Paper 235-28

A SAS/IML® Program for Mapping QTL in Line Crosses

Chenwu Xu, University of California at Riverside, Riverside, CA
Shizhong Xu, University of California at Riverside, Riverside, CA

ABSTRACT

We developed a SAS® program to implement QTL interval mapping and composite interval mapping for complex traits in line crosses. The program consists of four macros. The first macro (QTLPROB) calculates the conditional probabilities of QTL genotypes at any putative position of the genome given observed marker information. This macro handles missing and partially informative markers using the multipoint method. The second macro (QTLMAP) provides estimates of QTL effects and test statistic for any putative position of the genome. Maximum likelihood method and likelihood ratio test are used in the analysis. The third macro (THRESHOLD) determines the critical value used to declare statistical significance using the approximate method of Piepho. Piepho's method is simple and fast because it requires no permutation resamplings of the data. The last macro (GRAPH) graphes the result to visually identify the QTL position on the corresponding chromosome. Three properties of the program are unique compared to other commonly used software packages for QTL mapping: (1) we incoprorated a unified QTL mapping strategy that is aimed to handle a four-way cross family but treat F_2 and backcross (BC) as special cases, (2) the program facilitates QTL mapping for complex binary traits using the generalized linear model approach, and (3) the program is capable of computing the standard errors of estimated QTL effects using Louis observed information matrix.

INTRODUTION

Over the last 10 years, there has been a great deal of interests in the development of methodology to map polygenes or quantitative trait loci (QTL) relative to a known marker map in population derived from inbred line crosses. Many software packages are available for QTL mapping using C, FORTRAN, PASCAL or JAVA languages at many platforms such as Macintosh, Windows and Unix. The most commonly used software packages for QTL mapping include Mapmaker/QTL, QTL Cartographer, MapQTL, MQTL, Mapmanager QTX and so on.

However, mapping populations that can be handled by the software packages mentioned above must be derived from the cross of two inbred lines. The drawback of these designs is that the statistical inference space is quite narrow and thus results from one cross can not be generalized to other crosses derived from different inbred lines. Xu (1996, 1998) proposed the four-way cross design of QTL mapping, intended to increase the statistical inference space and the opportunity for detecting more QTL. This design involves four inbred lines and allows simultaneous estimation and test of several QTL effects. If one effect is significant, the QTL is declared as significant. Therefore, the power of QTL detection can be increased. In addition, the method can detect more QTL because it simultaneously tests the segregations of two crosses. There has been no software package developed so far for four-way cross mapping and this package is the first attempt of such a kind. In addition, we show that the four-way cross design is a general design with F_2 and BC treated as special cases.

Many complex diseases, such as disease resistances, show a binary or dichotomous phenotype, but do not follow a simple Mendelian pattern of inheritance. Mapping loci for such binary traits usualy requires a quite different method (McIntyre et al. 2001). The software packages mentioned previously do not handle binary disease mapping. We invoke in this paper the threshold model to map disease, assuming that the binary phenotype is controlled by an underlying variable called the liability. The liability is a continuous quantitative trait except that it is not observable. The connection between the liability and the observed discrete phenotypes is modeled by the probit function. Under this generalized linear model, mapping disease loci has been formulated as mapping loci for liability and thus for quantitative traits. In this study, we incorporated QTL mpping for both quantitative traits and binary disease traits in the same set of programs.

Although variances of estimated genetic effects are easily calculated using the least squares method (Haley and Knott 1992), there are no straightforward methods to obtain such estimates if a maximum likelihood method implemented via the EM algorithm is utilized. To calculate the variances and covariances of EM estimates, complicated methods must be resorted (Kao and Zeng 1997). In this study, we developed a Monte Carlo method to obtain the observed information matrix of Louis (1982). Based on the Louis matrix, we can easily calculate the variance-covariance matrix of EM estimates.

The main consideration for us to code the program using SAS instead of other computer languages such as C++ or FORTRAN is to take advantage of the diversified and well tested SAS data steps, SAS procedures and SAS macros. In addition, more and more scientists are using SAS for their data analyses and are familiar with the SAS languages. Almost all research institutes and companies have purchased the software and distribute site license to their users.

STATISTICAL METHODS

The four-way cross model

Let L_1 and L_2 be the two inbred lines initiating the first cross and L_3 and L_4 be the inbred lines intiating the second cross. Denote the QTL genotypes of L_1 and L_2 by $Q_1^mQ_1^m$ and $Q_2^mQ_2^m$, respectively, and the genotypes of L_3 and L_4 by $Q_1^fQ_1^f$ and $Q_2^fQ_2^f$, respectively. The genetic constitution of the four-way cross population will consists of four genotypes: $Q_1^mQ_1^f$, $Q_1^mQ_2^f$, $Q_2^mQ_1^f$ and $Q_2^mQ_2^f$, with equal frequency. Let G_{kl} be the value of genotype $Q_k^mQ_l^f$ and it can be expressed by the following linear model:

$$G_{kl} = \mu + a_k^m + a_l^m + d_{kl} \quad (1)$$

where μ is the population mean, a_k^m and a_l^f are the effects of the *k*th allele of the father and the *l*th allele of the mother, respectively, and d_{kl} is the dominance effect, for *k*, *l*=1, 2. Note that there are only four possible genotypes in the progeny, but we have nine parameters in the model. Therefore, we must impose some restrictions to the parameters to make the model estimable. One such a model with restriction is

$$\begin{cases} G_{11} = \mu + a^m + a^f + d \\ G_{12} = \mu + a^m - a^f - d \\ G_{21} = \mu - a^m + a^f - d \\ G_{22} = \mu - a^m - a^f + d \end{cases} \quad (2)$$

The eight genetic effects in model (1) have been reduced to three genetic effects in model (2). In matrix notation, the above model can be expressed as $\mathbf{G} = \mathbf{H}\mathbf{b}$, where

$$\mathbf{G} = \begin{bmatrix} G_{11} \\ G_{12} \\ G_{21} \\ G_{22} \end{bmatrix}, \ \mathbf{H} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \text{ and } \mathbf{b} = \begin{bmatrix} \mu \\ a^m \\ a^f \\ d \end{bmatrix}.$$

Let $\mathbf{H}_i$ be the *i*th row of matrix $\mathbf{H}$, then $G_{11} = \mathbf{H}_1\mathbf{b}$, $G_{12} = \mathbf{H}_2\mathbf{b}$, $G_{21} = \mathbf{H}_3\mathbf{b}$ and $G_{22} = \mathbf{H}_4\mathbf{b}$.

We now describe the linear model for a particular individual. An individual can take one of the four possible genotypes, and thus the linear model for individual *j* is

$$y_j = \mathbf{X}_j\mathbf{b} + \varepsilon_j \tag{3}$$

where $\mathbf{X}_j = \mathbf{H}_1$ if individual *j* takes the first genotype $Q_1^m Q_1^f$ and $\mathbf{X}_j = \mathbf{H}_2$ if *j* takes the second genotype $Q_1^m Q_2^f$ and so on, and ε_j is the residual error distributed as N(0, σ_e^2). Model (3) is a general linear model (GLM) with missing value in $\mathbf{X}_j$ because the genotype of *j* is not observable.

The next step of the GLM analysis with missing value is to infer the probabilities of QTL genotypes conditional on the marker information, denoted by $\Pr(\mathbf{X}_j = \mathbf{H}_k \mid \mathbf{I}_M)$ for $k = 1, \cdots, 4$ where $\mathbf{I}_M$ represents marker information (Rao and Xu 1998).

Maximum likelihood estimation (MLE)

If $\mathbf{X}_j$ were observed for every individual, the MLE of the parameters could be found explicitly in a single step using the following equations:

$$\begin{cases} \hat{\mathbf{b}} = \left[\sum_{j=1}^{n} \mathbf{X}_j^T\mathbf{X}_j\right]^{-1}\left[\sum_{j=1}^{n} \mathbf{X}_j^T y_j\right] \\ \hat{\sigma}_e^2 = \frac{1}{n}\sum_{j=1}^{n}(y_j - \mathbf{X}_j\hat{\mathbf{b}})^2 \end{cases} \tag{4}$$

In the case where $\mathbf{X}_j$ is missing but the distribution of $\mathbf{X}_j$ is given, the EM algorithm can be adopted to take advantage of the above equations. The EM equations simply replace all the terms related to $\mathbf{X}_j$ by their expectations, *i.e.*,

$$\begin{cases} \hat{\mathbf{b}} = \left[\sum_{j=1}^{n} E(\mathbf{X}_j^T\mathbf{X}_j)\right]^{-1}\left[\sum_{j=1}^{n} E(\mathbf{X}_j^T y_j)\right] \\ \hat{\sigma}_e^2 = \frac{1}{n}\sum_{j=1}^{n} E(y_j - \mathbf{X}_j\mathbf{b})^2 \end{cases} \tag{5}$$

The expectations are obtained conditional on both marker information and the phenotypic value y_j. The connection between the phenotype and the QTL genotype is through the three parameter values, but the parameters are what we are trying to find. Therefore, we need iterations on equation (5) by providing some initial values of the parameters to start the iteration. This is the EM algorithm. The E-step is to find the expectations and the M-step is to invoke euqtion (5) for iterations.

Recall that the probability of $\mathbf{X}_j$ conditional on marker information is denoted by $\Pr(\mathbf{X}_j = \mathbf{H}_k \mid \mathbf{I}_M)$. This probability may be called the prior probability. After incorporating the phenotypic value, we obtain the posterior probability, denoted by

$$\Pr(\mathbf{X}_j = \mathbf{H}_k \mid \mathbf{I}_M, y_j) = \frac{\Pr(\mathbf{X}_j = \mathbf{H}_k \mid \mathbf{I}_M) f(y_j - \mathbf{H}_k\mathbf{b})}{\sum_{i=1}^{4} \Pr(\mathbf{X}_j = \mathbf{H}_i \mid \mathbf{I}_M) f(y_j - \mathbf{H}_i\mathbf{b})}$$

where $f(y_j - \mathbf{H}_i\mathbf{b}) = \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp[-\frac{1}{2\sigma_e^2}(y_j - \mathbf{H}_i\mathbf{b})^2]$

The expectations are actually obtained using the posterior probabilities rather than the prior probabilities. Therefore,

$$\sum_{j=1}^{n} E(\mathbf{X}_j^T\mathbf{X}_j) = \sum_{j=1}^{n}\left(\sum_{i=1}^{4} \Pr(\mathbf{X}_j = \mathbf{H}_i \mid \mathbf{I}_M, y_j)\mathbf{H}_i^T\mathbf{H}_i\right),$$

$$\sum_{j=1}^{n} E(\mathbf{X}_j^T y_j) = \sum_{j=1}^{n}\left(\sum_{i=1}^{4} \Pr(\mathbf{X}_j = \mathbf{H}_i \mid \mathbf{I}_M, y_j)\mathbf{H}_i^T y_j\right),$$

$$\sum_{j=1}^{n} E(y_j - \mathbf{X}_j\mathbf{b})^2 = \sum_{j=1}^{n}\left(\sum_{i=1}^{4} \Pr(\mathbf{X}_j = \mathbf{H}_i \mid \mathbf{I}_M, y_j)(y_j - \mathbf{H}_i\mathbf{b})^2\right)$$

Likelihood ratio test statistic

Define the log-likelihood value evaluated at the MLE of parameters as

$$L(\hat{\mathbf{b}}, \hat{\sigma}_\varepsilon^2) = \sum_{j=1}^{n} \log[\sum_{i=1}^{4} \Pr(\mathbf{X}_j = \mathbf{H}_i \mid \mathbf{I}_M) f(y_j - \mathbf{H}_i\hat{\mathbf{b}})].$$

This is also called the likelihood value under the full model. We need the likelihood values under various restricted models to test various hypotheses.

The overall null hypothesis is no effect of QTL at the locus of interest, denoted by $H_0 : a^m = a^f = d = 0$ or $H_0 : \mathbf{L}\mathbf{b} = \mathbf{0}$, where $\mathbf{L} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$. If we solve the MLE of the parameters under the restriction of $\mathbf{L}\mathbf{b} = \mathbf{0}$ and evaluate the likelihood value at the solutions with this restriction, we have $L(\hat{\mu}, \hat{\sigma}_e^2) = L(\hat{\mathbf{b}}, \hat{\sigma}_e^2 \mid \mathbf{L}\mathbf{b} = \mathbf{0})$.

The likelihood ratio test statistic is

$$\Lambda = -2[L(\hat{\mu}, \hat{\sigma}_e^2) - L(\hat{\mathbf{b}}, \hat{\sigma}_e^2)] = -2[L(\hat{\mathbf{b}}, \hat{\sigma}_e^2 \mid \mathbf{L}\mathbf{b} = \mathbf{0}) - L(\hat{\mathbf{b}}, \hat{\sigma}_e^2)] \tag{6}$$

Various other test statistics can be defined by redefining the $\mathbf{L}$ matrix. To test the hypothesis of $H_1 : a^m = 0$, we define $\mathbf{L}_1 = [0 \ \ 1 \ \ 0 \ \ 0]$. The likelihood ratio test statistic is $\Lambda_1 = -2[L(\hat{\mathbf{b}}, \hat{\sigma}_e^2 \mid \mathbf{L}_1\mathbf{b} = 0) - L(\hat{\mathbf{b}}, \hat{\sigma}_e^2)]$. To test the hypothesis of $H_2 : a^f = 0$, we define $\mathbf{L}_2 = [0 \ \ 0 \ \ 1 \ \ 0]$ and use $\Lambda_2 = -2[L(\hat{\mathbf{b}}, \hat{\sigma}_e^2 \mid \mathbf{L}_2\mathbf{b} = 0) - L(\hat{\mathbf{b}}, \hat{\sigma}_e^2)]$. Similarly, we use $\Lambda_3 = -2[L(\hat{\mathbf{b}}, \hat{\sigma}_e^2 \mid \mathbf{L}_3\mathbf{b} = 0) - L(\hat{\mathbf{b}}, \hat{\sigma}_e^2)]$ to test the hypothesis of $H_3 : d = 0$ where $\mathbf{L}_3 = [0 \ \ 0 \ \ 0 \ \ 1]$.

We essentially generate the test statistics for the entire genome from one end to the other with one or two cM increment to form test statistic profiles. Using the SAS graphical procedures, we can visualize the test statistic profiles and identify the QTL locations and effects.

Variance-covariance matrix of EM estimates

Most QTL mapping software packages do not porovide estimates of the standard errors of estimated parameters because the EM algorithm does not offer a straightforward way to calculate the errors. In this section, we introduce a simple Monte Carlo method to calculate the variance-covariance matrix of the estimated parameters.

Let $\boldsymbol{\theta}=(\mathbf{b},\sigma_e^2)$ be the vector of parameters, $S(\boldsymbol{\theta},\mathbf{X})$ and $\mathrm{B}(\boldsymbol{\theta},\mathbf{X})$ be the first and second partial derivatives of the complete-data log-likelihood,

$$S(\boldsymbol{\theta},\mathbf{X})=\begin{bmatrix}\frac{1}{\sigma_e^2}\left(\sum_{j=1}^{n}\mathbf{X}_j^T\mathbf{X}_j\mathbf{b}-\sum_{j=1}^{n}\mathbf{X}_j^T y_j\right)\\ \frac{1}{2\sigma_e^4}\sum_{j=1}^{n}(y_j-\mathbf{X}_j\mathbf{b})^2-\frac{n}{2\sigma_e^2}\end{bmatrix} \quad (7)$$

$$B(\boldsymbol{\theta},\mathbf{X})=\begin{bmatrix}\frac{1}{\sigma_e^2}\sum_{j=1}^{n}\mathbf{X}_j^T\mathbf{X}_j & \frac{1}{\sigma_e^4}\left(\sum_{j=1}^{n}\mathbf{X}_j^T y_j-\sum_{j=1}^{n}\mathbf{X}_j^T\mathbf{X}_j\mathbf{b}\right)\\ \frac{1}{\sigma_e^4}\left(\sum_{j=1}^{n}\mathbf{X}_j^T y_j-\sum_{j=1}^{n}\mathbf{X}_j^T\mathbf{X}_j\mathbf{b}\right)^T & \frac{1}{\sigma_e^4}\left(\frac{1}{\sigma_e^2}\sum_{j=1}^{n}(y_j-\mathbf{X}_j\mathbf{b})^2-\frac{n}{2}\right)\end{bmatrix} \quad (8)$$

By complete-data log-likelihood, we mean the log-likelihood function found as if the $\mathbf{X}$ were observed. The observed information matrix of Louis (1982) evaluated at $\hat{\boldsymbol{\theta}}$ is

$$I(\hat{\boldsymbol{\theta}})=E_{\hat{\boldsymbol{\theta}}}\left\{B(\hat{\boldsymbol{\theta}},\mathbf{X})\right\}-E_{\hat{\boldsymbol{\theta}}}\left\{S(\hat{\boldsymbol{\theta}},\mathbf{X})S^T(\hat{\boldsymbol{\theta}},\mathbf{X})\right\} \quad (9)$$

where the expectation are taken with respect to the missing data, $\mathbf{X}_j$, using the posterior distribution of $\mathbf{X}$ in which $\boldsymbol{\theta}$ is substituted by $\hat{\boldsymbol{\theta}}$. The first expectation in equation (9) has an explicit form and is easy to evaluate. The explicit form of the second expectation, however, is very complicated (Kao and Zeng 1997). Fortunately, it can be easily evaluated using Monte Carlo integration by sampling $\mathbf{X}$ from its posterior distribution, i.e.,

$$E_{\hat{\boldsymbol{\theta}}}\{S(\hat{\boldsymbol{\theta}},\mathbf{X})S^T(\hat{\boldsymbol{\theta}},\mathbf{X})\}\approx\frac{1}{N}\sum_{i=1}^{N}S(\hat{\boldsymbol{\theta}},\mathbf{X}^{(i)})S^T(\hat{\boldsymbol{\theta}},\mathbf{X}^{(i)}) \quad (10)$$

where $\mathbf{X}^{(i)}$ is the *i*th sample of $\mathbf{X}$ for *i*=1, $\cdots$, *N* and *N* is a large number, say 1000. Although Monte Carlo integration itself is computationally tedious, it is not so in this situation because the integration is only required once at the very last iteration of the EM process. Given the observed information matrix, the variance-covatiance matrix of $\hat{\boldsymbol{\theta}}$ is calculated as

$$Var(\hat{\boldsymbol{\theta}})=I^{-1}(\hat{\boldsymbol{\theta}}) \quad (11)$$

Extension to F_2 and BC populations

The four-way cross model is a general model from which the F_2 and BC models can be expressed as special cases. Let us first consider a BC population. The genotypes of the two parents of the BC family can be defined as $Q_1^mQ_2^f\times Q_1^mQ_1^m$ or $Q_1^mQ_2^f\times Q_2^fQ_2^f$, depending on which inbred line is used as the tester. The constitution of genotypes of the mating pair may be called the mating type. Let us assume that $Q_1^mQ_2^f\times Q_2^fQ_2^f$ is the mating type for the BC family. A progeny from this mating type can take one of the four possible genotypes: $Q_1^mQ_2^f$, $Q_1^mQ_2^f$, $Q_2^fQ_2^f$ and $Q_2^fQ_2^f$. Note that the first and the second genotypes are not distinguishable, and neither are the third and the fourth. If we use the same notation as that of the four-way cross for the four genotypic values, we have $G_{11}=G_{12}$ and $G_{21}=G_{22}$. The genetic effects defined in the notation of a four-way cross are $a^m=G_{11}+G_{12}-G_{21}-G_{22}$, $a^f=G_{11}-G_{12}+G_{21}-G_{22}=0$ and $d=G_{11}-G_{12}-G_{21}+G_{22}=0$. Therefore, we can use the same four-way cross model for the BC mapping with the restriction of $a^f=d=0$. This can be acomplished by searching for the MLE of the four-way cross model with $\mathbf{Lb}=\mathbf{0}$, where $\mathbf{L}=\begin{bmatrix}0&0&1&0\\0&0&0&1\end{bmatrix}$.

All marker genotypes are considered as either partially informative (when typed) or non-informative (when missing), and thus the same multipoint method can be used to infer the QTL genotype of a putative position using all markers.

Let us now consider an F_2 population. The genotypes of the two parents of the F_2 family can be defined as $Q_1^mQ_2^f\times Q_1^mQ_2^f$. A progeny from this mating type can take one of the four possible genotypes: $Q_1^mQ_1^m$, $Q_1^mQ_2^f$, $Q_2^fQ_1^m$ and $Q_2^fQ_2^f$. Note that the second and the third genotypes are not distinguishable. If we use the same notation as that of the four-way cross for the four genotypic values, we have $G_{12}=G_{21}$. The genetic effects defined in the four-way cross are $a^m=G_{11}+G_{12}-G_{21}-G_{22}=G_{11}-G_{22}$, $a^f=G_{11}-G_{12}+G_{21}-G_{22}=G_{11}-G_{22}$ and $d=G_{11}-G_{12}-G_{21}+G_{22}=G_{11}-2G_{12}+G_{22}$. Therefore, we can use the same four-way cross model for the F_2 mapping with the restriction of $a^m=a^f$. This can be acomplished by searching for the MLE of the four-way cross model with $\mathbf{Lb}=\mathbf{0}$, where $\mathbf{L}=[0\ \ 1\ \ -1\ \ 0]$. A marker genotype is considered as fully informative if it is homozygous. A heterozygous genotype is considered as partially informative because we cannot tell between the second and the third genotypes. The same multipoint method can be used to infer the QTL genotype of a putative position.

Recall that the design matrix for the linear model in the four-way cross is denoted by $\mathbf{X}_j=\begin{bmatrix}X_{1j}&X_{2j}&X_{3j}&X_{4j}\end{bmatrix}$ for the *j*th individual. The coefficient of each genetic effect $(i.e.,X_{2j},X_{3j},X_{4j})$ takes one of two possible values, 1 and -1, with an equal probability. Therefore, they all have a zero expectation and a unit variance, and are orthogonal between each other. Therefore, the total genetic variance due to the three effects in a four-way cross can be expressed as

$$\sigma_G^2=Var(X_{2j})(a^m)^2+Var(X_{3j})(a^f)^2+Var(X_{4j})(d)^2$$
$$=(a^m)^2+(a^f)^2+(d)^2.$$

When extended to the BC family, X_{3j} and X_{4j} have vanished from the model. The coefficient left in the model is X_{2j}, which takes value 1 for heterozygote and -1 for homozygote. In the traditional BC model, however, the coefficient is defines as 1 for heterozygote and 0 for homozygote, which leads to an expectation of 1/2 and a variance of 1/4. Therefore, when the traditional BC model is compared with our extended BC model, we should take into consideration the scale difference. The estimated effect of the extended BC model would be half the effect of the traditional BC model.

When extended to the F_2 family, X_{2j} and X_{3j} have been combined because $a^m=a^f=a$. Therefore , the coefficient of the

additive effect is $X_{2j}+X_{3j}$, with a zero expectation and a variance of 2. This means that the coefficient of the additive effect is defined as -2 for one homozytote, 0 for the heterozygote and 2 for the other homozygote. In the traditional F_2 model, however, the coefficient of the additive effect is defined as 0 for one homozygote, 1 for the heterozygote and 2 for the other homozygote. In such a scale, the expectation of the additive coefficent is 1 and the variance is 1/2. Therefore, when the traditional F_2 model is compared with our extended F_2 model, we should take into consideration the scale difference. The estimated additive effect of the traditional F_2 model would be twice the effect of the extended F_2 model. The coefficient of the dominance effect in the extended F_2 model is defined as 1 for the homozygote and -1 for the heterozygote, whereas in the traditional F_2 model, this coefficient is defined as 1 for the heterozygote and 0 for the homozygote. Therefore, the estimated dominance effect in the extended F_2 model should be half the effect of the traditional F_2 model with an opposite sign.

Extension to complex binary traits

We took a generalized linear model approach and choose the probit as the link function between the QTL effects and the binary phenotype. The probit model is defined as

$$\Pr(w_j=1\mid \mathbf{X}_j,\mathbf{b})=\Pr(y_j>0\mid \mathbf{X}_j,\mathbf{b}) =\int_0^{\infty}\phi(y_j-\mathbf{X}_j\mathbf{b})dy_j=1-\Phi(\mathbf{X}_j\mathbf{b}) \quad (12)$$

where $\phi(y_j-\mathbf{X}_j\mathbf{b})$ and $\Phi(\mathbf{X}_j\mathbf{b})$ are the standardized normal density function and standardized normal cumulative function, respectively.

$$y_j=\mathbf{X}_j\mathbf{b}+e_j \quad (13)$$

is the latent variable assumed to be normally distributed with mean $\mathbf{X}_j\mathbf{b}$ and variance 1. For convenience, the probability of observed binary phenotype w_j is described by the Bernoulli distribution,

$$\Pr(w_j\mid \mathbf{X}_j,\mathbf{b})=[1-\Phi(\mathbf{X}_j\mathbf{b})]^{w_j}[\Phi(\mathbf{X}_j\mathbf{b})]^{1-w_j} \quad (14)$$

The overall log-likelihood function of the entire mapping population is

$$L(\boldsymbol{\theta})=\sum_{j=1}^{n}\log[\Pr(w_j\mid \mathbf{X}_j,\mathbf{b})].$$

The only difference between mapping binary traits and mapping quantitative trait is that y_j is also missing in the binary trait mapping. We can invoke the same EM algorithm by replacing y_j by $\hat{y}_j$, the expectation of y_j conditional on w_j, $\mathbf{X}_j$ and $\mathbf{b}$.

$$\hat{y}_j=\mathrm{E}(y_j\mid w_j,\mathbf{X}_j,\mathbf{b})=\mathbf{X}_j\mathbf{b}+(2w_j-1)\frac{\phi(\mathbf{X}_j\mathbf{b})}{\Phi[(1-2w_j)\mathbf{X}_j\mathbf{b}]} \quad (15)$$

and the posterior probability of QTL genotypes by

$$\Pr(\mathbf{X}_j=\mathbf{H}_k\mid \mathbf{I}_M,w_j)=\frac{\Pr(\mathbf{X}_j=\mathbf{H}_k\mid \mathbf{I}_M)\Pr(w_j\mid \mathbf{H}_k,\mathbf{b})}{\sum_{i=1}^{4}\Pr(\mathbf{X}_j=\mathbf{H}_i\mid \mathbf{I}_M)\Pr(w_j\mid \mathbf{H}_i,\mathbf{b})}.$$

Approximate threshold value for significance test

Consider the high computational workload of pemutation test encountered in practical QTL mapping, we adopted a quick method proposed by Piepho (2001) to compute approximate thresholds for QTL detection. The method can control the genome-wise type I error rates of test for QTL detection and use the variation and correlation of the likelihood-ratio (LR) test statistics across the genome locations. The upper bound of the genome-wise type I error rate is estimated by

$$\gamma=m\Pr(\chi_k^2>C)+(\sum_{i=1}^{m}V_i)C^{(1/2)(k-1)}e^{(-C/2)}2^{(-k/2)}/\Gamma(k/2) \quad (16)$$

where γ is the genome-wise type I error rate, C is the critical threshold value for LR test statistic, k is the number of genetic effects for the putative QTL, m is the number of chromosomes and V_i is the value of V for the ith chromosome, which is defined as

$$V_i=|\sqrt{T(0)}-\sqrt{T(\rho_1)}|+|\sqrt{T(\rho_1)}-\sqrt{T(\rho_2)}|+\cdots+|\sqrt{T(\rho_{L-1})}-\sqrt{T(\rho_L)}| \quad (17)$$

where $T(\rho)$ is the LR test statistic at the putative QTL position ρ in centimorgans (cM). $\rho_1\cdots\rho_L$ are the successive turning points of $\sqrt{T(\rho)}$. Formula (17) can be calculated simply by taking the absolute differences between successive square roots of $T(\rho)$ on the fine grid, e.g., between 1 cM and 2 cM, and sum these across the chromosome. The unknown critical value C in formula (16) is solved numerically using the bisection procedure.

EXECUTION OF THE SAS PROGRAM

Preparation for mapping

We coded four SAS macros: QTLPROB, QTLMAP, THRESHOLD and GRAPH, which are saved in one file named qtlmap.sas for QTL mapping. The QTLPROB macro is used to calculate the conditional probabilities of QTL genotypes at all positions (with 1 cM interval) along the entire genome for all individuals. This macro reads data from two files: the map file and the marker genotype file. The map file stores the marker map information, including the chromosome identification numbers, the marker names and the positions of the markers. The marker genotype file stores the genotypes for all individuals at all markers in the order specified in the map file. The contents of the two files are shown in Figure 1 for the map file and Figure 2 for the marker genotype file. The QTLPROB macro requires three arguments: *ns*, *nchr* and *nmark*, where *ns* defines the number of sibs (sample size), *nchr* is the number of chromosomes and *nmark* is the total number of markers on entire genome.

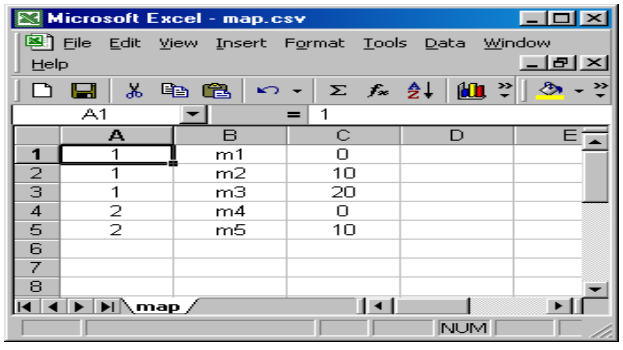

Figure 1 The map file

In the map file, column A through column C store the chromosome id, the marker name and the marker position measured in cM relative to the position of the first marker within each chromosome. In the marker genotype file, column A through column C store the individual id, the father id and the

mother id, respectively. Columns D and E contain the allelic forms of the two alleles of the first marker carried by all individuals. The alleles are ordered as paternal followed by

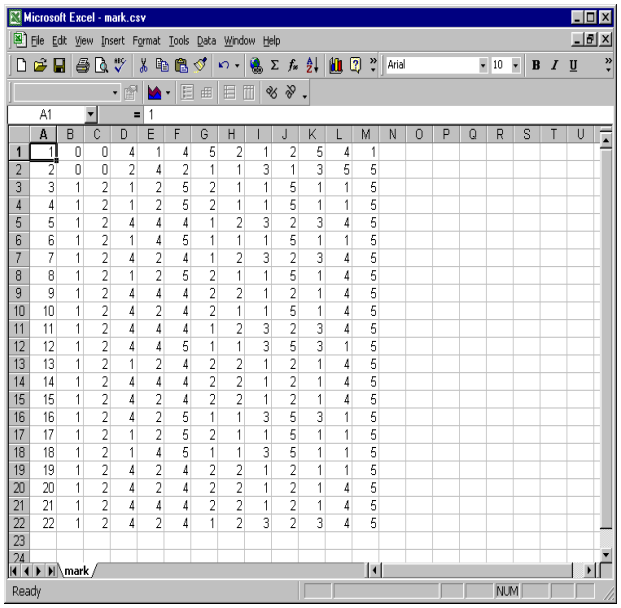

Figure 2 The marker genotype file

maternal. Columns F and G contain the allelic forms of the two alleles of the second marker carried by all individuals, and so on.

It is noted that the first two lines in marker genotype file are genotypic information of father ($L_1 \times L_2$) and mother ($L_3 \times L_4$) in four-way cross design, respectively. In addtion, if the mating type is $Q_1^m Q_2^f \times Q_2^f Q_2^f$ for the BC family, the allelic forms of the two alleles of each marker are 1, 2 for father $Q_1^m Q_2^f$ and 2, 2 for mother $Q_2^f Q_2^f$, respectively. Similarly, the allelic forms of two alleles of each marker are all 1, 2 for father as well as for mother. The allelic forms of two alleles of each marker are 1, 1 for one homozygote, 2, 2 for the other homozygote and 1, 2 for heterozygote for each progeny in BC and F_2 family. Based on this format, BC and F_2 family can be incorporated with four-way cross design.

The second SAS macro is called QTLMAP, which is to implement the EM algorithm for QTL mapping. This macro also contains a PROC IML module that calculates the variance-covariance matrix of the EM estimates of parameters. We do not recommend users to call this module before obtaining the mapping result. This is because we only need to report the variance-covariance matrix of estimated parameters at positions with evidence of QTL. Therefore, we may go back to the program to calculate the variance-covariance matrix of the estimated parameters only at those positions that have reached the threshold value of the test statistic. This macro reads a data file containing the phenotypic values and a SAS data set created by the QTLPROB macro described earlier. The format of the trait file is shown in Figure 3. The first column is the individual id and the second column is the actual phenotypic value of the trait. The macro QTLMAP requires three arguments: *type*, *ns* and *nchr* for interval mapping and four arguments: *type*, *ns*, *nchr* and *nmark* for composite interval mapping. Where *type* is a character variable taking one of three valid values, 'fw' for four-way cross, 'f2' for F_2 and 'bc' for BC. Variables *ns*, *nchr* and *nmark* are the sample size, the number of chromosomes and the number of markers, respectively. The mapping result will be saved in an external file. The folder or directory used for saving the result file and the filename may be specified by the users. The result file contains the following variables as shown in Figure 4. Column A

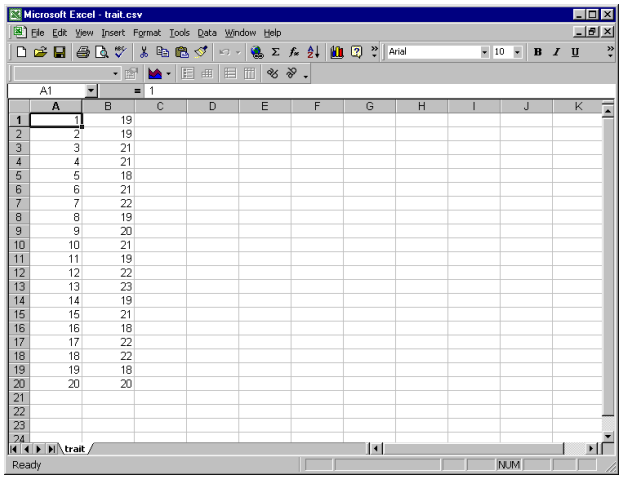

Figure 3 The trait file

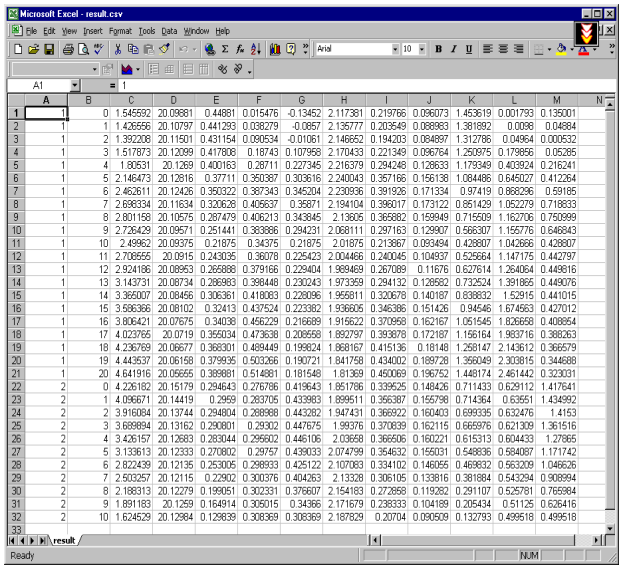

Figure 4 The result file

through column J are the chromosome id, position, likelihood ratio test statistic, $\mu, a^m, a^f, d, \sigma_e^2$, the genetic variance of QTL, and the heritability of QTL, respectively. The last three Columns are the test statistics for null hypotheses $H_1 : a^m = 0, H_2 : a^f = 0$ and $H_3 : d = 0$, respectively.

The third macro, called THRESHOLD, is to find the approximate threshold (*C*) using the quick method proposed by Piepho (2001). The value of *C* is used to test the significance of QTL at putative position. The macro THRESHOLD requires two arguments: *nchr*, the number of chromosomes, and *df*, the degree of freedom for null hypothesis test. The *df*=3, 2 and 1 for four-way cross, F_2 and BC population, respectively.

The last macro GRAPH graphes the result to visually declear the QTL position on the corresponding chromosome and needs one argument, *nchr*, the number of chromosomes.

The main program calling the macros

The main program that is saved in one file named mp.sas contains two blocks. The first block specify the file references for the external data files. The first three files are user data files and the last file is the result file. The second block is to call the macros to implement the QTL mapping.

Suppose that the three data files shown in Figure 1-3 are all saved in 'c:\qtlmap' folder and the result file is also saved in the same folder, the main program is shown as follows:

```
filename mark 'c:\qtlmap\mark.csv';
filename map 'c:\qtlmap\map.csv';
filename trait 'c:\qtlmap\trait.csv';
filename result 'c:\qtlmap\result.csv';

%QTLPROB(ns=20,nchr=2,nmark=5)
%QTLMAP(type='fw',ns=20,nchr=2)
%THRESHOLD(nchr=2,df=3)
%GRAPH(nchr=2)
```

The statement

```
%QTLPROB(ns=20,nchr=2,nmark=5)
```

is to call macro QTLPROB for calculating the multipoint conditional probabilities of QTL at all positions of the genome.

The statement

```
%QTLMAP(type='fw',ns=20,nchr=2)
```

is to call macro QTLMAP. We usually need to put the result in an external file for future use. There are 13 variables in the result file for four-way cross. However, there are only 11 variables (chromosome id, position, test statistic, μ, a, d, σ_e^2, the genetic variance of QTL, and the heritability of QTL, the two test statistics for null hypotheses $H_1: a = 0, H_2: d = 0$) for F_2 mapping and 8 variables (chromosome id, position, test statistic, μ, a, σ_e^2, the genetic variance of QTL, and the heritability of QTL) for BC mapping.

The statement

```
%THRESHOLD(nchr=2,df=3)
```

is to call macro THRESHOLD for calculating the threshold value at the genome-wise error rate.

The statement

```
%GRAPH(nchr=2)
```

is to call macro GRAPH to chart the QTL mapping profiles for each chromosome.

All arguments used in these macros are *type*='fw', *ns*=20, *nchr*=2, *nmark*=5 and *df*=3, respectively. This indicate that the mapping population is four-way cross family, the sample size is 20, the number of chromosomes is 2, the number of markers on this two chromosomes is 5 and the three genetic effects a^m, a^f and d will be estimated for four-way cross design, respectively.

The program is easy to run. User only needs to prepare the three external files previously mentioned, *i.e.* the marker genotypic file, the map file and the trait file. After running the program in SAS, the result file will be automatically created in the assigned folder and the QTL mapping profile will be charted in SAS Graph Window. From the profile, we can visually identify if significant QTLs have been detected and where they are located. Note that the three data files and the result file are all saved in Microsoft Excel format with comma delimited.

We provide two examples for demonstrating our program. Example 1 is a data set simulated for a four-way cross with 500 individuals. It includes one chromosome with 11 evenly spaced markers. The length of this chromosome is 100 cM. A single QTL is located at position 25 cM on the chromosome. The second example is also a data set simulated for a four-way cross but with a binary trait of 200 individuals. It includes one chromosome with 11 evenly spaced markers. The length of this chromosome is 100 cM. A single QTL is located at position 25 cM on the chromosome.

The source codes of the program and the sample data sets can be downloaded from our website http://www.statgen.ucr.edu.

The program is written in SAS V8.2 and run in both Windows and UNIX. To scan a genome of size 100 cM in 1 cM increment with 100 F_2 individuals, the program just takes about one minute in a Pentium 4 PC.

REFERENCES

Haley, C S and Knott, S A. 1992. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity, 69: 315-324

McIntyre, M, Coffman C J and Doerge R W. 2001. Detection and localization of a single binary trait locus in experimental populations. Genetical Research, 78: 79-92

Xu, S. 1996. Mapping quantitative trait loci using four-way crosses. Genet. Res., 68: 175-181

Xu, S. 1998. Iteratively reweighted least squares mapping of quantitative trait loci. Behavior Genetics 28:341-355

Kao, C H and Zeng Z B. 1997.General formulas for obtaining the MLEs and the asymptotic variance-covariance matrix in mapping quantitative trait loci when using the EM algorithm. Biometrics, 53: 653-665

Louis, T A. 1982. Finding the observed information matrix when using the EM algorithm. J of Royal Statistical Society Series B, 44: 226-233

Rao, S and Xu, S. 1998. Mapping quantitative trait loci for ordered categorical traits in four-way crosses. Heredity, 81: 214-224

Piepho, H P. 2001. A quick method for computing approximate thresholds for quantitative trait loci detection. Genetics, 157: 425-432

ACKNOWLEDGMENTS

This research was supported by the National Institutes of Health Grant GM55321 and the USDA National Research Initiative Competitive Grants Program 00-35300-9245 to SX.

CONTACT INFORMATION

Shizhong Xu
Department of Botany and Plant Sciences
University of California, Riverside, CA 92521
Phone: (909)787-5898
Fax: (909)787-4337
Email: xu@genetics.ucr.edu