

# Linear and generalized linear models for the detection of QTL effects on within-subject variability

DÖRTE WITTENBURG<sup>1</sup>, VOLKER GUIARD<sup>1</sup>, FRIEDRICH LIESE<sup>2</sup>  
AND NORBERT REINSCH<sup>1\*</sup>

<sup>1</sup>Forschungsinstitut für die Biologie landwirtschaftlicher Nutztiere, Wilhelm-Stahl-Allee 2, 18196 Dummerstorf, Germany

<sup>2</sup>Universität Rostock, Institut für Mathematik, Universitätsplatz 1, 18051 Rostock, Germany

(Received 27 March 2007 and in revised form 25 September 2007)

## Summary

Quantitative trait loci (QTLs) may affect not only the mean of a trait but also its variability. A special aspect is the variability between multiple measured traits of genotyped animals, such as the within-litter variance of piglet birth weights. The sample variance of repeated measurements is assigned as an observation for every genotyped individual. It is shown that the conditional distribution of the non-normally distributed trait can be approximated by a gamma distribution. To detect QTL effects in the daughter design, a generalized linear model with the identity link function is applied. Suitable test statistics are constructed to test the null hypothesis  $H_0$ : No QTL with effect on the within-litter variance is segregating versus  $H_A$ : There is a QTL with effect on the variability of birth weight within litter. Furthermore, estimates of the QTL effect and the QTL position are introduced and discussed. The efficiency of the presented tests is compared with a test based on weighted regression. The error probability of the first type as well as the power of QTL detection are discussed and compared for the different tests.

## 1. Introduction

Quantitative genetic analyses of body weight data in snails (Ros *et al.*, 2004) suggest genetic differences in variability. Therefore, QTLs (quantitative trait loci) may affect not only the mean of a certain character but also its variability. The analysis of QTL effects on between-subject variability of a normally distributed trait was investigated by Weller & Wyler (1992). They mentioned uniformity of flowering time of plants as an example of potential economic importance, in particular when crops are harvested mechanically. Some phenotypes are repeated several times by the same individual, such as the size and weight of tomatoes from the same panicle of a tomato plant. Uniformity of such repeated phenotypes may also be genetically controlled and affected by the individual's genotype.

In multiparous species, birth weight of newborns from the same litter may be regarded as a special case of a repeated phenotype of the mother. The difference

compared with the tomato example is that the phenotype of the newborns is not only under maternal control but is also affected by the father's genetic contribution. Höggberg & Rydhmer (2000) and Damgaard *et al.* (2003) considered the within-litter standard deviation of piglet birth weight and attributed it to the dam of each litter as a maternal trait. A low within-litter uniformity was considered as an effect which was unfavourable for sow productivity. Heritability estimates for the character were 10% (Höggberg & Rydhmer, 2000) and 8% (Damgaard *et al.*, 2003). The same trait has also been studied in rabbits (Bolet *et al.*, 2005).

In this article it is assumed that the QTL affects the within-litter variability of a mother's progeny, i.e. in contrast to Weller & Wyler (1992) the focus is on within-subject variability. This offers the opportunity to construct a test for  $H_0$ : No QTL with effect on the within-litter variance is segregating versus  $H_A$ : There is a QTL with effect on the variability of birth weight within litter. A daughter design is considered, where genotyped females are paternal half-sibs. The sample

\* Corresponding author. e-mail: reinsch@fhn-dummerstorf.de

variances of birth weights within litter are the traits to which our model is fitted. First, the QTL effect on the within-litter variance is described. Then it is shown that a generalized linear model (GLM) can be applied for QTL mapping. This GLM is contrasted with a weighted regression approach in terms of power of QTL detection by numerical simulation. Inclusion of sex effects, different experimental designs and further fields of application are part of the discussion.

**2. Methods**

(i) *QTL effect on the within-litter variance*

It is assumed that a population of pigs has two alleles at the QTL denoted by  $Q$  and  $q$ . We consider a fixed number  $N$  of sires in our study, which are drawn by chance from the population. Every sire is mated with  $n$  unrelated dams. We pick out one daughter per mating and consider her offspring’s birth weight as a multiple measurement. We assume that piglet birth weights are independently and identically distributed within one litter. The birth weight consists of a fixed litter mean, the normally distributed mendelian sampling effect  $N(0, \frac{1}{2}\sigma_{polygene}^2)$  and the additive QTL effect, which is dependent on the piglet’s genotype, with variance  $\sigma_{QTL}^2$  and the normally distributed random deviation  $N(0, \sigma_e^2)$ .

The sample variance of weights at birth within one litter, that is the secondary observation, is taken as a trait for every daughter amounting to  $Nn$  observations. Daughters having inherited the QTL allele  $Q$  from the presumed heterozygous sire feature uniformity of birth weights. Daughters with a paternal  $q$  allele show an increased variability of birth weight. In this case, the residual deviation of piglet birth weight is multiplicatively inflated by the factor  $c_* \in (0, \infty)$ . Thus, from the breeder’s perspective, the positive effect of the QTL (the lower within-litter variance) is inherited with the QTL allele  $Q$ . A detailed description of the model for piglet birth weight and a further outline on the distribution of the traits are given in Appendix A.

The within-litter variance, i.e. the sample variance  $S_{i,j}^2$  of birth weights within one litter, depends on the paternal QTL allele of the daughter  $i \in \{1, \dots, N\}$ ,  $j \in \{1, \dots, n\}$ . The indicator function  $\mathbf{1}_{\{Q\},i,j}$  takes the value 1 if the daughter  $i,j$  has inherited the allele  $Q$  and 0 otherwise. Later, in Sections 2(ii) and 2(iv), the probability  $Pr(\mathbf{1}_{\{Q\},i,j}=1)$  is determined conditional on the observed flanking marker alleles. The conditional expectation of  $S_{i,j}^2$  given the inherited paternal QTL allele is

$$\begin{aligned} \mathbb{E}(S_{i,j}^2 | \mathbf{1}_{\{Q\},i,j}=1) &= \frac{1}{2}\sigma_{polygene}^2 + \sigma_{QTL}^2 + \sigma_e^2, \\ \mathbb{E}(S_{i,j}^2 | \mathbf{1}_{\{Q\},i,j}=0) &= \frac{1}{2}\sigma_{polygene}^2 + \sigma_{QTL}^2 + c_*^2\sigma_e^2. \end{aligned} \tag{1}$$

The value  $\tau^2 := \frac{1}{2}\sigma_{polygene}^2 + \sigma_e^2$  summarizes the variance of the normally distributed effects of piglet birth weight under the condition that the sow has inherited the QTL allele  $Q$ . Similarly,  $\tau_*^2 := \sigma_{polygene}^2 + (c_*\sigma_e)^2$  includes the modified residual variance component. Set

$$c^2 := \frac{\mathbb{E}(S_{i,j}^2 | \mathbf{1}_{\{Q\},i,j}=0)}{\mathbb{E}(S_{i,j}^2 | \mathbf{1}_{\{Q\},i,j}=1)} = \frac{\tau_*^2 + \sigma_{QTL}^2}{\tau^2 + \sigma_{QTL}^2}. \tag{2}$$

The parameter  $c^2$  is the ratio of the within-litter variance if the daughter  $i,j$  has inherited the QTL allele  $q$  and the within-litter variance if the daughter has inherited the allele  $Q$ . If the QTL effect on the within-litter variance actually exists, then the sample variance depends on the inherited paternal QTL allele and the ratio  $c^2$  is different from 1. Otherwise  $c^2$  is equal to 1.

(ii) *Generalized linear model*

The sires may have the marker genotype of kind  $m_{l,1}, m_{l,2}$ , where  $l \in \{0, 1, \dots, \kappa\}$  denotes the marker position on the chromosome. The sire’s two marker alleles are denoted by  $m_{l,1}$  on his paternal chromosome and  $m_{l,2}$  on his maternal chromosome for every marker position. It is not possible to determine which sire is heterozygous or homozygous at the QTL *a priori*. After the sires are genotyped, we suppose that all daughters are fully informative. Therefore, we need only to consider the paternal allele of daughters. The recombination rates are calculated by Haldane’s mapping function. We consider intervals flanked by markers  $M_l$  and  $M_{l+1}$  with realizations  $m_{l,r}m_{l+1,s}$ , where the subscripts  $r, s \in \{1, 2\}$  specify the sire’s flanking marker alleles transmitted to the daughter. The transmission probability of the QTL allele  $Q$  at position  $d \in \{0, 1, \dots, \delta\}$  is a function of the flanking markers  $M_l M_{l+1}$  and the paternal QTL allele. Let  $T_{i,j}$  denote the random variable, which is realized by the respective transmission probability  $t_{i,j,d}$  depending on the observed flanking marker alleles per daughter  $i,j$  at position  $d$ .

First, one presumed QTL position  $d \in \{0, 1, \dots, \delta\}$  is investigated. The observed value per daughter  $i,j$  is the realized sample variance  $s_{i,j}^2$  of the piglet birth weights within one litter,  $i = 1, \dots, N, j = 1, \dots, n$  and  $s^2 = (s_{1,1}^2, s_{1,2}^2, \dots, s_{N,n}^2)^T$ . As a result of Appendix A, the distribution of the sample variance  $S_{i,j}^2$  is approximated by a gamma distribution. Note that a gamma distributed random variable has the expectation  $\mu$  and variance  $\mu^2 \frac{\phi}{w}$  with dispersion parameter  $\phi$  and weight  $w$ .

Our aim is to fit a GLM (McCullagh & Nelder, 1989) to the sample variances. To distend we introduce a multiplicative model. If the sire’s genotype is  $Qq$ , then for  $r, s \in \{1, 2\}$  the conditional expectation

of  $S^2_{i,j}$  given the observed marker alleles  $\{M_l = m_{l,r}, M_{l+1} = m_{l+1,s}\}$  at position  $d$  is

$$\begin{aligned} \mu_{i,j,d} &= \mathbb{E}(S^2_{i,j} | M_l = m_{l,r}, M_{l+1} = m_{l+1,s}; d) \\ &= t_{i,j,d}(\sigma^2_{QTL} + \tau^2) + (1 - t_{i,j,d})(\sigma^2_{QTL} + \tau_*^2) \\ &= \sigma^2_{QTL} + \tau_*^2 + t_{i,j,d}(\tau^2 - \tau_*^2) \\ &=: u_i + b_i t_{i,j,d}. \end{aligned} \tag{3}$$

The mean value  $u_i$  per sire  $i \in \{1, \dots, N\}$  is

$$u_i = \sigma^2_{QTL} + \tau_*^2 \tag{4}$$

and the parameter  $b_i$  describes the relation between the observed trait  $s^2_{i,j}$  per daughter and the inherited paternal QTL allele, i.e.

$$\begin{aligned} b_i &= (\tau^2 + \sigma^2_{QTL}) - (\tau_*^2 + \sigma^2_{QTL}) \\ &= (1 - c^2)(\tau^2 + \sigma^2_{QTL}). \end{aligned} \tag{5}$$

In view of (3) the sample variance is described by the model

$$S^2_{i,j} = \{u_i + b_i T_{i,j}\} \cdot \varepsilon_{i,j}. \tag{6}$$

The  $\varepsilon_{i,j}$  are independently gamma distributed random variables with expectation 1. The weights are defined by  $w_{i,j} = \frac{n_{i,j}-1}{2}$ , where  $n_{i,j}$  denotes the litter size of daughter  $i, j$ . The identity link function is used to obtain the linear predictor  $\eta_{i,j,d} = \mu_{i,j,d}$ . The parameter vector  $\beta$  consists of the regression coefficients

$$\beta = (u_1, \dots, u_N, b_1, \dots, b_N)^T. \tag{7}$$

The application of GLM theory leads to estimates of the expectations  $\mu_d = (\mu_{1,1,d}, \mu_{1,2,d}, \dots, \mu_{N,n,d})^T$  as well as the vector  $\beta$  in (7) at position  $d \in \{0, 1, \dots, \delta\}$ . Consequently, it is possible to construct an appropriate test statistic to check the local null hypothesis  $H_{0,d}$ : *There exists no QTL at position  $d$  affecting the within-litter variance*, which is equivalent to

$$\begin{aligned} H_{0,d}: & \quad b_1 = \dots = b_N = 0 \quad \text{or equivalently} \quad \mu_d = \mu^0 \\ \text{vs } H_{A,d}: & \quad b_k \neq b_l \quad \text{for some } k \neq l. \end{aligned} \tag{8}$$

With (3) the log-likelihood function  $\ell$  of the modelled gamma distributed random vector  $S^2 = (S^2_{1,1}, \dots, S^2_{N,n})^T$  can be expressed in terms of  $\mu_d$  at position  $d$ . It holds

$$\begin{aligned} \ell(s^2, \mu_d, \phi_d) &= \\ & \sum_{i=1, j=1}^{N, n} \left\{ \frac{w_{i,j}}{\phi_d} \left[ -\frac{s^2_{i,j}}{u_i + b_i t_{i,j,d}} - \ln(u_i + b_i t_{i,j,d}) \right] \right. \\ & \left. + \zeta(s^2_{i,j}, \phi_d) \right\} =: l(s^2, \beta, \phi_d). \end{aligned} \tag{9}$$

The  $\zeta(s^2_{i,j}, \phi_d)$  summarizes those components where  $\mu_{i,j,d}$  does not appear and  $\beta$  is the vector (7). The

estimate  $\hat{\beta}_{N,n,d} = (\hat{u}_{n,1,d}, \dots, \hat{u}_{n,N,d}, \hat{b}_{n,1,d}, \dots, \hat{b}_{n,N,d})^T$  may be obtained by iterative procedures (McCullagh & Nelder, 1989) as implemented in the ‘glm’ function of the R program (R Development Core Team, 2005). Using (4) and (5) the parameter  $c^{-1}$  is estimated for each sire  $i \in \{1, \dots, N\}$  at the detected QTL position  $\hat{d}$  by

$$\widehat{c^{-1}}_i = \sqrt{\frac{\hat{b}_{n,i,\hat{d}}}{\hat{u}_{n,i,\hat{d}}} + 1}.$$

For asymptotic investigations of the estimator  $\hat{\beta}_{N,n,d}$ , some special matrices are needed. The design matrix  $\mathcal{X}_d$ , which contains the transmission probabilities at position  $d$ , is

$$\mathcal{X}_d = \begin{pmatrix} 1 & 0 & \dots & t_{1,1,d} & 0 & \dots \\ 1 & 0 & \dots & t_{1,2,d} & 0 & \dots \\ \vdots & \vdots & \dots & \vdots & \vdots & \dots \\ 1 & 0 & \dots & t_{1,n,d} & 0 & \dots \\ 0 & 1 & \dots & 0 & t_{2,1,d} & \dots \\ \vdots & \vdots & \dots & \vdots & \vdots & \dots \\ 0 & 1 & \dots & 0 & t_{2,n,d} & \dots \\ \vdots & \vdots & \dots & \vdots & \vdots & \dots \end{pmatrix}. \tag{10}$$

Let  $\mathcal{W}$  be the diagonal weight matrix with elements  $w_{i,j} = \frac{n_{i,j}-1}{2}$ . Then it follows from (9) that the Fisher information matrix  $\mathcal{I}(\beta, d)$  of the conditional distribution at position  $d$  is

$$\begin{aligned} \mathcal{I}(\beta, d) &= -\mathbb{E}[\nabla \nabla^T l(S^2, \beta, \phi_d) | M_l = m_{l,r}, M_{l+1} \\ &= m_{l+1,s}; d] = \frac{1}{\phi_d} \mathcal{X}_d^T \mathcal{W} \mathcal{X}_d. \end{aligned}$$

Moreover, let  $\mathcal{G}_d \in \mathbb{R}^{2N \times 2N}$  be the root of  $\mathcal{I}^{-1}(\beta, d)$  defined by  $\mathcal{G}_d^T \mathcal{G}_d = \mathcal{I}^{-1}(\beta, d)$ . Under some conditions (Fahrmeir & Kaufmann, 1985), which can be shown to be satisfied (see Supplementary Appendixes), the ML estimator  $\hat{\beta}_{N,n,d}$  is asymptotically normal, i.e.

$$\mathcal{G}_d^{-T} (\hat{\beta}_{N,n,d} - \beta) \rightarrow_{\mathcal{D}} N(0, \mathcal{I}_{2N}) \quad \text{as } n \rightarrow \infty, \tag{11}$$

where  $\rightarrow_{\mathcal{D}}$  denotes the convergence in distribution.

Under the null hypothesis (8) the model reduces to  $S^2_{i,j} = u_i \cdot \varepsilon_{i,j}$ . It can be shown that the dispersion parameter  $\phi_d^0 = \phi^0$  is approximately 1 under the null hypothesis for every position  $d$  (see Supplementary Appendixes). Thus, the likelihood function of  $S^2$  at  $\phi^0 = 1$  is constant for every position  $d$  under  $H_{0,d}$ .

(iii) *Test statistics in the GLM*

Four different types of tests statistics differing, among others, in their treatment of the dispersion

parameter are described in detail in the following. Three of them are later compared via simulation (Section 3).

The estimate of the expectation vector  $\mu_d$  is defined by  $\hat{\mu}_{N,n,d} = \mathcal{X}_d \hat{\beta}_{N,n,d}$ . With the log-likelihood function  $\ell(s^2, \hat{\mu}_{N,n,d}, \phi_d)$  in (9) the scaled deviance  $D^*$  for a fixed value of the dispersion parameter  $\phi_d$  is defined by (McCullagh & Nelder, 1989)

$$D^*(s^2, \hat{\mu}_{N,n,d}, \phi_d) := 2[\ell(s^2, s^2, \phi_d) - \ell(s^2, \hat{\mu}_{N,n,d}, \phi_d)].$$

The deviance  $D$  is characterized by  $D(s^2, \hat{\mu}_{N,n,d}) := \phi_d D^*(s^2, \hat{\mu}_{N,n,d}, \phi_d)$ . Under the null hypothesis  $H_{0,d}$  and for a fixed value  $\phi_d$  the likelihood ratio is asymptotically  $\chi^2$ -distributed with  $N$  degrees of freedom (Fahrmeir & Tutz, 1994)

$$2[l(S^2, \hat{\beta}_{N,n,d}, \phi_d) - l(S^2, \hat{\beta}_{N,n}^0, \phi_d)] \xrightarrow{H_{0,d}} \mathcal{D}\chi_N^2$$

as  $n \rightarrow \infty$ , (12)

where  $\hat{\beta}_{N,n}^0$  is the ML estimate under the null hypothesis. With the log-likelihood function  $\ell(s^2, \mu_d, \phi_d) = l(s^2, \beta, \phi_d)$  in (9), the statement (12) is equivalent to

$$D^*(S^2, \hat{\mu}_{N,n}^0, \phi_d) - D^*(S^2, \hat{\mu}_{N,n,d}, \phi_d) \xrightarrow{H_{0,d}} \mathcal{D}\chi_N^2$$

as  $n \rightarrow \infty$ . (13)

If  $\phi_d = 1$  is satisfied, then

$$D(S^2, \hat{\mu}_{N,n}^0) - D(S^2, \hat{\mu}_{N,n,d}) \xrightarrow{H_{0,d}} \mathcal{D}\chi_N^2 \quad \text{as } n \rightarrow \infty. \quad (14)$$

The generalized Pearson estimator for the dispersion parameter  $\phi_d$  at position  $d \in \{0, 1, \dots, \delta\}$  is defined by (e.g. Fahrmeir & Tutz, 1994)

$$\hat{\phi}_{N,n,d} = \frac{1}{Nn - 2N} \sum_{i=1, j=1}^{N, n} w_{i,j} \left( \frac{s_{i,j}^2 - \hat{\mu}_{N,n,i,j,d}}{\hat{\mu}_{N,n,i,j,d}} \right)^2.$$

This estimator is consistent and approximately  $\chi^2$ -distributed (Fahrmeir & Tutz, 1994). If  $\phi_d$  is replaced by a consistent estimator in (12), then this statement remains valid. According to Jørgensen (1987) it holds

$$\frac{1}{\hat{\phi}_{N,n,d}} [D(S^2, \hat{\mu}_{N,n}^0) - D(S^2, \hat{\mu}_{N,n,d})] \xrightarrow{H_{0,d}} \mathcal{D}\chi_N^2$$

as  $n \rightarrow \infty$ . (15)

Similarly, the deviance estimator  $\tilde{\phi}_{N,n,d}$  is usually applied to estimate the dispersion parameter,

$$\tilde{\phi}_{N,n,d} = \frac{D(s^2, \hat{\mu}_{N,n,d})}{Nn - 2N}.$$

Note that  $\tilde{\phi}_{N,n,d}$  is not necessarily consistent. Using this deviance estimator the distribution of the

left-hand term in (13) is approximated by the  $F$ -distribution with  $N$  and  $Nn - 2N$  degrees of freedom (Jørgensen, 1987),

$$\frac{D(S^2, \hat{\mu}_{N,n}^0) - D(S^2, \hat{\mu}_{N,n,d})}{D(S^2, \hat{\mu}_{N,n,d})} \frac{Nn - 2N}{N} \xrightarrow{H_{0,d}} \mathcal{D}F_{N, Nn - 2N}. \quad (16)$$

To test the local null hypothesis in (8)  $H_{0,d}: \mu_d = \mu^0$  there are at least four natural test statistics:

according to (13)

$$L_d^* = D^*(s^2, \hat{\mu}_{N,n}^0, \phi_d) - D^*(s^2, \hat{\mu}_{N,n,d}, \phi_d),$$

according to (14)  $L_d = D(s^2, \hat{\mu}_{N,n}^0) - D(s^2, \hat{\mu}_{N,n,d})$ ,

according to (15)

$$\hat{L}_d = \frac{1}{\hat{\phi}_{N,n,d}} [D(s^2, \hat{\mu}_{N,n}^0) - D(s^2, \hat{\mu}_{N,n,d})],$$

according to (16)

$$\tilde{F}_d = \frac{D(s^2, \hat{\mu}_{N,n}^0) - D(s^2, \hat{\mu}_{N,n,d})}{D(s^2, \hat{\mu}_{N,n,d})} \frac{Nn - 2N}{N}.$$

If  $\phi_d = 1$  is not fulfilled, then the test based on  $\hat{L}_d$  and  $\tilde{F}_d$  can be expected to have more QTL detection power. The threshold value under local investigations is given by the 95% quantile of the corresponding distribution in (13), (14), (15) and (16) of the test statistics  $L_d^*$ ,  $L_d$ ,  $\hat{L}_d$  and  $\tilde{F}_d$ , respectively.

We now consider the global hypothesis testing problem  $H_0$ : *There exists no QTL on the chromosome with effect on the within-litter variance*, which is equivalent to

$$H_0: b_1 = \dots = b_N = 0 \text{ or equivalently } \mu_d = \mu^0$$

for  $d = 0, 1, \dots, \delta$

$$\text{vs } H_A: b_k \neq b_l \text{ for some } k \neq l. \quad (17)$$

The following test statistics are appropriate:

$$L^* = \max_{d \in \{0, 1, \dots, \delta\}} L_d^*, \quad (18)$$

$$L = \max_{d \in \{0, 1, \dots, \delta\}} L_d, \quad (19)$$

$$\hat{L} = \max_{d \in \{0, 1, \dots, \delta\}} \hat{L}_d, \quad (20)$$

$$\tilde{F} = \max_{d \in \{0, 1, \dots, \delta\}} \tilde{F}_d. \quad (21)$$

The null hypothesis (17) is rejected for large values of the corresponding test statistic. The theoretical distribution of the presented test statistics is unknown because of marker dependencies. Thus, to find the threshold we use the permutation test approach (Churchill & Doerge, 1994). Properties of these tests will be given in Section 3.

Furthermore, the QTL is estimated by maximum likelihood at that position, where the value of the test statistic is maximal, e.g.

$$\hat{d} \in \arg \max_{d \in \{0,1, \dots, \delta\}} L_d^* \tag{22}$$

(iv) *Weighted regression*

Using ideas from Haley & Knott (1992) we introduce a weighted regression model and construct the test statistic for the global null hypothesis in (17). Later, in Section 3, the two approaches will be compared by computer simulations. Ros *et al.* (2004) recommended the use of a log-transformation on skew distributed traits (Box & Cox, 1964). Applying the logarithm of the sample variances the data are approximated by a normal distribution (see Supplementary Appendixes) and a linear model (LM) is constructed. Note that under the log-transformation the multiplicative effect on the within-litter variance becomes additive.

Considering the observed marker alleles  $m_{l,r}, m_{l+1,s}$  of individual  $i, j$  and the parameter  $c^2$  in (2) the conditional expectation given the flanking marker alleles is

$$\begin{aligned} E(\ln S_{i,j}^2 | M_l = m_{l,r}, M_{l+1} = m_{l+1,s}, ; d) \\ \approx \ln(\tau_*^2 + \sigma_{QTL}^2) - t_{i,j,d} \ln c^2. \end{aligned} \tag{23}$$

Thus, the LM at a fixed position  $d \in \{0,1, \dots, \delta\}$  is defined by

$$\ln S_{i,j}^2 = u_i + b_i T_{i,j} + \varepsilon_{i,j}. \tag{24}$$

Here the  $\varepsilon_{i,j}$  are normally distributed random variables with expectation null and  $T_{i,j}$  are the transmission probabilities as explained in Section 2(ii). The parameter  $u_i$  is the mean value per family and  $b_i$  describes the linear connection between the observations  $\ln s_{i,j}^2$  and the inherited paternal QTL allele expressed by the individual transmission probabilities. A standard (weighted) regression analysis has been carried out in order to estimate the parameter vector  $\beta$  similar to (7) for every position  $d$  on the chromosome. For weighted regression we refer to Seber (1977). To check the local null hypothesis  $H_{0,d}$  in (8), a test statistic  $F_d$  is constructed, which is a function of the residual sum of squares of the full and reduced model. This statistic is approximately  $F$ -distributed under  $H_{0,d}$  with  $N$  and  $Nn - 2N$  degrees of freedom (Seber, 1977). To test the global null hypothesis  $H_0$  in (17) the permutation test is used again to determine the threshold value. The suitable test statistic is

$$F = \max_{d \in \{0,1, \dots, \delta\}} F_d. \tag{25}$$

Under the assumption that the sire  $i \in \{1, \dots, N\}$  has the genotype  $Qq$ , it follows from (23) that  $b_i = -\ln c^2$ .

Thus, the parameter  $c^{-1}$  is estimated at the detected QTL position  $\hat{d}$  by

$$\widehat{c_{n,i}^{-1}} = \sqrt{\exp\{\widehat{b}_{n,i,\hat{d}}\}} \quad \text{for } i=1, \dots, N.$$

**3. Simulation studies**

When genotyping the individuals of the population, we set markers at intervals of 10 centiMorgan (cM) on a chromosome of length 100 cM ( $\delta=99$ ). We thus have 11 markers at our disposal ( $\kappa=10$ ) and the local test statistics are evaluated in steps of 1 cM. Under the null hypothesis  $H_0$ , no QTL is segregating in the population. To model the alternative hypothesis we placed a single QTL at position 25 cM (between the third and fourth marker). In the simulation study we used  $N=4$  sires and  $n=200$  daughters per sire. The litter size is Poisson distributed (Thomson, 2003) with a mean value of 10. The transformed weight at birth  $X_{i,j,k}$  in (A.4) is simulated. The standard deviation of piglet birth weight is assumed to be 320 g (e.g. Roehe, 1999). Similar to Roehe (1999) the residual variance is about 40% of the phenotypic variance,  $\sigma_e^2 = (200 \text{ g})^2$ . The direct polygenic variance is about 9% of the phenotypic variance,  $\sigma_{polygene}^2 = (96.8 \text{ g})^2$ . The value of the transformed additive QTL effect  $\widehat{G}_{i,j,k}$  is listed in Table 2 and depends on the piglet's genotype. Because the variance of the additive QTL effect takes about 1–3% of the phenotypic variance (e.g. Bidanel *et al.*, 2001), the additive value is  $a=61$  g. The factor  $c_*$  varies from 1 to 1.4 by 0.1. The gene frequency is assumed to be one-half. Covariances between the maternal effects and the direct effects of the piglet are neglected. The marker alleles  $m_{l,r}, l=0,1, \dots, 10$ , are drawn by chance according to the recombination rates. The simulation was repeated 100 times for every investigated factor  $c_*$ . Ten thousand permutations of the first simulated sample variances were used to determine the chromosome-wise threshold value. This critical value was also applied for the following 99 repeated simulations of sample variances. The simulations and tests were carried out with 'glm' and self-written functions of the R program (R Development Core Team, 2005).

Fig. 1a shows the density estimates of the conditional distributions of  $S_{i,j}^2$  obtained from the application 'density' with Gaussian kernel in R. This figure shows how the densities depend on the inherited paternal QTL allele as it is pointed out in Appendix A.

(i) *Results based on the GLM*

Examples of the results of analysing the simulated sample variances with use of the GLM theory are shown for the factor  $c_* = 1.2$  ( $c = 1.177$ ) in Figs. 1b

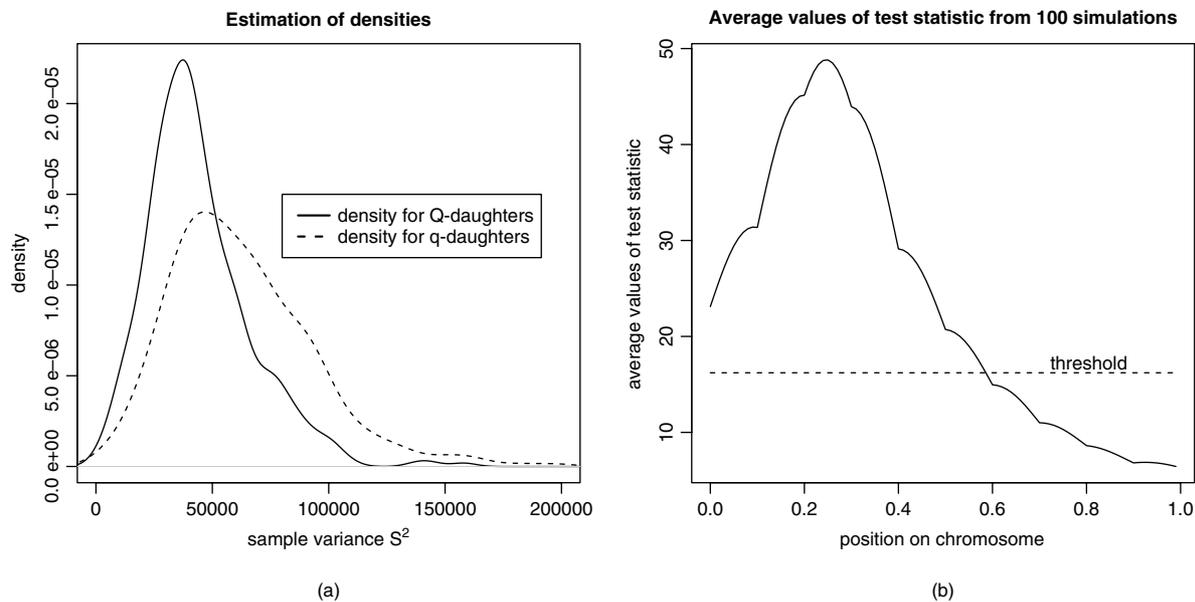


Fig. 1. QTL simulated at 25 cM,  $c_* = 1.2$ . (a) Estimation of densities separated by paternal QTL allele  $Q$  and  $q$ ; (b) average values of test statistic  $\hat{L}$  based on the GLM and 100 repetitions.

and 2a, b. Fig. 1b displays the average values of the test statistic  $\hat{L}$  in (20). The maximum of these values of the test statistic is attained at about 25 cM, where the QTL was actually simulated. Fig. 2a is a histogram of detected QTL positions (22) if the null hypothesis is rejected. One can see that the estimated positions closely surround the correct position and deviate by only about 5–10 cM. In Fig. 2b it is conspicuous that the estimates of  $\hat{c}_{n,i}^{-1}$  are split into three groups. Depending on the linkage of marker and QTL alleles the values fluctuate around the parameters  $c$  or  $c^{-1}$  for heterozygous sires with genotype  $qQ$  and  $Qq$ , respectively. Otherwise, in case of homozygotes, the estimates vary about 1. When test statistic  $\hat{F}$  is applied, the results are similar to the application of  $\hat{L}$  as shown in Table 3.

#### (ii) Results based on the LM

Evaluating the simulated sample variances with use of the test statistic  $F$  in (25), the results differ slightly from the case of applying the test statistics  $\hat{L}$  and  $\hat{F}$ . For the weighted regression model we used a maximization as in (22) to estimate the QTL position. Fig. 2c gives a histogram of the estimated QTL positions if  $H_0$  is rejected. The fluctuation around the simulated QTL position at 25 cM is similar to the GLM. Fig. 2d displays a histogram of the estimator  $\hat{c}_{n,i}^{-1}$  and shows again that the estimator has a mixed distribution.

Table 3 summarizes the results achieved using the test statistics  $L$ ,  $\hat{L}$ ,  $\hat{F}$  and  $F$  in (19)–(21) and (25), respectively. The empirical QTL detection power is determined by the relative frequency of rejecting the null hypothesis. It is obvious that with increasing

factor  $c_*$  the empirical power increases. The empirical global power should not exceed the value of 90%, because 10% of the repeated simulations created exclusively homozygous sires at random (all  $N=4$  sires are homozygous) and therefore a QTL effect on the within-litter variance could not be detected. Under the null hypothesis  $H_0$  ( $c=1$ ) and chromosome-wise investigations the  $\alpha$  level of 5% holds for the verified test statistics except statistic  $\hat{F}$ . This may be due to the approximative permutation test with only 10 000 re-samples.

For  $c_*$  values 1.2 and larger values, where the QTL detection power is already very high, all tests perform equally well. However, for  $c_*=1.1$  or equivalent  $c=1.088$  the GLM clearly outperforms the weighted regression approach and provides an extra gain of 12% empirical global power. From Table 3 we can see that the tests based on the GLM provide a higher empirical QTL detection power than tests based on weighted regression. Similar results are given under local investigations at  $d=25$  cM (see Table 3).

## 4. Discussion

### (i) LM versus GLM

Applying the test based on the LM requires far less computing time compared with the tests based on the GLM (2 hours vs 17 hours for simulating the dataset and running the tests on a PC with a 3 GHz Intel processor). The surplus of computing time for the GLM is, however, not large compared with the total time and costs usually required for QTL experiments. The benefit of the GLM with respect to empirical

Table 1. Calculation of genotype frequencies within litter; presumed gene frequency  $p_Q=0.5$ ; dominance effect is omitted;  $a$  denotes the additive value

1	2	3	4	5.1			6	7
				Fraction of piglets with genotype				
	Realization of $B$	Genotype daughter $i, j$	Genotype male $i, j$	$QQ$	$Qq$	$qq$	$\mu_{QTL,m}$	Frequency of $b_m$ under condition of $\mathbf{1}_{\{Q\},i,j} \rightarrow$ weights in Table 2
$Q$ from sire $\mathbf{1}_{\{Q\},i,j}=1$	$b_1$	$QQ$	$QQ$	1	0	0	$a$	$p_Q p_Q^2 = \frac{1}{8}$
	$b_2$		$Qq, qQ$	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}a$	$p_Q 2p_Q(1-p_Q) = \frac{1}{4}$
	$b_3$		$qq$	0	1	0	0	$p_Q(1-p_Q)^2 = \frac{1}{8}$
	$b_4$	$Qq$	$QQ$	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}a$	$(1-p_Q)p_Q^2 = \frac{1}{8}$
	$b_5$		$Qq, qQ$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	0	$(1-p_Q)2p_Q(1-p_Q) = \frac{1}{4}$
	$b_6$		$qq$	0	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}a$	$(1-p_Q)(1-p_Q)^2 = \frac{1}{8}$
$q$ from sire $\mathbf{1}_{\{Q\},i,j}=0$	$b_7$	$qQ$	$QQ$	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}a$	$p_Q p_Q^2 = \frac{1}{8}$
	$b_8$		$Qq, qQ$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	0	$p_Q 2p_Q(1-p_Q) = \frac{1}{4}$
	$b_9$		$qq$	0	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}a$	$p_Q(1-p_Q)^2 = \frac{1}{8}$
	$b_{10}$	$qq$	$QQ$	0	1	0	0	$(1-p_Q)p_Q^2 = \frac{1}{8}$
	$b_{11}$		$Qq, qQ$	0	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}a$	$(1-p_Q)2p_Q(1-p_Q) = \frac{1}{4}$
	$b_{12}$		$qq$	0	0	1	$-a$	$(1-p_Q)(1-p_Q)^2 = \frac{1}{8}$

Table 2. Calculation of various expectations of the transformed QTL effect  $\tilde{G}_{i,j,k,m} = G_{i,j,k} - \mathbb{E}(G_{i,j,k} | B = B_m)$ ; dominance effect is omitted;  $a$  denotes the additive value and  $\sigma_{QTL}^2$  denotes the variance of the QTL effect  $G_{i,j,k}$

1	2	3.1			4	5	6
		Value of QTL effect $\tilde{G}_{i,j,k,m}$ of piglets with genotype					
	Realization of $B$	$QQ$	$Qq$	$qq$	$\mathbb{E}\tilde{G}_{i,j,k,m}^2 = \tilde{\sigma}_{QTL,m}^2$	$(\mathbb{E}\tilde{G}_{i,j,k,m})^2 = \tilde{\sigma}_{QTL,m}^4$	$\mathbb{E}\tilde{G}_{i,j,k,m}^4 = \tilde{\mu}_{4,m}$
$Q$ from sire $\mathbf{1}_{\{Q\},i,j}=1$	$b_1$	0	-	-	0	0	0
	$b_2$	$\frac{1}{2}a$	$-\frac{1}{2}a$	-	$\frac{1}{4}a^2$	$\frac{1}{16}a^4$	$\frac{1}{16}a^4$
	$b_3$	-	0	-	0	0	0
	$b_4$	$\frac{1}{2}a$	$-\frac{1}{2}a$	-	$\frac{1}{4}a^2$	$\frac{1}{16}a^4$	$\frac{1}{16}a^4$
	$b_5$	$a$	0	$-a$	$\frac{3}{2}a^2$	$\frac{1}{4}a^4$	$\frac{1}{2}a^4$
	$b_6$	-	$\frac{1}{2}a$	$-\frac{1}{2}a$	$\frac{1}{4}a^2$	$\frac{1}{16}a^4$	$\frac{1}{16}a^4$
Weighted sum					$\frac{1}{4}a^2 = \sigma_{QTL}^2$	$\frac{3}{32}a^4 = \frac{3}{2}\sigma_{QTL}^4$	$\frac{5}{32}a^4 = \frac{5}{2}\sigma_{QTL}^4$
$q$ from sire $\mathbf{1}_{\{Q\},i,j}=0$	$b_7$	$\frac{1}{2}a$	$-\frac{1}{2}a$	-	$\frac{1}{4}a^2$	$\frac{1}{16}a^4$	$\frac{1}{16}a^4$
	$b_8$	$a$	0	$-a$	$\frac{1}{2}a^2$	$\frac{1}{4}a^4$	$\frac{1}{2}a^4$
	$b_9$	-	$\frac{1}{2}a$	$-\frac{1}{2}a$	$\frac{1}{4}a^2$	$\frac{1}{16}a^4$	$\frac{1}{16}a^4$
	$b_{10}$	-	0	-	0	0	0
	$b_{11}$	-	$\frac{1}{2}a$	$-\frac{1}{2}a$	$\frac{1}{4}a^2$	$\frac{1}{16}a^4$	$\frac{1}{16}a^4$
	$b_{12}$	-	-	0	0	0	0
Weighted sum					$\frac{1}{4}a^2 = \sigma_{QTL}^2$	$\frac{3}{32}a^4 = \frac{3}{2}\sigma_{QTL}^4$	$\frac{5}{32}a^4 = \frac{5}{2}\sigma_{QTL}^4$

global power is obvious when  $c_* = 1 \cdot 1$ ; the use of the test statistics  $\hat{L}$  and  $\hat{F}$  should be favoured. The presented test statistics remain suitable in the case of a varying number of daughters per sire. To avoid a loss of QTL detection power, a decreased number of daughters per sire may be adjusted for by considering more half-sib families. It may be interesting to

mention that using the log-link instead of the identity link function leads to slightly less QTL detection power (Wittenburg, 2005).

Thomson (2003) and Lange & Whittaker (2001) specified the detection of a QTL for non-normal traits. Thomson (2003) proposed a model for non-normal data types using normal-based profile

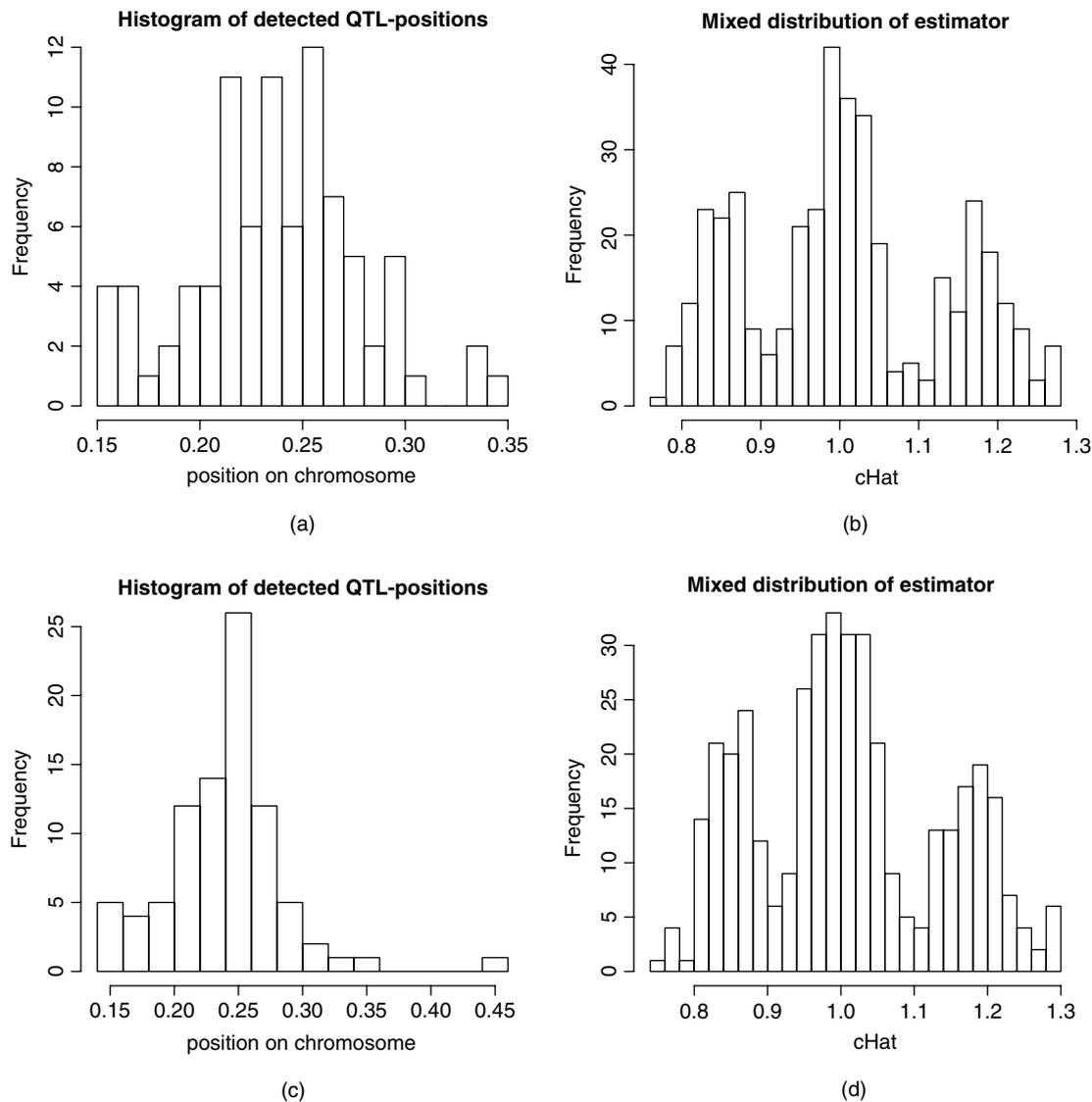


Fig. 2. QTL simulated at 25 cM,  $N=4$  sires,  $n=200$  daughters,  $c^*=1.2$  ( $c=1.177$ ). (a) Detected QTL positions based on the GLM with test statistic  $\hat{L}$ ; (b) histogram of estimator  $c_{n,i}^{-1}$  based on the GLM; (c) detected QTL positions based on the LM with test statistic  $F$ ; (d) histogram of estimator  $c_{n,i}^{-1}$  based on the LM.

log-likelihood and solving the generalized estimating equations. These methods of parameter estimation are comparable to the techniques described above. The essential difference is, that the present work employs the approximate gamma distribution of  $S_{i,j}^2$  and develops the appropriate profile log-likelihood for the estimation of the QTL position parameter  $d$ .

Standard regression interval mapping has been proved to be a robust method for non-normally distributed continuous traits in comparison with non-parametric approaches (Rebaï, 1997). Kadarmideen *et al.* (2000) found similar QTL detection power with LM and GLM for binary traits. However, Yin & Zhang (2006) demonstrated for ordinal data that the GLM outperformed the LM in terms of QTL detection power. Therefore, our results provide another

example in which a GLM should be preferred for QTL mapping.

#### (ii) Analysis of heterogeneous variances

Several authors have dealt with heterogeneity of variances. Mixed Gaussian models for the within-litter standard deviation have already been mentioned (Högberg & Rydhmer, 2000; Damgaard *et al.*, 2003). Foulley *et al.* (1990) put forward a log-linear model for residual variances in order to identify sources of heterogeneity, an idea which has been further extended by SanCristobal *et al.* (1993), Foulley & Quaas (1995) and SanCristobal-Gaudy *et al.* (1998) by including heterogeneous genetic components of variance and random factors affecting variances. A Bayesian approach jointly considering genetic effects

Table 3. Summary of simulation results (10% of repetitions with exclusive homozygous sires); power\_p\_emp denotes the empirical pointwise power evaluated at the simulated QTL position at 25 cM with use of tabulated quantiles of the  $\chi^2$ - and F-distribution; power\_g\_emp is the empirical global power; mean\_detec is the average of detected QTL positions and variance\_detec is the sample variance of estimated positions; the statistic F is based on the LM and L,  $\hat{L}$ ,  $\tilde{F}$  are based on the GLM

		Simulated factor of random deviation $c_*$				
Test statistic		1·0	1·1	1·2	1·3	1·4
power_p_emp	$F$	0·06	0·66	0·90	0·90	0·90
	$L$	0·06	0·74	0·90	0·90	0·90
	$\hat{L}$	0·07	0·73	0·90	0·90	0·90
	$\tilde{F}$	0·06	0·73	0·90	0·90	0·90
power_g_emp	$F$	0·04	0·50	0·88	0·90	0·90
	$L$	0·02	0·52	0·87	0·90	0·90
	$\hat{L}$	0·03	0·62	0·88	0·90	0·90
	$\tilde{F}$	0·06	0·62	0·89	0·90	0·90
mean_detec	$F$	0·4400	0·3568	0·2552	0·2674	0·2647
	$L$	0·4620	0·3555	0·2732	0·2665	0·2630
	$\hat{L}$	0·4619	0·3556	0·2727	0·2656	0·2647
	$\tilde{F}$	0·4620	0·3555	0·2732	0·2665	0·2630
variance_detec	$F$	0·1074	0·0643	0·0158	0·0230	0·0176
	$L$	0·1086	0·0600	0·0280	0·0206	0·0164
	$\hat{L}$	0·1088	0·0600	0·0280	0·0207	0·0163
	$\tilde{F}$	0·1086	0·0600	0·0280	0·0206	0·0164

on mean and residual variance was developed by Sorensen & Waagepetersen (2003). Other extensions comprise mean–variance relationships (Foulley, 2004) in models allowing also for effects of explanatory variables on variance components. The latter approach in particular would presumably be worthwhile to be investigated as an alternative to the methods presented.

(iii) Ambiguity of the parameter  $c^2$

The parameter  $c^2$  denotes the ratio of within-litter variance of  $q$ -daughters compared with  $Q$ -daughters. If the parameter  $c$  is significantly different from 1, it is ambiguous whether the within-litter variance is affected by a raised residual variance, by an enlarged polygenic variance or by an increased QTL variance (see equation (1)). When a non-significant result is observed, a constant within-litter variance could also be generated by, for example, an increased residual variance and decreased polygenic variance. Therefore, it should be kept in mind that the parameter  $c^2$  is just a cumulative effect for any changes in the components of the within-litter variance.

(iv) Granddaughter design

The applied model for the daughter design can be extended, with some modifications, to the granddaughter design. Consider a fixed number of

grandsires, which are mated with unrelated granddams of the population. We select one son per mating and these sons are mated with unrelated dams. One daughter per mating is chosen to analyse the within-litter variance. For each granddaughter we have to calculate the sample variance of observed birth weights within one litter. We may assign the sample variances pooled over all daughters as observation for every sire and apply the techniques of Section 2 with adjusted degrees of freedom (number of total piglets minus number of daughters) in the matrix of weights.

(v) Sex effect

Up to now it has been assumed that no sex effect occurs on the piglet’s mean value or on the variability of phenotypes. But the expected phenotypic value of male piglets may be larger than for female piglets. To consider such an effect, the model (A.1) in Appendix A has to be adjusted. Three different scenarios are possible: (i) No sex effect on the mean and variability of the phenotypic value exists. Therefore, the observed value per daughter consists of the sample variance of all weights at birth within one litter (degrees of freedom: litter size minus 1). (ii) A sex effect acts on the expected phenotypic value but not on its variability. Thus, the observed value per daughter is the sample variance of birth weights pooled over male and female progeny (degrees of

freedom: litter size minus 2). (iii) The sex affects the variability of the phenotypic value. Thus, there are two observed values per litter: the respective sample variances of male and female progeny. In this case it is possible to test a QTL effect as well as a sex by QTL interaction.

(vi) *Other fields of application*

A second series of simulations was started with multiple measurements taken from the genotyped individuals themselves (i.e. the daughters), such as the withers height of cows in a daughter design. In this case the theory simplifies, because the phenotypic value depends only on the individual's own genotype and not on any paternal QTL allele in progeny. Thus, the model for the primary trait consists only of the mean value within daughter  $i, j$  and the normally distributed residual deviation, which is modified by  $c \in (0, \infty)$  if the daughter has inherited the QTL allele  $q$ . The value  $c$  appears directly in this model. Therefore, the parameter  $c^2$  still denotes the ratio of variance within individual of  $q$ -daughters to  $Q$ -daughters. Note that in this case the distribution of  $S_{i,j}^2$  is exactly gamma.

In the simulated example, the withers height of cows was measured 10 times in a daughter design with  $N=4$  sires and  $n=200$  daughters per sire. The simulated QTL effect of  $c=1.2$  was detected in 94% of the repetitions with use of the test statistic  $F$  and  $\hat{L}$  (six repetitions created exclusively homozygous sires). When the withers height was measured three times, the simulated QTL effect of  $c=1.3$  was detected in 70% of the repetitions based on the test statistic  $F$  and in 85% based on the statistic  $\hat{L}$  (seven repetitions with only homozygous sires).

In plants and laboratory animals, a panel of recombinant inbred lines (RILs; e.g. Broman, 2005) can be produced for QTL mapping purposes. As all members of a certain RIL share the same genotype but may vary in their phenotype, RILs can serve as a well-suited tool for mapping QTL effects on within-genotype variability. Essentially a panel of RILs can be treated with the methods presented in this article, when the data are analysed as a single half-sib family (or backcross) in the same way as in the cow example.

There are possible applications in plants which closely resemble the repeated measurements of withers height of cows. For example, one could examine some characteristic of tomato fruits as a multiple measurement of a tomato plant. Again, in this application the theory simplifies because of lack of a paternal genetic effect on the fruits (tomatoes are almost purely maternal tissue). Moreover, the sample variances per panicle may be pooled over all panicles to generate one observation per tomato plant and to

consider some effect of panicle on the phenotypic mean of the fruits.

(vii) *Gene frequency*

The GLM (6) and LM (24) do not include a parameter for the gene frequency  $p_Q$  of the QTL allele  $Q$ ; in the simulations the gene frequency was assumed to be one-half. Looking at the components which are affected by  $p_Q$ , the ratio  $c^2$  in (2) is obvious. The parameter  $c^2$  depends on the variance of the transformed QTL effect  $\tilde{G}_{i,j,k}$  in (A.2) of Appendix A, which is calculated on the basis of known genotype frequencies within one litter (Table 2). Thus, if  $p_Q$  differs essentially from one-half, the condition  $\mathbb{E}(\tilde{G}_{i,j,k} | \mathbf{1}_{\{Q\},i,j}=1) = \mathbb{E}(\tilde{G}_{i,j,k} | \mathbf{1}_{\{Q\},i,j}=0) = \sigma_{QTL}^2$  is no longer satisfied. Consequently, the ratio  $c^2$  deviates from 1 even though a QTL effect on the within-litter variance does not exist. To consider the consequences of a gene frequency being different from one-half, the simulations under the null hypothesis ( $c_*=1$ ) have been repeated for the statistics  $\hat{L}$  in (20) and  $F$  in (25) using gene frequencies of 0.10 and 0.25. Without giving the detailed simulation results we remark that the  $\alpha$  level of 5% was always maintained in both tests. This can be referred to a relatively small variance of the additive QTL effect in comparison with the other variance components (see equation (1)). Hence in most practical applications with a possible paternal QTL effect on progeny (piglets), a gene frequency different from 0.5 can be neglected and the theory provided may still serve as a good approximation. In all other cases (tomatoes, withers height of cows) where repeated measurements are taken from either the genotyped individuals themselves or from purely maternal tissues, our theory is exact and unaffected by the allele frequency at the QTL.

### Appendix A. Distribution of the sample variance

The phenotypic value  $Y_{i,j,k}$  of piglet within one litter is described by the following model consisting of independent components:

$$Y_{i,j,k} = \mu_{i,j} + A_{i,j,k} + G_{i,j,k} + \mathbf{1}_{\{Q\},i,j} E_{i,j,k} + c_*(1 - \mathbf{1}_{\{Q\},i,j}) E_{i,j,k}, \quad (\text{A.1})$$

where  $i$  indicates the sire,  $j$  the daughter per sire and  $k$  the piglet. The constant litter mean is denoted by  $\mu_{i,j}$ . The random components are the mendelian sampling effect  $A_{i,j,k} \sim N(0, \frac{1}{2}\sigma_{polygene}^2)$ , the additive QTL effect  $G_{i,j,k}$  and the random deviation  $E_{i,j,k} \sim N(0, \sigma_e^2)$ .

The indicator function  $\mathbf{1}_{\{Q\},i,j}$  takes the value 1 if the daughter  $i, j$  inherits the QTL allele  $Q$  at the unknown QTL position from the sire  $i$ . In the case of

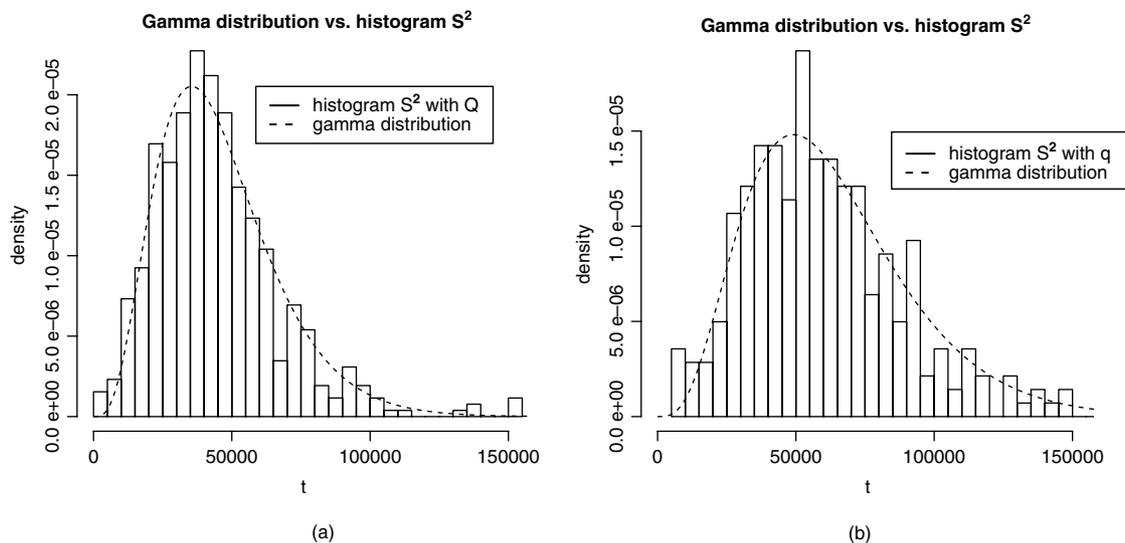


Fig. 3. (a) Histogram of simulated values  $s_{i,j}^2$  when inheriting  $Q$  versus density of a gamma distribution with parameters in (A.6); (b) histogram of simulated values  $s_{i,j}^2$  when inheriting  $q$  versus density of a gamma distribution with parameters in (A.6), where  $\tau^2$  is replaced by  $\tau_*^2$  and  $c_* = 1.2$ .

inheriting  $q$ , the random deviation  $E_{i,j,k}$  of model (A.1) is modified by the factor  $c_* \in (0, \infty)$ . Let  $\mathbf{1}_{\{Qq, qQ\}, i, j, k}$  be the indicator function with value 1 if the piglet has a heterozygous genotype,  $\mathbf{1}_{\{QQ\}, i, j, k}$  and  $\mathbf{1}_{\{qq\}, i, j, k}$  in the case of genotype  $QQ$  and  $qq$ , respectively. The additive QTL effect  $G_{i,j,k}$  depends on the piglet's genotype and has a three-point distribution. In the absence of a dominance effect its probabilities are

$$\begin{aligned} Pr(G_{i,j,k} = -a) &= Pr(\mathbf{1}_{\{qq\}, i, j, k} = 1), \\ Pr(G_{i,j,k} = -0) &= Pr(\mathbf{1}_{\{Qq, qQ\}, i, j, k} = 1), \\ Pr(G_{i,j,k} = a) &= Pr(\mathbf{1}_{\{QQ\}, i, j, k} = 1), \end{aligned}$$

where  $a$  is some unknown constant which, in view of (A.1), is called the additive value. The probability that the piglet has a special genotype, e.g.  $Pr(\mathbf{1}_{\{QQ\}, i, j, k} = 1)$ , is only allocatable in combination with the unobservable parental genotypes. Therefore, an additional random variable  $B$  is required, which denotes the random combination of QTL alleles of the daughter  $i, j$  and her associated male (mating types) as described in columns 3 and 4 of Table 1. The realizations of  $B$  are denoted by  $b_m, m = 1, \dots, 12$  (column 2 of Table 1). The conditional distributions of  $G_{i,j,k}$  given  $B$  lead to the distribution of the additive QTL effect  $G_{i,j,k}$ . For a fixed index  $m$  the phenotypic values of the offspring within one litter are independently and identically distributed. Because  $\mu_{QTL, m} = \mathbb{E}(G_{i,j,k} | B = b_m) \neq 0$  (for some exceptions see Table 1), we set

$$\tilde{G}_{i,j,k} = G_{i,j,k} - \mathbb{E}(G_{i,j,k} | B), \tag{A.2}$$

which now satisfies  $\mathbb{E}(\tilde{G}_{i,j,k} | B = b_m) = 0 \quad \forall m$ . If  $A_\theta = \{\mathbf{1}_{\{Q\}, i, j} = \theta\}, \theta \in \{0, 1\}$ , then

$$\begin{aligned} \mathbb{E}(\tilde{G}_{i,j,k}^2 | A_1) &= \sum_{m=1}^6 Pr(B = b_m | A_1) \mathbb{E}(\tilde{G}_{i,j,k}^2 | A_1 \cap \{B = b_m\}) \\ &= \sum_{m=1}^6 Pr(B = b_m | A_1) \tilde{\sigma}_{QTL, m}^2 = \frac{1}{4} a^2, \\ \mathbb{E}(\tilde{G}_{i,j,k}^2 | A_0) &= \sum_{m=7}^{12} Pr(B = b_m | A_0) \tilde{\sigma}_{QTL, m}^2 = \frac{1}{4} a^2, \end{aligned} \tag{A.3}$$

where the different values of  $\tilde{\sigma}_{QTL, m}^2 = \mathbb{E}(\tilde{G}_{i,j,k}^2 | A_\theta \cap \{B = b_m\})$  are listed in Table 2, column 4 and the corresponding row  $b_m$ . The variance of the additive QTL effect is defined by  $\sigma_{QTL}^2 := \frac{1}{4} a^2$ . We see from (A.3) that  $\mathbb{V}(\tilde{G}_{i,j,k} | A_\theta) = \sigma_{QTL}^2, \theta \in \{0, 1\}$ .

To eliminate  $\mu_{i,j}$  that appears in (A.1), we introduce  $X_{i,j,k}$  by

$$\begin{aligned} X_{i,j,k} &= Y_{i,j,k} - \mu_{i,j} - \mathbb{E}(G_{i,j,k} | B) \\ &= A_{i,j,k} + \tilde{G}_{i,j,k} + \mathbf{1}_{\{Q\}, i, j} E_{i,j,k} + c_*(1 - \mathbf{1}_{\{Q\}, i, j}) E_{i,j,k}. \end{aligned} \tag{A.4}$$

If  $n_{i,j}$  denotes the litter size of daughter  $i, j$ , then the sample variance is

$$\begin{aligned} S_{i,j}^2 &= \frac{1}{n_{i,j} - 1} \sum_{k=1}^{n_{i,j}} (X_{i,j,k} - \bar{X}_{i,j, \cdot})^2 \\ \text{with } \bar{X}_{i,j, \cdot} &= \frac{1}{n_{i,j}} \sum_{k=1}^{n_{i,j}} X_{i,j,k}. \end{aligned} \tag{A.5}$$

The within-litter variance depends on the paternal QTL allele of the daughter  $i, j$ . The conditional

expectation of  $X_{i,j,k}$  and  $S_{i,j}^2$  given the events  $A_\theta$ ,  $\theta \in \{0, 1\}$ , are

$$\begin{aligned} \mathbb{E}(X_{i,j,k}|A_\theta) &= \mathbb{E}(A_{i,j,k} + \tilde{G}_{i,j,k} + \mathbf{1}_{\{Q\},i,j}E_{i,j,k} \\ &\quad + c_*(1 - \mathbf{1}_{\{Q\},i,j})E_{i,j,k}|A_\theta) = 0, \\ \mathbb{E}(S_{i,j}^2|A_1) &= \mathbb{V}(X_{i,j,k}|A_1) \\ &= \mathbb{E}([A_{i,j,k} + \tilde{G}_{i,j,k} + \mathbf{1}_{\{Q\},i,j}E_{i,j,k} \\ &\quad + c_*(1 - \mathbf{1}_{\{Q\},i,j})E_{i,j,k}]^2|A_1) \\ &= \frac{1}{2}\sigma_{polygene}^2 + \sigma_{QTL}^2 + \sigma_e^2 =: \tau^2 + \sigma_{QTL}^2, \\ \mathbb{E}(S_{i,j}^2|A_0) &= \mathbb{V}(X_{i,j,k}|A_0) \\ &= \frac{1}{2}\sigma_{polygene}^2 + \sigma_{QTL}^2 + c_*^2\sigma_e^2 =: \tau_*^2 + \sigma_{QTL}^2. \end{aligned}$$

Because the investigated sample variance includes the non-normally distributed variable  $\tilde{G}_{i,j,k}$ , the conditional distribution of  $S_{i,j}^2$  is not a  $\chi^2$  distribution. Therefore, an approximation with a gamma distribution  $\Gamma_{\mu, \nu_{i,j}}$  with expectation  $\mu$  and variance  $\frac{\mu^2}{\nu_{i,j}}$  will be considered.

**Assertion 1** *The conditional variance of  $S_{i,j}^2$  in (A.5) given  $\{\mathbf{1}_{\{Q\},i,j}=1\}$  is*

$$\begin{aligned} \mathbb{V}(S_{i,j}^2|\mathbf{1}_{\{Q\},i,j}=1) &= \frac{2}{n_{i,j}-1} \\ &\times \left[ (\tau^2 + 2\sigma_{QTL}^2)\tau^2 + \sigma_{QTL}^4 \left( \frac{n_{i,j}}{4} + \frac{1}{4} + \frac{1}{n_{i,j}} \right) \right]. \end{aligned}$$

*An analogous statement holds under the condition of  $\{\mathbf{1}_{\{Q\},i,j}=0\}$  if  $\tau_*^2$  is used instead of  $\tau^2$ .*

**Proof:** See Supplementary Appendixes.

Fig. 3a and b show a simulated histogram of  $S_{i,j}^2$ . Paternal QTL alleles  $Q$  and  $q$  were distinguished. The figures suggest approximating the conditional distribution of  $S_{i,j}^2$  in (A.5) by a two-parameter gamma distribution  $\Gamma_{\mu, \nu_{i,j}}$  with

$$\begin{aligned} \mu &= \tau^2 + \sigma_{QTL}^2, \\ \nu_{i,j} &= \frac{n_{i,j}-1}{2} \frac{(\tau^2 + \sigma_{QTL}^2)^2}{(\tau^2 + 2\sigma_{QTL}^2)\tau^2 + \sigma_{QTL}^4 \left( \frac{n_{i,j}}{4} + \frac{1}{4} + \frac{1}{n_{i,j}} \right)}. \end{aligned} \tag{A.6}$$

Note that the presented model (A.1) is an all-purpose model that covers two assumptions. First, similar to the ideas of Hill & Zhang (2004), the QTL may affect both the phenotypic mean and its variability. In this case, the distribution of the daughter's trait is approximated by a gamma distribution as shown above. Second, the QTL affects only the variability of the phenotype. Then the general model (A.1) simplifies and the distribution of the within-subject sample variance is exactly gamma.

The research project was financially supported by the H. Wilhelm Schaumann Stiftung.

## References

Bidanel, J.-P., Milan, D., Iannuccelli, N., Amigues, Y., Boscher, M.-Y., Bourgeois, F., Caritez, J.-C., Gruand, J., le Roy, P., Lagant, H., Quintanilla, R., Renard, C., Gellin, J., Ollivier, L. & Chevalet, C. (2001). Detection of quantitative trait loci for growth and fatness in pigs. *Genetics Selection Evolution* **33**, 289–309.

Bolet, G., Garreau, H., Joly, T., Theau-Clement, M., Hurtaud, J. & Bodin, L. (2005). Genetic homogenization of birth weights in rabbits: evolution of the characteristics of the genital tract after two generations of selection. In *Book of Abstracts of the 56th Annual Meeting of the EAAP*. Wageningen Academic Publishers, p. 80.

Box, G. E. P. & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society B* **26**, 211–252.

Broman, K. W. (2005). The genomes of recombinant inbred lines. *Genetics* **169**, 1133–1146.

Churchill, G. A. & Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963–971.

Damgaard, L. H., Rydhmer, L., Løvendahl, P. & Grandinson, K. (2003). Genetic parameters for within-litter variation in piglet birth weight and change in within-litter variation during suckling. *Journal of Animal Science* **81**, 604–610.

Fahrmeir, L. & Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics* **13**, 342–368.

Fahrmeir, L. & Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. New York: Springer.

Foulley, J.-L. (2004). Including mean–variance relationships in heteroskedastic mixed linear models: theory and application. URL <http://interstat.statjournals.net> (retrieved 30 May 2007).

Foulley, J. L., Gianola, D., SanCristobal, M. & Im, S. (1990). A method for assessing extent and sources of heterogeneity of residual variances in mixed linear models. *Journal of Dairy Science* **73**, 1612–1624.

Foulley, J. L. & Quaas, R. L. (1995). Heterogeneous variances in gaussian linear mixed models. *Genetics Selection Evolution* **27**, 211–228.

Haley, C. S. & Knott, S. A. (1992). A simple regression method for mapping quantitative trait loci in lines crosses using flanking markers. *Heredity* **69**, 315–324.

Hill, W. G. & Zhang, X.-S. (2004). Effects on phenotypic variability of directional selection arising through genetic differences in residual variability. *Genetical Research* **83**, 121–132.

Högberg, A. & Rydhmer, L. (2000). A genetic study of piglet growth and survival. *Acta Agricultura Scandinavica, Section A, Animal Science* **50**, 300–303.

Jørgensen, B. (1987). Exponential dispersion models. *Journal of the Royal Statistical Society B* **49**, 127–162.

Kadarmideen, H. N., Janss, L. L. G. & Dekkers, J. C. M. (2000). Power of quantitative trait locus mapping for polygenic binary traits using generalized and regression interval mapping in multi-family half-sib designs. *Genetical Research* **76**, 305–317.

Lange, C. & Whittaker, J. C. (2001). Mapping quantitative trait loci using generalized estimating equations. *Genetics* **159**, 1325–1337.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.

- R Development Core Team (2005). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. URL <http://www.R-project.org> (retrieved 24 October 2005).
- Rebaï, A. (1997). Comparison of methods for regression interval mapping in QTL analysis with non-normal traits. *Genetical Research* **69**, 69–74.
- Roehe, R. (1999). Genetic determination of individual birth weight and its association with sow productivity traits using Bayesian analyses. *Journal of Animal Science* **77**, 330–343.
- Ros, M., Sorensen, D., Waagepetersen, R., Dupont-Nivet, M., SanCristobal, M., Bonnet, J. C. & Mallard, J. (2004). Evidence for genetic control of adult weight plasticity in the snail *Helix aspersa*. *Genetics* **168**, 2089–2097.
- SanCristobal, M., Foulley, J. L. & Manfredi, E. (1993). Inference about multiplicative heteroskedastic components of variance in a mixed linear gaussian model with an application to beef cattle breeding. *Genetics Selection Evolution* **25**, 3–30.
- SanCristobal-Gaudy, M., Elsen, J. M., Bodin, L. & Chevalet, C. (1998). Prediction of the response to a selection for canalisation of a continuous trait in animal breeding. *Genetics Selection Evolution* **30**, 423–451.
- Seber, G. A. F. (1977). *Linear Regression Analysis*. New York: Wiley.
- Sorensen, D. & Waagepetersen, R. (2003). Normal linear models with genetically structured residual variance heterogeneity: a case study. *Genetical Research* **82**, 207–222.
- Thomson, P. C. (2003). A generalized estimating equations approach to quantitative trait locus detection of non-normal traits. *Genetics Selection Evolution* **35**, 257–280.
- Weller, J. I. & Wyler, A. (1992). Power of different sampling strategies to detect quantitative trait loci variance effects. *Theoretical and Applied Genetics* **83**, 582–588.
- Wittenburg, D. (2005). Lineare und verallgemeinerte lineare Modelle zum Nachweis von QTL-Effekten auf die Varianz wiederholter Messungen. Diploma thesis, Universität Rostock.
- Yin, Z.-J. & Zhang, Q. (2006). Mapping quantitative trait loci for ordinal traits using the generalized linear model in half-sib designs. *Animal Research* **55**, 245–255.