Theor Appl Genet (2009) 118:609–617
DOI 10.1007/s00122-008-0924-y

ORIGINAL PAPER

Robust Bayesian mapping of quantitative trait loci using Student-*t* distribution for residual

Xin Wang · Zhongze Piao · Biye Wang · Runqing Yang · Zhixiang Luo

Received: 13 November 2007 / Accepted: 24 October 2008 / Published online: 20 November 2008

Abstract In most quantitative trait loci (QTL) mapping studies, phenotypes are assumed to follow normal distributions. Deviations from this assumption may affect the accuracy of QTL detection, leading to detection of false positive QTL. To improve the robustness of QTL mapping methods, we replace the normal distribution assumption for residuals in a multiple QTL model with a Student-*t* distribution that is able to accommodate residual outliers. A Robust Bayesian mapping strategy is proposed on the basis of the Bayesian shrinkage analysis for QTL effects. The simulations show that Robust Bayesian mapping approach can substantially increase the power of QTL detection when the normality assumption does not hold and applying it to data already normally distributed does not influence the result. The proposed QTL mapping method is applied to mapping QTL for the traits associated with physics–chemical characters and quality in rice. Similarly to the simulation study in the real data case the robust approach was able to detect additional QTLs when compared to the traditional approach. The program to implement the method is available on request from the first or the corresponding author.

Xin Wang and Zhongze Piao contributed equally to this study.

Communicated by M. Sillanpää.

X. Wang · B. Wang · R. Yang (✉)
School of Agriculture and Biology, Shanghai Jiaotong University, 200240 Shanghai, China
e-mail: runqingyang@sjtu.edu.cn

Z. Piao
Crop Breeding and Cultivation Research Institute,
Shanghai Academy of Agricultural Sciences,
201106 Shanghai, China

Z. Luo
Rice Research Institute, Anhui Academy of Agricultural
Sciences, 230036 Hefei, China

Introduction

Most quantitative trait loci (QTL) mapping methods such as least-squares-based, maximum likelihood-based or Bayes-based ones require the common assumption of normally distributed phenotypes. These approaches are not appropriate for the analysis of the phenotypes that are known to violate the normality assumption, because many desirable properties of the normal distribution cannot be fully utilized and deviations from normality are likely to affect the accuracy of QTL detection.

For continuous non-normally distributed traits, a classical mapping approach is to convert the trait into an approximately normal variable by applying a mathematical transformation (Sokal and Rohlf 1995). Box–Cox transformation, as a general formula, has been therefore used in QTL mapping analysis (Yang et al. 2006). Diao and Lin (2005) have plugged the true transformation function completely unspecified into the variance-components model for robust mapping QTL in human outbred population. A simple approach is applying parametric methods, such as the least-squares-based method that has legendary robustness, to directly analyze non-normally data. People have used different types of theoretical distributions to simulate non-normally distributed phenotypes and showed that robustness of parametric QTL mapping methods to non-normally distributed phenotypes is difficult to establish (e.g. Jansen 1992; Rebaï 1997; Hackett 1997; Coppieters et al. 1998). In addition, the appropriate likelihood function can also be established on any known non-normal distributions. For instance, Jansen (1992) presented a general mixture model for mapping QTL which uses the distributional properties of the data by fitting a generalized linear model; the Cox's proportional hazards model is an adequate model for mapping survival times (e.g. Symons et al. 2002; Diao et al. 2004).The distribution-free nonparametric approach had commonly been used for locating the loci of non-normal traits. Kruglyak and Lander (1995) described a nonparametric interval mapping approach based upon the Wilcoxon rank-sum test applicable to backcross designs, they demonstrated by the example of an exponential distribution that the non-parametric test would outperform parametric ones. The approach has been extended by the Coppieters et al. (1998) for half-sib pedigrees in outbred populations. Elsen and coworkers (Elsen et al. 1999; Goffinet et al. 1999; Mangin et al. 1999) presented heteroskedastic models for QTL detection in livestock populations. Furthermore, rank-based statistical methodologies have been synoptically proposed for quantitative trait locus mapping (Zou et al. 2003). When the data is non-normal, assuming that the distributions of the random effects and of the residuals are Gaussian makes inferences vulnerable to the presence of outliers (Pinheiro et al. 2001). Some symmetric and long-tailed distributions, such as the Student-*t* distribution (Rogers and Tukey 1972; Dempster et al. 1980; Lange et al. 1989), have been therefore suggested for robust estimation. Fernandez and Steel (1998) applied the method of inverse scaling of the probability density function on the left and on the right side of the distribution to a symmetric heavy-tailed distribution, thereby simultaneously capturing heavy tails and skewness. Rohr and Hoeschele (2002) have incorporated the Fernandez and Steel's approach into a Bayesian QTL mapping, developing a Robust Bayesian QTL mapping method, which allows for non-normal, continuous distributions of phenotypes within QTL genotypes, via skewed Student-*t* distributions of residual errors in the analysis. Additionally, Feenstra and Skovgaard (2004) have demonstrated that the two- (or more) component model may fit to the data much better than the single-component model within the framework of maximum likelihood.

The objective of the study is to develop a robust mapping strategy that uses the Student-*t* distribution to characterize residual error in multiple QTL model, and to investigate the robustness of mapping QTL under the framework of Bayesian shrinkage mapping by a series of simulations and a real data analysis.

Method

Genetic model

For simplicity, we only consider a backcross population derived from two inbred line. However, the method can be applied to other experimental designs, such as recombination inbred lines, F_2 design, and four-way crosses. The phenotypes and molecular marker data are collected from n individuals. Assume that there are q quantitative trait loci responsible for a trait of interest and no interactions between each others, the phenotypic value y_i of individual i can be then described by the following multiple QTL model:

$$y_i = \mu + \sum_{j=1}^{q} x_{ij} b_j + \varepsilon_i \tag{1}$$

where μ is the population mean, b_j for $j = 1,\ldots, q$, is the additive effect of the jth QTL. Variable x_{ij} is a genotype indicator variable for individual i at locus j and defined as 1 for one genotype and -1 for the other genotype; and ε_i is a random environmental error, which is assumed as a heavy-tailed Student-*t* distribution to cover much more outliers caused by non-normal distributed phenotypes.

For convenience to contrast with the normal model, we factorize ε_i into $\frac{e_i}{\sqrt{w_i}}$, where $e_i \sim N(0, \sigma^2)$ and $w_i \sim \text{Gamma}\left(\frac{\mathrm{d}f}{2}, \frac{\mathrm{d}f}{2}\right)$, here, the d*f* is the degree of freedom of Student-*t* distribution, which is a measure of the tail behavior. The smaller the d*f*, the heavier were the tails of the distribution. Apparently, the normal model is a particular case of (1), obtained by taking $w_i = 1$, for all i.

Likelihood function

The probability distribution of the phenotype data conditional on all parameters is called the likelihood. According to model (1), the conditional density of all phenotypes, given the parameters, is

$$p(y|\mu, b, \sigma^2, w, x) \propto (\sigma^2)^{-\frac{n}{2}} \left(\prod_{i=1}^{n} w_i \right)^{1/2} \times \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} w_i \left(y_i - \mu - \sum_{j=1}^{q} x_{ij} b_j \right)^2 \right]$$

where $y = \{y_i\}$, $x = \{x_{ij}\}$, $b = \{b_j\}$ and $w = \{w_i\}$ for $i = 1, 2,\ldots, n$ and $j = 1, 2,\ldots, q$.

Prior distribution and joint posterior density

For the population mean μ, there is a little prior knowledge about the values. Its prior distribution is represented by assuming $p(\mu) \propto$ constant. Following the Bayesian shrinkage estimation (Wang et al. 2005), the prior knowledge about the each QTL regression effect b_j can be imaged as various evaluations from different researchers. These results of evaluations are considered as $b_j \sim N(0,\ \sigma_j^2)$, $\sigma_j^2 \sim IC[v_b,\ (v_b s_b)^{-1}]$ for $j = 1,\ 2,\ldots,\ q$, where v_b and s_b are prior given as hyper-parameters. A scaled inverse-chi-square distribution with hyper-parameters v_e and s_e will be adopted as prior for σ^2, i.e., $\sigma^2 \sim IC[v_e,\ (v_e s_e)^{-1}]$. As pointed in above genetic model, the prior distribution of w_i, given df, is Gamma$\left(\frac{\mathrm{d}f}{2}, \frac{\mathrm{d}f}{2}\right)$. We adopt a flat prior for df, yielding: $p(\mathrm{d}f) \propto \mathrm{d}f^{-2}$. The position of jth QTL $p(\lambda_j) = \frac{1}{d_j}$, where d_j is the length of the sampling interval where the jth QTL resides.

The joint posterior density of all unknown parameters is then:

$$p(\mu, b, \sigma^2, w, \mathrm{d}f, x, \lambda|y, m) = p(y|\mu, b, \sigma^2, w, x)p(w|\mathrm{d}f)p(\mathrm{d}f)$$
$$p(x|\lambda, m)p(\lambda)p(\mu)p(b|\sigma_b^2)p(\sigma_b^2|v_b, s_b)p(\sigma^2|v_e, s_e) \qquad (2)$$

where m is the known marker information; $\lambda = \{\lambda_j\}$ and $\sigma_b^2 = \{\sigma_j^2\}$ for $j = 1,\ 2,\ldots,\ q$.

Posterior distribution and MCMC sampling

In order to implement Bayesian estimation via the Markov Chain Monte Carlo (MCMC), the marginal posterior distributions of all parameters need to be derived from the above joint posterior density (2) by fixing other parameters.

The fully conditional posterior density of the population mean μ, given all other parameters, can be shown to be a normal distribution with mean $\hat{\mu} = \left(\sum_{i=1}^n w_i\right)^{-1} \sum_{i=1}^n w_i(y_i - \sum_{j=1}^q x_{ij}b_j)$, and variance $\hat{\sigma}_0^2 = \left(\sum_{i=1}^n w_i\right)^{-1} \sigma^2$. Conditionally, on all other parameters, the QTL effects are mutually independent. In particular, the density of the fully conditional posterior distribution of b_j is normal with mean $\hat{b}_j = \left(\sigma^2\sigma_j^{-2} + \sum_{i=1}^n w_i x_{ij}^2\right)^{-1} \sum_{i=1}^n w_i x_{ij}(y_i - \mu - \sum_{k\neq j}^q x_{ik}b_k)$, and variance $\hat{\sigma}_j^2 = \left(\sigma^2\sigma_j^{-2} + \sum_{i=1}^n w_i x_{ij}^2\right)^{-1}\sigma^2$, for $j = 1, 2,\ldots, q$. The fully conditional posterior distribution of the variance σ_j^2 of each QTL effect is a scaled inverse-chi-square with parameters $v_b + 1$ and $(v_b + 1)s_b + b_j^2$. For the residual variance σ^2, the corresponding fully conditional distribution is also a scaled inverse-chi-square with parameters $v_e + n$ and $(v_e + n)s_e + \sum_{i=1}^n w_i(y_i - \mu - \sum_{j=1}^q x_{ij}b_j)^2$. Note that w_i can be interpreted as a "weight" assigned to in the analysis. For each element of w, the density is:

$$p\left(w_i|\mu, b, \sigma_j^2, \sigma^2, \mathrm{d}f, y\right) \propto w_i^{(1+\mathrm{d}f-2)/2}$$
$$\times \exp\left\{-\frac{w_i}{2}\left[\mathrm{d}f + \frac{1}{\sigma^2}\sum_{i=1}^n\left(y_i - \mu - \sum_{j=1}^q x_{ij}b_j\right)^2\right]\right\},$$

which corresponds to a Gamma distribution with parameters $\frac{1+\mathrm{d}f}{2}$ and $2[\mathrm{d}f + \frac{1}{\sigma^2}\sum_{i=1}^q (y_i - \mu - \sum_{j=1}^q x_{ij}b_j)^2]^{-1}$. Using the prior for d$f$ stated above, the fully conditional posterior density of df is:

$$p\left(\mathrm{d}f|\mu, b, \sigma_j^2, \sigma^2, y, w\right) \propto \left[2^{\frac{\mathrm{d}f}{2}}\Gamma\left(\frac{\mathrm{d}f}{2}\right)\right]^{-n} \mathrm{d}f^{\frac{n\mathrm{d}f}{2}-2}$$
$$\times \exp\left[-\frac{\mathrm{d}f}{2}\sum_{i=1}^n (w_i - \ln w_i)\right]$$

The distribution does not have a closed form but a Metropolis–Hastings or rejection sampling step (Ripley 1987) can be embedded in the MCMC scheme to obtain draws for df.

The conditional posterior distribution of the position of a QTL also has no explicit form. Therefore, the general Metropolis–Hastings (Metropolis et al. 1953; Hastings 1970) algorithm is required to sample λ. Since the genotype of QTL (x) depends on the QTL position (λ), we decide to sample $\{\lambda_j, x_j\}$ jointly as a block but proceed with the sampling with one locus at a time. Each locus is sampled from a variable interval (Wang et al. 2005; Zhang and Xu 2005) whose boundaries are the positions of adjoining QTL. The prior distribution of λ_j can be written as

$$p(\lambda_j) = U\left(\lambda_j; \lambda_{j-1}, \lambda_{j+1}\right) = 1/\left(\lambda_{j+1} - \lambda_{j-1}\right),$$

where λ_{j-1} and λ_{j+1} are the positions of the left and the right QTL. Let $\lambda_j^{(t)}$ be the current position of the locus of interest and $x_j^{(t)} = [x_{1j} \cdots x_{nj}]^T$ be the genotype array of all individuals at the locus. We first sampled a new position for the QTL, called the proposed position and denoted by $\lambda_j^* = \lambda_j + \delta$, where δ is sampled from $U(-s,\ s)$ and s is a small positive number (tuning parameter) such as 1 cM. For the new position, we simulate the genotypes for all individuals, denoted by x_j^*. We then use the M–H rule to decide whether λ_j^* should be accepted or not. If λ_j^* is accepted, we update both the position and the genotype using $\lambda_j^{(t+1)} = \lambda_j^*$ and $x_j^{(t+1)} = x_j^*$. Otherwise, the old values of λ_j and x_j are carried over so that $\lambda_j^{(t+1)} = \lambda_j^{(t)}$ and $x_j^{(t+1)} = x_j^{(t)}$. Detailed formula of the M–H acceptance rule can be found in Wang et al. (2005) and Zhang and Xu (2005).

Genotypes of missing markers were generated randomly in each iteration on the basis of the probability inferred jointly from the nearest non-missing flanking markers and the phenotype. The probability from the markers is treated as the prior probability. After incorporation of the marker (QTL) effects through the phenotype, the probability becomes the posterior probability, which is used to generate the missing marker genotype. See Wang et al. (2005)

for details. In summary, the MCMC process is described in the following steps:

(1) Initialize all variables with some legal values or values sampled from their prior distributions.
(2) Update the population mean μ.
(3) Update the genetic effects b_j ($j = 1, 2,\ldots, q$) for each QTL.
(4) Update the variance σ_j^2 ($j = 1, 2,\ldots, q$)of each QTL effect.
(5) Update the residual variance σ^2.
(6) Update the degree of freedom d*f*.
(7) Update the "weight" w_i ($i = 1, 2,\ldots, n$).
(8) Update the QTL position λ_j ($j = 1, 2,\ldots, q$) and the genotypes for each QTL.
(9) Impute the genotypes of missing markers.
(10) Repeat steps (2)–(9) until the Markov chain reaches a desirable length.

Post MCMC analysis

The product of MCMC sampling is a realized sample of all unknown variables drawn from the joint posterior distribution. In practice those results should be interpreted in a different way. In conventional Bayesian mapping (e.g. Sillanpää and Arjas 1998, 1999; Yi and Xu 2000; Wang et al. 2005), the marginal posterior distribution of QTL position can be depicted via plotting the number of hits by QTL in a short segment (say a 1 cM segment), called a bin, against the genome position where the bin is located. The curve is called the QTL intensity profile.

In addition to the QTL intensity profile, there is an alternative profile to present the result of MCMC, which is the U test statistic profile denoted by $U = \frac{b(\lambda)}{\sqrt{V(\lambda)}}$ (Yang and Xu 2007), where $b(\lambda)$ is the average effect of QTL for the bin located at position λ and $V(\lambda)$ is the corresponding sample variance for the QTL effect at position λ. U follows a standard normal distribution. The critical value is 1.96 for declaring statistical significance at position λ at the significant level of 0.05. Hereinafter, the U statistics is used to claim the presence of QTL.

Simulations

We simulate 61 equally spaced codominant markers on a single large chromosome of length 600 cM for a backcross population with sample size of 150 and 300. Ten QTL are put along the genome. The total genetic variance contributed by all 10 QTL was 45.06, where the proportion of phenotypic variance contributed by an individual QTL ranged from 0.40 to 34.0%. The population mean and the environmental (residual) variance were set at $\mu = 5.0$ and $\sigma^2 = 2.0$.

In all Bayesian estimation, the initial number of QTL $q = 15$, that is, each evenly covers 40 cM of the genome, which is empirically determined according to Bayesian shrinkage mapping for single trait (Wang et al. 2005). The actual values for the hyper parameters take $v_b = 0$, $v_e = 0$, $s_b = 1$ and $s_e = 1$. The initial values of all variables are sampled from their prior distributions. The MCMC is run for 6,000 cycles as burn-in period (deleted) and then for additional 60,000 cycles after the burn-in. Note that here the length of the burn-in is judged by visually inspecting the plots of some samples across rounds and is set to make enough cycles for ensuring the MCMC convergence. The chain is then thinned to reduce serial correlation by saving one observation in every 40 cycles. The posterior sample contains 1,500 (60,000/40 = 1,500) observations for the

Table 1 Statistical power of QTL detection (%) obtained with Robust method and Traditional method

Sample size	d*f*	Method	QTL no.									
			1	2	3	4	5	6	7	8	9	10
150	1	Robust	100	100	70	80	15	70	35	100	30	80
		Traditional	100	100	10	35	0	15	5	100	0	25
	5	Robust	100	100	95	100	30	95	100	95	40	100
		Traditional	100	100	75	45	0	45	50	100	15	75
	15	Robust	100	100	100	100	45	95	100	100	55	100
		Traditional	100	100	90	100	0	95	95	100	35	100
300	1	Robust	100	100	75	100	25	75	60	100	30	85
		Traditional	100	100	15	50	10	25	5.0	100	15	35
	5	Robust	100	100	100	100	35	100	95	100	45	100
		Traditional	100	100	80	60	20	60	75	100	20	100
	15	Robust	100	100	100	100	45	100	100	100	60	100
		Traditional	100	100	90	100	40	100	100	100	40	100

d*f* is the degree of freedom given in simulation, which in other Table is the same

Table 2 Mean estimates and standard deviations (in parentheses) of QTL positions detected with Robust method and Traditional method

Sample size	d*f*	Method	QTL no.									
			1	2	3	4	5	6	7	8	9	10
		True position	23	56	148	193	267	332	390	476	522	574
150	1	Robust	23.0 (3.4)	52.3 (3.5)	143.5 (3.8)	193.2 (3.6)	267.3 (5.0)	332.0 (6.7)	394.5 (2.5)	477.2 (3.4)	524.7 (4.1)	572.3 (5.3)
		Traditional	23.9 (3.6)	55.9 (5.4)	149.4 (4.5)	192.0 (5.4)	–	331.0 (8.1)	392.5 (7.2)	476.8 (5.2)	–	577.3 (7.7)
	5	Robust	22.7 (2.5)	54.7 (2.4)	146.5 (2.8)	193.2 (2.1)	266.0 (4.9)	329.0 (2.0)	390.5 (2.7)	475.2 (1.8)	520.4 (5.3)	573.8 (3.0)
		Traditional	23.3 (3.8)	55.0 (4.5)	146.1 (3.5)	194.5 (2.4)	–	328.3 (2.3)	390.6 (4.2)	476.3 (2.8)	528.7 (6.8)	573.8 (5.4)
	15	Robust	22.4 (2.0)	52.6 (1.1)	144.3 (2.9)	193.8 (2.1)	267.3 (4.3)	329.7 (2.2)	390.4 (1.7)	474.4 (1.5)	524.6 (3.6)	574.1 (6.7)
		Traditional	23.0 (1.8)	55.3 (1.5)	144.8 (3.9)	193.7 (4.0)	–	328.4 (4.8)	391.1 (1.5)	475.3 (1.6)	522.0 (4.0)	575.9 (4.3)
300	1	Robust	22.7 (1.3)	55.1 (1.5)	149.5 (2.9)	193.1 (2.9)	267.1 (3.6)	331.7 (4.3)	391.3 (2.6)	477.3 (3.9)	520.0 (3.2)	576.2 (3.2)
		Traditional	25.3 (3.4)	54.4 (2.8)	154.7 (4.2)	192.4 (3.7)	269.0 (4.2)	332.0 (5.8)	400.0 (7.0)	478.5 (4.6)	517.3 (4.2)	575.1 (7.6)
	5	Robust	24.1 (1.4)	55.6 (1.4)	148.0 (2.7)	193.0 (1.4)	266.1 (3.2)	329.8 (2.3)	390.9 (1.8)	475.8 (1.1)	522.8 (3.8)	575.3 (3.3)
		Traditional	23.1 (1.8)	53.7 (1.8)	146.2 (4.1)	193.3 (3.3)	265.6 (3.6)	329.6 (3.7)	390.9 (2.1)	474.8 (1.4)	521.9 (4.3)	573.4 (5.3)
	15	Robust	22.7 (1.8)	54.4 (1.2)	148.7 (2.2)	192.7 (1.2)	266.4 (3.1)	332.5 (3.4)	390.8 (1.6)	475.3 (1.3)	523.3 (3.3)	574.5 (1.3)
		Traditional	22.3 (1.9)	53.9 (2.6)	145.7 (4.0)	193.1 (3.1)	270.0 (3.2)	330.7 (5.3)	389.6 (3.7)	475.3 (5.1)	522.5 (6.1)	574.6 (6.3)

Table 3 Mean estimates and standard deviations (in parentheses) of QTL effects obtained with Robust method and Traditional method

Sample size	d*f*	Method	QTL no.									
			1	2	3	4	5	6	7	8	9	10
		True effect	3.00	4.00	1.44	2.20	−0.44	1.40	−1.30	2.50	0.70	−1.60
150	1	Robust	3.06 (0.28)	4.38 (0.99)	1.55 (0.02)	1.85 (0.82)	1.04 (0.24)	1.63 (0.98)	−1.63 (0.69)	2.85 (0.43)	0.83 (0.90)	−1.34 (0.49)
		Traditional	2.15 (0.51)	3.11 (0.93)	1.28 (0.28)	1.64 (0.90)	–	1.76 (1.22)	−1.10 (0.72)	2.13 (0.59)	–	−1.33 (0.69)
	5	Robust	3.03 (0.17)	3.61 (0.66)	1.32 (0.17)	1.74 (0.20)	−0.66 (0.05)	1.31 (0.16)	−1.23 (0.13)	2.35 (0.22)	0.76 (0.10)	−1.50 (0.13)
		Traditional	2.78 (0.29)	3.66 (0.64)	1.29 (0.26)	1.55 (0.25)	–	1.15 (0.35)	−1.17 (0.16)	2.21 (0.27)	0.86 (0.23)	−1.31 (0.15)
	15	Robust	3.03 (0.16)	3.83 (0.18)	1.46 (0.16)	2.00 (0.23)	−0.36 (0.15)	1.45 (0.13)	−1.21 (0.12)	2.39 (0.16)	0.69 (0.09)	−1.41 (0.17)
		Traditional	2.54 (0.37)	3.55 (0.29)	1.33 (0.18)	1.9 2 (0.31)	–	1.06 (0.20)	−1.10 (0.21)	2.07 (0.17)	0.79 (0.21)	−1.24 (0.25)
300	1	Robust	2.28 (0.40)	3.66 (0.35)	1.15 (0.18)	1.97 (0.31)	−0.84 (0.34)	1.09 (0.11)	−1.01 (0.15)	2.13 (0.29)	0.94 (0.01)	−1.38 (0.18)
		Traditional	1.41 (0.43)	3.45 (0.84)	1.88 (0.62)	2.76 (0.69)	−1.06 (0.64)	1.92 (0.26)	−1.96 (0.32)	3.19 (0.41)	0.50 (0.83)	−1.38 (0.63)
	5	Robust	3.03 (0.36)	3.81 (0.19)	1.39 (0.18)	2.07 (0.14)	−0.56 (0.19)	1.32 (0.10)	−1.29 (0.09)	2.37 (0.17)	0.72 (0.08)	−1.52 (0.15)
		Traditional	2.39 (0.07)	3.87 (0.10)	1.26 (0.31)	1.92 (0.15)	−0.50 (0.05)	1.11 (0.10)	−1.05 (0.12)	2.20 (0.13)	0.71 (0.19)	−1.48 (0.13)
	15	Robust	2.99 (0.36)	3.93 (0.23)	1.40 (0.16)	1.96 (0.19)	−0.34 (0.20)	1.53 (0.13)	−1.25 (0.07)	2.41 (0.15)	0.70 (0.06)	−1.57 (0.07)
		Traditional	2.48 (0.21)	3.93 (0.25)	1.28 (0.22)	1.83 (0.42)	−0.51 (0.38)	1.27 (0.11)	−1.11 (0.10)	2.25 (0.14)	0.69 (0.19)	−1.37 (0.11)

Table 4 Statistical power of QTL detection (%) for log normal and normal data detected with Robust method and Traditional method

Sample size	Data	Method	QTL no.									
			1	2	3	4	5	6	7	8	9	10
150	Log Normal	Robust	100	100	65	70	10	65	35	85	25	65
		Traditional	65	75	5	25	0	10	0	45	0	25
	Normal	Robust	100	100	100	100	50	100	100	100	65	100
		Traditional	100	100	100	100	10	90	100	100	40	95
300	Log Normal	Robust	100	100	75	80	15	70	55	90	30	75
		Traditional	80	85	10	40	5	20	5	75	5	30
	Normal	Robust	100	100	100	100	55	100	100	100	85	100
		Traditional	100	100	100	100	35	100	100	100	55	100

post-MCMC analysis. The simulation experiment is replicated 40 times for statistical power evaluation. QTL parameters are calculated by averaging posterior estimates from those simulations in which significant QTL is detected.

Real data

A 162 F_{10} recombinant inbred lines (RILs) derived from the hybrids of Dasanbyeo (a Korean tongil type rice) × TR22183 (a Chinese japonica variety), had been designed for mapping QTL for traits associated with physics–chemical characters and quality in rice. On the basis of the population, the framework linkage map of 1467.5 cM containing 208 SSR and STS markers has been constructed. This map consists of the 17 largest linkage groups (LG) for each parental map.

Results

Simulated data

We conducted three simulation experiments to demonstrate the flexibility of the Robust Bayesian mapping proposed here. In the first simulation experiment, we sampled residual error from t distribution with degree of freedom $\mathrm{d}f = 1$, 5 and 15, respectively, generating phenotype values according to model (1). Those data are analyzed by adopting Robust Bayesian mapping (Robust method) and traditional Bayesian mapping as if residual were normally distributed (Traditional method), respectively. The statistical powers of QTL detection with both methods are given in Table 1. In general, Robust method can detect more QTL than Traditional method if the residual error subjects to heavy-tailed t distribution, especially with lower degree of freedom. Both the methods are able to accurately estimate positions and effects of QTL detected (see Tables 2 and 3). Estimates of degree of freedom given 1, 5, and 15 are 1.51 ± 0.55, 7.01 ± 2.12 and 17.79 ± 5.46 for sample 150, and 1.21 ± 0.35, 6.12 ± 2.00 and 16.13 ± 4.41 for sample 300, respectively. As seen, the Robust method can better fit the non-normal data by accurately estimating the degree of freedom in Student-t distribution. In the second simulation experiment, we simulated residual errors with log-normal distribution. Mapping results from Robust and Traditional method were listed in Tables 4, 5, and 6, respectively. Apparently, Robust method is superior to Traditional method in the terms of either the statistical powers of QTL detection or estimation of QTL parameters, although both methods perform a little lower statistical powers of QTL detection and lower estimation accuracy of QTL parameters for log-normal simulated data than t distribution data.

In the final simulation experiment, we demonstrate that applying the Robust Bayesian analysis for data already normally distributed will not harm the result. We generate normally distributed phenotypes by sampling residuals from normal distribution and analyzed them with both the Robust method and Traditional method. The simulation results shows that the Robust method does not harm the result if the data are already normally distributed (see Tables 4, 5 and 6). Student-t distribution with 30 degrees of freedom usually has been treated as the normal distribution. The degree of freedom is 20.6, estimated from simulated normal data with Robust method. The possible reason is that the estimation of the degree of freedom is closely related to the sample size (Jamrozik et al. 2004). When we additionally simulate a backcross population with 1,000 individuals, the estimate of the degree of freedom is closed to 50 (result not shown).

Real data

We analyzed the data with both the robust method and traditional method procedure. Using Bayesian analysis, we assumed a total of 70 QTL across the whole genome. The initial value of each unknown parameter was taken same as

Table 5 Mean estimates and standard deviations (in parentheses) of QTL positions for log normal and normal data detected with Robust method and Traditional method

Sample size	Data	Method	QTL no.									
			1	2	3	4	5	6	7	8	9	10
		True position	23	56	148	193	267	332	390	476	522	574
150	Log Normal	Robust	24.1 (3.4)	56.8 (4.4)	149.7 (5.1)	191.3 (5.6)	268.5 (7.1)	334.1 (8.9)	391.6 (7.6)	474.2 (5.5)	526.2 (7.9)	571.7 (6.4)
		Traditional	23.9 (5.9)	57.6 (6.7)	150.2 (7.5)	190.1 (7.2)	–	334.3 (10.4)	–	478.9 (7.3)	–	578.9 (10.8)
	Normal	Robust	23.6 (2.9)	55.9 (1.4)	150.3 (3.1)	189.3 (1.6)	267.3 (4.1)	333.0 (3.9)	390.1 (2.6)	471.1 (1.0)	518.0 (3.2)	570.4 (2.4)
		Traditional	22.6 (3.0)	55.6 (1.7)	149.1 (3.5)	189.1 (1.2)	267.0 (4.2)	333.9 (2.4)	390.8 (1.2)	471.6 (1.2)	521.5 (2.6)	570.0 (1.8)
300	Log Normal	Robust	22.4 (3.0)	57.1 (3.1)	149. 7(5.2)	191.5 (5.4)	269.8 (7.9)	334.8 (8.1)	392.8 (6.9)	478.4 (4.9)	526.1 (6.9)	576.1 (5.5)
		Traditional	23.7 (5.7)	56.1 (4.4)	148.4 (6.5)	191.7 (7.6)	269.0 (8.2)	335.6 (9.9)	393.9 (6.8)	470.1 (6.3)	528.4 (9.9)	577.6 (9.8)
	Normal	Robust	21.9 (1.0)	55.5 (1.1)	149.0 (2.2)	190.8 (1.4)	266.8 (2.9)	332.6 (1.1)	391.2 (1.9)	474.5 (0.9)	520.9 (1.9)	572.0 (1.5)
		Traditional	20.7 (2.7)	55.1 (0.4)	146.1 (1.5)	189.7 (1.6)	265.0 (1.2)	332.4 (1.9)	392.9 (2.2)	473.4 (1.3)	522.4 (3.9)	577.4 (1.8)

Table 6 Mean estimates and standard deviations (in parentheses) of QTL effects for log normal and normal data detected with Robust method and Traditional method

Sample size	Data	Method	QTL no.									
			1	2	3	4	5	6	7	8	9	10
		True effect	3.00	4.00	1.44	2.20	−0.44	1.40	−1.30	2.50	0.70	−1.60
150	Log Normal	Robust	3.09 (0.31)	4.06 (0.25)	1.32 (0.26)	2.11 (0.23)	−0.49 (0.16)	1.33 (0.19)	−1.47 (0.20)	2.42 (0.31)	0.64 (0.17)	−1.72 (0.28)
		Traditional	2.91 (0.53)	3.94 (0.56)	1.30 (0.41)	1.92 (0.42)	–	1.64 (0.38)	–	2.66 (0.42)	–	−1.81 (0.32)
	Normal	Robust	2.94 (0.45)	3.95 (0.19)	1.30 (0.17)	2.01 (0.20)	−0.46 (0.05)	1.22 (0.10)	−1.45 (0.12)	2.45 (0.23)	0.69 (0.07)	−1.69 (0.18)
		Traditional	2.93 (0.19)	3.90 (0.20)	1.34 (0.13)	1.97 (0.12)	−0.67 (0.25)	0.94 (0.18)	−1.27 (0.14)	2.33 (0.20)	0.78 (0.14)	−1.51 (0.22)
300	Log Normal	Robust	3.07 (0.23)	4.04 (0.22)	1.48 (0.20)	2.15 (0.21)	−0.48 (0.22)	1.35 (0.18)	−1.39 (0.19)	2.54 (0.26)	0.79 (0.16)	−1.65 (0.21)
		Traditional	2.74 (0.46)	3.96 (0.51)	1.39 (0.36)	1.94 (0.35)	−0.57 (0.34)	1.72 (0.40)	−1.60 (0.28)	2.61 (0.38)	0.79 (0.28)	−1.74 (0.27)
	Normal	Robust	2.95 (0.15)	3.98 (0.15)	1.41 (0.10)	2.09 (0.11)	−0.34 (0.20)	1.26 (0.10)	−1.21 (0.14)	2.57 (0.16)	0.76 (0.08)	−1.62 (0.11)
		Traditional	2.60 (0.12)	3.98 (0.10)	1.35 (0.16)	1.98 (0.15)	−0.57 (0.04)	1.16 (0.13)	−1.01 (0.08)	2.44 (0.14)	0.60 (0.08)	−1.48 (0.07)

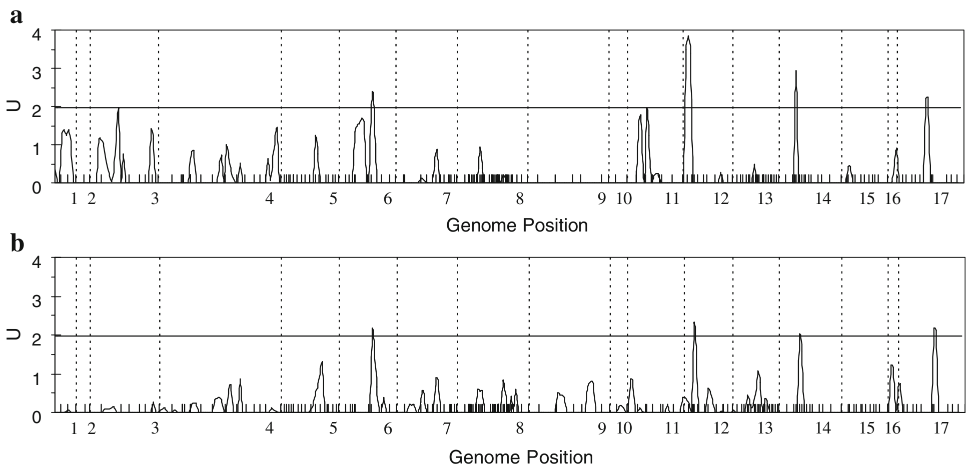

Fig. 1 The *U* test-statistic profiles for QTL mapping from the rice data analysis: **a** the one generated by the Robust method; **b** the one drawn from the traditional mapping analysis. The *horizontal reference lines* in the both plot are the critical value of 1.96 for the significance test. The genome consists of 17 linkage groups that are separated by the *vertical dotted lines*. The 17 linkage groups are drawn in scales proportional to their lengths. Positions of the markers are indicated by the *ticks* on the horizontal axis

in the simulation study. The mapping results from 13 of 21 traits of interest support the robust method. In the following, we take breakdown viscosity (BDV) as an example to compare the mapping results from two kinds of Bayesian mapping methods. BDV, that is used to describe rice paste profile characteristic, is an important parameter for the cooking and eating quality (Bao and Xia 1999).

The *U* statistic profile for the Robust method and Traditional method procedure method are depicted in Fig. 1. Apparently, Robust method is not only able to detect all QTL detected by the Traditional method procedure, but also it detected two more QTL than Traditional method procedure. The comparative results of the position and genetic effect of QTL detected from both methods were exhibited in Table 7.

Table 7 Estimated QTL positions and effects obtained from robust method and traditional method for BDV in Rice

QTL No.	Robust		Traditional	
	LG-position	Effect	LG-position	Effect
1	3–45.6	−0.027 (0.014)	–	–
2	6–52.6	−0.022 (0.011)	6–52.6	−0.017 (0.001)
3	11–32.0	−0.015 (0.007)	–	–
4	12–5.5	0.031 (0.008)	12–13.0	0.027 (0.012)
5	14–28.7	0.026 (0.009)	14–32.5	0.029 (0.012)
6	17–48.7	−0.041 (0.009)	17–58.7	−0.058 (0.023)

Discussion

On the basis of the Bayesian shrinkage mapping, we develop a robust mapping strategy for analyzing continuous non-normal quantitative traits, by replacing the normal distribution for residuals in multiple QTL model with a Student-*t* distribution. Compared with Bayesian shrinkage mapping for normal trait, the robust mapping strategy additionally has sample "weight" w_i with a Gibbs sampler and the degree of freedom d*f* with a Metropolis–Hastings algorithm in the MCMC process. However, it does not significantly increase computing time on solving QTL parameters. The flexibility of the Robust Bayesian mapping for either non-normal or normal data demonstrated by the simulations can compensate for the expense of two additional sampling. Hence, it is recommendable to apply the robust mapping strategy to the practice of mapping QTL.

Rohr and Hoeschele (2000) first implemented a Robust Bayesian method to mapping QTL. Their study is different from ours in that: (1) their mapping analysis is aimed at outbred population whereas ours is at line cross; (2) their proposed method was based on single QTL model whereas ours is multiple QTL model and (3) they used skewed Student-*t* distributions to describe residual phenotypes in the analysis whereas we adopted a student-*t* distribution. In single QTL model, it seems to be reasonable to assume that residuals follow skewed Student-*t* distributions, because the "skewness" may absorb the effects of other QTL on phenotypes. However, no "skewness" is possible necessary for multiple QTL model.

When the phenotypes deviate from normality, Student-*t* distribution is capable of accommodating much more abnormal residuals by thick tails, improving the robustness inference of QTL parameters. Except for the most commonly used Student-*t* distribution, there may be also many thick-tailed distributions available for Robust Bayesian mapping of QTL, such as a class of robust distributions, known as normal/independent (Andrews and Mallows 1974; Lange and Sinsheimer 1993). These distributions have been used in multivariate linear regression models (Liu 1996) and linear mixed model (Stranden and Gianola 1999; Rosa et al. 2003, 2004, within a Bayesian framework. It will be easy to apply those distributions to robust mapping QTL because multiple QTL model is also linear.

In addition, the Robust Bayesian mapping strategy proposed here can be further extended to more complex experimental population, such as multiple line crosses and outbred population and more complex QTL models including epistatic effects between QTLs.

Acknowledgments The research was supported by the Chinese National Natural Science Foundation Grant 30471236 to RY.

References

Andrews DF, Mallows CL (1974) Scale mixtures of normal distributions. J Roy Stat Soc Ser B 36:99–102

Bao JS, Xia YW (1999) Genetic control of the paste viscosity characteristics in indica rice (*Oryza sativa* L.). Theor Appl Genet 98:1120–1124

Coppieters W, Kvasz A, Farnir F, Arranz JJ, Grisart B, Mackinnon M, Georges M (1998) A rank-based nonparametric method for mapping quantitative trait loci in outbred half-sib pedigrees: application to milk production in a granddaughter design. Genetics 149:1547–1555

Dempster AP, Laird NM, Rubin DB (1980) Iteratively reweighted least squares for linear regression when errors are normal/independent distributed. In: Krishnaiah PR (ed) Multivariate analysis. North-Holland, Amsterdam

Diao G, Lin DY (2005) A powerful and robust method for mapping quantitative trait loci in general pedigrees. Am J Hum Genet 77:97–111

Diao G, Lin DY, Zou F (2004) Mapping quantitative trait loci with censored observations. Genetics 168:1689–1698

Elsen JM, Mangin B, Goffinet B, Boichard D, Le RP (1999) Alternative models for QTL detection in livestock I. General introduction. Genet Sel Evol 31:213–224

Feenstra B, Skovgaard IM (2004) A quantitative trait locus mixture model that avoids spurious LOD score peaks. Genetics 167:959–965

Fernandez C, Steel M (1998) On Bayesian modeling of fat tails and skewness. J Am Statist Assoc 93:359–371

Goffinet B, Le RP, Boichard D, Elsen JM, Mangin B (1999) Alternative models for QTL detection in livestock III. Heteroskedastic model and models corresponding to several distributions of the QTL effect. Genet Sel Evol 31:341–350

Hackett CA (1997) Model diagnostics for fitting QTL models to trait and marker data by interval mapping. Heredity 79:319–328

Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97–109

Jamrozik J, Stranden I, Schaeffer LR (2004) Random regression test-day models with residuals following a Student's-*t* distribution. J Dairy Sci 87:699–705

Jansen RC (1992) A general mixture model for mapping quantitative trait loci by using molecular markers. Theor Appl Genet 85:252–260

Kruglyak L, Lander ES (1995) A nonparametric approach for mapping quantitative trait loci. Genetics 139:1421–1428

Lange K, Sinsheimer JS (1993) Normal/independent distributions and their applications in robust regression. J Am Stat Assoc 2:175–198

Lange KL, Little RJA, Taylor JMG (1989) Robust statistical modelling using the *t*-distribution. J Am Stat Assoc 84:881–896

Liu C (1996) Robust Bayesian multivariate linear regression with incomplete data. J Am Stat Assoc 435:1219–1227

Mangin B, Goffinet B, Le RP, Boichard D, Elsen JM (1999) Alternative models for QTL detection in livestock II. Likelihood approximations and sire marker genotype estimation. Genet Sel Evol 31:225–237

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equations of state calculations by fast computing machines. J Chem Phys 21:1087–1091

Pinheiro JC, Liu CH, Wu YN (2001) Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate *t* distribution. J Comput Graph Stat 10:249–276

Rebaï A (1997) Comparison of methods for regression interval mapping in QTL analysis with non-normal traits. Genet Res 69:69–74

Ripley B (1987) Stochastic simulation. Wiley, New York

Rogers WH, Tukey JW (1972) Understanding some long-tailed distributions. Stat Neerl 26:211–226

Rohr PV, Hoeschele I (2002) Bayesian QTL mapping using skewed Student-*t* distributions. Genet Sel Evol 34:1–21

Rosa GJM, Gianola D, Padovani CR (2004) Bayesian longitudinal data analysis with mixed models and thick-tailed distributions using MCMC. J Appl Stat 7:855–873

Rosa GJM, Padovani CR, Gianola D (2003) Robust linear mixed models with normal/independent distributions and Bayesian MCMC implementation. Biom J 5:573–590

Sillanpää MJ, Arjas E (1998) Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. Genetics 148:1373–1388

Sillanpää MJ, Arjas E (1999) Bayesian mapping of multiple quantitative trait loci from incomplete outbred offspring data. Genetics 151:1605–1619

Sokal RR, Rohlf FJ (1995) Biometry: the principles and practice of statistics in biological research. W.H. Freeman, New York

Stranden I, Gianola D (1999) Mixed effects linear models with *t*-distributions for quantitative genetic analysis: a Bayesian approach. Genet Sel Evol 31:25–42

Symons RC, Daly MJ, Fridlyand J, Speed TP, Cook WD, Gerondakis S, Harris AW, Foote SJ (2002) Multiple genetic loci modify susceptibility to plasmacytoma-related morbidity in Eμ-v-abl transgenic mice. Proc Natl Acad Sci 99:11299–11304

Wang H, Zhang YM, Li X, Masinde GL, Mohan S, Baylink DJ, Xu S (2005) Bayesian shrinkage estimation of quantitative trait loci parameters. Genetics 170:465–480

Yang R, Xu S (2007) Bayesian shrinkage analysis of quantitative trait loci for dynamic traits. Genetics 176:1169–1185

Yang R, Yi N, Xu S (2006) Box–Cox transformation for QTL mapping. Genetica 128:133–143

Yi N, Xu S (2000) Bayesian mapping of quantitative trait loci for complex binary traits. Genetics 155:1391–1403

Zhang YM, Xu S (2005) Advanced statistical methods for detecting multiple quantitative trait loci. Recent Res Devel Genet Breed 2:1–23

Zou F, Yandell BS, Fine JP (2003) Rank-based statistical methodologies for quantitative trait locus mapping. Genetics 165:1599–1605