



## Mapping of quantitative trait loci using the skew-normal distribution\*

FERNANDES Elisabete<sup>†‡1,2,4</sup>, PACHECO António<sup>3</sup>, PENHA-GONÇALVES Carlos<sup>4</sup>

<sup>(1)</sup>Centre for Mathematics and Its Applications, IST-Technical University of Lisbon, 1049-001 Lisboa, Portugal)

<sup>(2)</sup>Department of Statistics and Operational Research, Faculty of Sciences, University of Lisbon, 1749-016 Lisboa, Portugal)

<sup>(3)</sup>Department of Mathematics and Centre for Mathematics and Its Applications, IST-Technical University of Lisbon, 1049-001 Lisboa, Portugal)

<sup>(4)</sup>Gulbenkian Institute of Science, P-2781-901 Oeiras, Portugal)

<sup>†</sup>E-mail: ebfernandes@fc.ul.pt

Received Sept. 19, 2007; revision accepted Sept. 27, 2007

**Abstract:** In standard interval mapping (IM) of quantitative trait loci (QTL), the QTL effect is described by a normal mixture model. When this assumption of normality is violated, the most commonly adopted strategy is to use the previous model after data transformation. However, an appropriate transformation may not exist or may be difficult to find. Also this approach can raise interpretation issues. An interesting alternative is to consider a skew-normal mixture model in standard IM, and the resulting method is here denoted as skew-normal IM. This flexible model that includes the usual symmetric normal distribution as a special case is important, allowing continuous variation from normality to non-normality. In this paper we briefly introduce the main peculiarities of the skew-normal distribution. The maximum likelihood estimates of parameters of the skew-normal distribution are obtained by the expectation-maximization (EM) algorithm. The proposed model is illustrated with real data from an intercross experiment that shows a significant departure from the normality assumption. The performance of the skew-normal IM is assessed via stochastic simulation. The results indicate that the skew-normal IM has higher power for QTL detection and better precision of QTL location as compared to standard IM and nonparametric IM.

**Key words:** Interval mapping (IM), Quantitative trait loci (QTL), Skew-normal distribution, Expectation-maximization (EM) algorithm

doi:10.1631/jzus.2007.B0792

Document code: A

CLC number: Q78; TP31

### INTRODUCTION

Mapping genetic loci affecting quantitative traits (called quantitative trait loci or QTL) in plants and animals is an important issue with a broad range of applications. Lynch and Walsh (1998) provided a review of statistical methods for detecting and locating QTL in experimental crosses. The interval mapping (IM), here denoted by standard IM, pioneered by Lander and Botstein (1989) and generalized to multiple loci by Kao *et al.* (1999), was the first approach

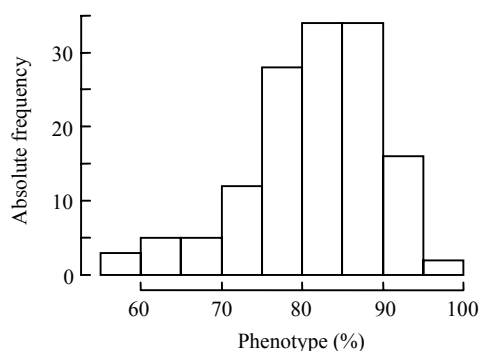
based on the fact that a QTL is located in an interval flanked by two genetic markers with observed genotypes and known positions. This approach, as most of the QTL mapping methods, makes use of the assumption that the quantitative phenotype follows a normal distribution with equal variance in both parental strains (Kruglyak and Lander, 1995).

Many phenotypes of interest, however, are not normally distributed. An example is the data on the percentage of CD19+/CD69+ B-cells that were stimulated with bacterial lipopolysaccharide (LPS) in 139 F<sub>2</sub> mice from an intercross experiment between the C57BL/6 and BALB/c mouse strains (Rodo *et al.*, 2006). The histogram in Fig.1 shows that this phenotype follows a highly skewed distribution.

Therefore, the assumption of normality many times is unreal and can occult important characteristic

<sup>‡</sup> Corresponding author

\* Project supported in part by Foundation for Science and Technology (FCT) (No. SFRD/BD/5987/2001), and the Operational Program Science, Technology, and Innovation of the FCT, co-financed by the European Regional Development Fund (ERDF)



**Fig.1 Histogram of phenotype percentage of CD19+/CD69+ B-cells that were stimulated with bacterial lipopolysaccharide (LPS) in 139 intercross mice**

of the model. A major reason for this assumption is certainly the unrivalled mathematical tractability of the normal distribution (Azzalini and Capitanio, 1999). A problem is that if this assumption is violated, then false detection of a major locus effect may occur (Morton, 1984).

The most commonly adopted method to achieve normality involves transforming the data using, for example, the Box-Cox transformation (Draper and Smith, 1998). Although this method may give reasonable empirical results, it should be avoided if a more suitable theoretical model can be found (Azzalini and Capitanio, 1999).

Moreover, an appropriate transformation may not exist or may be difficult to find. Even if a good transformation is found, the effect of outliers may be still too great (Kruglyak and Lander, 1995). Also this approach changes the original unit of the data, which implies a careful interpretation on the results, and the transformation involves an extra parametric assumption (Zou *et al.*, 2003). Therefore, it is more realistic and helpful to analyze the data on the original scale.

It is true that we can also use the nonparametric interval mapping based on the Kruskal-Wallis test statistic (Broman, 2003), here called as nonparametric IM, when the data are not normal. However, power to detect genes is lost by using nonparametric IM, particularly if the sample size is small and there is much missing marker data.

As result recent proposals have been made based on replacing the assumption of normality by a weaker assumption that the quantitative variable has a "smooth" density that may be skewed (Dalla Valle, 2004). In particular, the skew-normal model has been

used to extend the usual symmetric normal model. Advantages of using such model include estimation efficiency, as well as easiness of interpretation (Arellano-Valle *et al.*, 2005).

In this paper, we describe an interesting alternative approach for QTL mapping, here denoted as skew-normal interval mapping, or skew-normal IM. This method, which is similar to standard IM, assumes that the quantitative phenotype follows a skew-normal distribution for each QTL genotype.

Therefore, in section METHODS, after a brief summary of the main probabilistic properties of the skew-normal distribution, we define the skew-normal IM. One of the inference problems associated to this model is the singularity of the Fisher information matrix when skewness is absent, as is the case of the normal distribution. Thus, the maximum likelihood estimators (MLEs) are obtained by an expectation-maximization (EM) algorithm (Dempster *et al.*, 1977) under Azzalini (1985)'s centred parametrization which overcomes this inference problem.

In section RESULTS, we illustrate the skew-normal IM with one intercross data set for which the quantitative phenotype follows a highly skewed distribution. QTL mapping methods (skew-normal IM, standard IM, and nonparametric IM) are compared with respect to important performance criteria, such as power and efficiency. The performance of the proposed procedures is assessed via computer simulation. In section DISCUSSION, we discuss the practical utility of the proposed method.

## METHODS

### Skew-normal distribution

The skew-normal distribution was first named by Azzalini (1985) but its appearance in the literature dates back to Roberts (1966). This distribution is mathematically tractable and able to reflect varying degrees of skewness, with the normal distribution as its special case (Pewsey, 2000).

**Definition** A random variable  $Y$  follows a skew-normal distribution with location parameter  $\beta \in \mathbb{R}$ , scale parameter  $\omega > 0$ , and skewness parameter  $\lambda \in \mathbb{R}$ , if its density function is given by

$$f(y; \beta, \omega, \lambda) = \frac{2}{\omega} \phi\left(\frac{y - \beta}{\omega}\right) \Phi\left\{\lambda\left(\frac{y - \beta}{\omega}\right)\right\}, \quad (1)$$

where  $y \in \mathbb{R}$ , and  $\phi(\cdot)$  and  $\Phi(\cdot)$  denote the standard normal density and distribution functions of the univariate normal distribution, respectively.

Following Azzalini and Capitanio (1999), we use the notation  $Y \sim SN(\beta, \omega, \lambda)$  to denote this distribution for the direct parameterization presented. When  $\beta=0$  and  $\omega=1$ , it is reduced to the standard skew-normal distribution denoted by  $Z \sim SN(\lambda)$  (Arellano-Valle et al., 2005). Note that if  $Z \sim SN(\lambda)$  and  $Y = \beta + \omega Z$  then  $Y \sim SN(\beta, \omega, \lambda)$ . If  $\lambda=0$ , then the density of  $Z \sim SN(\lambda)$  is equivalent to the density of  $X \sim N(0, 1)$  and it becomes more skewed to the right as  $\lambda$  goes to  $+\infty$  or skewed to the left when  $\lambda$  goes to  $-\infty$ . When  $\lambda = +\infty$ , it is the standard positive half-normal distribution, i.e.,  $Z = |X|$ , where  $X \sim N(0, 1)$  and  $Z \sim HN(0, 1)$ ; when  $\lambda = -\infty$ , it is the standard negative half-normal distribution, i.e.,  $Z = -|X|$  (Pewsey, 2000).

The odd moments are easily derived from the moment-generating function given by Azzalini (1985). In particular, the mean value, the variance, and the coefficient of skewness of a random variable  $Y \sim SN(\beta, \omega, \lambda)$  are given by

$$\begin{aligned} E[Y] &= \beta + b\omega\delta, & Var[Y] &= (1 - b^2\delta^2)\omega^2, \\ \gamma &= (2b^2 - 1)b\delta^3(1 - b^2\delta^2)^{-3/2}, \end{aligned} \quad (2)$$

where  $\delta = \lambda/(1 + \lambda^2)^{1/2}$ ,  $\delta \in (-1, 1)$ ,  $b = (2/\pi)^{1/2}$ , and  $-0.9953 < \gamma < 0.9953$ . It follows that even moments of  $Z = (Y - \beta)/\omega \sim SN(\lambda)$  coincide with the standard normal ones (Arellano-Valle et al., 2005) and the coefficient of skewness for  $Z$  is that of  $Y$ . The following proposition is useful in the estimation of the parameters of this distribution (Henze, 1986; Azzalini, 1986).

**Proposition 1** If  $Z \sim SN(\lambda)$  and  $|\delta| < 1$ , then

$$Z = \delta |V_0| + \sqrt{1 - \delta^2} V_1, \quad (3)$$

where  $\delta = \lambda/(1 + \lambda^2)^{1/2}$ ,  $V_0$  and  $V_1$  are independent and identically distributed (i.i.d.)  $N(0, 1)$  random variables, and  $|V_0| \sim HN(0, 1)$ .

**Skew-normal interval mapping model**

For simplicity, we suppose  $n F_2$  progeny from an intercross between two inbred strains, but the results

extend easily to other kinds of crosses. Let  $y_i$  and  $\mathbf{m}_i$  denote the quantitative phenotype and the multipoint marker data, respectively, for individual  $i$ , where  $i = 1, \dots, n$ . The marker data include observed genotypes and known genetic marker maps.

Let  $g_i = 1, 2, 3$ , according to whether individual  $i$  has one of the three possible QTL genotypes  $aa$ ,  $ab$ , or  $bb$ , respectively, in this population, and  $j$  be the index of the QTL genotype. Then, an individual with genotype  $j$  at the putative QTL is assumed to have phenotype that follows a skew-normal distribution, i.e.,  $Y|g=j \sim SN(\beta_j, \omega, \lambda)$ , where  $j = 1, 2, 3$ .

Since the QTL genotypes will generally not be known, the phenotype distribution given the marker data is a mixture of three skew-normal distributions. Moreover, we may calculate  $p_{ij} = Pr(g_i = j | \mathbf{m}_i)$ , the conditional probability of QTL genotype  $j$  given the multipoint marker data  $\mathbf{m}_i$ , assuming some fixed position in the genome as the location of a putative QTL (Lynch and Walsh, 1998). Thus, the likelihood function for the parameter vector  $\theta = (\beta_1, \beta_2, \beta_3, \omega, \lambda)$  is given by

$$L(\theta | \mathbf{y}, \mathbf{m}) = \prod_{i=1}^n \sum_{j=1}^3 p_{ij} f(y_i; \beta_j, \omega, \lambda), \quad (4)$$

with  $p_{ij}$  defined as above and  $f(y_i; \beta_j, \omega, \lambda)$  being the density function of a skew-normal distribution, defined in Eq.(1) with parameters  $\beta_j$ ,  $\omega$ , and  $\lambda$ .

**Parameter estimation**

Closed-form expressions for the MLEs are not available for the skew-normal mixture model, nor for the normal mixture model. Therefore, estimation under the skew-normal IM method must be done numerically. Accordingly, a version of the EM algorithm (Dempster et al., 1977) is developed by treating the putative QTL as missing information.

Let  $z_{ij}$  be an unobserved variable,

$$z_{ij} = \begin{cases} 1, & \text{if the QTL genotype for individual } i \text{ is } j; \\ 0, & \text{otherwise.} \end{cases}$$

Then, the complete data likelihood function in Eq.(4) may be written as:

$$L_c(\theta | \mathbf{x}) = \prod_{i=1}^n \prod_{j=1}^3 [p_{ij} f(y_i; \beta_j, \omega, \lambda)]^{z_{ij}}, \quad (5)$$

where  $\mathbf{x}=(\mathbf{y},\mathbf{z},\mathbf{m})$  is a vector of complete data and  $Y|g\sim SN(\beta_j,\omega,\lambda)$  for  $j=1,2,3$ , all independent. The complete data log likelihood function is:

$$\ell_c(\boldsymbol{\theta}|\mathbf{x}) \propto \sum_{i=1}^n \sum_{j=1}^3 z_{ij} \log[f(y_i; \beta_j, \omega, \lambda)]. \quad (6)$$

If  $Y|g\sim SN(\beta_g,\omega,\lambda)$ , and we let  $V|g=(Y|g-\beta_g)/\omega$ , then  $V|g\sim SN(\lambda)$ , which jointly with Proposition 1 implies that:

$$V|g = \delta|V_0| + \sqrt{1-\delta^2}V_1, \quad (7)$$

where  $V_0$  and  $V_1$  are *i.i.d.*  $N(0,1)$  random variables and  $\delta=\lambda/(1+\lambda^2)^{1/2}$  (Arellano-Valle *et al.*, 2005), so that

$$Y|g = \beta_g + \omega\delta T|g + R, \quad (8)$$

where

$$\begin{aligned} T|g &= |V_0| \sim HN(0,1), \\ R &= \omega\sqrt{1-\delta^2}V_1 \sim N(0,\omega\sqrt{1-\delta^2}). \end{aligned} \quad (9)$$

These variables are all independent. Thus, the results in Eq.(8) and Eq.(9) imply that:

$$(Y|T, g) \sim N(\beta_g + \omega\delta t_g, \omega\sqrt{1-\delta^2}). \quad (10)$$

Under Eq.(9) and Eq.(10) it follows that the joint distribution of the latent variables  $t_{ij}=(T_i|g_i=j)$ , that we consider as the missing quantities, and  $Y_i|g_i=j$  is:

$$\begin{aligned} f(y_i, t_{ij}; \beta_j, \omega, \lambda) \\ = 2\phi(y_i; \beta_j + \omega\delta t_{ij}, \omega\sqrt{1-\delta^2}) \times \phi(t_{ij})I_{\{t_{ij}>0\}}, \end{aligned} \quad (11)$$

where  $I_{\{t_{ij}>0\}}$  is an indicator function with  $i=1,\dots,n$  and  $j=1,2,3$ , so that by independence the complete data log likelihood function in Eq.(6) can be written as:

$$\begin{aligned} \ell_c(\boldsymbol{\theta}|\mathbf{x}, \mathbf{t}) \propto \sum_{i=1}^n \sum_{j=1}^3 z_{ij} \log[\phi(y_i; \beta_j + \omega\delta t_{ij}, \omega\sqrt{1-\delta^2})] \\ + \sum_{i=1}^n \sum_{j=1}^3 z_{ij} \log[\phi(t_{ij})], \end{aligned} \quad (12)$$

so that,

$$\begin{aligned} \ell_c(\boldsymbol{\theta}|\mathbf{x}, \mathbf{t}) \propto \sum_{i=1}^n \sum_{j=1}^3 z_{ij} \log[\phi(y_i; \beta_j, \omega)] \\ + \sum_{i=1}^n \sum_{j=1}^3 z_{ij} \log[\phi(t_{ij}; \eta_{ij}, \tau)], \end{aligned} \quad (13)$$

where

$$\eta_{ij} = \delta \left( \frac{y_i - \beta_j}{\omega} \right) = \delta \frac{d_{ij}}{\omega}, \quad \tau = \sqrt{1 - \delta^2}. \quad (14)$$

Thus, we get

$$\begin{aligned} \ell_c(\boldsymbol{\theta}|\mathbf{x}, \mathbf{t}) \propto -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^3 z_{ij} \log[\omega^2(1-\delta^2)] \\ - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^3 z_{ij} \frac{[\omega^2 t_{ij}^2 - 2\delta\omega t_{ij} d_{ij} + d_{ij}^2]}{\omega^2(1-\delta^2)}. \end{aligned} \quad (15)$$

However, the direct parameterization of this family of distributions is inadequate for making inferences for the important case of the normal distribution because of the singularity of the Fisher information matrix when  $\lambda=0$  (Arellano-Valle *et al.*, 2005). Most significantly, there is no unique solution to the likelihood equations when the parent population is normal (Pewsey, 2006). As a means of avoiding this problem, we consider the centred parameterization of the distribution suggested by Azzalini (1985), that is:

$$Y = \beta + \omega Z = \mu + \sigma \left( \frac{Z - E[Z]}{\sqrt{Var[Z]}} \right), \quad (16)$$

where  $Y$  is a skew-normal variable, denoted by  $Y\sim SN_{cp}(\mu,\sigma,\gamma)$ , with mean  $E[Y]=\mu \in \mathbb{R}$  and variance  $Var[Y]=\sigma^2>0$ ,  $\gamma$  is the coefficient of skewness of  $Y$  and  $Z\sim SN(\lambda)$ . The cp subscript indicates, here and in its subsequent uses, that the centred parametrization is being referred to.

By inversion of the expressions in Eq.(2), the direct parameters are related to the centred ones according to

$$\begin{aligned} \beta &= \mu - \sigma c\gamma^{1/3}, \quad \omega = \sigma\sqrt{1+c^2\gamma^{2/3}}, \\ \lambda &= \frac{c\gamma^{1/3}}{\sqrt{b^2+c^2\gamma^{2/3}(b^2-1)}}, \quad \delta = \frac{c\gamma^{1/3}}{b\sqrt{1+c^2\gamma^{2/3}}}, \end{aligned} \quad (17)$$

where  $b=(2/\pi)^{1/2}$  and  $c=[2/(4-\pi)]^{1/3}$ . Replacing  $\beta_j$ ,  $\omega$ , and  $\lambda$  by these expressions in Eq.(15), the complete data log likelihood function for the parameter vector  $\theta_{cp}=(\mu_1,\mu_2,\mu_3,\sigma,\gamma)$  is given by

$$\ell_c(\theta_{cp} | \mathbf{x}, \mathbf{t}) \propto -\frac{n}{2} \log \left| \frac{\sigma^2 [b^2 + c^2 \gamma^{2/3} (b^2 - 1)]}{b^2} \right| - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^3 \frac{b^2 z_{ij} h_{ij}}{\sigma^2 [b^2 + c^2 \gamma^{2/3} (b^2 - 1)]} \tag{18}$$

where

$$h_{ij} = \sigma^2 (1 + c^2 \gamma^{2/3}) t_{ij}^2 - 2c\gamma^{1/3} \sigma b^{-1} t_{ij} d_{ij}^* + (d_{ij}^*)^2, \tag{19}$$

$$d_{ij}^* = (y_i - \mu_j) + \sigma c \gamma^{1/3}.$$

Assuming at iteration  $k+1$  we have estimates of the parameters  $\mu_j$ ,  $\sigma$ ,  $\gamma$ , where  $j=1,2,3$ . It follows from Eq.(18) that to implement the expectation-step or E-step it is necessary to calculate the following conditional expected value of the complete data log likelihood function given the observed phenotypes:

$$Q(\theta_{cp} | \hat{\theta}_{cp}^{(k)}, \mathbf{y}) = E[\ell_c(\theta_{cp} | \mathbf{y}, \mathbf{m}) | \hat{\theta}_{cp}^{(k)}]. \tag{20}$$

To do so, we calculate conditional expected values of  $z_{ij}$  given  $Y_i=y_i$  for each individual and for each of the three possible QTL genotypes:

$$z_{ij}^{(k+1)} = E[z_{ij} | y_i, \mathbf{m}_i, \hat{\theta}_{cp}^{(k)}] = Pr(z_{ij} = 1 | y_i, \mathbf{m}_i, \hat{\theta}_{cp}^{(k)}) = \frac{p_{ij} f(y_i; \hat{\mu}_j^{(k)}, \hat{\sigma}^{(k)}, \hat{\gamma}^{(k)})}{\sum_{j=1}^3 p_{ij} f(y_i; \hat{\mu}_j^{(k)}, \hat{\sigma}^{(k)}, \hat{\gamma}^{(k)})}. \tag{21}$$

Also, we calculate the conditional expected values of  $t_{ij}$  given  $Y_i=y_i$ . In order to obtain these conditional expected values, we consider the following lemma and proposition (Arellano-Valle *et al.*, 2005):

**Lemma 1** Let  $X \sim N(\eta, \tau)$ , then, for any real constant  $a$  it follows that

$$E[X | X > a] = \eta + \frac{\phi\left(\frac{a-\eta}{\tau}\right)}{1 - \Phi\left(\frac{a-\eta}{\tau}\right)} \tau,$$

$$E[X^2 | X > a] = \eta^2 + \tau^2 + \frac{\phi\left(\frac{a-\eta}{\tau}\right)}{1 - \Phi\left(\frac{a-\eta}{\tau}\right)} (\eta + a)\tau. \tag{22}$$

**Proposition 2** Let us consider  $Y|T \sim N(\beta + \omega\delta t, \omega(1-\delta^2)^{1/2})$  and  $T \sim HN(0, 1)$ . Then,

$$E[T | y, \theta] = E[X | X > 0],$$

$$E[T^2 | y, \theta] = E[X^2 | X > 0], \tag{23}$$

where  $X \sim N(\eta, \tau)$ ,  $\eta = \delta(y_i - \beta)/\omega$ ,  $\tau = (1 - \delta^2)^{1/2}$ ,  $\theta = (\beta, \omega, \lambda)$ , and  $\delta = \lambda / (1 + \lambda^2)^{1/2}$ . In particular,

$$E[T | y, \theta] = \eta + \frac{\phi(\eta/\tau)}{\Phi(\eta/\tau)} \tau,$$

$$E[T^2 | y, \theta] = \eta^2 + \tau^2 + \frac{\phi(\eta/\tau)}{\Phi(\eta/\tau)} \eta \tau. \tag{24}$$

Thus, considering the centred parametrization and Eq.(9) and Eq.(10), we get

$$t_{ij}^{(k+1)} = E[t_{ij} | y_i, \mathbf{m}_i, \hat{\theta}_{cp}^{(k)}] = \hat{\eta}_{ij}^{(k)} + \frac{\phi(\hat{\eta}_{ij}^{(k)} / \hat{\tau}^{(k)})}{\Phi(\hat{\eta}_{ij}^{(k)} / \hat{\tau}^{(k)})} \hat{\tau}^{(k)},$$

$$(t_{ij}^2)^{(k+1)} = E[t_{ij}^2 | y_i, \mathbf{m}_i, \hat{\theta}_{cp}^{(k)}] = (\hat{\eta}_{ij}^{(k)})^2 + (\hat{\tau}^{(k)})^2 + \frac{\phi(\hat{\eta}_{ij}^{(k)} / \hat{\tau}^{(k)})}{\Phi(\hat{\eta}_{ij}^{(k)} / \hat{\tau}^{(k)})} \hat{\eta}_{ij}^{(k)} \hat{\tau}^{(k)}, \tag{25}$$

where

$$\hat{\eta}_{ij}^{(k)} = \frac{b^{-1} c (\hat{\sigma}^{(k)})^{-1} (\hat{\gamma}^{1/3})^{(k)} (\hat{d}_{ij}^*)^{(k)}}{1 + c^2 (\hat{\gamma}^{2/3})^{(k)}},$$

$$\hat{\tau}^{(k)} = \frac{\sqrt{b^2 + c^2 (\hat{\gamma}^{2/3})^{(k)} (b^2 - 1)}}{b \sqrt{1 + c^2 (\hat{\gamma}^{2/3})^{(k)}}}, \tag{26}$$

with  $b=(2/\pi)^{1/2}$ ,  $c=[2/(4-\pi)]^{1/3}$ , and  $d_{ij}^*$  defined in Eq.(19).

In the maximization-step or M-step, we maximize the conditional expected value in Eq.(20), by taking the derivatives with respect to the parameters, setting the derivatives equal to zero and solving for  $\mu_j$ ,  $\sigma$ ,  $\gamma$ , where  $j=1,2,3$ . Thus, the MLEs are given by

$$\hat{\mu}_j^{(k+1)} = \frac{\sum_{i=1}^n z_{ij}^{(k)} y_i - c(\hat{\gamma}^{1/3})^{(k)} \hat{\sigma}^{(k)} \sum_{i=1}^n z_{ij}^{(k)} (t_{ij}^{(k)} / b - 1)}{\sum_{i=1}^n z_{ij}^{(k)}},$$

$$\hat{\sigma}^{(k+1)} = \sqrt{\frac{b^2 \sum_{i=1}^n \sum_{j=1}^3 z_{ij}^{(k)} h_{ij}^{(k)}}{2n[b^2 + c^2 (\hat{\gamma}^{2/3})^{(k)} (b^2 - 1)]}},$$

$$(\hat{\gamma}^{1/3})^{(k+1)} = \frac{\sqrt{1 - \frac{4b^2 (\hat{r}^{(k)})^2}{c^2 (b^2 - 1)}} - 1}{2\hat{r}^{(k)}}, \tag{27}$$

with  $b=(2/\pi)^{1/2}$ ,  $c=[2/(4-\pi)]^{1/3}$ ,  $h_{ij}$  and  $d_{ij}^*$  defined in Eq.(19), and

$$r = - \frac{\sum_{i=1}^n \sum_{j=1}^3 z_{ij} \left( \frac{\partial h_{ij}}{\partial \gamma^{1/3}} \right)}{\sum_{i=1}^n \sum_{j=1}^3 z_{ij} h_{ij}}. \tag{28}$$

Initial values for the EM algorithm may, for example, be obtained by taking the mean, variance, and coefficient of skewness, respectively, with the weights  $p_{ij}$ . We iterate until the estimates converge. The EM algorithm is performed at each position in the genome (in practice, at 1 cM steps). Finally, after calculating  $\hat{\mu}_j$ ,  $\hat{\sigma}$ , and  $\hat{\gamma}$  in Eq.(27), we convert these parameters into the direct parametrization by the relations in Eq.(17) and obtain the MLEs  $\hat{\beta}_j$ ,  $\hat{\omega}$ , and  $\hat{\lambda}$ .

Now, the null hypothesis of no QTL effect and a possible alternative hypothesis of the likelihood ratio (LR) test are:

$$H_0 : \beta_1 = \beta_2 = \beta_3 \text{ vs } H_a : \beta_1 \neq \beta_2 \vee \beta_1 \neq \beta_3 \vee \beta_2 \neq \beta_3.$$

The likelihood function under the null model is:

$$L(\theta_0 | \mathbf{y}, \mathbf{m}) = \prod_{i=1}^n f(y_i; \beta_0, \omega_0, \lambda_0), \tag{29}$$

where  $f(y_i; \beta_0, \omega_0, \lambda_0)$  is the density function of a skew-normal distribution with parameter vector  $\theta_0=(\beta_0, \omega_0, \lambda_0)$  defined in Eq.(1). As with standard IM,

the likelihood under  $H_0$  is calculated once, because the distribution in Eq.(29) does not depend on the genotype of the putative QTL. Under the null hypothesis, we also use a form of the EM algorithm to obtain MLEs of the parameters and consider the centred parameterization with parameters  $(\mu_0, \sigma_0, \gamma_0)$ . In this case, there is no QTL effect, so  $t_i$  and  $t_i^2$  are the unique unobserved variables, defined in similar form to the  $t_{ij}$  in Eq.(9), where  $i=1, \dots, n$  and  $j=1, 2, 3$ . We begin the EM algorithm by taking the method of moments (MM) estimates:

$$\hat{\mu}_0^{(0)} = \bar{y}, \quad \hat{\sigma}_0^{(0)} = s, \quad \hat{\gamma}_0^{(0)} = \sum_{i=1}^n (y_i - \bar{y})^3 / (n \times s^3), \tag{30}$$

where  $\bar{y}$  and  $s$  are the sample mean and sample standard deviation, respectively, and iterate until the estimates converge. Moreover, the test statistic follows, approximately, a  $\chi^2$  distribution with 2 degrees of freedom (DOF) under the null hypothesis of no QTL effect.

## RESULTS

### Example with real data

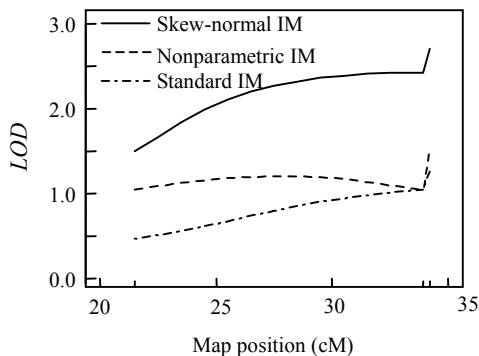
As described above (Fig.1), to search genes controlling the percentage of CD19+/CD69+ B-cells that were stimulated with bacterial LPS in mice, a genetic intercross experiment was performed between the C57BL/6 and BALB/c mouse strains (Rodo *et al.*, 2006). From this cross, 139 F<sub>2</sub> mice were generated. The mice were typed at a set of 98 markers on 19 chromosomes. We did not include markers on the X chromosome, because in this study design only two genotypic classes can be found for the X chromosome. We used the map distances (centiMorgan, cM) of the markers included in the database of the Whitehead Center for Genome Research (USA) (Mouse Genome Informatics, <http://www.informatics.jax.org/>).

We have applied the skew-normal IM, the standard IM, and the nonparametric IM to this data set. We have used a Bonferroni correction and declared significant linkage if the LOD scores exceeded the corresponding 95% genome-wide LOD thresholds, which for the three methods were:

$$\chi_{(1-\rho; 2)}^2 \times 0.217 = 3.29,$$

where  $\rho=\alpha/M$  is the significance level for each individual test,  $\alpha=0.05$  is the overall significance level for the entire experiment,  $M$  is the number of markers and  $LOD=\chi^2 \times 0.217$ . Since this correction is appropriate for tests using unlinked markers (such as those on different chromosomes), but tests involving linked markers are generally not independent (Lynch and Walsh, 1998). These thresholds are a conservative choice, the actual false positive rate is guaranteed to be smaller than  $\alpha$ . Thus, genome-wide LOD thresholds were also calculated (Churchill and Doerge, 1994; Zeng, 1994), using 1 000 permutation replicates. The estimated 95% genome-wide LOD thresholds for the three methods, skew-normal IM, standard IM, and nonparametric IM, were 2.5, 2.48, and 2.6, respectively.

In both methods, standard IM and nonparametric IM, no QTL was detected in this cross. However, the skew-normal IM detected one putative QTL, in chromosome 16. The LOD scores were calculated at every 1 cM and were plotted in Fig.2 for chromosome 16. LOD score curve produced by skew-normal IM and the corresponding LOD score curves from the standard IM and nonparametric IM were included for comparison.



**Fig.2** LOD score curves produced by the three QTL mapping methods, skew-normal IM, standard IM, and nonparametric IM, for data on the percentage of CD19+/CD69+ B-cells that were stimulated with bacterial lipopolysaccharide (LPS) in 139 intercross mice

The maximum LOD score was 2.70 corresponding to map position 34.2 cM on chromosome 16. For the centred parametrization, the ML estimates of the parameters were:  $\hat{\mu}_1 = 83.355$ ,  $\hat{\mu}_2 = 82.252$ ,  $\hat{\mu}_3 = 77.499$ ,  $\hat{\sigma} = 8.079$ , and  $\hat{\gamma} = -0.879$  (Eq.(27)).

Thus, the estimate of the negative skewness of the phenotype distribution was  $-0.879$ . The respective ML estimates of the parameters for the direct parametrization were:  $\hat{\beta}_1 = 93.608$ ,  $\hat{\beta}_2 = 92.507$ ,  $\hat{\beta}_3 = 87.75$ ,  $\hat{\omega} = 13.051$ , and  $\hat{\lambda} = -5.641$  (Eq.(17)).

Also, we obtained the estimate of the additive QTL effect ( $a_{sn}$ ) and the estimate of the dominance QTL effect ( $d_{sn}$ ) at the position of the maximum LOD score under the skew-normal IM:  $\hat{a}_{sn} = 0.5(\hat{\beta}_1 - \hat{\beta}_3) = 2.928$  and  $\hat{d}_{sn} = \hat{\beta}_2 - 0.5(\hat{\beta}_1 + \hat{\beta}_3) = 1.826$ , respectively.

For a single model, assuming that environmental and genetic effects are uncorrelated, we have the usual representation of the phenotypic variance as:

$$\sigma_p^2 = \sigma_G^2 + \sigma_E^2,$$

where  $\sigma_G^2 = (0.25a^2 + 0.5d^2)$  is the genotypic variance and  $\sigma_E^2$  is the environmental variance. The wide sense estimated heritability is 0.50. Thus, 50% of the proportion of phenotypic variation in a population is attributable to genetic variation among the individuals.

## Simulation

To investigate the power and precision of each of these QTL mapping methods, we simulated 200 intercross individuals under the skew-normal model. We also simulated one chromosome with 50 cM long covered by 6 markers equally spaced, with a marker spacing of 10 cM. A single bi-allelic QTL was placed at position 26 cM (between markers 3 and 4) of the chromosome.

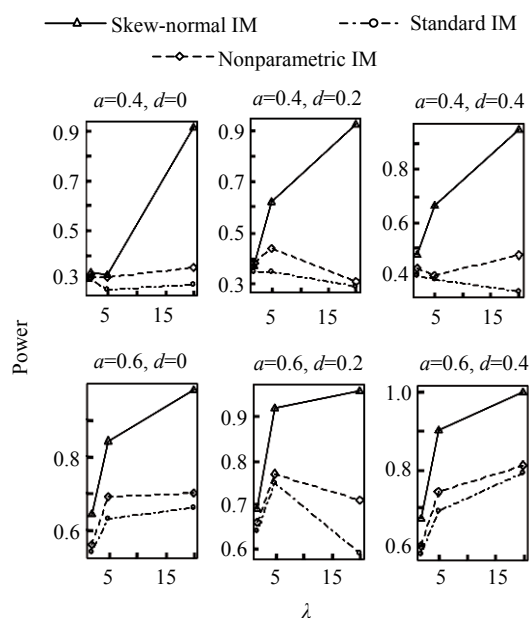
The QTL was taken to have additive and dominance effects. The location parameters were chosen so that  $\beta_1 = \beta + a_{sn}$ ,  $\beta_2 = \beta + d_{sn}$  and  $\beta_3 = \beta - a_{sn}$ , with  $\beta = 10$ . The scale parameter was  $\omega = 2$ . We considered the values  $\lambda = 2, 5, 20$ ,  $a_{sn} = 0, 0.4, 0.6$ , and  $d_{sn} = 0, 0.2, 0.4$ . Note that  $a_{sn} = 0$  and  $d_{sn} = 0$  correspond to the null hypothesis of no QTL.

For this case, 1 000 simulations were used to estimate the 95% genome-wide LOD thresholds. The simulated phenotype follows a skew-normal distribution and it was independent of the marker data. The LOD thresholds for the skew-normal IM, standard IM, and nonparametric IM methods appear in Table 1.

**Table 1** LOD thresholds produced by the three methods, skew-normal IM (SN), standard IM (Normal), and nonparametric IM (NP), for each simulated data set, under different values of the skewness parameter ( $\lambda=2,5,20$ )

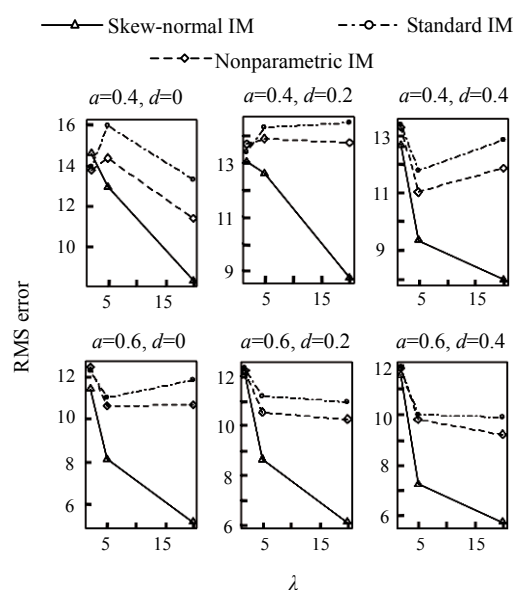
$\lambda$	LOD thresholds		
	SN	Normal	NP
2	2.07	2.01	1.96
5	2.21	2.01	2.00
20	2.21	1.97	1.93

One hundred replicates were conducted for each combination of the triple ( $a_{sn}, d_{sn}, \lambda$ ), except for the case  $a_{sn}=0$  and  $d_{sn}=0$ . The power of the three methods was calculated as the proportion of the simulation replicates for which the maximum LOD score exceeds the corresponding LOD threshold. The estimated power of the methods appears in Fig.3. The skew-normal IM had a higher power for QTL detection than the standard IM and the nonparametric IM, when the phenotype follows the skew-normal distribution, particularly for high values of  $\lambda$ .



**Fig.3** Estimated power to detect a QTL, based on 100 simulation replicates. Two hundred  $F_2$  individuals, one chromosome, and one QTL on the chromosome were simulated. Different values of the additive QTL effect ( $a_{sn}=0.4,0.6$ ) and dominance QTL effect ( $d_{sn}=0,0.2,0.4$ ), and different skewness parameters ( $\lambda=2,5,20$ ) were considered. Three methods were compared: skew-normal IM, standard IM, and nonparametric IM

We have also estimated the precision in locating the QTL by means of the root-mean-square (RMS) error of the estimated QTL position, among simulation replicates in which there was significant evidence for the presence of a QTL. Fig.4 contains the results on the precision of QTL location for the three methods. The skew-normal IM led to a greater precision of QTL location (smaller RMS error) compared to the other methods.



**Fig.4** Estimated root-mean-square (RMS) error of the estimated QTL location, based on 100 simulation replicates. Two hundred  $F_2$  individuals, one chromosome, and one QTL on the chromosome were simulated. Different values of the additive QTL effect ( $a_{sn}=0.4,0.6$ ) and dominance QTL effect ( $d_{sn}=0,0.2,0.4$ ), and different skewness parameters ( $\lambda=2,5,20$ ) were considered. Three methods were compared: skew-normal IM, standard IM, and nonparametric IM

**DISCUSSION**

Most QTL mapping methods, such as the standard IM, assume that the phenotype is modelled as a normal mixture distribution when a QTL is included in the model. In the case of  $F_2$  intercross populations, the model is a mixture of three components corresponding to the three different genotypes at the putative QTL. Under a null model of the non QTL effect, the phenotype follows a normal distribution. The normality assumption of the underlying distributions greatly simplifies the form of the likelihood.



However, it may be unrealistic, obscuring important characteristic of the model (Arellano-Valle *et al.*, 2005). Thus, in cases where the phenotype distribution deviates from a normal distribution, the standard IM may result in spurious LOD score peaks when in fact there is none QTL, i.e., false-positive results, in regions of low genotype information (Broman, 2003), e.g., widely spaced markers or much missing marker data.

Other methods for QTL mapping have been developed for cases where the phenotype distribution is not normal. If, for example, the phenotype follows a highly skewed distribution, the most commonly adopted method to achieve normality involves transforming the data using, for example, the Box-Cox transformation (Draper and Smith, 1998). However, an appropriate transformation may not exist or may be difficult to find, and the results obtained by this approach are more difficult of interpretation.

Also, the nonparametric IM is other alternative method (Kruglyak and Lander, 1995; Broman, 2003). However, power to detect genes is lost by using this method, particularly if the sample size is small and there is much missing marker data.

In this regard, the skew-normal distribution may fit much better to data than the normal distribution. Thus, an interesting alternative approach is to consider the IM under a skew-normal mixture distribution rather than a normal mixture distribution, here denoted as skew-normal IM.

The skew-normal model is an extension of the symmetric normal model, which incorporates asymmetry when some skewness is present in the data. The advantages of using such model include easiness of interpretation, as well as estimation efficiency (Arellano-Valle *et al.*, 2005). Also, the skew-normal distribution appears to attain a good compromise between mathematical tractability and shape flexibility. The skew-normal distribution shares a number of good properties with the normal distribution, such as that it is unimodal and the square of a standard random variable has a  $\chi^2$  distribution with one degree of freedom.

When dealing with the skew-normal distribution, the problems arise in the inferential steps. The estimation of the parameters is not easy (Azzalini, 1985). One of the problems is that the Fisher information matrix goes to singular as the skewness parameter  $\lambda$

goes to 0. Azzalini (1985) addressed this problem by proposing a different parameterization, named centred parameterization. Other problem is that the method of moments (MM) usually provides good initial estimates of the parameters  $\mu$  and  $\sigma$ , but it does not usually provide good initial estimate of the parameter  $\gamma$ , particularly if dimension  $n$  is small or  $|\gamma|$  is large (Pewsey, 2000). Thus, if the MM estimates are used as starting values, they may lead to local, rather than the global maximum of log likelihood. A standard approach is to use a grid of starting values in an attempt to ensure that the true global maximum is reached.

Moreover, in skew-normal IM, we could also consider an alternative skew-normal distribution with different scale and skewness parameters, but the resulting model may have smaller power and the estimation of the parameters becomes more complicated.

In the skew-normal mixture model, we assume that the three distributions have different location parameters  $\beta_j$  ( $j=1,2,3$ ), but equal scale  $\omega$  and skewness parameters  $\lambda$ . We have obtained the MLEs of the parameters by the EM algorithm for the distribution's centred parameterization. The EM algorithm is often found to be somewhat slow but fairly robust and easy to program.

There are two routines, `sn.em` and `sn.mle`, of the software `R/sn` (Azzalini, 1985), an add-on package for the general statistical software R (Ihaka and Gentleman, 1996), for fitting the skew-normal distribution using the ML estimation. The routine `sn.em` uses the EM algorithm and the direct parameterization, and the routine `sn.mle` employs gradient-based methods and the centred parameterization (Pewsey, 2000).

In our analysis, the standard IM and the nonparametric IM were implemented with functions of the QTL mapping software `R/qtl` (Broman, 2003), also an add-on package for the R. The skew-normal IM was implemented with new functions within the frameworks of `R/qtl` and `R/sn`. These functions are based on the existing routines `scanone` (`R/qtl`) and `sn.em` (`R/sn`). The function `scanone` performs a genome scan with a single QTL model. By default, it performs the standard IM and the nonparametric IM.

In order to compare the three QTL mapping methods, we have designed a simulation study. The results obtained indicate that the skew-normal IM has higher power for QTL detection and better precision

of QTL location than other methods, particularly when the skewness parameter is large.

To sum up, we may say that the skew-normal model should be preferred when the phenotype follows an asymmetric distribution.

## References

- Arellano-Valle, R.B., Ozan, S., Bolfarine, H., Lachos, V.H., 2005. Skew normal measurement error models. *Journal of Multivariate Analysis*, **96**(2):265-281. [doi:10.1016/j.jmva.2004.11.002]
- Azzalini, A., 1985. A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, **12**(2): 171-178.
- Azzalini, A., 1986. Further results on a class of distributions which includes the normal ones. *Statistica*, **46**(2): 199-208.
- Azzalini, A., Capitanio, A., 1999. Statistical applications of the multivariate skew-normal distribution. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **61**(3):579-602. [doi:10.1111/1467-9868.00194]
- Broman, K., 2003. Mapping quantitative trait loci in the case of a spike in the phenotype distribution. *Genetics*, **163**(3): 1169-1175.
- Churchill, G.A., Doerge, R.W., 1994. Empirical threshold values for quantitative trait mapping. *Genetics*, **138**(3): 963-971.
- Dalla Valle, A., 2004. The Skew-Normal Distribution. In: Genton, M.G. (Ed.), *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*. Chapman & Hall CRC, Boca Raton, FL.
- Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**(1): 1-38.
- Draper, N.R., Smith, H., 1998. *Applied Regression Analysis*, 3th Ed. John Wiley & Sons Inc., New York.
- Henze, N., 1986. A probabilistic representation of the skew-normal distribution. *Scandinavian Journal of Statistics*, **13**(4):271-275.
- Ihaka, R., Gentleman, R., 1996. R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, **5**(3):299-314. [doi:10.2307/1390807]
- Kao, C.H., Zeng, Z.B., Teasdale, R.D., 1999. Multiple interval mapping for quantitative trait loci. *Genetics*, **152**(3): 1203-1216.
- Kruglyak, L., Lander, E.S., 1995. A Nonparametric approach for mapping quantitative trait loci. *Genetics*, **139**(3): 1421-1428.
- Lander, E.S., Botstein, D., 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, **121**(1):185-199.
- Lynch, M., Walsh, B., 1998. *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Inc., Sunderland, MA, Massachusetts, USA.
- Morton, N.E., 1984. Trials of Segregation Analysis by Deterministic and Macro Simulation. In: Chakravarti, A. (Ed.), *Human Population Genetics: The Pittsburgh Symposium*. Van Nostrand Reinhold, New York, p.83-107.
- Pewsey, A., 2000. Problems of inference for Azzalini's skewnormal distribution. *Journal of Applied Statistics*, **27**(7):859-870. [doi:10.1080/02664760050120542]
- Pewsey, A., 2006. Modelling asymmetrically distributed circular data using the wrapped skew-normal distribution. *Environmental and Ecological Statistics*, **13**(3):257-269. [doi:10.1007/s10651-005-0010-4]
- Roberts, C., 1966. A correlation model useful in the study of twins. *Journal of the American Statistical Association*, **61**(316):1184-1190. [doi:10.2307/2283207]
- Rodo, J., Gonçalves, L.A., Demengeot, J., Coutinho, A., Penha-Gonçalves, C., 2006. MHC class II molecules control murine B cell responsiveness to lipopolysaccharide stimulation. *The Journal of Immunology*, **177**(7): 4620-4626.
- Zeng, Z.B., 1994. Precision mapping of quantitative trait loci. *Genetics*, **136**(4):1457-1468.
- Zou, F., Yandell, B.S., Fine, J.P., 2003. Rank-based statistical methodologies for quantitative trait locus mapping. *Genetics*, **165**(3):1599-1605.