

# Biased estimators of quantitative trait locus heritability and location in interval mapping

M Bogdan<sup>1,2</sup> and RW Doerge<sup>2,3</sup>

<sup>1</sup>Institute of Mathematics, Wrocław University of Technology, Wrocław, Poland; <sup>2</sup>Department of Statistics, Purdue University, West Lafayette, IN 47907, USA; <sup>3</sup>Department of Agronomy, Purdue University, West Lafayette, IN 47907, USA

In many empirical studies, it has been observed that genome scans yield biased estimates of heritability, as well as genetic effects. It is widely accepted that quantitative trait locus (QTL) mapping is a model selection procedure, and that the overestimation of genetic effects is the result of using the same data for model selection as estimation of parameters. There are two key steps in QTL modeling, each of which biases the estimation of genetic effects. First, test procedures are employed to select the regions of the genome for which there is significant evidence for the presence of QTL. Second, and most important for this demonstration, estimates of the genetic effects are reported only at the locations for which the evidence is maximal. We demonstrate that even when we know there is just one QTL present (ignoring

the testing bias), and we use interval mapping to estimate its location and effect, the estimator of the effect will be biased. As evidence, we present results of simulations investigating the relative importance of the two sources of bias and the dependence of bias of heritability estimators on the true QTL heritability, sample size, and the length of the investigated part of the genome. Moreover, we present results of simulations demonstrating the skewness of the distribution of estimators of QTL locations and the resulting bias in estimation of location. We use computer simulations to investigate the dependence of this bias on the true QTL location, heritability, and the sample size.

*Heredity* (2005) **95**, 476–484. doi:10.1038/sj.hdy.6800747; published online 28 September 2005

**Keywords:** quantitative trait locus; interval mapping; biased estimators; model selection

## Introduction

Interval mapping (Thoday, 1961; Lander and Botstein, 1989) is a well-known method for identifying and locating quantitative trait loci (QTL). Extended (Zeng, 1993, 1994) to include, and conditional on, additional regions of the genomes, interval mapping has been expanded to composite interval mapping, and both are currently enjoying a renewed popularity as statistical tools for identifying genomic regions associated with quantitative data (eg gene expression QTL or eQTL). Where QTL mapping was once an end point for experimentation, it is now a jumping-off point for exciting genomic applications (Jansen and Nap, 2001; Doerge, 2002; Wayne and McIntyre, 2002; Schadt *et al*, 2003). Since QTL are genomic regions associated with variation of a quantitative trait, it is useful to estimate effects of QTL and heritability, as well as their genetic map location for the purpose of narrowing in on important components involved in the genetic architecture of many experimental populations (Mackay, 2001).

In practice, it has been observed that the estimators of effects for QTL estimated via interval mapping, and its extensions (Zeng, 1993, 1994), can be severely inflated (see Beavis, 1994, 1998; Utz *et al*, 2000; Allison *et al*, 2002,

and references given there). It is already known that the bias in the estimation of QTL effects arises naturally from the fact that standard QTL mapping can be interpreted as a model selection procedure and that it uses the same data to choose the best model and estimate its parameters (Utz *et al*, 2000; Broman, 2001; Göring *et al*, 2001; Ball, 2001; Allison *et al*, 2002). There are two main steps in the model selection process each of which adds to the bias of the estimators of genetic effects. First, in performing the genome scan, all QTL are identified when the corresponding test statistic (log-odds score or likelihood-ratio test) exceeds a certain threshold. The bias resulting from using the same data for testing and estimation has been very well investigated (Broman, 2001; Göring *et al*, 2001; Allison *et al*, 2002; Xu, 2003). The second, and more subtle point, in QTL mapping relies on choosing the locations for which the evidence of a QTL is maximal, and then reporting only the estimates of the genetic effects obtained for these optimal locations. In QTL mapping, while many test statistics are significant, only the location that provides the largest test statistic value along with the estimated QTL effect(s) for that location is reported. Although Broman (2001) and Allison *et al* (2002) suggest that this latter step is also a source of the bias, they do not discuss this issue in detail.

In this work, we explore the second source of bias. Namely, we demonstrate via interval mapping that even when one QTL is assumed present (ie ignoring the testing step), model selection is still performed and the resulting estimator of the genetic effect is biased. Furthermore, the relative importance of the two sources

Correspondence: RW Doerge, Department of Statistics, Purdue University, 1399 Mathematical Sciences Building, West Lafayette, IN 47907-1399, USA. E-mail: doerge@purdue.edu

Received 29 June 2004; accepted 18 July 2005; published online 28 September 2005

of bias is investigated with respect to the dependence of bias of heritability estimators on true QTL heritability, sample size, and the length of the genome that is tested. We also demonstrate that the distribution of estimators for QTL location is skewed toward the middle of the chromosome under investigation, which in turn results in the bias of QTL location.

Although we explain the phenomenon of biased QTL effects and location via interval mapping, it is well known for cases involving multiple QTL that the estimators provided by simple interval mapping can still be biased because the effects of the other QTL are neglected (ie employing a single QTL model). To address this problem improved methods, like composite interval mapping (CIM) (Zeng, 1993, 1994), multiple QTL mapping (MQM) (Jansen, 1993), or multiple interval mapping (MIM) (Kao *et al*, 1999), have been developed. These methods search the entire genome and report QTL effects at the locations where the test statistics are largest and exceed a given threshold. The model selection process that is incorporated in these more advanced methods (eg CIM, MQM, MIM) is the same as that used for traditional interval mapping; therefore, the resulting estimators of QTL effects and heritability will also suffer from the bias issue that is the focus of this research.

## Methods

### Explanation of bias of heritability estimates resulting from model selection

Consider a backcross population where  $X_i$  denotes the QTL genotype for the  $i$ th individual, and allow it to take on two values (Kao *et al*, 1999):  $X_i = \frac{1}{2}$  if the  $i$ th individual is homozygous at a QTL and  $X_i = -\frac{1}{2}$  if it is heterozygous. We assume that the relationship between the quantitative trait value  $Y_i$  and a QTL genotype  $X_i$  is described by a normal regression model

$$Y_i = \mu + aX_i + \xi_i \quad (1)$$

where  $\mu$  and  $a$  are the overall mean and QTL effect parameters, respectively, and  $\xi_i$  is the error term (environmental noise), which is normally distributed with mean 0 and variance  $\sigma^2$ .

Interval mapping is based on an estimated genetic map and the assumption that a putative QTL is at a particular location. Each incremental location is flanked by a pair of markers that are related to each other via an *a priori* estimated recombination value  $r$ . The probability of recombination between a QTL and a left flanking marker is denoted by  $r_1$ , and when the proposed location of the QTL changes incrementally across the interval, say left to right, the range of  $r_1$  is restricted to  $0 \leq r_1 \leq r$ . After fixing the QTL at a specific location in the interval (which corresponds to fixing  $r_1$ ), the known flanking marker genotypes are used to assign conditional probabilities to the two possible QTL genotypes. We denote these probabilities (Lander and Botstein, 1989) by  $p_i = P(X_i = \frac{1}{2})$  and  $1 - p_i = P(X_i = -\frac{1}{2})$ . Because  $X_i$  can take on two values, the distribution of the trait values in a backcross population is not normal; instead, it is a mixture of two normal distributions with means  $\mu + \frac{1}{2}a$ , and  $\mu - \frac{1}{2}a$ , respectively (Doerge *et al*, 1997). The density

of this distribution is given by

$$f_i(y) = p_i \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y - \mu - 0.5a)^2}{2\sigma^2}\right) + (1 - p_i) \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y - \mu + 0.5a)^2}{2\sigma^2}\right) \quad (2)$$

There are no closed-form calculations for the maximum-likelihood estimators of parameters  $\mu$ ,  $a$ , and  $\sigma^2$  relative to  $r_1$ ; therefore, the likelihood

$$L(\mu, a, \sigma^2) = \prod_{i=1}^n f_i(Y_i) \quad (3)$$

may be maximized by employing the EM algorithm (Dempster *et al*, 1977). For an application of the EM algorithm to interval mapping, see Jansen and Stam (1994) and Kao and Zeng (1997). The maximization procedure is repeated at each increment through the interval, and then over a dense grid of locations, or marker intervals, covering the genome. At each location, the likelihood of the fitted model is recorded, and the location that maximizes the likelihood function is regularly used as a point estimator of QTL location with the corresponding estimate  $a$  taken as the point estimate for the QTL effect. Therefore, interval mapping can be viewed as a process that selects from statistical models corresponding to different QTL locations. Since the likelihood of a given model depends on the estimate of heritability (see Appendix A1), the model selection process chooses models with the highest heritability estimates, thus leading to overestimation of this quantity.

We demonstrate the overestimation of heritability in more detail by considering a situation where the markers are densely spaced and no model evaluations within the interval are performed. Essentially, the interval mapping process, in this extreme situation, reduces to a regression over markers and chooses the marker that results in the highest coefficient of determination  $R^2$  (ie explains the most phenotypic variation) as a candidate location for a QTL. To understand this, assume that a QTL is located at the first marker ( $\mathbf{M}_1$ ) such that the quantitative trait data are distributed according to the normal regression model

$$Y_i = \mu + aZ_{1i} + \xi_i \quad (4)$$

where  $Z_{1i}$  is the genotype of the  $i$ th individual at first marker  $\mathbf{M}_1$  and  $\xi_i \sim N(0, \sigma^2)$ . If the position of the QTL is known, then the problem of estimating the genetic effect  $a$  and heritability  $h^2$  (proportion of the phenotypic variance explained by the QTL) reduces to a standard regression problem. Following Utz *et al* (2000), we use an asymptotically unbiased estimator of heritability

$$\hat{h}_1^2 = R_1^2 - \frac{1 - R_1^2}{n - 2} = \frac{(n - 1)R_1^2 - 1}{n - 2} \quad (5)$$

where  $R^2$  is the coefficient of determination for the model with the genotype of the first marker  $\mathbf{M}_1$  as the predictor variable, and  $n$  is the sample size. Extending this idea, let  $\hat{h}_i^2$  denote the estimator of heritability if the  $i$ th marker is taken as a predictor variable. If we continue in this manner by selecting a model not within an interval, but rather only at the markers,  $k$  markers are considered and the marker yielding the highest  $R^2$  is

chosen. The corresponding estimator of the heritability is equal to

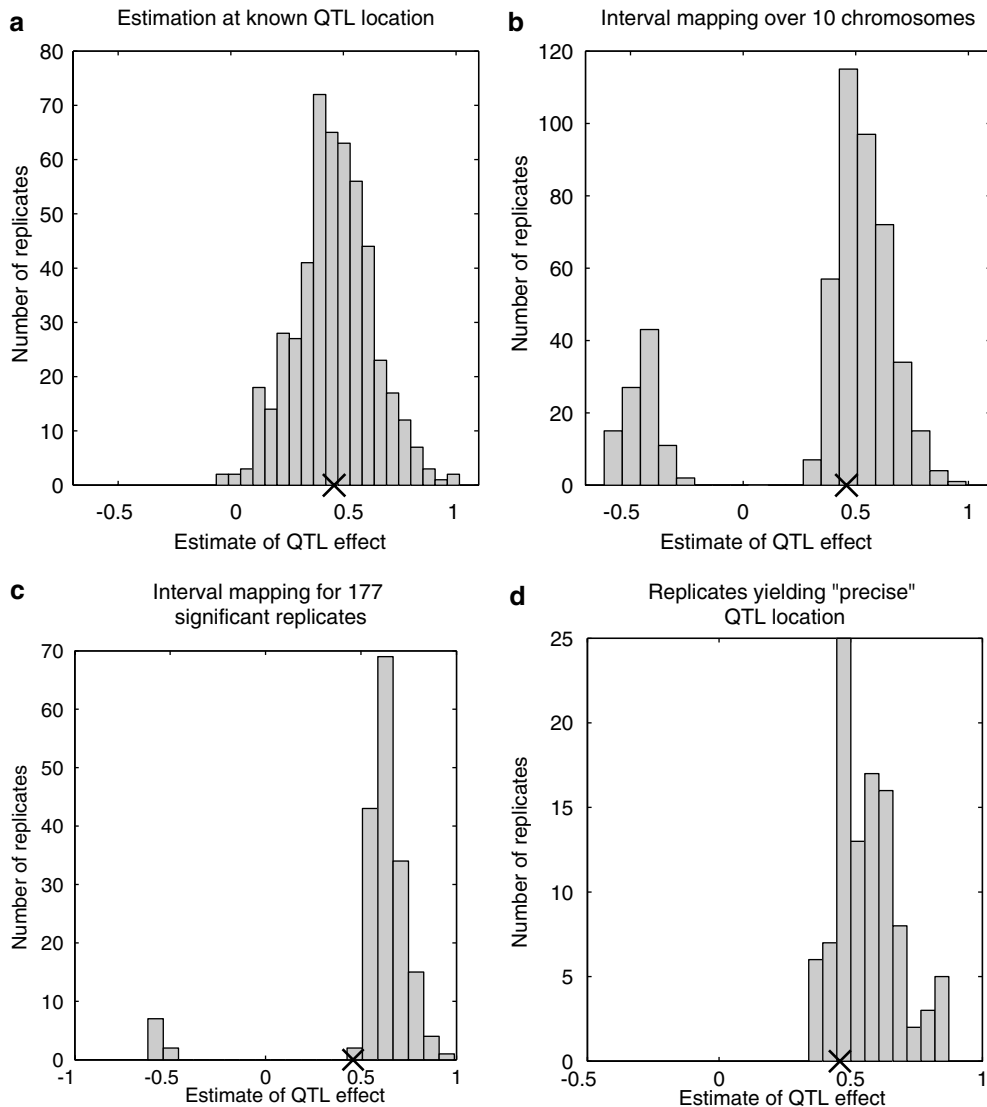
$$\hat{h}^2 = \max_{1 \leq i \leq k} \hat{h}_i^2 \geq \hat{h}_i^2 \quad (6)$$

For each marker  $i$ , the distribution of  $\hat{h}_i^2$  is continuous on the support  $[-(1/(n-2)), 1]$ . When  $k > 1$ , the probability that  $\hat{h}^2 > \hat{h}_i^2$  is larger than zero (proof of this is included in Appendix A1). Therefore, the distribution of the heritability estimate  $\hat{h}^2$ , when compared to the distribution of the approximately unbiased estimator  $\hat{h}_i^2$ , is shifted in the direction of larger values. There is a relatively large probability that  $\hat{h}^2 > \hat{h}_i^2$  when the true heritability is small. In other words, there is a greater chance that we will realize an estimated QTL effect larger than the correct one at the proper location in the genome. It is important to note that the conditional distribution of the estimator of heritability, conditioned on the event that we select the proper QTL location, is also shifted to the right (ie larger

values) with respect to the distribution of  $\hat{h}_i^2$ . This is because we select the correct QTL location only when the estimated heritability corresponding to the true model is large enough to outperform estimated heritabilities corresponding to other investigated locations, which may be large just by random chance. The impact of this is that the small values for the heritability estimates are effectively removed from the resulting conditional distribution (see Figure 1d). Further ramifications of this overestimation are carried through to the estimator of the QTL effect,  $\hat{a}_i$ , when viewed as a function of the coefficient of determination,

$$\hat{a}_i^2 = R_i^2 \frac{s_Y^2}{s_{Z_i}^2} \quad (7)$$

where  $s_{Z_i}^2$  and  $s_Y^2$  are the estimators of variance for  $Z_i$  (the genotype of putative QTL) and  $Y$  (the quantitative trait), respectively. Therefore, the estimators of the magnitude



**Figure 1** Histograms of the estimators of the QTL effect under three different estimation procedures. Results based on 500 replicates consisting of 200 individuals from backcross population. Trait data are generated according to model (1) with parameters  $\mu = 0$ ,  $a = 0.4588$ , and  $\sigma = 1$ , and the QTL is located 52.5 cM from the top end of chromosome 1. Panel (d) reports the results for replicates where the estimated QTL location was within 10 cM of the true value. 'X' marks the true value of QTL effect.

of QTL effects (their absolute values) reported from interval mapping are also positively biased. In this work, we demonstrate the bias of QTL effects assuming that the search is performed only at marker positions. When the genetic map is sparse and the search is performed at intermarker positions, the number of points at which the models are fitted is larger than the number of markers, and the problem of bias of QTL effects may be even greater.

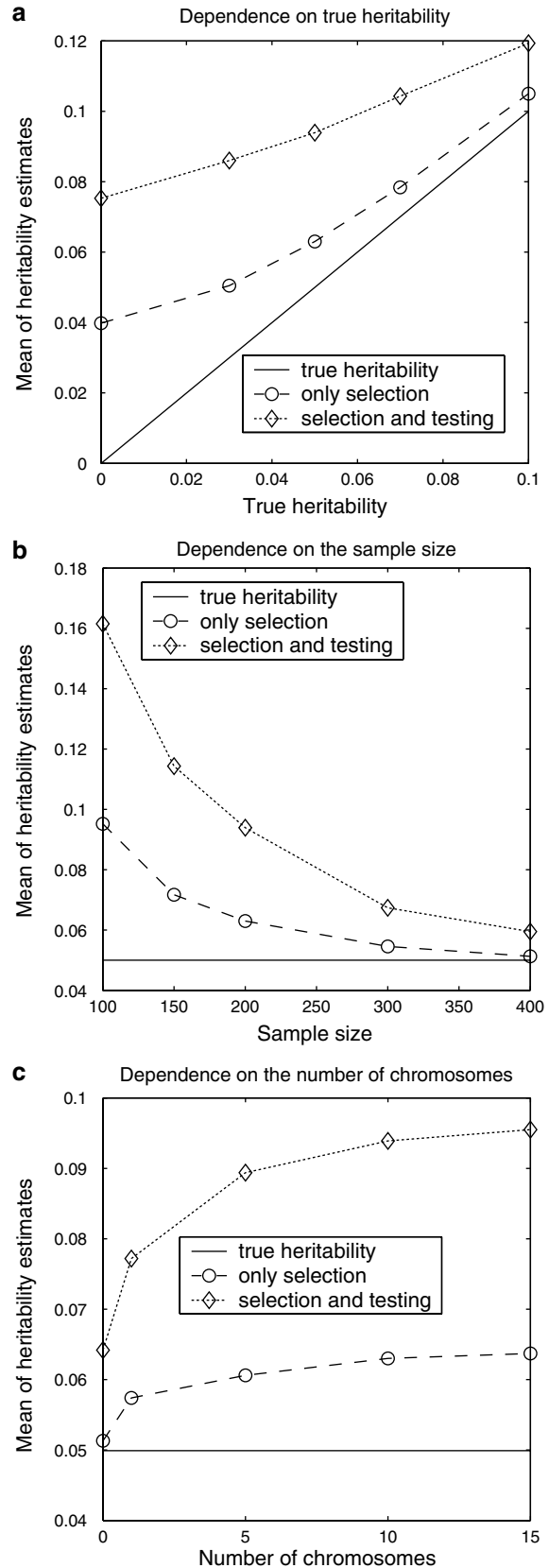
## Simulations

We investigate the bias in the estimators of QTL heritability and QTL location via a simulation study. Marker and QTL data (genotypes) were simulated on up to 15 chromosomes each of length 225 cM for a sample size that ranged from 100 to 500. The experimental system is a backcross design, and the markers are located every 15 cM. In experiments demonstrating the bias of heritability, a single QTL was fixed at 52.5 cM from the left (top) end of chromosome 1. For the purpose of illustrating the bias of location, we used different QTL locations spanning the whole of chromosome 1. For each combination of sample size and length of the investigated portion of the genome, thresholds for the likelihood ratio statistic to detect the presence of QTL at the significance level  $\alpha = 0.05$  were simulated using 1000 replicates from the null distribution with no QTL.

## Results

For the first investigation, marker and QTL information were simulated on 10 chromosomes for 200 backcross individuals. The trait data were simulated according to model (1) with  $\mu = 0$ ,  $a = 0.4588$ , and  $\sigma = 1$  ( $h^2 = 0.05$ ). A total of 500 replicate data sets were used to investigate the estimators for the QTL effect and heritability at the known (52.5 cM from the top end of chromosome 1) QTL location, and when the search was performed on all 10 chromosomes. To estimate heritability, we first estimated the genetic effect  $a$  and the coefficient of determination  $R^2$  (7), replacing the estimator of the variance for  $Z_i$  (the genotype of putative QTL) by the true value ( $\frac{1}{4}$ ). We then calculated an estimate of heritability according to equation (5). Our simulations strongly suggest that the bias of the resulting estimator at a proper QTL location is negligible; thus, the reported bias resulting from interval mapping is related only to the model selection process and testing.

Figure 1a illustrates that when the QTL location is known, the estimator of the QTL effect is approximately unbiased with a mean 0.4567 (true value 0.4588) and the corresponding mean heritability 0.0513 (true value 0.05). When the QTL location is assumed unknown, and the adjacent intervals searched using interval mapping, a histogram (Figure 1b) of the QTL effects illustrates no values near zero. This reflects that the estimated QTL effects achieve considerably large values randomly throughout the genome. Since the true QTL effect was relatively small, the probability that the estimated heritability will have a maximum at a location away from the true location is quite high. As illustrated (Figure 1b), it is also possible that the estimator of the QTL effect can be negative. Such a phenomenon usually

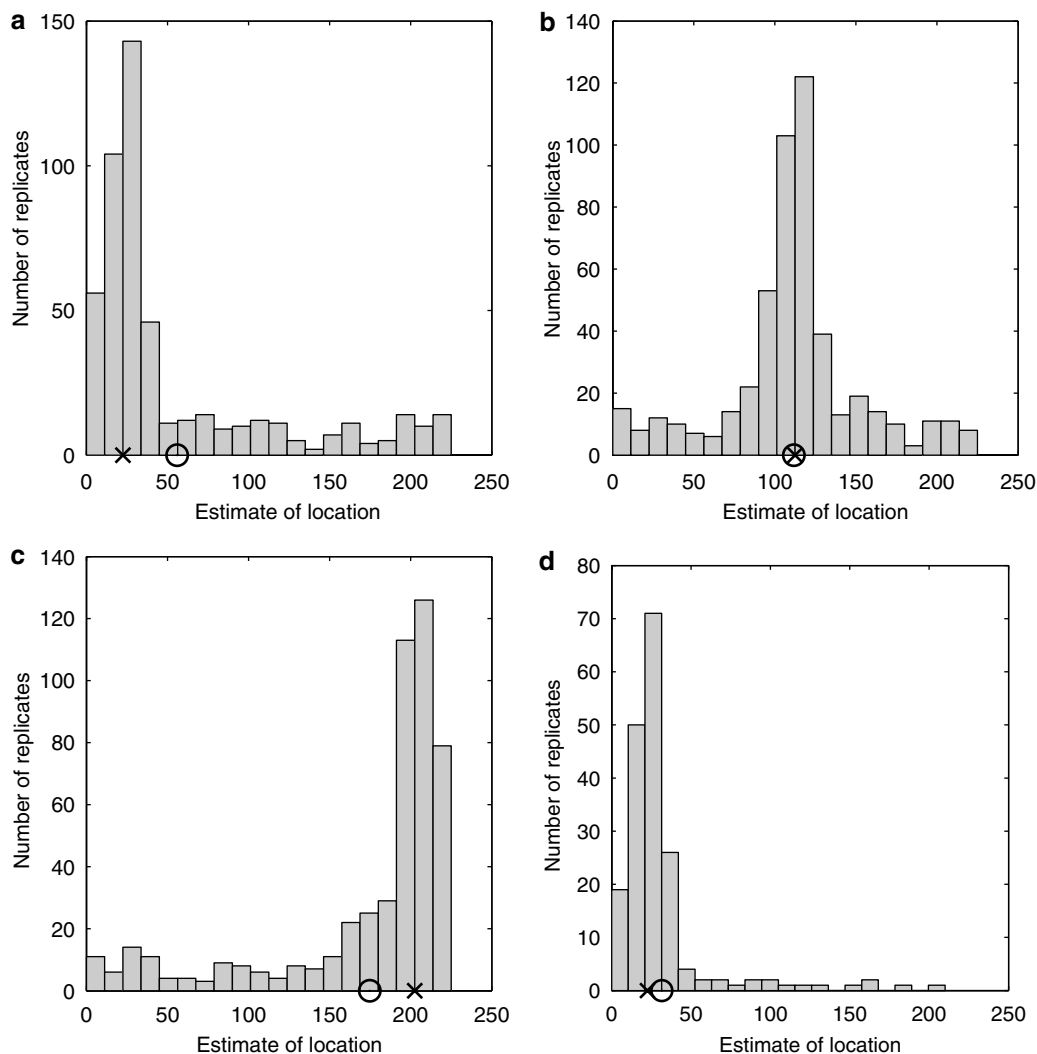


**Figure 2** Dependence of the bias of heritability estimates on true heritability, sample size, and the number of the investigated chromosomes. Panels (a) and (b) show the results of the search over ten 225 cM chromosomes. Panels (a) and (c), the sample size  $n = 200$ . In panels (b) and (c) heritability value  $h^2 = 0.05$ .

occurs when the estimated location is not near (unlinked) the true location, but as seen in Figure 1a this can also occur at the true location. Finally, when each location was tested for a QTL, and attention restricted to only significant replicates (of the 500), the distribution of the estimator of QTL effect is seen to move further away from zero (Figure 1c). In this situation, the value of the heritability estimate and the absolute value of the QTL effect estimate were always larger than 0.0612 (0.05 true value) and 0.4865 (0.4588 true value), respectively, indicating significant upward bias. Note that the threshold for the heritability estimate for significant replicates can be easily predicted by using the relationship between the heritability estimate and the likelihood ratio statistic (see Appendix A1). Furthermore (Figure 1d), when we conditioned on the event that the QTL was precisely located, the distribution of the estimate of the QTL effect is also shifted in the direction of large values. This is due to the fact that we picked the proper location only when the corresponding heritability estimate was large enough

to exceed the heritability estimates at other locations, thus effectively eliminating small values from the resulting distribution. For this simulation, the mean of the corresponding QTL effect estimates was 0.5647 (true value 0.4588).

The dependence of bias of heritability estimators on true QTL heritability, sample size and the length of the investigated part of the genome was explored and is illustrated in Figures 2a–c. Figures 2a and b illustrate that the bias of heritability estimators diminishes over increasing true QTL heritability and the sample size. This can be explained by the fact that both the accuracy of estimating QTL location and the power to detect QTL increase when heritability and sample size increase. For very small heritabilities, the bias resulting from the joint estimation of location and genetic effect is comparable with the bias resulting from testing; however, for larger heritabilities, the bias introduced by the testing procedure is almost three times larger than the bias resulting from the choice of optimal location. Figure 2c demon-

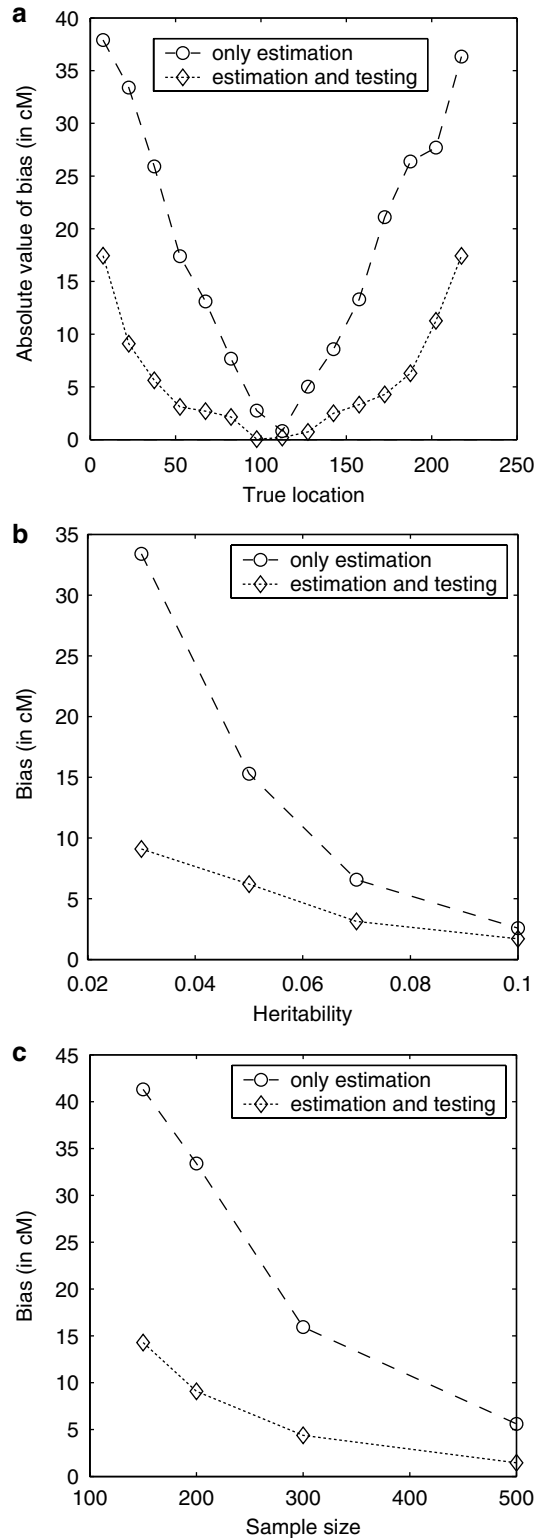


**Figure 3** Histograms of the estimators of QTL location resulting from the search over a 225 cM chromosome. Panels (a)–(c) are based on all 500 replicates. Panel (d) is based on 187 significant replicates. Heritability is  $h^2 = 0.03$  and the sample size is 200. Means of estimators of QTL location in panels (a)–(d) are marked by ‘O’, and are equal to 55.9, 111.7, 174.8, and 31.6, respectively. True values of QTL location are marked by ‘X’ (22.5, 112.5, 202.5, 22.5, respectively).

strates the dependence of bias of heritability estimators on the number of investigated chromosomes. The value 0 on the  $x$ -axis (ie 0 chromosomes) corresponds to estimating the heritability at the true QTL location. As shown in Figure 2c, the bias of QTL heritability increases when the length of the investigated part of the genome increases (ie number of chromosomes). An increase of bias due to the choice of optimal location results from the fact that the probability of obtaining a large heritability estimate in a position different from QTL location increases when we increase the search area. Additionally, the threshold values for the likelihood ratio test statistic that are used to detect the presence of QTL increase as the investigated portion of the genome increases. This results in a decrease of power, as well as an increase in bias due to testing (see Appendix A1 for the relationship between the likelihood ratio statistic and the heritability estimate).

Figures 3 and 4 demonstrate the bias involved in estimating QTL location. When the QTL heritability is low, there is a relatively large chance of making an error in the QTL location. Furthermore, if the QTL is located close to one end of the chromosome, then the number of tested positions (and the margin of error) in the direction toward the other end of the chromosome is much larger (Figure 3). Figure 3 is based on simulations with a fixed heritability value  $h^2=0.03$  and sample size 200. Figures 3a–c present histograms of the estimators of QTL location resulting from interval mapping over a 225 cM chromosome based on 500 replicates. Figure 3d illustrates the histogram of the estimator of QTL location based on only significant replicates. As seen in Figure 3b, when the QTL is located in the middle of the chromosome, the distribution of the QTL location estimates is symmetric and a bias in QTL location is not observed. However, when the QTL is close to one end of the chromosome (Figure 3a and c), the distribution of estimators of QTL location is skewed toward the opposing end of the chromosome, and therefore the means of these estimators are shifted in this direction. This effect (Figure 3d) is also observed when we restrict attention to only those replicates that produce significant results. Finally, Figure 3 demonstrates that the bias in QTL location is directly related to a very low accuracy of locating QTL with small heritability. The standard deviation of estimators of QTL location that are reported in Figure 3 is approximately equal to 50 cM when testing is not employed and 30 cM when we use testing and restrict the attention to only significant replicates.

Figure 4 demonstrates the dependence of bias of QTL location on true location, QTL heritability, and sample size. The bias resulting from the estimation procedure is computed as the difference between the mean of location estimators over all 500 replicates and the true location. To compute the bias resulting from using both estimation and testing, we used only significant replicates. As demonstrated in Figure 4a, the absolute value of the bias of location is equal to 0 if the QTL is in the middle of the chromosome and increases symmetrically with increasing distance from the center of the chromosome. Furthermore, the bias of the estimators of QTL locations obtained for only significant replicates is much lower than in the case when we do not use testing (Figure 4). However, even when we apply



**Figure 4** Dependence of the bias of location estimators on true location, heritability, and sample size. The QTL search is performed over a 225 cM chromosome. In panels (a) and (b), the sample size is 200. In panels (a) and (c), the heritability is 0.03, and in panels (b) and (c) the true QTL location is equal to 22.5 cM.

testing, and the sample size and heritability are small, the bias at locations close to the end of the chromosome can reach 15 cM.

## Discussion

Interval mapping (Lander and Botstein, 1989) is a popular statistical method for locating QTL. However, as demonstrated by our simulations, the estimators of QTL locations and heritability resulting from this procedure can be severely biased. We explain this phenomenon by observing that inherent in the interval mapping framework lies the issue of a constantly changing model that ultimately affects the accuracy of the parameter estimates. To further exacerbate this problem, a direct consequence of both increased genome size and marker number is an increase in the number of models considered. Therefore, as the genome and/or marker number increases, the number of models increases, and hence the effect of bias on the estimators resulting from interval mapping increases.

The search for a solution to remedy the bias problem in QTL mapping is ongoing (Utz *et al*, 2000; Ball, 2001; Allison *et al*, 2002). We have great hope that this problem can be solved by developing appropriate Bayesian methods. The advantages of Bayesian approaches are that they give the investigator the chance to employ expert, prior knowledge to the problem at hand, keep track of many possible models, and quantify the uncertainty related to the model choice by reporting posterior probabilities. The quantities of interest (ie parameters) can then be estimated by computing the mean or the mode of the related posterior distribution (for the tutorial on Bayesian model averaging, see Hoeting *et al*, 1999). Because Bayesian methods avoid concentrating on only one model, their results do not suffer from the biased estimates of QTL effects that are addressed in this work. This is not to say that we expect Bayesian methods to provide a solution to the bias of location of weak QTL which do not result from the model selection procedure. With this issue aside, it is important to realize the impact that the prior distributions have on the outcome of Bayesian methods (Clyde, 1999). By comparison to standard, classical estimates, Bayesian estimators are shrunken in the direction specified by the prior distribution, which in itself is a different form of bias that can be severe if the prior is selected inappropriately. Although Bayesian methods are becoming more mainstream in the QTL mapping community (Satagopan *et al*, 1996; Uimari and Hoeschele, 1997; Stephens and Fisch, 1998; Yi and Xu, 2000; Sen and Churchill, 2001; Vogl and Xu, 2002; Yi and Xu, 2002; Yi *et al*, 2003a; Xu, 2003; Kilpikari and Silanpää, 2003; Yi *et al*, 2003b; Bogdan *et al*, 2004; Jannink and Fernando, 2004; van de Ven, 2004; Zhang *et al*, 2005; Hayashi and Awata, 2005), we feel that there is a need for a more systematic investigation of the influence that the choice of priors has on the resulting estimates. Our hope is that work toward solving the bias problem will continue, but our larger hope is that readers will be aware that bias exists in the estimators rendered from QTL mapping methodologies and will view their results with a cautious eye.

## Acknowledgements

We thank Professors Witold Klonecki, Dan Nettleton, Friedrich Utz, Krzysztof Bogdan, and Przemyslaw

Biecek as well as two anonymous referees for helpful comments and suggestions. This work was funded by a USDA-IFAFs (00-52100-9615) grant to RWD.

## References

- Allison DB, Fernandez JR, Heo M, Zhu S, Etzel C, Beasley TM *et al* (2002). Bias in estimates of quantitative trait locus effect in genome scans: demonstration of the phenomenon and a methods-of-moments procedure for reducing bias. *Am J Hum Genet* **70**: 575–585.
- Ball RD (2001). Bayesian methods for quantitative trait loci mapping based on model selection: approximate analysis using the Bayesian information criterion. *Genetics* **159**: 1351–1364.
- Beavis WD (1994). The power and deceit of QTL experiments: lessons from comparative QTL studies. In: *Proceedings of the Forty-Ninth Annual Corn & Sorghum Industry Research Conference*. American Seed Trade Association: Washington, DC. pp 250–266.
- Beavis WD (1998). QTL analyses: power, precision, and accuracy. In: Paterson AH (ed) *Molecular Dissection of Complex Traits*. CRC Press: New York. pp 145–162.
- Bogdan M, Ghosh JK, Doerge RW (2004). Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci. *Genetics* **167**: 989–999.
- Broman KW (2001). Review of statistical methods for QTL mapping in experimental crosses. *Lab Anim* **30**: 44–52.
- Clyde M (1999). Comment for 'Bayesian model averaging: a tutorial' by Hoeting *et al*. *Stat Sci* **14**: 401–404.
- Dempster AP, Laird NM, Rubin DB (1977). Maximum likelihood from incomplete data via EM algorithm. *J R Stat Soc Ser B* **39**: 1–38.
- Doerge RW (2002). Mapping and analysis of quantitative trait loci in experimental populations. *Nat Genet Rev* **3**: 43–52.
- Doerge RW, Zeng Z-B, Weir BS (1997). Statistical issues in the search for genes affecting quantitative traits in experimental populations. *Stat Sci* **12**: 195–219.
- Göring HHH, Terwilliger JD, Blangero J (2001). Large upward bias in estimation of locus-specific effects from genomewide scans. *Am J Hum Genet* **69**: 1357–1369.
- Hayashi T, Awata T (2005). Bayesian mapping of QTL in out bred F2 families allowing inference about whether F0 grandparents are homozygous or heterozygous at QTL. *Heredity* **94**: 326–338.
- Hoeting JA, Madigan D, Raftery AE, Volinsky CT (1999). Bayesian model averaging: a tutorial. *Stat Sci* **14**: 382–401.
- Jannink J, Fernando RL (2004). On the Metropolis–Hastings acceptance probability to add or drop a quantitative trait locus in Markov Chain Monte Carlo-based Bayesian Analyses. *Genetics* **166**: 641–643.
- Jansen RC (1993). Interval mapping of multiple quantitative trait loci. *Genetics* **135**: 205–211.
- Jansen RC, Stam P (1994). High resolution of quantitative traits into multiple loci via interval mapping. *Genetics* **136**: 1447–1455.
- Jansen RC, Nap J-N (2001). Genetical genomics: the added value from segregation. *Trends Genet* **17**: 388–391.
- Kao C-H, Zeng Z-B (1997). General formulas for obtaining the MLEs and the asymptotic variance–covariance matrix in mapping quantitative trait loci when using the EM algorithm. *Biometrics* **53**: 653–665.
- Kao C-H, Zeng Z-B, Teasdale RD (1999). Multiple interval mapping for quantitative trait loci. *Genetics* **152**: 1203–1216.
- Kilpikari R, Silanpää MJ (2003). Bayesian analysis of multilocus association in quantitative and qualitative traits. *Genetic Epidemiol* **25**: 122–135.
- Lander ES, Botstein D (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**: 185–199.

Mackay TFC (2001). Quantitative trait loci in *Drosophila*. *Nat Rev Genet* **2**: 11–20.

Satagopan JM, Yandell BS, Newton MA, Osborn TC (1996). Bayesian model determination for quantitative trait loci. *Genetics* **144**: 805–816.

Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V *et al* (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**: 297–302.

Sen S, Churchill GA (2001). A statistical framework for quantitative trait mapping. *Genetics* **159**: 371–387.

Stephens DA, Fisch RD (1998). Bayesian analysis of quantitative trait locus data using reversible jump Markov chain Monte Carlo. *Biometrics* **54**: 1334–1347.

Thoday JM (1961). Location of polygenes. *Nature* **191**: 368–370.

Uimari P, Hoeschele I (1997). Mapping linked quantitative trait loci using Bayesian analysis and Markov chain Monte Carlo algorithms. *Genetics* **146**: 735–743.

Utz HF, Melchinger AE, Schön CC (2000). Bias and sampling error of the estimated proportion of genotypic variance explained by quantitative trait loci determined from experimental data in maize using cross validation and validation with independent samples. *Genetics* **154**: 1839–1849.

van de Ven R (2004). Reversible-jump Markov Chain Monte Carlo for quantitative trait loci mapping. *Genetics* **167**: 1033–1035.

Vogl C, Xu S (2002). QTL analysis in arbitrary pedigrees with incomplete marker information. *Heredity* **89**: 339–334.

Wayne M, McIntyre LM (2002). Combining mapping and arraying: an approach to candidate gene identification. *Proc Natl Acad Sci USA* **99**,23: 14903–14906.

Xu S (2003). Theoretical basis of the Beavis effect. *Genetics* **165**: 2259–2268.

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

Yi N, Xu S (2002). Mapping quantitative trait loci with epistatic effects. *Genet Res* **79**: 185–198.

Yi N, George V, Allison DB (2003a). Stochastic search variable selection for identifying multiple quantitative trait loci. *Genetics* **164**: 1129–1138.

Yi N, Xu S, Allison DB (2003b). Bayesian model choice and search strategies for mapping interacting quantitative trait loci. *Genetics* **165**: 867–883.

Zeng Z-B (1993). Theoretical basis of precision mapping of quantitative trait loci. *Proc Natl Acad Sci USA* **90**: 10972–10976.

Zeng Z-B (1994). Precision mapping of quantitative trait loci. *Genetics* **136**: 1457–1468.

Zhang M, Montooth KL, Wells MT, Clark AG, Zhang D (2005). Mapping multiple quantitative trait loci by Bayesian classification. *Genetics* **169**: 2305–2318.

$R^2 = 1 - \exp(-LRT/n)$ , and the heritability estimate

$$\hat{h}^2 = \frac{(n-1)(1 - \exp(-LRT/n)) - 1}{n-2} \quad (A.1)$$

This representation can be used to compute the threshold for a heritability estimate to detect the presence of QTL. For example, in our first simulation, the simulated threshold value for the likelihood ratio test statistic at the significance level  $\alpha=0.05$  was equal to 13.45, and the corresponding heritability estimate given by equation (A.1),  $h^2 = 0.0603$ . These computations, based on a regression on markers, agree well with the minimal value of the heritability estimate for significant replicates resulting from interval mapping, which in our experiment was equal to 0.0612.

The probability that the heritability estimate will obtain its maximum at the wrong marker is larger than zero

**Proposition 1** *Let the QTL be located at marker 1, and let marker 2 be located at any other position in the genome. Let  $h_1^2$  and  $h_2^2$ , be estimators of heritability resulting from fitting a simple regression model (4) with predictor variables specified by genotypes of markers 1 and 2, respectively. If the sample size  $n \geq 3$ , then*

$$P(h_2^2 > h_1^2) > 0$$

(ie probability that the maximum of heritability estimates will be obtained at the wrong marker is larger than zero).

**Proof** Let  $Z_1 = (Z_{11}, Z_{12}, \dots, Z_{1n})$  and  $Z_2 = (Z_{21}, Z_{22}, \dots, Z_{2n})$  be the genotypes at markers 1 and 2, respectively. Consider the  $Z_{1i}$  as i.i.d. random variables with  $P(Z_{1i} = 1/2) = P(Z_{1i} = -1/2) = 1/2$ . Assume that the same holds for  $Z_{2i}$ . Denote the probability of recombination between markers 1 and 2 as  $c \in (0, 1/2]$ . Therefore,  $P(Z_{2i} = 1/2 | Z_{1i} = 1/2) = P(Z_{2i} = -1/2 | Z_{1i} = -1/2) = (1-c)$  and  $P(Z_{2i} = -1/2 | Z_{1i} = 1/2) = P(Z_{2i} = 1/2 | Z_{1i} = -1/2) = c$ ,  $i = 1, \dots, n$ . Letting  $\mathbf{1} = (1, 1, \dots, 1) \in \mathbf{R}^n$ ,  $(c, c, \dots, c)$  is written as  $c$  instead of  $c_1$ .

Let  $A$  be the event that  $\{Z_1 \neq \text{const}, Z_2 \neq \text{const}, Z_1 \neq Z_2 \text{ and } Z_1 \neq -Z_2\}$ . It is easy to check that  $P(A) > 0$ . Denote the standard inner product in  $\mathbf{R}^n$  as  $\langle \cdot, \cdot \rangle$  (ie  $\langle y, z \rangle = \sum_{i=1}^n y_i z_i$ ), and let  $|\cdot|$  denote the Euclidean norm in  $\mathbf{R}^n$  (ie  $|y| = \sqrt{\sum_{i=1}^n y_i^2}$ ). The square of the sample correlation coefficient between vectors  $y$  and  $z$  is given by

$$\rho^2(y, z) = \frac{\langle y - \bar{y}, z - \bar{z} \rangle^2}{|y - \bar{y}|^2 |z - \bar{z}|^2}$$

It is straightforward to verify that on  $A$  the vector  $Z_2 - \bar{Z}_2$  is linearly independent of  $Z_1 - \bar{Z}_1$  (ie cannot be represented as  $k(Z_1 - \bar{Z}_1)$ , where  $k \in \mathbf{R}$ ).

Let  $\xi \sim N(0, \sigma^2 \mathbf{1}_{n \times n})$  be independent of  $Z_1$  and  $Z_2$ . According to model (4), we define the vector of trait values as  $Y = \mu + aZ_1 + \xi$ . Let  $R = |\xi|$  and  $X = (1/|\xi|)\xi$ , so that  $\xi = RX$ . Observe that  $X$  is uniformly distributed on the unit sphere,  $R^2/\sigma^2$  is  $\chi^2$  distributed with  $n$  degrees of freedom, and that  $X$  and  $R$  are independent. It holds

$$\frac{\rho^2(Y, Z_2)}{\rho^2(Y, Z_1)} = \frac{((a/R)\langle Z_1, Z_2 - \bar{Z}_2 \rangle + \langle X, Z_2 - \bar{Z}_2 \rangle)^2 |Z_1 - \bar{Z}_1|^2}{((a/R)|Z_1 - \bar{Z}_1|^2 + \langle X, Z_1 - \bar{Z}_1 \rangle)^2 |Z_2 - \bar{Z}_2|^2}$$

## Appendix A1

Relationship between the likelihood ratio statistic and the heritability estimate

Consider the likelihood ratio statistic to detect a QTL at a marker position

$$LRT = 2 \ln \frac{L(\hat{\mu}, \hat{a}, \hat{\sigma}_0^2)}{L(\bar{Y}, 0, \hat{\sigma}_0^2)}$$

where  $L$  is the likelihood given by equation (3),  $\bar{Y} = (1/n) \sum_{i=1}^n Y_i$  and  $\hat{\sigma}_0^2 = (1/n) \sum_{i=1}^n (Y_i - \bar{Y})^2$ . Simple calculations yield  $LRT = n \ln(\text{SST}/\text{SSE})$ , where SST and SSE are the total sum of squares and the sum of residuals from regression of the trait data on marker genotypes. Thus,  $LRT = n \ln 1/(1-R^2)$ , which results in



Since  $Z_1 - \bar{Z}_2$  and  $Z_2 - \bar{Z}_2$  are linearly independent (a.s.) on  $A$ , there exists  $X_0$  on the unit sphere in  $\mathbf{R}^n$  such that  $\langle X_0, Z_2 - \bar{Z}_2 \rangle \neq 0$  and  $\langle X_0, Z_1 - \bar{Z}_1 \rangle = 0$ . Therefore,  $\rho^2(Y, Z_2) / \rho^2(Y, Z_1)$  can be arbitrarily large, provided  $R$  is large and  $X$  is in a sufficiently small neighborhood of  $X_0$ . Since the distribution of  $\xi$  is

multivariate normal and independent of  $Z_1$  and  $Z_2$ , the probability of such event is larger than zero and it immediately holds that

$$P(h_2^2 > h_1^2) = P\left(\frac{\rho^2(Y, Z_2)}{\rho^2(Y, Z_1)} > 1\right) > 0 \quad \square$$