aim of our method is to find the haplotype that contributes to this mixture of phenotypes, which is calculated as a linear combination of the phenotypes. As a measure for evaluating the strength of the association, we introduced an area under the receiver operating characteristic (ROC) curve (Ransohoff and Feinstein 1978; Toetteson and Begg 1988). The ROC curve is a plot to show the tradeoff between sensitivity and specificity, and work in many theoretical aspects of this curve has been carried out (Pepe 1997, 1998, 2000). The accuracy of a diagnostic test can be evaluated by the area under the ROC curve (AUC) (McClish 1989). The AUC value varies from 0 to 1, being close to 1 when the diagnostic test has a high degree of accuracy. In our algorithm, the set of the mixture of phenotypes is divided into two parts: sets for subjects with and without a haplotype. Then, our algorithm finds the haplotype so as to maximize the AUC value between the two sets of the mixture of phenotypes. Since our algorithm is a very simple model, the results can be easily understood; further, several quantitative phenotypes are available in our model.

In this report, we explain the ROC curve and AUC, and propose their application to QTL analysis. The algorithm for QTL analysis based on the maximization of AUC is also explained. The results of numerical examples serve to illustrate the properties and the validity of our method. The results of the application of the method to real genotype and phenotype data for diabetics are also shown. The advantages of our algorithm are also compared to those of the generalized linear model (GLM). As another comparison with a machine learning model, results obtained from the neural network model are given.

Materials and methods

We aim to extract the genotypic value associated with a haplotype from several quantitative phenotypic values. Let $\mathbf{x}_j$, $j = 1, \dots, p$, be a quantitative phenotypic value as an n -dimensional vector with n observations.

We assume that the genotypic value $\mathbf{z}$ can be written as a linear combination of quantitative phenotypic values,

$$\mathbf{z} = \sum_{j=1}^p \beta_j \mathbf{x}_j, \quad (1)$$

where β_j , $j = 1, \dots, p$, is the coefficient of $\mathbf{x}_j$.

A problem in this case concerns the estimation of the coefficient vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$. Since this study aims to efficiently estimate the haplotype associated with quantitative phenotypes, the coefficient $\boldsymbol{\beta}$ should be estimated so as to divide the individuals clearly into two parts: individuals with a haplotype and individuals without a haplotype. A problem of this type is resolved by the introduction of the concept of AUC.

Maximization of the area under the ROC curve

The ROC curve is a plot to show the tradeoff between sensitivity and specificity, and is used to evaluate a diagnostic test. The accuracy of a diagnostic test can be evaluated by the AUC. The AUC value varies from 0 to 1, being close to 1 when the diagnostic test has a high degree of accuracy.

It is well known that the AUC can be represented by the standard normal distribution function when two test results, with and without a condition, are normally distributed (McClish 1989). In this section, we will demonstrate the estimation of the haplotype associated with quantitative phenotypic values by introducing the concept of the AUC, and an algorithm for extracting genotypic values from several quantitative phenotypic values will be proposed.

In order to apply the AUC, we introduced the dominant model for a haplotype. In other words, we divided the individuals in the data into two parts: individuals with a haplotype (G_1) and those without a haplotype (G_2). Now we want to estimate coefficient $\boldsymbol{\beta}$ so as to maximize the AUC value between two genotypic values $\mathbf{z}|G_1$ and $\mathbf{z}|G_2$ when the genotypic value $\mathbf{z}$ is divided into two parts, the sets $\mathbf{z}|G_1$ and $\mathbf{z}|G_2$ correspond to individuals with a particular haplotype and without the haplotype, respectively. However, it is complicated to describe the function of AUC in terms of $\boldsymbol{\beta}$ without assuming quantitative phenotypic values $\mathbf{x}_j$, $j = 1, \dots, p$. For simplicity, in this study, we assumed that each $\mathbf{x}_j$, $j = 1, \dots, p$, is independently identically distributed (i.i.d.) and that $\mathbf{x}_j$, given the condition G_k , $k = 1, 2$, is normally distributed,

$$\mathbf{x}_j|G_k \sim N(\mu_{j,k}, \sigma_{j,k}^2) \quad \mu_{j,k} = E(\mathbf{x}_j|G_k), \quad \sigma_{j,k}^2 = \text{Var}(\mathbf{x}_j|G_k). \quad (2)$$

From Eq. 2, given the above condition, the variable $\mathbf{z}$ is then normally distributed,

$$\mathbf{z}|G_k \sim N\left(\sum_{j=1}^p \beta_j \mu_{j,k}, \sum_{j=1}^p \beta_j^2 \sigma_{j,k}^2\right) =: N(\mu_{\mathbf{z},k}(\boldsymbol{\beta}), \sigma_{\mathbf{z},k}^2(\boldsymbol{\beta})). \quad (3)$$

For simplicity of notation of the AUC, we express functions $a(\beta)$ and $b(\beta)$ as

$$a(\beta) = \frac{\mu_{z,2}(\beta) - \mu_{z,1}(\beta)}{\sigma_{z,2}(\beta)} \quad \text{and} \quad b(\beta) = \frac{\sigma_{z,1}(\beta)}{\sigma_{z,2}(\beta)}, \quad (4)$$

respectively. Note that each function takes a scalar value. Under these assumptions, AUC $F(\beta)$ as a function of β can be written as

$$F(\beta) := \Phi\left(\frac{a(\beta)}{\sqrt{1 + b^2(\beta)}}\right). \quad (5)$$

Therefore, the coefficient β is estimated so as to maximize $F(\beta)$, and the magnitude of association between the haplotype and phenotypes will be evaluated by the value of AUC $F(\hat{\beta})$, where $\hat{\beta}$ is the estimator of β . A detailed description of the estimation of the haplotype is given below.

Updating the algorithm by the Quasi-Newton method

The algorithm for estimating coefficient β so as to maximize the value of AUC $F(\beta)$ is shown in Eq. 5. For updating the parameter β , the Quasi-Newton method is used,

$$\beta^{[t+1]} = \beta^{[t]} - \left(f(\beta^{[t]})'\right)^{-1} f(\beta^{[t]}), \quad (6)$$

where $\beta^{[t]}$ is the updated parameter vector at iteration time t . Here, the p -dimensional vector $f(\beta)$ is derived from AUC $F(\beta)$ in terms of β as follows:

$$\begin{aligned} \underline{f}(\beta) &:= \frac{d}{d\beta} F(\beta) \\ &= \phi\left(\frac{a(\beta)}{\sqrt{1 + b^2(\beta)}}\right) \frac{a'(\beta)(1 + b^2(\beta)) - \frac{1}{2}a(\beta)(b^2(\beta))'}{(1 + b^2(\beta))^{3/2}}. \end{aligned}$$

The function $f(\beta)'$ is also obtained by differentiating $\underline{f}(\beta)$ in terms of β as follows:

$$\begin{aligned} \underline{f}(\beta)' &= \text{diag}(\Delta)^{-1} \text{diag}(f(\beta + \Delta) - f(\beta)), \\ \Delta &= (\Delta, \dots, \Delta)^T, \end{aligned}$$

where Δ is a small constant. Here,

$$\begin{aligned} \underline{a}'(\beta) &= \left[-\sigma_{z,2}^2(\beta)(\mu_2 - \mu_1) - (\mu_{z,2}(\beta) - \mu_{z,1}(\beta))(\Sigma_{x,2}\beta - \underline{\sigma}_{y,2}) \right] / \sigma_{z,2}^3(\beta), \\ \underline{b^2}(\beta)' &= 2(\Sigma_{x,1}\beta - \underline{\sigma}_{y,1})\sigma_{z,2}^2(\beta) - 2(\Sigma_{x,2}\beta - \underline{\sigma}_{y,2})\sigma_{z,1}^2(\beta) / \sigma_{z,2}^4(\beta), \end{aligned}$$

where

$$\begin{aligned} \sigma_{z,k}^2(\beta) &= \sigma_{y,k}^2 + \beta^T \Sigma_{x,k} \beta - 2\beta^T \underline{\sigma}_{y,k}, \\ \underline{\sigma}_{y,k} &= (\sigma_{1y,k}, \dots, \sigma_{py,k}), \\ \Sigma_{x,k} &= \text{diag}(\sigma_{1,k}^2, \dots, \sigma_{p,k}^2), \\ \underline{\mu}_k &= (\mu_{1,k}, \dots, \mu_{p,k})^T. \end{aligned}$$

It is necessary to iterate the updating algorithm (Eq. 6) when the convergence of β is observed. Hereafter, we describe this method as maximization of area under the ROC curve (*MARC*).

Test of AUC

In order to test whether the discrimination between patients with and without a haplotype is significant, the following hypothesis testing is considered:

For the null hypothesis H_0 , $\text{AUC} = 0.5$,
and for the alternative hypothesis H_1 , $\text{AUC} \neq 0.5$.

Note that the value of AUC becomes 0.5 in the case of two identical distributions. Thus, the test statistic can be written as

$$\frac{F(\hat{\beta}) - 0.5}{\sqrt{\text{Var}[F(\hat{\beta})]}}$$

(Zhou et al. 2002). Here,

$$\begin{aligned} \text{Var}[F(\beta)] &= g_1(\beta)^2 \text{Var}[a(\beta)] + g_2(\beta)^2 \text{Var}[b(\beta)] \\ &\quad + 2g_1(\beta)g_2(\beta) \text{Cov}[a(\beta), b(\beta)], \\ g_1(\beta) &= \frac{\exp\left(-a(\beta)^2/2(1 + b(\beta)^2)\right)}{\sqrt{2\pi(1 + b(\beta)^2)}} \\ g_2(\beta) &= \frac{a(\beta)b(\beta) \exp\left(-a(\beta)^2/2(1 + b(\beta)^2)\right)}{\sqrt{2\pi(1 + b(\beta)^2)^3}} \\ \text{Var}[a(\beta)] &= \frac{n_2(a(\beta)^2 + 2) + 2n_1b(\beta)^2}{2n_1n_2}, \\ \text{Var}[b(\beta)] &= \frac{(n_1 + n_2)b(\beta)^2}{2n_1n_2}, \\ \text{Cov}[a(\beta), b(\beta)] &= \frac{a(\beta)b(\beta)}{2n_1}. \end{aligned}$$

It is known that this statistic has a standard normal distribution.

Algorithm for haplotype estimation

In this section, we propose the manner in which the haplotype associated with quantitative phenotypes is estimated by using the *MARC* method. Here, we assume that the genotypes and p quantitative phenotypes $x_1, \dots, x_p$ are observed as the data.

Let H denote a set of all possible haplotypes from the genotypes at the locus; and an element of H is denoted as h_i .

Haplotype estimation (HE)-step

HE-step 1 The individuals in the data are divided into two parts: individuals with a haplotype h_i (G_1) and without a haplotype h_i (G_2).

HE-step 2 By using the *MARC* method, the coefficient β is estimated so as to maximize the value of $AUC F(\beta)$ between $z|G_1$ and $z|G_2$, as above.

For all $h_i \in H$, coefficient $\hat{\beta}$ and the value of $AUC F(\hat{\beta})$ are calculated. Further, a haplotype with the maximum AUC value is the best haplotype associated with phenotypes $x_1, \dots, x_p$.

Model selection (MS)-step

MS-step 1 Consider a model that omits the single term x_j . For example, a model omitting a single term from the model in Eq. 1 can be written as

$$z = \sum_{i \neq j} \beta_i x_i, \quad i = 1, \dots, p.$$

MS-step 2 Calculate $\hat{\beta}$ and $F(\hat{\beta})$ by HE-steps 1 and 2 for each such model.

If the value of AUC from these models is larger than that from the previous model, it is selected as the best model, following which MS-steps 1 and 2 are followed until the AUC value decreases. Hereafter, this operation with these two steps based on *MARC* may be described as *QTLmarc*.

Software

Algorithms *MARC* and *QTLmarc* are implemented by R language. For the usage of our algorithm, please contact the corresponding author.

Results

Numerical examples

In this section, we will describe the results of two numerical applications.

Plane for partitioning

First, we will explain the meaning of coefficient β in our algorithm. We generated 1,000 samples, which are divided into two sets of 500 each: groups G_1 and G_2 . Variables X_1 and X_2 for each group are normal random numbers, given as follows:

$$\begin{aligned} X_1|G_1 &\sim N(2, 1^2), & X_1|G_2 &\sim N(-2, 1^2), \\ X_2|G_1 &\sim N(3, \sqrt{2}^2), & X_2|G_2 &\sim N(-3, \sqrt{2}^2). \end{aligned}$$

The variable z is defined as

$$z = \beta_0 \mathbf{1} + \beta_1 X_1 + \beta_2 X_2, \quad \mathbf{1} = (1, \dots, 1)^T.$$

We then estimate the coefficients

$$\beta = (\beta_0, \beta_1, \beta_2)^T$$

so as to maximize the AUC for $z|G_1$ and $z|G_2$ by *MARC*. The results of the estimation of β are as follows:

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} -0.5000 \\ -1.3957 \\ 6.9527 \end{pmatrix}. \quad (7)$$

In order to explain the meaning of β , the scatter plot of the samples is illustrated with axis X_1 and X_2 , as shown in Fig. 1. In Fig. 1, $\circ$ and $*$ represent points belonging to groups G_1 and G_2 , respectively. In addition, from the results of β in Eq. 7, we plot the following line

$$-0.5000 - 1.3957X_1 + 6.9527X_2 = 0 \quad (8)$$

in Fig. 1. From this figure, it appears that the line represented by Eq. 8 divides the two groups effectively. Therefore, this result indicates that the coefficient vector β represents the normal vector of the plane for partitioning. However, since the coefficient β represents only the direction of the plane, the values of β are not unique solutions, and change according to the initial value of β in our algorithm. Further, it can be noted that the ratio of the appropriate value for β is unique.

Simulation of haplotype estimation

Next, we will demonstrate the estimation of the haplotype associated with several quantitative phenotypic values using our algorithm *QTLmarc*. Here, we will use simulation data generated by the R language. The simulation data has genotypes at three bi-allelic loci (alleles A or a) and two quantitative phenotypic values x_1 and x_2 for every 1,000 samples. If haplotype “aaa” is observed in a subject, phenotype x_1 for the subject is normally distributed with mean 1 and variance 1; conversely, x_1 corresponding to an individual without haplotype “aaa” is normally distributed with mean 0 and variance 1. Similarly, if haplotype “aaA” is observed in a subject, phenotype x_2 for the subject is normally distributed with mean 1 and variance 2; conversely, x_2 corresponding to an individual without haplotype “aaA” is normally distributed with a mean 0 and variance 1. In this experiment, since genotype data are available in our algorithm, it was reconstructed from the haplotype data.

Table 1 shows the results of the association between haplotypes and the two phenotypes x_1 and x_2 obtained by *QTLmarc*. The table also shows the calculation of the P value for evaluating the significance of the AUC value as described in the [Materials and methods](#) section. From Table 1, when the condition for partitioning is haplotype “aaA,” the value of AUC for haplotype “aaA” is optimum and the P value is small sufficiently ($P = 1.44 \times 10^{-15}$); followed by “aaa.” This result is natural in view of the manner of generation of samples. Here, it might be asked whether the signs and absolute values of

Fig. 1 Scatter of sample data and a line represented by the estimated coefficients based on maximization of area under the receiver operating characteristic (ROC) curve (*MARC*)

the regression coefficients of haplotype “aaA” and “aaa” are sufficiently different; this is because the values are simulated so that they are positive and their absolute values are similar to each other. Here, it should be noted that it is important to compare ratio β_2/β_1 of coefficients β_1 and β_2 for haplotype “aaA” with that for haplotype “aaa.” From Table 1, the ratios for haplotypes “aaA” and “aaa” become 9.4240 and 0.5328, respectively. These results indicate that the signs of these ratios are positive, i.e., the gradients of the lines for partitioning (in this case, the plane for partitioning becomes the line because of the two phenotypes) are identical. This result is natural because phenotypes x_1 and x_2 , corresponding to the data with haplotypes “aaA” and “aaa,” respectively, have larger values than those corresponding to the data without the haplotypes “aaA” and “aaa.” Also, the ratio 9.4240 for the haplotype “aaA” is larger than 1, implying that the haplotype “aaA” contributes more to phenotype x_2 than to phenotype x_1 . This result is natural since the phenotype x_2 is generated so as to be normally distributed with a mean of 1 for individuals with haplotype “aaA.” Similarly, the ratio 0.5382 for haplotype

“aaa” is smaller than 1; this result implies that haplotype “aaa” contributes more to phenotype x_1 than to phenotype x_2 . This result is also natural since phenotype x_1 is generated so as to be normally distributed with a mean of 1 for individuals with haplotype “aaa.”

Moreover, as another comparison, it can be seen that the ratio and AUC value for haplotype “aaA” are larger than those for haplotype “aaa”, despite phenotypes x_1 and x_2 being normally distributed with a mean of 1. Here, it is noted that the variance of phenotype x_2 for haplotype “aaA” is larger than that of phenotype x_1 for haplotype “aaa.” The estimation of the coefficient depends not only on the mean of the data but also the variance, since the AUC function $F(\beta)$ depends on μ and σ in Eqs. 4 and 5. Since the variance of phenotype x_2 for the haplotype “aaA” is larger than that of phenotype x_1 for the haplotype “aaa,” the AUC value for the former is thus larger than that for the latter.

The result of this experiment indicates that *QTLmarc* can evaluate the association between haplotype and phenotype, taking account of not only the mean of the phenotypes but also the variance of phenotypes.

Table 1 Results obtained from QTLmarc [determination of quantitative trait locus (QTL) based on maximization of area under the receiver operating characteristic (ROC) curve (*MARC*)] using simulation data. *AUC* Area under the ROC curve

Haplotype	Number of carriers	β_1	β_2	AUC	<i>P</i> value
aaA	231	1.1831	11.1489	0.6599	1.44E−15
aaa	130	−0.7231	−0.3892	0.6168	2.14E−06
AaA	400	−0.1082	4.1647	0.5765	1.40E−05
AAA	764	0.7528	2.0335	0.5553	3.08E−03
AAa	391	−0.4486	0.8180	0.5500	1.42E−03
aAA	443	−0.1292	0.5527	0.5451	1.09E−02
Aaa	225	−0.8354	−0.3161	0.5442	3.19E−02
aAa	228	3.3956	−1.5597	0.5392	4.61E−02

Analysis of real data

In order to evaluate the validity and effectiveness of our algorithm, we analyzed data from diabetics. This data is identical to that in Shibata et al. (2004). The data includes genotypes of the Calpine-10 gene (three bi-allelic loci with alleles 1 and 2) and quantitative phenotypes such as height, weight, age, body mass index (BMI), blood glucose levels (BS), hemoglobin A1c (HbA1c), immunoreactive insulin (IRI) level, and insulinogenic index (II). The phenotype BS includes four phenotypes—BS at 0 min (BS0), 30 min (BS30), 60 min (BS60), and 120 min (BS120). Similarly, the phenotype IRI also includes four phenotypes—IRI0, IRI30, IRI60, and IRI120. The exact data has been published by Iwasaki et al. (2005).

Normality of phenotypes

In our algorithm, it is assumed that a quantitative phenotype is normally distributed. Therefore it is necessary to evaluate whether the quantitative phenotype we used was normally distributed. If it is not normally distributed, the evaluation of the results obtained from our algorithm may be complicated. Figures 2, 3, 4, and 5

show the quantile–quantile (QQ) plots of 14 quantitative phenotypes that appear in the data for diabetics. If the phenotype is normally distributed, the QQ plot is a straight line. From these figures, it is natural to assume that eight phenotypes—weight, BMI, 4 BSs, IRI0, and HbA1c—are normally distributed. Therefore, henceforth, we will use the above phenotypes for analysis.

Comparison with *QTLhaplo*

To evaluate the validity of our algorithm, we compared the results obtained from *QTLmarc* with those from *QTLhaplo* (Shibata et al. 2004). Based on the result obtained from *QTLhaplo*, Shibata et al. suggested that the association of haplotype 112 with BS30 and BS60 was significant. To evaluate this result, we explored the haplotype associated with the two phenotypes—BS30 and BS60—by using *QTLmarc*. Table 2 shows the results obtained from the *QTLmarc* for these phenotypes. In Table 2, we also calculated the *P* value to evaluate the significance of the AUC value.

From Table 2, haplotype 112 gives the optimum AUC value significantly ($P = 0.0081$), and this result is equivalent to that obtained from *QTLhaplo*. Shibata et al. (2004) also stated that the means of BS30 and

Fig. 2 Quantile–quantile plot for four phenotypes—age, height (*Ht*), weight (*Wt*), and body mass index (*BMI*)

Fig. 3 Quantile–quantile plot for four blood glucose (BS) levels

BS60 for patients with haplotype 112 were larger than those without haplotype 112. In our study, this result is expressed by the sign of the coefficients for BS30 and BS60. In Table 2, the coefficients for BS30 and BS60 are positive. This implies that *QTLmarc* gives equivalent results to *QTLhaplo* since, in our study, the means of BS30 or BS60 for patients with haplotype 112 are larger than those without haplotype 112.

Best model by *QTLmarc*

We also analyzed the manner in which combinations of phenotypes give rise to the optimum AUC value and the optimum haplotype associated with their phenotype. In this study, we used eight phenotypes—weight, BMI, 4 BSs, IRI0, and HbA1c—as quantitative phenotypes. Note that, for simplicity of understanding of coefficients, all phenotypes were standardized; in other words, the phenotypes were changed to variables with a mean = 0 and variance = 1. Unfortunately, it serves no purpose to use the variable “Age” as a quantitative phenotype in our algorithm since Age is not normally distributed, as shown in Fig. 2; further, Age may not be termed a phenotype. However, it is well known that age is a risk factor in diabetes (Iwasaki et al. 2005). In order to solve this problem, we divided the data into

two parts: patients aged 50 years and over, and those aged less than 50 years. Our aim in this study was to estimate the haplotype and the combination of phenotypes so as to maximize the AUC value for each data set—the entire data, data corresponding to age ≥ 50 years, and data corresponding to age < 50 years.

The results of the best models for all three single nucleotide polymorphisms (SNPs) tested are shown in Table 3. Table 3 includes three results corresponding to three data sets—the entire data, data corresponding to age ≥ 50 years and data corresponding to age < 50 years. In Table 3, the number of candidates with a haplotype is given in the column “Number of carriers”. In Table 3, we now concentrate on the result corresponding to patients aged ≥ 50 years. Haplotype 111 gives the minimum P value ($P = 6.10 \times 10^{-7}$), followed by haplotypes 112 and 222. The AUC value for haplotype 111 is 0.6954. It is well known that haplotype 111 has a high diabetes risk in older Japanese subjects (Iwasaki et al. 2005). On the other hand, the association of haplotype 112 with BS60 is stronger than any other phenotype since the coefficient of BS60 is the largest. Moreover, since all coefficients are positive, this result indicates that four phenotypes—BMI, BS0, BS60, and BS120—corresponding to patients with haplotype 112 tend to be more highly rep-

Fig. 4 Quantile–quantile plot for four immunoreactive insulin (IRI) levels

resented than those without haplotype 112. Note that these results are equivalent to those obtained from *QTLhaplo*. Therefore, *QTLmarc* can estimate haplotypes important for diabetes. Moreover, the result obtained from *QTLmarc* indicates that haplotype 111 is associated with four phenotypes: BMI, BS0, BS60, and BS120; further, based on the magnitude of coefficients, BS60 gives the best association. *QTLmarc* can be applied to several quantitative phenotypes, and it is possible to explore the magnitude of association between the haplotypes and phenotypes.

Comparison with the generalized linear model

We now compare the predictor based on *QTLmarc* to the GLM. In this study, diabetes data corresponding to patients aged ≥ 50 years is applied.

The GLM is a non-linear model for the mean and the GLM employed in this study can be written as

$$g(p_i) = \beta_0 + \sum_{j=1}^q \beta_j x_{ij}$$

in the case of a no interaction model. Here, x_{ij} , $j=1, \dots, q$, is the i -th element of the phenotypes x_j . In the GLM, the expectation $p_i = E[y_i]$ is the probability of the i -th patient with a haplotype, and y_i is the binary variable for

the patient with or without the haplotype. The function $g(p)$ is known as the link function. For the binary variable, it is natural to choose the logit function,

$$g(p) = \log(1 - p)/p,$$

as the link function. Note that the GLM is also known as the logistic model when the response variable y is binary.

Table 4 shows the result obtained from the GLM. In the GLMs, the response variable y is the binary variable for a patient with or without a haplotype, one of eight haplotypes. The explanatory variables were selected from eight phenotypes—weight, BMI, 4 BSs, IRI0, and HbA1c—based on the AIC minimization (Akaike 1973). In Table 4, the coefficients of the explanatory variables are given when all AIC values for the eight GLMs are simultaneously minimized. In Table 4, explanatory variables contributing significantly to the model are indicated. In other words, the P value for deviance, which represents the magnitude of contribution of the phenotype to y , is smaller than 0.05 (Chambers et al. 1992). The goodness of GLM is also evaluated by the AUC value in Table 4 (Qin 2000; Copas et al. 2002). From Table 4, in the GLM for haplotype 111, BS60 and BS120 are significantly associated with haplotype 111; this result is equivalent to the result obtained from *QTLmarc*, as seen in Table 3.

Fig. 5 Quantile-quantile plot for phenotypes hemoglobin A1c (*HbA1c*) and insulinogenic index (*II*)

Our aim in this study was to find a model that gives a meaningful result—a model that can express the discrimination with or without a haplotype. However, it is difficult to compare the AUC values from GLM with those of *QTLmarc* since the minimization of AIC and the maximization of AUC are not equivalent as criteria of parameter estimation and model selection. In fact, the haplotypes, given the minimum of AIC and the maximum of AUC, are different from those in Table 4. Thus, as an evaluation of discriminability by the GLM, the haplotypes can be used to evaluate the goodness-of-fit by using the predicted values obtained from the GLM. The GLM is a model used to predict the probability that the response variable takes the value 1 when the response variable is the binary 0 or 1. In this study, the GLM predicts the probability of the appearance of a patient with a haplotype by using phenotype data. In Table 4, the mean of the probability for the existence of a patient with or without the haplotype is given. Table 4

shows that the GLMs, except the GLM for haplotype 111, are not suitable for this data since the sum of the two probabilities deviates significantly from 1. In the GLM, it is assumed that the response variable has a binomial distribution. Therefore, it is necessary for the sum of two probabilities corresponding to a carrier or non-carriers of the haplotype to be equal to 1. These results indicate that the GLM may be not suitable for estimating the haplotype associated with several phenotypes for this data.

Discussion

The advantage of *QTLmarc* is that several phenotypes are available for estimating the associated haplotype; further, the strength of association between the phenotypes and the haplotype and the validity of the model can be evaluated by the AUC value. However, the possibility of the estimation of the haplotype associated with several phenotypes by a machine learning model is of interest. This possibility will be explored by the analysis of the data for the diabetics aged ≥ 50 years by a neural network model.

Neural network model

The neural network (NN) model used in this study is a three-layer feed-forward neural network comprising one input layer with four neurons, one output layer with one neuron, and one hidden layer with three neurons, as shown in Fig. 6. The neurons on the hidden and output layers are all deterministic, i.e., the output y_i is a non-linear function of the given input x_{ij} , $j=1, \dots, 4$, as

$$y_i = f\left(\sum_{k=1}^3 h_k f\left(\sum_{j=1}^4 w_{jk} x_{ij}\right)\right),$$

where w_{jk} and h_k , $j=1, \dots, 4$, $k=1, \dots, 3$, are the connection weights. The function $f(x)$ is known as an activation function and a typical example is the sigmoid function,

$$f(x) = 1/(1 + \exp(-x)).$$

Table 2 Association between haplotypes and two phenotypes—BS30 and BS60—as determined using *QTLmarc*

Haplotype	Number of carriers	Phenotype		AUC	P value
		BS30	BS60		
212	10	2.9513	4.9981	0.6161	0.0928
211	11	1.7647	2.4510	0.6038	0.0973
222	9	-6.3477	-4.4496	0.5990	0.1342
112	127	3.2970	1.8822	0.5784	0.0081
122	95	2.1306	1.4030	0.5599	0.0416
221	28	5.8188	1.2599	0.5599	0.1408
111	144	-4.5639	-4.2530	0.5331	0.1621
121	235	0.1344	-2.9106	0.5209	0.3133

Table 3 Best models obtained from *QTLmarc* when the single nucleotide polymorphism (SNP) loci SNP-43, Indel-19, and SNP-63 are used. *HbA1c* Hemoglobin A1c, *BMI* body mass index

Haplotype	Number of carriers	Phenotype			AUC	P value	
		BS0	BS30	HbA1c			
Entire data							
222	9	4.3876	2.2663	5.1740	0.7228	0.0034	
211	11	2.5881	1.4408	4.6395	0.6785	0.0127	
212	10	−1.9731	−1.0969	−2.2769	0.6444	0.0423	
221	28	0.6542	1.0809	2.1091	0.6235	0.0091	
122	95	−0.3814	0.5715	−0.2421	0.6073	0.0005	
112	127	−0.1198	21.1590	−1.1809	0.5813	0.0062	
111	144	1.5230	−0.7395	0.0690	0.5806	0.0066	
121	235	−0.3404	0.0252	1.2484	0.5354	0.2206	
Haplotype	Number of carriers	Phenotype				AUC	P value
		BMI	BS0	BS60	BS120		
Data from patients aged ≥50 years							
221	9	7.1367	9.4493	−26.2790	−8.4942	0.7116	0.0041
222	4	1.3692	3.4324	−3.2579	−4.0592	0.7087	0.0026
111	64	0.1532	−0.6446	1.0344	−0.5304	0.6954	0.0000
211	4	−0.2473	2.1815	−1.5759	0.3523	0.6687	0.0509
112	54	1.1523	0.8362	3.1134	0.1982	0.6489	0.0004
212	3	−0.8632	0.2621	−0.2221	−0.7076	0.6254	0.2396
122	42	−0.0572	0.0499	−8.9660	−0.1150	0.6224	0.0058
121	108	−18.2788	2.2698	0.9149	−10.2325	0.5988	0.0837
Haplotype	Number of carriers	Phenotypes			AUC	P value	
		BS30	BS120	HbA1c			
Data from patients aged < 50 years							
222	5	0.2863	0.5299	0.6173	0.7865	0.0206	
211	7	−4.5734	−12.2539	−12.7785	0.6631	0.0655	
212	7	−3.4828	−14.2790	−7.6141	0.6183	0.1430	
112	73	2.5544	2.2208	−1.8872	0.6000	0.0095	
122	53	0.7968	0.9244	−1.3017	0.5977	0.0136	
221	19	−7.4192	−12.3016	−14.7500	0.5929	0.0731	
111	80	1.3982	2.5157	−5.2779	0.5443	0.1613	
121	127	−1.4195	−0.2743	−0.9468	0.5392	0.2594	

Therefore, the output y_i varies from 0 to 1. In this experiment, y_i is the binary 0 or 1 for the i -th patient with or without haplotype 111. The result obtained from the NN model is shown in Fig. 6 and Fig. 7 shows the ROC curve and the AUC value for evaluating the goodness-of-

fit of the model. Figure 7 shows that the discrimination obtained from the NN model is better than that determined by *QTLmarc* in terms of the AUC value. (the AUC by NN is 0.8981 and AUC by *QTLmarc* is 0.6954). However, as shown in Fig. 6, the results obtained from

Table 4 Results of the best generalized linear models (GLMs) explaining the association between each haplotype and phenotypes for data from diabetic patients aged ≥ 50 years. Six phenotypes are selected by minimization of AIC

Haplotype	Number of carriers	Intercept	Phenotype						AIC	AUC	Predicted probability	
			BMI	BS0	BS30	BS60	BS120	HbA1c			Non-carrier	carrier
212	3	-3.7590	0.2894	-0.1502	-0.3963	0.0632	0.4418	-0.7601	40.2117	0.7507	0.0165	0.0341
222	4	-3.8055	-0.1932	-0.7961	-0.4085	0.4320	0.9490	-0.9102	41.4962	0.8031	0.0135	0.0940
211	4	-3.4838	0.2994	-0.7166	-0.2411	0.7109	-0.2706	-0.8674	44.2789	0.7271	0.0176	0.0612
221	9	-2.9488	-0.2738	-0.4436	-1.0408	1.5379 ^a	0.1287	-0.6048	66.6565	0.8053	0.0360	0.1466
121	108	2.2292	-0.4876	0.1835	-0.0837	0.1544	-0.1105	-0.5136	104.6632	0.6563	0.8506	0.8901
122	42	-0.8324	0.0962	-0.2290	0.0988	0.5748	-0.2166	0.0432	165.3384	0.6619	0.3086	0.3702
111	64	0.0514	0.3474	-0.5259	0.0335	0.7983 ^a	-0.5611 ^a	0.1332	169.1515	0.7022	0.4442	0.5854
112	54	-0.5748	0.3954	-0.1541	0.2278 ^a	0.6335	-0.3486	0.2299	169.2884	0.6981	0.3762	0.4997

^aExplanatory variable contributing significantly to the model

Fig. 6 Results obtained from the neural network (NN) model: the response variable is binary—patient with or without a haplotype. The explanatory variables are BMI, BS0, BS60, and BS120

the NN model are too complicated to explain the association between phenotypes and haplotypes. Therefore, in order to explore the association between phenotypes and haplotypes, *QTLmarc* is more suitable than the NN model with regard to simplicity.

It is also noteworthy whether discrimination based on the maximization of AUC is equivalent to the linear discrimination method (Fisher 1936). However, it is known that the AUC value calculated using estimated parameters based on the minimization of squared error is not always equal to the maximum AUC value (Cortes et al. 2004; IBM 2004). A comparison of the results obtained from *QTLmarc* with those obtained from the linear discrimination method would be worthwhile; however, this is beyond the scope of this study.

The reader may wonder how the effect of correlation between phenotypes is applied to *MARC* or *QTLmarc*. As a simple solution, the multiplicative phenotypes correlated with each other are added into the model as a new phenotype. It is reasonable to represent the interaction between variables as the multiplicative variables in various statistical models.

Another issue is whether estimation of a haplotype by *QTLmarc* is not in fact HE in the original sense. In *QTLmarc*, two haplotypes of an individual (or diplotype) are not estimated; however, whether an individual becomes a carrier or non-carrier of a haplotype as the combination of alleles is estimated. Since, in the original sense, HE implies that the diplotype configuration is

unambiguously determined, it might be misunderstood that *QTLmarc* can determine the diplotype configuration of an individual from genotype data. However, our aim in this study is not the estimation of the haplotype but the evaluation of the association between a haplotype and several quantitative phenotypes. If the reader needs to evaluate the association between quantitative phenotypes and haplotype in the original sense, it is possible to use the haplotype estimated by other algorithms, for example, *LDSUPPORT* (Kitamura et al. 2002), to evaluate the association between haplotype and phenotype. In *MARC*, diplotype data can be applied instead of genotype data to evaluate the association between haplotype and several quantitative phenotypes.

We conclude this paper by mentioning a theoretical problem that is yet to be resolved. In the *MARC* method, it is assumed that the observed phenotype is normally distributed. However, we cannot expect all observed phenotypes to be normally distributed. To solve such a problem, we can calculate the AUC value in a non-parametric manner (Zhou et al. 2002). Moreover, the *MARC* is only available for a binary-discrimination problem. Therefore, our method is not applicable to the Genotype model, wherein the behavior of phenotypes changes due to the number of haplotypes. Such an expansion is mathematically easy; however, the results of the model are difficult to understand. These two problems warrant further investigation.

Fig. 7 ROC curve and AUC value for partitioning by the NN model: the response variable is binary—patient with or without a haplotype; the explanatory variables are BMI, BS0, BS60, and BS120

Acknowledgements This work was supported by grants from New Energy and Industrial Technology Development Organization.

References

- Akaike J (1973) Information theory and an extension of the maximum likelihood principle. In: The 2nd International symposium on information theory. Tsahkadsor, Armenian SSR, Hungary, 2–8 September 1971, pp 267–281
- Chambers JM, Hastie TJ (eds) (1992) Statistical models. Wadsworth and Brooks Gale, Pacific Grove, CA
- Copas JB, Corbett P (2002) Overestimation of the receiver operating characteristic curve for logistic regression. *Biometrika* 89:315–331
- Cortes C, Mohri M (2004) AUC optimization vs. error rate minimization. *Advances in neural information processing systems* (NIPS 2003). MIT Press, Cambridge, MA
- Fisher RA (1918) The correlation between relatives on the supposition of Mendelian inheritance. *Trans R Soc Edinburgh* 52:399–433
- Fisher RA (1936) The use of multiple measurements in taxonomic problems. *Ann Eugen* 7:179–188
- IBM TJ Watson Research Center (2004) Model selection via the AUC. In: Proceedings of the 21st international conference on machine learning, Banff, Canada
- Iwasaki N, Horikawa Y, Tsuchiya T, Kitamura Y, Nakamura T, Tanizawa Y, Oka Y, Hara K, Kadowaki T, Awata T, Honda M, Yamashita K, Oda N, Yu L, Yamada N, Ogata M, Kamatani N, Iwamoto Y, Del Bosque-Plata L, Hayes MG, Cox NJ, Bell GI (2005) Genetic variants in the calpain-10 gene and the development of type 2 diabetes in the Japanese population. *J Hum Genet* 2:92–98
- Kitamura Y, Moriguchi M, Kaneko H, Morisaki H, Morisaki T, Toyama K, Kamatani N (2002). Determination of probability distribution of diplotype configuration (diplotype distribution) for each subject from genotypic data using the EM algorithm. *Ann Hum Genet* 66:183–193
- McClish DK (1989) Analyzing a portion of the ROC curve. *Med Decis Making* 9:190–195
- Pepe MS (1997) A regression modelling framework for receiver operating characteristic curve in medical diagnostic testing. *Biometrika* 84:595–608
- Pepe MS (1998) Three approaches for regression analysis of receiver operating characteristic curves for continuous test results. *Biometrics* 54:124–135
- Pepe MS (2000) Interpretation, estimation and regression for ROC curves. *Biometrics* 56:352–359
- Qin J (2000) Using logistic regression procedures for estimating receiver operating characteristic curves. *Biometrika* 90:585–596
- Ransohoff DJ, Feinstein AR (1978) Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med* 299:926–930
- Sebastiani P, Ramoni MF, Nolan V, Baldwin CT, Steinberg MH (2005) Genetic dissection and prognostic modeling of overt stroke in sickle cell anemia. *Nat Genet* 37:435–400
- Shibata K, Ito T, Kitamura Y, Iwasaki N, Tanaka H and Kamatani N (2004) Simultaneous estimation of haplotype frequencies and quantitative trait parameters: applications to the test of association between phenotype and diplotype configuration. *Genetics* 168:525–539
- Thomas A, Camp NJ (2004) Graphical modeling of the joint distribution of alleles at associated loci. *Am J Hum Genet* 74:1088–1101
- Toesteson AAN, Begg CB (1988) A general regression methodology for ROC curve estimation. *Med Decis Making* 8:204–215
- Yi N, George V, Allison DB (2003). Stochastic search variable selection for Identifying multiple quantitative trait loci. *Genetics* 164:1129–1138
- Zhou XH, Obuchowski NA, McClish DK (2002) Statistical methods in diagnostic medicine. Wiley, New York