Genet. Res., Camb. (2010), **92**, *pp.* 39–53.
doi:10.1017/S0016672310000042

Composite interval mapping to identify quantitative trait loci for point-mass mixture phenotypes

SANDRA L. TAYLOR[1]* AND KATHERINE S. POLLARD[2]
[1] *Biostatistics Graduate Group, University of California, Davis, CA 95616, USA*
[2] *Gladstone Institutes and Department of Epidemiology and Biostatistics, University of California, San Francisco, CA 94158, USA*

(*Received 17 August 2009 and in revised form 30 November 2009; first published online 3 March 2010*)

Summary

Increasingly researchers are conducting quantitative trait locus (QTL) mapping in metabolomics and proteomics studies. These data often are distributed as a point-mass mixture, consisting of a spike at zero in combination with continuous non-negative measurements. Composite interval mapping (CIM) is a common method used to map QTL that has been developed only for normally distributed or binary data. Here we propose a two-part CIM method for identifying QTLs when the phenotype is distributed as a point-mass mixture. We compare our new method with existing normal and binary CIM methods through an analysis of metabolomics data from *Arabidopsis thaliana*. We then conduct a simulation study to further understand the power and error rate of our two-part CIM method relative to normal and binary CIM methods. Our results show that the two-part CIM has greater power and a lower false positive rate than the other methods when a continuous phenotype is measured with many zero observations.

1. Introduction

Quantitative trait loci (QTL) mapping is the process of identifying genetic loci that contribute to variation in a trait and estimating the individual allelic effects. Lander & Botstein's (1989) interval method for mapping QTL greatly improved the estimation of QTL locations and effects over the earlier single-marker methods and laid the foundation for most modern statistical QTL mapping methods. In their approach, the observed phenotype is modelled as a mixture of normal distributions (one for each genotype) with the mixing proportions corresponding to the probability of each genotype at locations between two markers. At specified positions throughout the genome, commonly every 1 centiMorgan (cM), a LOD score is used to test the null hypothesis of no QTL present versus the alternative that a QTL is present. Regions where the LOD score exceeds a specified significance threshold are regarded as indicating the presence of a QTL. The location of the candidate QTL in a significant region is estimated by the location of the peak LOD score.

While still widely used, interval mapping can yield biased estimates of the location and effect of a QTL when there are multiple linked QTL (Knott & Haley, 1992; Martinez & Curnow, 1992). Several methods have been implemented to improve upon interval mapping. These include (1) composite interval mapping (CIM) (Zeng, 1994) and (2) multiple interval mapping (MIM) (Kao *et al.*, 1999). Both approaches seek to account for phenotype variation caused by other linked QTL.

CIM extends the interval mapping method by using additional markers (beyond the two that flank the interval of interest) as covariates in the linear model underlying the LOD score calculation. The additional markers are intended to account for the influence of QTLs not located in the interval being evaluated but that may be linked to the hypothetical QTL. By accounting for linked QTLs, CIM increases the power to detect a QTL and increases precision of the estimated location relative to interval mapping (Jensen, 1993; Zeng, 1994). In MIM, multiple QTLs are

* Corresponding author: One Shields Avenue, Department of Statistics, University of California, Davis, CA 95616, USA. Tel: +1 (916) 248 1963. Fax: +1 (530) 752 7099. e-mail: sltaylor@ucdavis.edu

modelled jointly. A selection procedure is used (forward, backward or stepwise) to add and delete QTL from the model. A final model is identified based on a selection criterion such as Akaike information criterion or Bayesian information criterion. In addition to modelling multiple QTLs, epistasis also can be modelled with MIM (Kao *et al*., 1999).

As originally proposed, interval mapping, CIM and MIM all assume that the phenotype is normally distributed. Subsequently, interval mapping was extended to analyse phenotypes with a wide variety distributions, including binary traits (Xu & Atchley, 1996; Deng *et al*., 2006), ordinal traits (Hackett & Weller, 1995), counts (Thomson, 2003), phenotypes with spikes (Broman, 2003) and censored data (Diao *et al*., 2004). All of these methods use parametric likelihood approaches. Non-parametric (Kruglyack & Lander, 1995; Fine *et al*., 2004), semi-parametric (Zou *et al*., 2002) and empirical likelihood (Huang *et al*., 2007) methods also have been implemented. Extensions of CIM and MIM to traits that are not normally distributed have received less attention. Xu & Atchley (1996) extended CIM to binary traits, and Li & Chen (2009) recently adapted MIM to analyse phenotypes with a spike in their distributions.

Data with spikes are a mixture of two components, a point-mass at one value and quantitative observations at values other than the point-mass value. The point-mass value is commonly the smallest value (e.g. zeros that reflect the absence of a compound or concentrations below a detection limit) or largest value (e.g. survival time of animals that survived to the end of an experiment). The distribution of data with spikes can be characterized by (i) the proportion of point-mass values and (ii) the distribution of the continuous component. We refer to such data-generating distributions as *point-mass mixtures*. These distributions arise in many different settings including high-throughput genomic experiments and cancer studies. A common feature of data from all of these applications is that there is information in both the continuous and the discrete components.

Point-mass mixture data present a challenge for QTL mapping methods that assume the phenotype is normally distributed and identify QTL based on a difference in phenotype means among the genotypes. Treating the point-mass as part of the continuous distribution violates the underlying assumptions of the method (i.e. normally distributed phenotypes). Alternatively, the point-mass observations could be dropped and only continuous observations analysed, but this approach discards meaningful information and reduces power. Thus, it is important to employ methods specific to point-mass mixture data that account for the separate contributions of each component of the mixture.

Several researchers have addressed QTL mapping for point-mass mixtures. Broman (2003) proposed a two-part interval mapping procedure for point-mass mixtures that jointly tests for differences between genotypes in the proportion of observations in the point-mass and in the means of the continuous component. Through simulations, he found this procedure to perform better than binary interval mapping and normal interval mapping methods that used only the observations not in the point-mass. However, a non-parametric, rank-based method generally was more powerful than the two-part interval mapping method. Jin *et al*. (2007) presented a framework for analysing QTL based on a semi-parametric, exponential tilt model and showed that this model could be used to analyse point-mass mixtures. Finally, Li & Chen (2009) recently extended MIM to accommodate point-mass mixtures. They showed that their two-part MIM model is more effective at identifying QTL than Broman's two-part interval mapping method when the trait's heritability is moderate or high, while the two-part interval mapping method is superior at low heritability.

Here, we develop and evaluate a CIM method for identifying QTLs when the phenotype is distributed as a point-mass mixture. In the first section, we present our two-part CIM procedure. We then demonstrate its use and compare it to existing normal and binary CIM methods through an analysis of metabolomics data from recombinant inbred lines (RILs) of *Arabidopsis thaliana*. Finally, we conduct a simulation study to further understand the method's power and error rate relative to normal and binary CIM methods.

2. Methods

(i) *Two-part CIM*

Data for mapping QTLs consist of a set of markers on a known genetic map and quantitative phenotype values for n individuals. Let y_i represent the quantitative phenotype value and M_i the set of marker genotypes at known locations for individual i, with m_{ik} the genotype of individual i at marker k. Without loss of generality, we assume a point-mass at zero and let $z_i=0$ if $y_i=0$ and $z_i=1$ if $y_i>0$.

Consider a putative QTL located at a fixed chromosome position. For simplicity, we consider an experimental population consisting of two genotypes (e.g. backcross and RIL), but the method can be directly extended to other experimental populations. In this setting, each individual has one of two genotypes at each location. Let $g_i=\{0, 1\}$ signify the genotype of individual i at the putative QTL. The g_i are unknown, but we assume that the QTL genotype

https://doi.org/10.1017/S0016672310000042 Published online by Cambridge University Press

probabilities, $p_i(j) = \Pr(g_i = j | M_i)$, where $j = \{0, 1\}$ signifies the genotype of individual i, can be calculated from the observed marker data at any position using the estimated recombination fraction.

Point-mass mixture data consist of two parts (1) a binary component representing the proportion of observations in the point-mass and (2) a continuous component consisting of observations not at the point-mass. For the binary component of the phenotype, represented by z_i, let $\pi(j) = \Pr(z_i = 1 | g_i = j)$. We assume the continuous component of the phenotype is normally distributed with mean μ_j and variance σ^2, i.e. $y_i | (g_i = j, z_i = 1) \sim \text{Normal}(\mu_j, \sigma^2)$. Then, for our two genotype set up, the likelihood function of the observed data is

$$L = \prod_{i=1}^{n} [p_i(1)\{(1-\pi(1))\}^{1-z_i}\{\pi(1) f_i(1)\}^{z_i} + p_i(0)\{(1-\pi(0))\}^{1-z_i}\{\pi(0) f_i(0)\}^{z_i}]. \quad (1)$$

If the genotype at the putative QTL were known, the likelihood would be

$$L = \prod_{i=1}^{n} [p_i(1)(1-\pi(1))^{1-z_i}\{\pi(1) f_i(1)\}^{z_i}]^{g_i} \times [p_i(0)(1-\pi(0))^{1-z_i}\{\pi(0) f_i(0)\}^{z_i}]^{1-g_i}, \quad (2)$$

where $f_i(1)$ and $f_i(0)$ are normal density functions for the random variable y_i with mean μ_1 and μ_0, respectively, and common variance σ^2.

We extend the two-part interval mapping method to CIM by adding markers as covariates in the likelihood function used to estimate $\pi(j)$, μ_j and σ. We estimate these parameters with maximum likelihood estimates (MLEs) via an Expectation–Maximization (EM) algorithm as follows. Because the likelihood (2) factors, it can be maximized separately for the binary component parameter, $\pi(j)$, and the continuous component parameters, μ_j and σ. Following Zeng (1994), we model the continuous component as

$$y_i | (g_i = j, z_i = 1) = b_o + b^* g_i + \sum_{k \neq l, l+1} b_k m_{ik} + e_i, \quad (3)$$

where b^* is the effect of the putative QTL, g_i is the genotype of individual i at the putative QTL, b_k is the effect of the kth marker outside the interval $(l, l+1)$, m_{ik} is the genotype of individual i at the kth marker and b_o is the intercept. The e_i are assumed to be independent and identically distributed normal random variables with mean 0 and standard deviation σ^2. Thus,

$$\mu_1 = b_o + b^* + \sum_{k \neq l, l+1} b_k m_{ik}, \quad (4)$$

$$\mu_0 = b_o + \sum_{k \neq l, l+1} b_k m_{ik}. \quad (5)$$

Maximizing the likelihood with respect to b_o, b^*, b_k and σ yields the following estimators:

$$\hat{b}^* = (\mathbf{Y} - \mathbf{M}\hat{\mathbf{B}})'\hat{\mathbf{P}}/\hat{c}, \quad (6)$$

$$\hat{\mathbf{B}} = (\mathbf{M}'\mathbf{M})^{-1}\mathbf{M}'(\mathbf{Y} - \hat{\mathbf{P}}\hat{b}^*), \quad (7)$$

$$\hat{\sigma}^2 = [(\mathbf{Y} - \mathbf{M}\hat{\mathbf{B}})'(\mathbf{Y} - \mathbf{M}\hat{\mathbf{B}}) - \hat{c}\hat{b}^{*2}]/n_c, \quad (8)$$

where n_c is the number of individuals with non-zero phenotypes, $\mathbf{Y}$ is an $n_c \times 1$ vector of non-zero phenotype values, $\mathbf{M}$ is a $(n_c \times (t-1))$ matrix of t covariate markers (m_{ik}), $\hat{\mathbf{P}}$ is an $n_c \times 1$ vector specifying the posterior probabilities $(\hat{P}_i)$ of $g_i = 1$ and $\hat{c} = \sum_{i=1}^{n_c} \hat{P}_i(1)$.

For the binary component of the likelihood, we model the probability that $z_i = 1$ conditional on the marker data with a logistic model. Estimates for $\pi(j)$ are obtained through solving

$$\pi(1) = \frac{\exp\left\{\gamma_o + \gamma^* + \sum_{k \neq l, l+1} \gamma_k m_{ik}\right\}}{1 + \exp\left\{\gamma_o + \gamma^* + \sum_{k \neq l, l+1} \gamma_k m_{ik}\right\}}, \quad (9)$$

$$\pi(0) = \frac{\exp\left\{\gamma_o + \sum_{k \neq l, l+1} \gamma_k m_{ik}\right\}}{1 + \exp\left\{\gamma_o + \sum_{k \neq l, l+1} \gamma_k m_{ik}\right\}}, \quad (10)$$

where γ^* is the effect of the putative QTL, γ_k is the effect of the kth marker outside the interval $(l, l+1)$, m_{ik} is the genotype of the kth marker and γ_o is the intercept. The Newton–Raphson algorithm is used to estimate γ_o, γ^* and γ_k as detailed in Xu & Atchley (1996).

Results from the two maximization procedures are combined in the E-step of the EM algorithm to estimate the posterior probabilities that $g_i = 1$ at the putative QTL by

$$\hat{P}_i = \begin{cases} \dfrac{p_i(1)(1-\hat{\pi}(1))}{p_i(1)(1-\hat{\pi}(1)) + p_i(0)(1-\hat{\pi}(0))}, & \text{if } y_i = 0, \\ \dfrac{p_i(1)\hat{\pi}_i(1) f_i(1)}{p_i(1)\hat{\pi}(1) f_i(1) + p_i(0)\hat{\pi}(0) f_i(0)}, & \text{if } y_i > 0. \end{cases} \quad (11)$$

To summarize, the EM algorithm is initiated by setting $P_{ij}^0 = p_{ij}$. Next, we obtain estimates for $\pi(j)$, μ_j and σ through equations (4)–(10). Then, posterior probabilities are calculated via equation (11) using updated parameter estimates. The last two steps are iterated until the log likelihood converges.

Based on the EM-estimated likelihood, we calculate a LOD score to test the joint null hypothesis, H_o: $\pi_1 = \pi_0 \cap \mu_1 = \mu_0$ as follows. First, we need a second set of parameter estimates under a null model

https://doi.org/10.1017/S0016672310000042 Published online by Cambridge University Press

with no QTL. The MLEs for π_o, μ_o and σ are estimated directly with the following equations from Zeng (1994) and Xu & Atchley (1996):

$$\mu_o = b_o + \sum_{k \neq l, l+1} b_k m_{ik}, \tag{12}$$

$$\hat{\mathbf{B}} = (\mathbf{M}'\mathbf{M})^{-1}\mathbf{M}'\mathbf{Y}, \tag{13}$$

$$\hat{\sigma}^2 = [(\mathbf{Y} - \mathbf{M}\hat{\mathbf{B}})'(\mathbf{Y} - \mathbf{M}\hat{\mathbf{B}})]/n_c, \tag{14}$$

$$\pi_o = \frac{\exp\left\{\gamma_o + \sum_{k \neq l, l+1} \gamma_k m_{ik}\right\}}{1 + \exp\left\{\gamma_o + \sum_{k \neq l, l+1} \gamma_k m_{ik}\right\}}. \tag{15}$$

Then, the LOD score for testing the null hypothesis at a putative QTL position is $\mathrm{LOD} = \log_{10}[L(\hat{\pi}(1), \hat{\pi}(0), \hat{\mu}_1, \hat{\mu}_0, \hat{\sigma})/L(\hat{\pi}_o, \hat{\mu}_o, \hat{\sigma})]$.

(ii) *Covariate selection*

Zeng (1994) evaluated an approach to selecting covariates from the set of all markers other than those directly flanking the interval of interest. He compared this method to one that uses all unlinked markers. In practice, a user-specified number of markers (t) is commonly used, typically three to six [see e.g., R/qtl (Broman *et al.*, 2003) and QTL Cartographer (Basten *et al.*, 2005)]. In these programmes, covariate markers are selected through a step-wise selection procedure and the t most significant markers for explaining phenotypic variation are retained as covariates.

We use a forward step-wise selection process based on deviance tests to rank markers. First, the significance of each marker for explaining the phenotypic variation is tested individually as follows. We use logistic regression to model $\mathrm{logit}(z_i) = \alpha_o + \alpha_k m_{ik} + \varepsilon_{ik}$, where α_k is the effect of genotype at the kth marker. Similarly, simple linear regression is used to model $y_i|(z_i = 1) = \beta_o + \beta_1 m_{ik} + \varepsilon_{ik}$. The full model deviance is the sum of the deviances from the logistic regression and linear regression components. For the null model, we model $\mathrm{logit}(z_i) = \alpha_o + \varepsilon_{ik}$ and $y_i|(z_i = 1) = \beta_o + \varepsilon_{ik}$. The null deviance is again the sum of the deviances from these two models. The significance of marker k is evaluated using the null deviance minus the full model deviance, which follows a chi-square distribution with two degrees of freedom when the null hypothesis is true. We proceed by adding the most significant marker to both the full and null models and repeat the testing procedure. The procedure is terminated once the user-specified number of markers has been identified. This selection process is conducted genome-wide and is not prohibitively computationally intensive.

3. Example

To determine if the two-part CIM can improve detection of QTLs when data are distributed as point-mass mixtures, we applied the method to metabolomics data from the model plant *A. thaliana*. We used metabolite concentration measurements from Rowe *et al.*'s (2008) study on RIL of *A. thaliana*. The 210 RILs were created from crossing the Bayreuth-0 (*Bay*) and Shahdara (*Sha*) inbred lines of *A. thalina* (Loudet *et al.*, 2002). Each RIL was genotyped at 95 markers on five chromosomes with an established genetic map (Loudet *et al.*, 2002). Gas chromatography time-of-flight mass spectrometry was used to obtain the concentrations of 557 metabolites in each RIL. The concentration of each metabolite is a quantitative trait for which we seek to identify QTLs. Of the 557 metabolites, 121 had no point-mass observations and were excluded from this analysis. In addition, we dropped five RILs with large numbers of missing phenotype values from the analysis. Thus, 436 metabolites were analysed based on observations from 205 RILs.

The proportion of observations at zero was highly variable across metabolites, ranging from less than 1 % to more than 95 %. On average, the proportion of observations at zero was 48 %. The distribution of the continuous observations tended to be strongly right-skewed suggesting the use of a logarithmic transformation to meet the normality assumption of the two-part CIM method. However, zero values pose a challenge when applying a logarithmic transformation. We employed a commonly used approach to address this problem that enabled us to use a log transformation with the normal CIM. Specifically, we substituted 0·001 for all zero values before transforming the data. This substitution was only used for analysing data with the normal CIM. Zero values were retained for analysis with the two-part and binary CIM methods. We compared this approach to an alternative method in which we treated zero values as missing and analysed only continuous observations with the normal CIM.

We analysed the *A. thaliana* metabolomics data using normal CIM, binary CIM and our new two-part CIM method. All analyses were conducted in R version 2.7.2 (R Development Team, 2008). Normal CIM was implemented through the R package R/qtl (Broman *et al.*, 2003). We modified functions from the R/qtl package to implement the binary CIM. We used $t = 5$ covariate markers, excluding markers within 10 cM the markers flanking the interval of interest. Significance thresholds for identifying QTL were determined based on a permutation null distribution. For each metabolite, we generated the null distribution by permuting phenotype values 1000 times relative to the matrix of RIL genotypes. This is a

https://doi.org/10.1017/S0016672310000042 Published online by Cambridge University Press

common method for determining significance in quantitative genetics studies (Broman *et al*., 2003; Basten *et al*., 2005). QTLs were identified as peaks in the LOD score profile that exceeded the 95th percentile of the permutation distribution. For significant peaks, 1·5 LOD support intervals were determined.

(i) *Results*

Of the 436 metabolites, no method detected a QTL for 306 metabolites and all methods predicted at least one QTL for 38 metabolites. The two-part CIM method identified at least one QTL for the most metabolites (92), while the binary CIM method identified the fewest metabolites with QTLs (66). The normal CIM identified 75 metabolites with at least one QTL. For 38 metabolites, only the two-part CIM method identified a QTL. Binary and normal CIM methods were the only methods to detect QTLs in 15 and 12 metabolites, respectively.

To better understand the characteristics of the metabolites and the performance of the three methods, we calculated the means of the continuous components and proportions of point-mass observations for *Bay* and *Sha* alleles at the marker closest to predicted QTLs for each method. Metabolites for which all methods predicted QTLs tended to have large differences in proportions of point-mass observations. Metabolites with QTLs predicted only by the binary CIM either had large point-mass proportions ($>80\%$) or large differences between the point-mass proportions of the two alleles ($>40\%$) but similar means for the continuous observations. For these metabolites, the difference in proportions was not large enough to result in a rejection of the null hypothesis with the two-part CIM given the small difference in means.

Metabolites for which only the normal CIM predicted QTLs (15 QTLs for 12 metabolites) had variable characteristics. For two predicted QTLs, one of the alleles had no continuous observations at the closest marker; thus precluding an assessment of the differences between the alleles. Three predicted QTLs had *dissonant differences* with very large point-mass proportions ($>80\%$). Based on the simulation results (see below), some of these predicted QTLs could be false positives. Dissonant differences occur when the genotype with the larger mean has a larger proportion of zero values so that the binary and continuous effects are in opposite directions. One additional QTL showed dissonant differences but had very small point mass proportions ($<5\%$). The remaining nine QTLs were for traits with *consonant differences*, meaning that the genotype with the larger mean has a smaller proportion of zero values. For these QTLs, the means of the continuous components were similar between alleles but the point-mass proportions differed. The simulation study (see below) showed the normal CIM to have slightly greater power in this situation compared to the two-part CIM.

Metabolites for which only the two-part CIM detected QTLs tended to have moderate differences in means and point-mass proportions. The two-part CIM detected more QTLs resulting from dissonant differences than the normal CIM. Of the 41 QTLs detected for 38 metabolites, 16 showed dissonant differences and 25 showed consonant differences.

When zero values were treated as missing values for analysis with the normal CIM, slightly fewer metabolites were identified as having QTLs than when 0·001 was substituted for the zero values (72 versus 75). However, more metabolites had QTLs only identified by the normal CIM (22 versus 12). These metabolites had small differences in the point-mass proportions between the two alleles. By dropping the zero values, differences in the means between the alleles were enhanced since the zero observations were not included in estimating the means and standard deviation. Because the point-mass proportions differed by only a small amount; the binary CIM did not predict any QTLs. The two-part CIM uses information about the means and the point-mass proportions to determine if the two groups differ. In this case, for these metabolites, the difference in means was not large enough to reject the null hypothesis given the small difference in point-mass proportions.

4. Simulations

We conducted two simulation studies to evaluate and compare the two-part CIM method to normal and binary CIM methods when applied to data distributed as point-mass mixtures. The first study consisted of simulating a single QTL, while the second considered multiple QTLs. The focus of the simulation studies was to characterize the power and error rates of these methods.

In the first study, we simulated a backcross population consisting of 200 individuals with a single chromosome, 100 cM long with 22 equally spaced markers, which yielded a marker density similar to the genetic map used in the *A. thaliana* analysis. To simulate data from a point-mass mixture distribution, first the number of observations in the point-mass was determined by randomly sampling from a binomial (n, p) distribution, where n is the number of individuals and p is the target (i.e. average) proportion of observations in the point-mass. For the homozygote, we simulated average point-mass proportions of $P=0{\cdot}05$, $0{\cdot}30$, $0{\cdot}50$, $0{\cdot}70$ and $0{\cdot}95$. Continuous observations were then randomly generated for all individuals not in the point-mass using a normal distribution with mean 6 and standard deviation 0·6.

https://doi.org/10.1017/S0016672310000042 Published online by Cambridge University Press

These values were motivated by the point-mass proportions and means and standard deviations of the log transformed continuous values observed in the *A. thalinana* metabolomics data.

A single QTL with additive effects on the mean and/or the proportion of observations in the point-mass was simulated at 30 cM. We considered additive effects on the mean of $\pm 0{\cdot}6$, $\pm 1{\cdot}2$ and $\pm 2{\cdot}4$ (i.e. 10, 20 and 40% of the mean value in homozygotes). Similarly, we considered additive effects on the point-mass proportions that were -10, -20 and -40% of the homozygote proportion. We simulated data for which only the proportion in the point-masses differed, only the means of the continuous components differed, and both the means and proportions differed. For simulations in which both the mean and point-mass proportions differed, we considered both consonant differences and dissonant differences. To simulate null datasets (no QTL), we generated data in which the point-mass proportions and means of the continuous components were the same for both genotypes. For each combination of parameters, 1000 datasets were simulated.

Because CIM was developed specifically to analyse data with multiple linked QTL, we also conducted simulations with two QTLs on one or two chromosomes. We again simulated a backcross population of 200 individuals either with one or with two chromosomes, each 100 cM long with 22 equally spaced markers. We fixed the average proportion of observations in the point-mass for the homozygote at 0·5. Continuous observations were randomly generated from a normal distribution with mean 6 and standard deviation 0·6. The additive effect of one QTL (QTL 1) on the mean and point-mass proportion was set at $\pm 0{\cdot}6$ and $\pm 0{\cdot}05$, respectively, and the additive effect of the second QTL (QTL 2) was twice as large. We investigated various combinations of effects of the two QTLs on the point-mass proportion and mean of the continuous component. We also varied the distance between the two QTLs as summarized in Table 1. For each combination of mean and proportion differences, 1000 datasets were generated.

For both the one and two QTL simulations, we applied the three CIM methods to each simulated dataset and tested for QTLs every 1 cM. Zero values were retained as zeros in applying the normal CIM, because a log transformation was not used. In each LOD score calculation, three markers selected through our forward selection process (see above) were used as covariates, with markers within 10 cM of the markers flanking the interval being evaluated dropped. Significance thresholds were derived by simulating 1000 null datasets for each homozygote point-mass proportion. The 95th percentile of the maximum LOD scores from the permuted datasets was used as the QTL significance threshold for all datasets with the same homozygote point-mass proportion. Predicted QTLs were identified as peaks in the LOD score profile that exceed the significance threshold. For all predicted QTLs, 1·5 LOD support intervals were defined for the estimated QTL location.

Because more than one QTL was predicted in some datasets, we calculated power as the percentage of datasets with at least one predicted QTL with a LOD support interval that encompassed a true QTL location. As a measure of the false positive rate, we calculated the percentage of datasets with at least one predicted QTL that also had at least one predicted QTL whose LOD support interval did not cover the true QTL location. Thus, a dataset with two predicted QTLs, one with a LOD support interval covering the true QTL location and one with a LOD support interval not covering the true QTL location, is counted as identifying the true QTL for the power measure but also contributes to the false positive calculation.

(i) *Results for single QTL simulations*

The two-part CIM performed better than both normal and binary CIM methods, having the highest power and lowest false positive rate for almost all simulations. For most simulations, the binary CIM had low power and a high percentage of false positives, particularly when the additive effect on the proportion was small (Supplemental Table 1). Because the binary CIM performed poorly across the simulated conditions, we focus on comparing the two-part and normal CIM methods. Detailed results are provided in Supplemental Table 1 and summarized below.

The two-part CIM had the highest power and lowest false positive rate for almost all simulations with substantially better performance in some cases. For many simulations, power of the two-part CIM exceeded 80% whereas power of the normal CIM was usually less than 60% (Figs 1 and 2). The two methods also differed in the percentage of false positives. The percentage of false positives for the two-part CIM was more stable, typically near 30%, and usually much lower than for the normal CIM. The percentage of false positives for the normal CIM varied considerably among simulations and commonly exceeded 60% (Figs 1 and 2). The only situation for which the normal CIM performed somewhat better than the two-part CIM was when only the point-mass proportions differed (Supplemental Table 1).

The power of both methods was influenced by the proportion of observations in the point-mass, albeit in different ways. When only the means differed, power of the two-part CIM decreased as the point-mass proportion increased. This pattern occurred because

https://doi.org/10.1017/S0016672310000042 Published online by Cambridge University Press

Table 1. *Additive effects and locations of QTLs for simulations of two QTLs on one or two chromosomes*

QTL No.	Chr.	QTL location	Mean effect	Prop. effect	Description
Two chromosomes					
1	1	30	1·2	−0·1	QTLs on different chromosomes. Consonant differences
2	2	69	2·4	−0·2	
1	1	30	−1·2	−0·1	QTLs on different chromosomes. Dissonant differences
2	2	69	−2·4	−0·2	
One chromosome					
1	1	30	1·2	−0·1	QTLs far apart on same chromosome. Consonant
2	1	69	2·4	−0·2	differences. QTL effects in the same direction.
1	1	30	−1·2	−0·1	QTLs far apart on same chromosome. Dissonant
2	1	69	−2·4	−0·2	differences QTL effects in the same direction.
1	1	30	−1·2	0·1	QTLs far apart on same chromosome. QTL effects in
2	1	69	2·4	−0·2	opposite directions
1	1	30	1·2	−0·1	QTLs far apart on same chromosome. QTL effects in
2	1	69	−2·4	0·2	opposite directions
1	1	30	1·2	−0·1	QTLs moderately distant on same chromosome.
2	1	53	2·4	−0·2	Consonant differences. QTL effects in the same direction.
1	1	30	−1·2	−0·1	QTLs moderately distant on same chromosome.
2	1	53	−2·4	−0·2	Dissonant differences QTL effects in the same direction.
1	1	30	−1·2	0·1	QTLs moderately distant on same chromosome. QTL
2	1	53	2·4	−0·2	effects in opposite directions
1	1	30	1·2	−0·1	QTLs moderately distant on same chromosome.
2	1	53	−2·4	0·2	QTL effects in opposite directions
1	1	30	1·2	−0·1	One interval between QTLs. Consonant differences.
2	1	40	2·4	−0·2	QTL effects in the same direction.
1	1	30	−1·2	−0·1	One interval between QTLs Dissonant differences.
2	1	40	−2·4	−0·2	QTL effects in the same direction.
1	1	30	−1·2	0·1	One interval between QTLs. QTL effects in opposite
2	1	40	2·4	−0·2	directions
1	1	30	1·2	−0·1	One interval between QTLs. QTL effects in
2	1	40	−2·4	0·2	opposite directions
1	1	30	1·2	−0·1	QTLs in adjacent intervals. Consonant differences.
2	1	35	2·4	−0·2	QTL effects in the same direction.
1	1	30	−1·2	−0·1	QTLs in adjacent intervals. Dissonant differences.
2	1	35	−2·4	−0·2	QTL effects in the same direction.
1	1	30	−1·2	0·1	QTLs in adjacent intervals. QTL effects in
2	1	35	2·4	−0·2	opposite directions
1	1	30	1·2	−0·1	QTLs in adjacent intervals. QTL effects in
2	1	35	−2·4	0·2	opposite directions

Chr = chromosome on which the QTL was located. QTL Location = position of the QTL. Mean Effect and Prop. Effect = additive effects of the QTL on the mean of the continuous component of the distribution and the proportion of observations in the point-mass at zero, respectively.

as the point-mass proportion increased, the number of observations in the continuous component decreased; thus, the power to detect differences in the mean based only on the continuous observations decreased. For the normal CIM, power also tended to decrease with increasing point-mass proportions for datasets in which the proportions do not differ. In this case, the large number of zeros reduced the overall mean differences. However, for consonant differences, when the point-mass observations enhanced the differences in means between the genotypes, the power of normal CIM to detect a given mean difference increased as the proportion of zeros increased. Interestingly, power of the normal CIM could be high for dissonant differences when the point-mass proportion and the difference between genotypes were large. At large point-mass proportions, there were few continuous observations. A large difference in point-mass proportions created a large difference in overall means, often in the opposite direction to the difference of the means of the continuous observations.

The greatest differences between the two-part and normal methods occurred for simulated datasets with dissonant differences. For these datasets, the genotype with the larger proportion of observations in the point-mass had a larger mean for observations greater than zero. With dissonant differences, the overall means of the two genotypes were brought closer together by the point-mass observations. Since the normal CIM tests for a difference in the overall means

https://doi.org/10.1017/S0016672310000042 Published online by Cambridge University Press

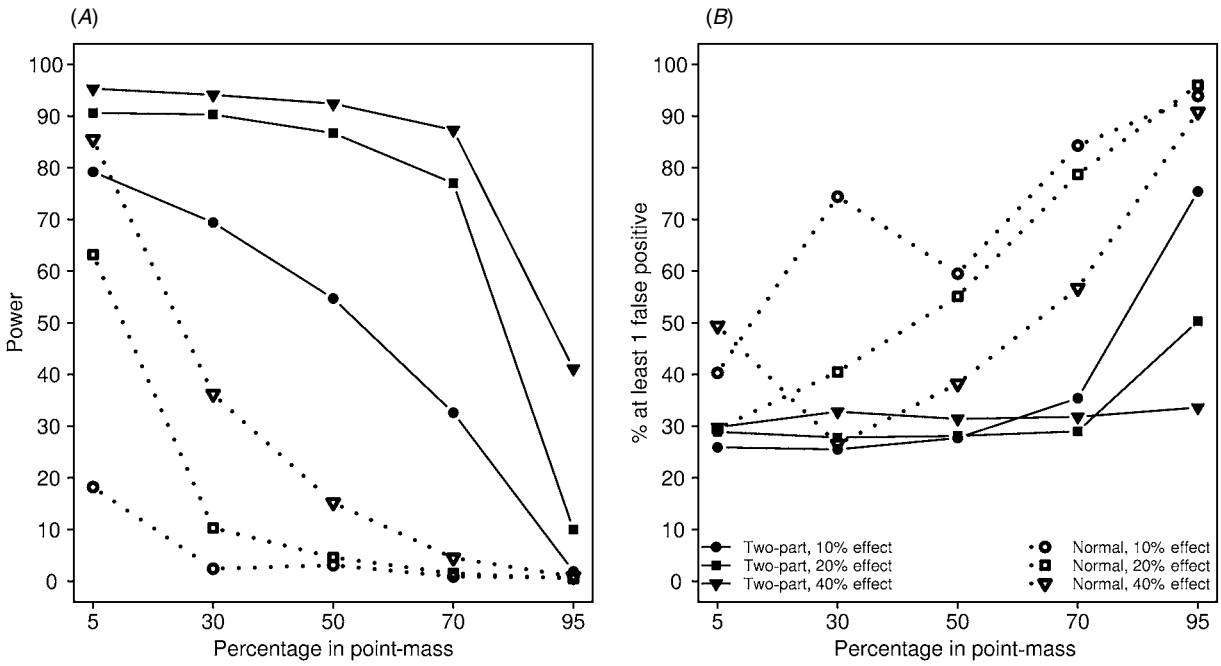

Fig. 1. Results of applying normal CIM and two-part CIM to simulated point-mass mixture datasets for a backcross with one QTL. The point-mass proportions did not differ between genotypes but the mean of the continuous component did. The additive effect of the QTL on the mean was 10, 20 and 40% of the mean of continuous observations in homozygotes. Power, defined as the percentage of datasets for which the 1·5 LOD support interval of at least one predicted QTL covered the true QTL, is shown in (*A*). The percentage of false positives out of the total number of datasets for which at least one QTL was predicted is shown in (*B*).

between the two genotypes, its power to detect QTLs was reduced. This effect was most noticeable at intermediate point-mass proportions (Fig. 2). The normal CIM also had a much higher percentage of false positives than the two-part CIM (Fig. 2).

(ii) *Results for two QTL simulations*

When we applied the three methods to simulated datasets with two QTLs, the two-part CIM continued to have the best performance, more frequently identifying at least one of the QTLs. As with the single-QTL simulations, the binary CIM performed poorly relative to the other two methods, never detecting either of the QTLs in more than 60% of the datasets for each simulation scenario (Fig. 3). Because of the poor performance of the binary CIM, we again focus on comparing normal and two-part CIM methods. Detailed results are provided in Supplemental Table 2 and summarized below.

The two-part CIM always had greater power than the normal CIM when the two QTLs were located on two different chromosomes (Supplemental Table 2). This method correctly identified both QTLs in more than 90% of the datasets and at least one of the QTLs in all datasets regardless of whether the QTL effects were consonant or dissonant. In comparison, the normal CIM detected both QTLs in only 24% of the datasets when both QTLs had consonant effects on the mean and point-mass proportions, and in only 1% of the datasets when the effects were dissonant.

When both QTLs were on the same chromosome, the number of QTLs predicted by each method was influenced by the distance between the QTLs and the direction of the effects. The two-part CIM method predicted two QTLs for most simulated datasets when the QTLs were more than 20 cM apart (Fig 4). For these simulations, the normal CIM commonly predicted only one QTL, but frequently failed to predict any QTLs when the QTL effects were dissonant. As the distance between the QTLs decreased, the two-part CIM method predicted only one QTL in more of the simulated datasets. The normal CIM continued to predict one QTL in most datasets when both QTLs had consonant differences, but often did not predict any QTL when the effects of both QTLs were dissonant or in opposite directions.

For most simulations with both QTLs on the same chromosome, the two-part CIM remained more powerful than the normal CIM (Fig. 3). The two-part CIM detected at least one QTL in at least 89% of the datasets for all but two simulations whereas the normal CIM detected at least one QTL in fewer than 90% of the datasets for all but two simulations. The one simulation scenario for which the two-part CIM had lower power than the normal CIM was when the QTLs were located with one interval between them (i.e. at 30 and 40 cM) and both QTLs had consonant

https://doi.org/10.1017/S0016672310000042 Published online by Cambridge University Press

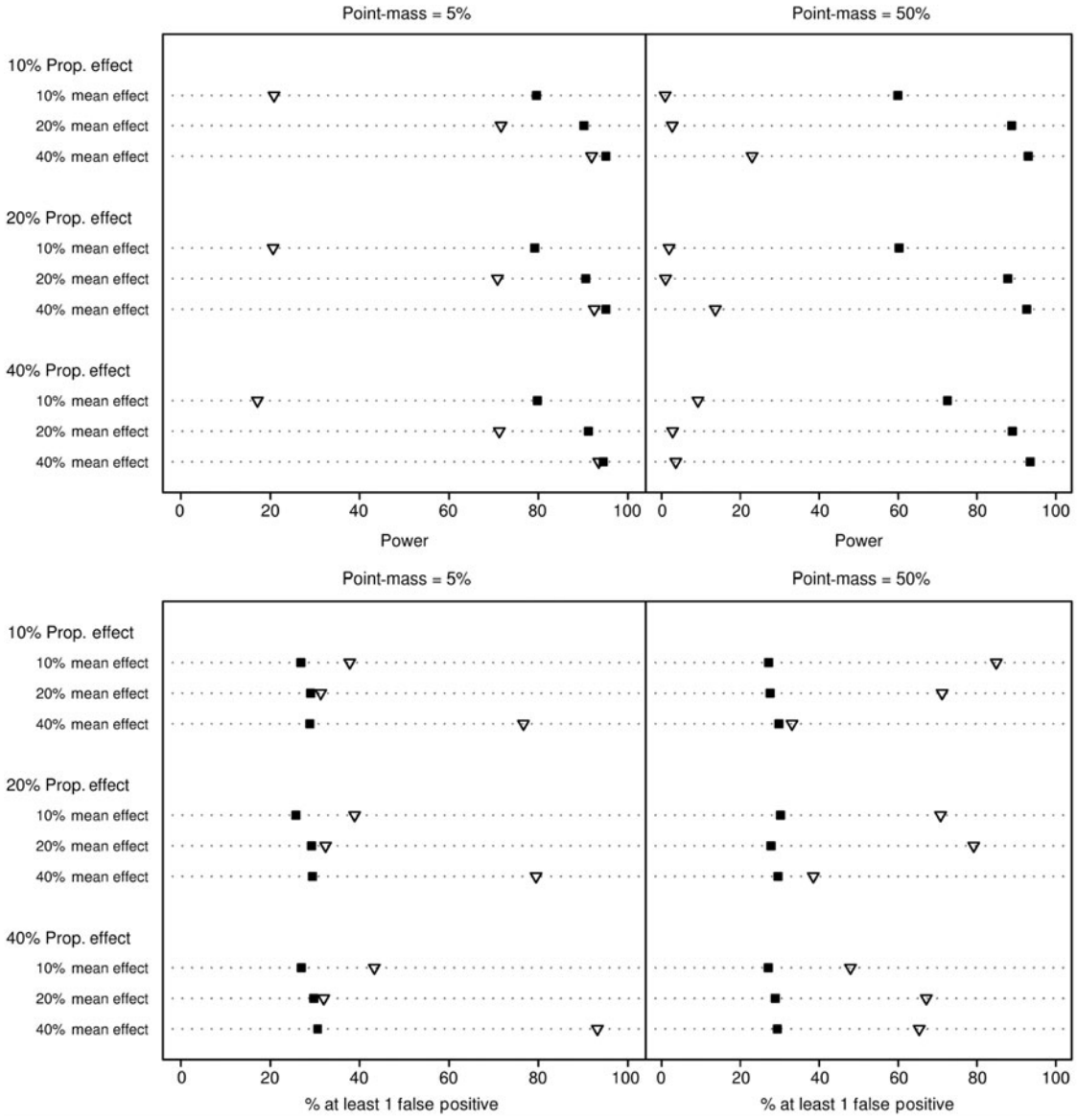

Fig. 2. Power and percentage of datasets with at least one false positive for simulated datasets for which both the point-mass proportions and means of the continuous component differed and the differences were dissonant. Power was calculated as the percentage of datasets for which the 1·5 LOD support interval of at least one predicted QTL covered the true QTL. Percentage of datasets with at least one false positive was calculated as the percentage of datasets with at least one QTL whose 1·5 LOD support interval did not cover the true QTL out of the total number of datasets for which at least one QTL was predicted. Black squares represent the two-part CIM and open triangles show the normal CIM. Vertical axis shows the simulated effects on the mean and point-mass proportions. Results are shown for simulations for which the homozygote point-mass proportions were 5 and 50%.

effects. For these simulations, both methods often predicted a single QTL between the two true QTLs. However, the two-part CIM method tended to have a steeper LOD score profile than the normal CIM method such that the support intervals were shorter and did not cover either of the true QTLs.

The percentages of datasets with at least one false positive were generally similar for the two-part and normal CIM methods and typically 20–40% (Fig. 5). For a few simulations, the two-part CIM had a higher percentage of datasets with at least one false positive, notably simulations when the QTLs were at 30 and 40 cM. As noted above, the two-part CIM tended to predict a single QTL between these points with a short LOD support interval that did not cover either QTL and thus, yielded a false positive. For other simulations, two-part CIM method predicted more datasets with greater than two QTLs than the normal CIM method (Fig. 4). These excessive QTLs often corresponded to smaller peaks on the shoulders of main peaks that were identified as peaks by the automated process we used to identify QTLs.

https://doi.org/10.1017/S0016672310000042 Published online by Cambridge University Press

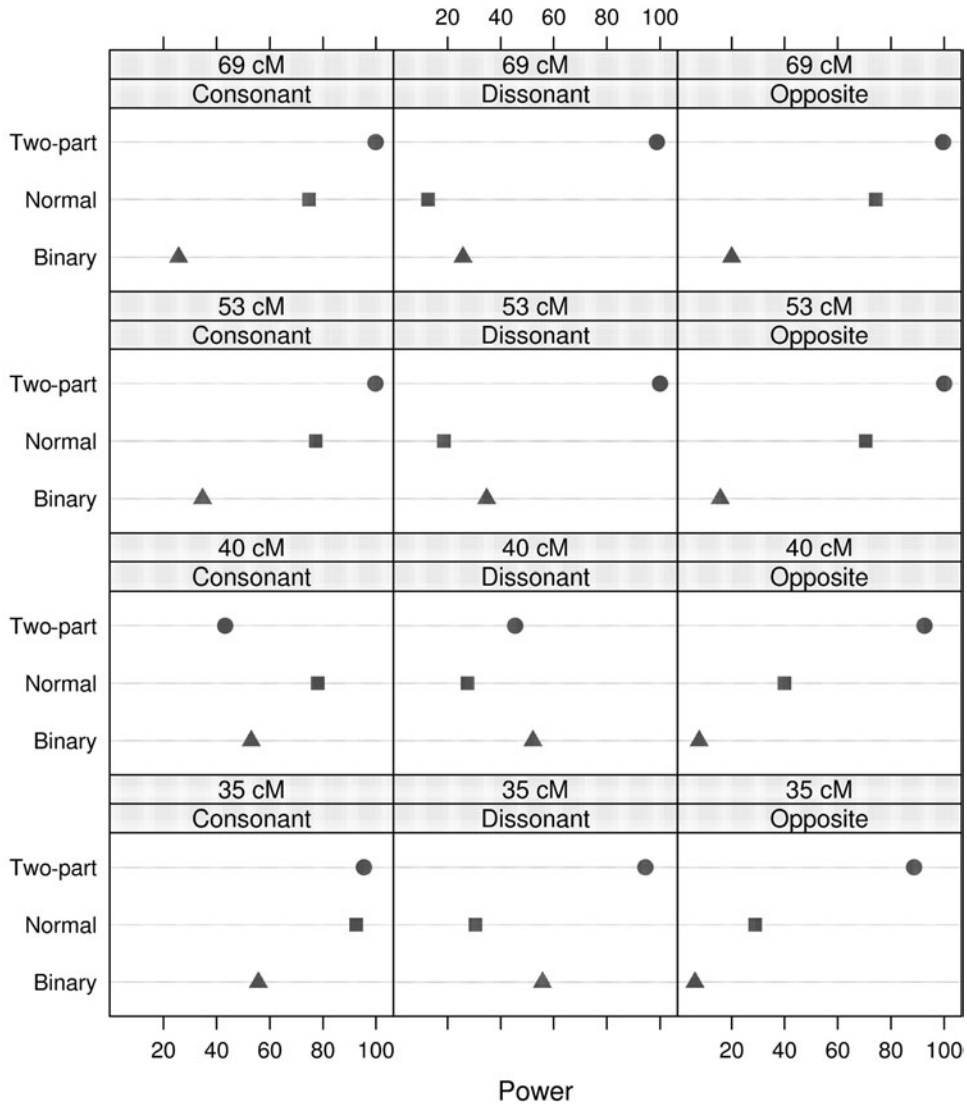

Fig. 3. Power of normal, binary and two-part CIM applied to simulated point-mass mixture datasets for a backcross with two QTLs. The first QTL was located at 30 cM and the second QTL was located at 69, 53, 40 and 35 cM. Power is the percentage of datasets for which the 1·5 LOD support interval of at least one predicted QTL covered at least one of the true QTLs. Results are shown for consonant effects (QTL 1: proportion effect = −0·1, mean effect = 1·2, QTL 2: proportion effect = −0·2, mean effect = 2·4), dissonant effects (QTL 1: proportion effect = −0·1, mean effect = −1·2, QTL 2: proportion effect = −0·2, mean effect = −2·4) and opposite effects (QTL 1: proportion effect = 0·1, mean effect = −1·2, QTL 2: proportion effect = −0·2, mean effect = 2·4).

(iii) *Comparison with interval mapping*

Next, we compared our two-part CIM method to two-part interval mapping (Broman, 2003). We applied two-part interval mapping to simulations with both QTLs on the same chromosome with the second QTL located at 69, 53, 40 or 35 cM. Effects of the first QTL were −0·1 on the point-mass proportion and 1·2 on the mean of the continuous component. For the second QTL, the effects were −0·2 and 2·4 on the proportion and mean, respectively.

The two-part CIM yielded smoother LOD profile curves with narrower and more well-defined peaks than the interval mapping method (Fig. 6). When the second QTL was at 53 or 69 cM, the LOD score profile of the two-part CIM method typically produced two clear peaks at or near the true QTL locations. In contrast, the LOD score profile of the interval mapping method usually had one large peak near the second QTL. Unlike the CIM method, the LOD score profile from interval mapping tended to gradually climb to the maximum LOD score but with much of the profile exceeding the significance threshold. Further, while the LOD score profile for the CIM method was relatively smooth between the two peaks, the LOD score profile for the interval mapping method often had small peaks between markers. These characteristics obscured the signal of the smaller effect QTL. When the two QTLs were 10 cM or closer, neither method was effective at discerning

https://doi.org/10.1017/S0016672310000042 Published online by Cambridge University Press

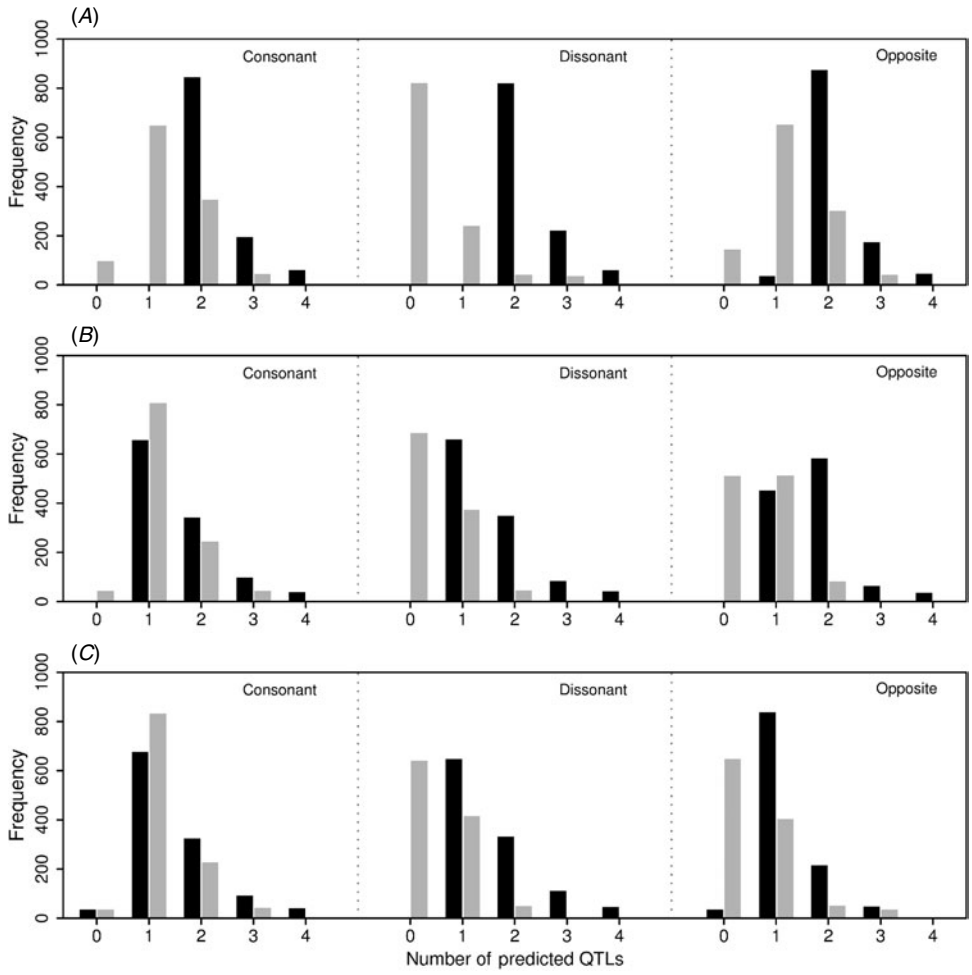

Fig. 4. Number of QTLs predicted by normal and two-part CIM methods for 1000 simulated dataset with two QTLs. The first QTL was at 30 cM and the second was at 69 cM (*A*), 40 cM (*B*) and 35 cM (*C*). Results are shown for consonant effects (QTL 1: proportion effect = −0·1, mean effect = 1·2, QTL 2: proportion effect = −0·2, mean effect = 2·4), dissonant effects (QTL 1: proportion effect = −0·1, mean effect = −1·2, QTL 2: proportion effect = −0·2, mean effect = −2·4) and opposite effects (QTL 1: proportion effect = 0·1, mean effect = −1·2, QTL 2: proportion effect = −0·2, mean effect = 2·4).

both QTLs. Nevertheless, the two-part CIM method produced narrower and more well-defined peaks than the interval mapping method.

5. Discussion

CIM was developed to improve upon interval mapping, which can yield biased estimates of the location and effect of a QTL when there are multiple linked QTLs. Although CIM accounts for the effects of multiple QTL, it is a single QTL model, testing for the presence of a QTL in one interval at a time. Subsequent methods, such as MIM (Kao *et al*., 1999) and Bayesian mapping methods (Satagopan *et al*., 1996), were developed to explicitly model and test for the presence of multiple QTLs. However, both MIM and Bayesian methods are computationally intensive. For example, using the R packages qtl and qtlbim, we compared processing times for CIM, MIM and Bayesian interval mapping when applied to 10 simulated traits with three QTLs in a backcross population of 200 individuals. The processing time for MIM averaged 20 times longer than for CIM, while BIM averaged nearly 80 times longer. Many researchers are now conducting QTL mapping studies using 'omics data, consisting of thousands of traits as phenotypes. The computational requirements of MIM and Bayesian methods limit the broad application of these approaches to these very large datasets.

Interval mapping is commonly used to identify which chromosomes carry QTL, with further studies focusing on determining the number and location of QTLs. In our comparison of CIM to interval mapping, interval mapping was effective at identifying the general location of the largest effect QTL on a chromosome but did not clearly indicate the presence of more than one QTL. CIM was effective at identifying more than one QTL on a chromosome when the QTLs were not very close together and yielded narrower and more well-defined peaks than interval mapping.

https://doi.org/10.1017/S0016672310000042 Published online by Cambridge University Press

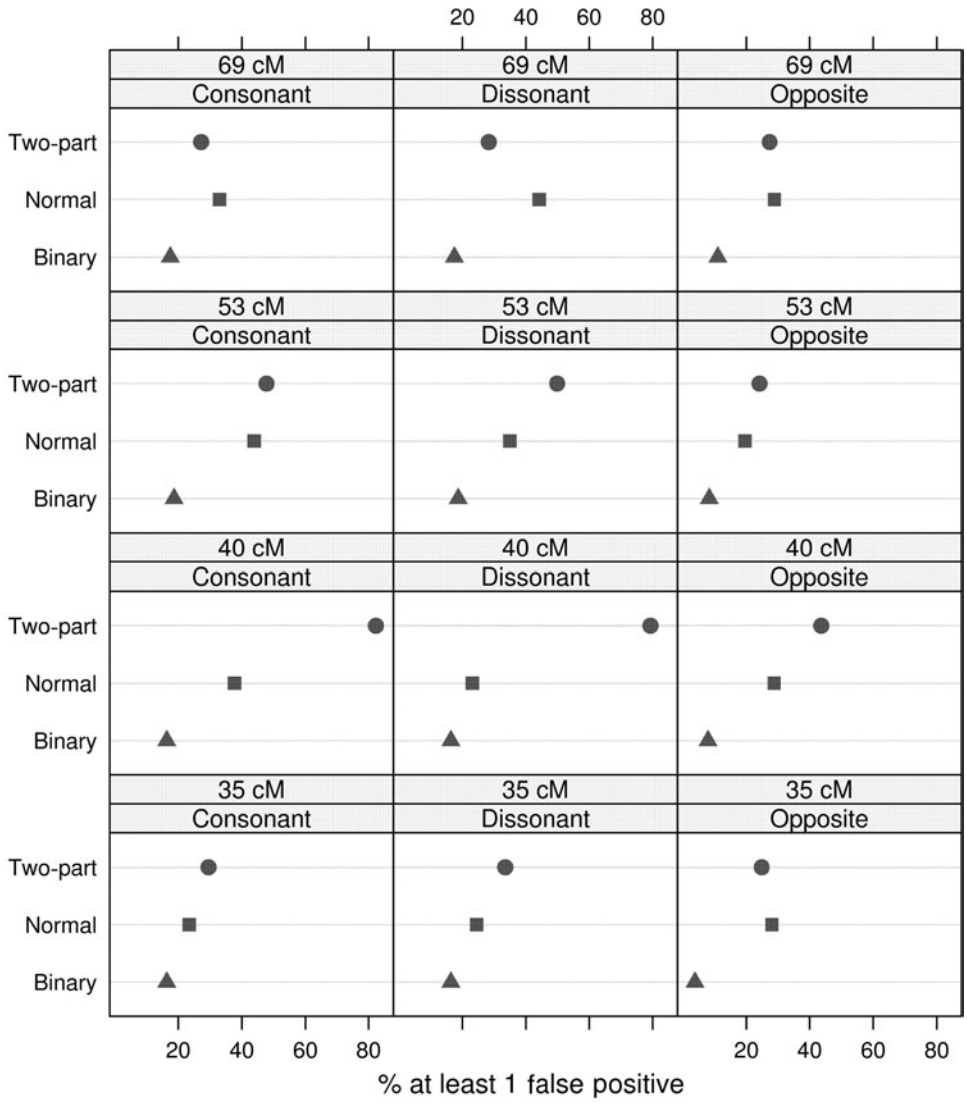

Fig. 5. False positives rates for normal, binary and two-part CIM applied to simulated point-mass mixture datasets for a backcross with two QTLs. The first QTL was located at 30 cM and the second QTL was located at 69, 53, 40 and 35 cM. The percentage with at least 1 false positive is the percentage of datasets with at least one predicted QTL whose 1·5 LOD support interval did not cover a true QTL out of the total number of datasets for which at least one QTL was predicted. Results are shown for consonant effects (QTL 1: proportion effect = −0·1, mean effect = 1·2, QTL 2: proportion effect = −0·2, mean effect = 2·4), dissonant effects (QTL 1: proportion effect = −0·1, mean effect = −1·2, QTL 2: proportion effect = −0·2, mean effect = −2·4) and opposite effects (QTL 1: proportion effect = 0·1, mean effect = −1·2, QTL 2: proportion effect = −0·2, mean effect = 2·4).

Many methods are available to researchers for mapping QTLs, each with advantages and disadvantages. CIM offers a computationally tractable approach to QTL mapping that can play an integral role in a comprehensive QTL mapping study. Results of an initial CIM analysis can be used to identify a small number of traits to further investigate with multiple QTL methods. Further, in MIM, many different models are evaluated to identify the best fitting model in terms of number and location of QTLs. Results from CIM can be used to identify starting models and thereby reduce the search space for the MIM procedure (see e.g. Basten *et al.*, 2005). Bayesian methods take a fundamentally different approach than the other QTL mapping methods. While these methods can be the only method used for mapping QTL in a dataset, Bayesian methods have been recommended as a complementary analysis following application of a standard method, such as CIM (Wang *et al.*, 2007). Thus, CIM remains an important and commonly used technique in the suite of methods available for QTL mapping.

Marker density has been increasing, particularly for intensely studied model organisms. Interval mapping methods, including CIM, will remain important tools for organisms with sparse genetic maps. However, for organisms with very dense marker maps, QTL mapping can be conducted by analysing the data at

https://doi.org/10.1017/S0016672310000042 Published online by Cambridge University Press

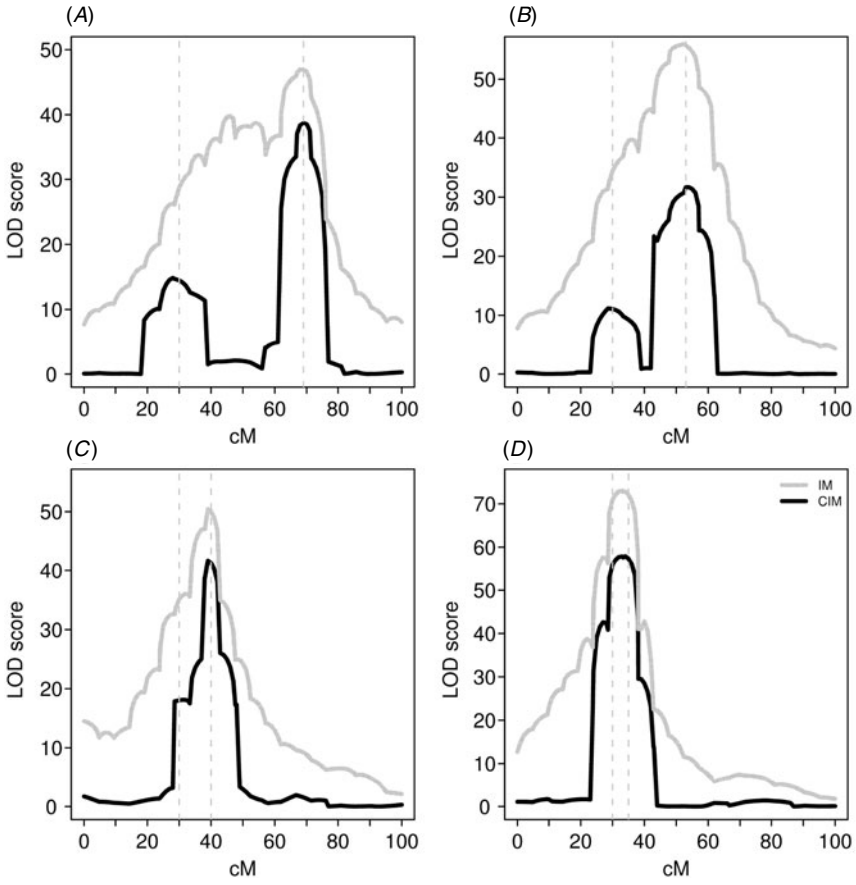

Fig. 6. Representative LOD score profiles from applying two-part interval mapping (IM) and two-part CIM (CIM) to simulated point-mass mixture datasets with two QTLs. Results are from one simulated dataset with the first QTL located at 30 cM and the second QTL was located at 69 (*A*), 53 (*B*), 40 (*C*), and 35 cM (*D*). Dashed, vertical lines show the true QTL locations.

marker locations using techniques such as Haley & Knott's (1992) regression method. For point-mass mixtures, the likelihood factors into two parts (binary and continuous components) at markers, and our two-part CIM method reduces to a two-part statistic that jointly tests for a difference in means and a difference in proportions.

Two-part tests are advantageous over conducting two separate single-part tests, i.e., one for a difference in means and one for a difference in point-mass proportions. When a QTL affects both the point-mass proportion and the mean of the continuous component, Broman (2003) showed that two-part interval mapping is more powerful than the single-part methods, binary and normal likelihood interval mapping using only the continuous observations. Similarly, our results showed the two-part CIM method to have higher power and a lower false positive rate as compared to the binary CIM method and the normal CIM when 0 values were dropped from the analysis. We expect a similar result when testing at markers with a two-part statistic as compared to conducting two separate single-part tests. Finally, conducting separate analyses would require analysing the data twice, whereas using a two-part test necessitates only one pass through the data, thus reducing computation time and avoiding multiple testing issues.

Increasingly researchers are conducting QTL mapping in metabolomics and proteomics studies. These data are often distributed as a point-mass mixture, consisting of a spike at zero in combination with continuous non-negative measurements. Similar distributions arise in other settings. CIM methods in commonly used software packages implement Zeng's (1994) method, which assumes the data are normally distributed and is therefore inappropriate with point-mass mixtures.

In this paper, we presented a new CIM method based on two-part statistics that accounts for differences in point-mass proportions as well as differences in means of the quantitative measurements between genotypes. Our simulations showed that this novel two-part CIM method performs better than the normal CIM and binary CIM methods when data are distributed as a point-mass mixture. The two-part CIM has higher power, as well as lower and more stable false positive rates compared to these methods. By using information from both the binary and

https://doi.org/10.1017/S0016672310000042 Published online by Cambridge University Press

continuous components, the two-part CIM is able to detect smaller differences than are the other methods. Our approach also has a substantial advantage over the normal CIM when the difference in the means and point-mass proportions are in opposite directions, as dissonant differences act to obscure differences in the overall means.

Multiple problems arise in applying the normal CIM method to data distributed as a point-mass mixture. Because of the spike in the data, even modest numbers of point-mass observations severely violate the normal model assumption. Additionally, with 'omics data, the non-zero observations are often highly right-skewed, further violating the normality assumption of the normal CIM model. Commonly, a log transformation is applied to right-skewed 'omics data'. But zero values cannot be log transformed. Using the normal CIM then requires (1) dropping the zero values and only analysing the non-zero observations, (2) employing an *ad hoc* approach such as adding a small amount to the zero in order to log transform the data, or (3) analysing the untransformed data including the zero values.

None of these approaches is desirable. With dropping zero values, informative data are eliminated and the effective sample size is reduced. Further, by not taking into account the zero values, the results could be biased or misleading. While the *ad hoc* approach of adding a small amount to each zero allows use of a log transformation, the data could still substantially violate the normality assumption because of the spike in the distribution. Our simulations showed that normal CIM performed poorly when the point-mass proportion was large, a condition that would not be improved with this approach. Finally, untransformed data, retaining the zeros, could severely violate model assumptions. Our simulations suggest that the normal CIM has low power and high false positives rates in this case as well.

Point-mass mixtures with the point-mass values at zero can be modelled differently depending on whether the zeros are considered true zeros and/or assumed to result from left-censoring. Zero values caused by censoring represent values from the left tail of a continuous distribution and would provide information about the mean of the continuous distribution if they were not truncated. True zero values not resulting from truncation do not provide information about the continuous component of the distribution.

In our two-part CIM method, we modelled the zero values as true zeros that did not arise as a result of censoring of the continuous distribution. A more complicated model could combine 'true zeros' reflecting the true absence of the metabolite and zeros resulting from metabolite values below detection limits (i.e. censored values from a continuous distribution) (Moulton & Halsey, 1995; Moulton *et al.*, 2002). This model could be particularly appropriate for metabolomics data in which some zeros may reflect the absence of a compound while other zeros may result from metabolites having concentrations below the detection limit.

In this work, we jointly considered the binary and continuous component when selecting covariate markers, i.e., we assumed that that the covariates affect both the probability of being in the point-mass and the level of the continuous values. The model could be generalized by allowing the two components to have different marker covariates and testing at a given genome position could assess the effects on both components or one or the other (see e.g. Moulton *et al.*, 2002). Alternatively or in addition to, as suggested by Moulton *et al.* (2002), a proportionality parameter could be introduced to model the relative contribution of covariates to the probability of being in the point-mass and the magnitude of the continuous observations. Greater flexibility in model structure could better reflect the underlying biological processes. For example, an allele at one QTL could block a pathway creating a metabolite and result in true zeros while other QTLs could regulate the amount of the metabolite produced. More refined modelling could facilitate elucidating these effects. Finally, markers selected as covariates could represent QTL that act epistatically with a QTL at a testing location. Integrating epistatic effects into CIM methods, including normal CIM, could be a promising area for future improvements to the methods.

The two-part CIM method described in this article has been implemented in R as a companion package, twopartqtl, to the QTL mapping package R/qtl and is available for download from www.r-project.org.

The authors are grateful to Dr Daniel Kliebenstein for providing experimental data and to Dr Kyoungmi Kim for comments on an earlier version of this manuscript.

References

Basten, C. J., Weir, B. S. & Zeng, Z.-B. (2005). QTL Cartographer. A reference Manual and Tutorial for QTL Mapping. http://statgen.ncsu.edu/qtlcart/manual.pdf.

Broman, K. W. (2003). Mapping quantitative trait loci in the case of a spike in the phenotype distribution. *Genetics* **163**, 1169–1175.

Broman, K. W., Wu, H., Sen, S. & Churchill, G. A. (2003). R/qtl: QTL mapping in experimental crosses. *Bioinformatics* **19**, 889–890.

Deng, W., Chen, H. & Li, Z. (2006). A logistic regression mixture model for interval mapping of genetic trait loci affecting binary phenotypes. *Genetics* **172**, 1349–1358.

Diao, G., Lin, D. Y. & Zou, F. (2004). Mapping quantitative trait loci with censored observations. *Genetics* **168**, 1689–1698.

Fine, J. P., Zou, F. & Yandell, B. S. (2004). Nonparametric estimation of the effects of quantitative trait loci. *Biostatistics* **5**, 501–513.

Hackett, C. A. & Weller, J. I. (1995). Genetic mapping of quantitative trait loci for traits with ordinal distributions. *Biometrics* **51**, 1252–1263.

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

Huang, C.-Y., Qin, J. & Zou, F. (2007). Empirical likelihood-based inference for genetic mixture models. *The Canadian Journal of Statistics* **34**, 563–574.

Jensen, R. C. (1993). Interval mapping of multiple quantitative trait loci. *Genetics* **135**, 205–211.

Jin, C., Fine, J. P. & Yandell, B. S. (2007). A unified semiparametric framework for quantitative trait loci analyses, with application to spike phenotypes. *Journal of the American Statistical Association* **102**, 56–67.

Kao, C.-H., Zeng, Z.-B. & Teasdale, R. D. (1999). Multiple interval mapping for quantitative trait loci. *Genetics* **152**, 1203–1216.

Knott, S. A. & Haley, C. S. (1992). Aspects of maximum likelihood methods for the mapping of quantitative trait loci in line crosses. *Genetical Research* **60**, 139–151.

Kruglyak, L. & Lander, E. S. (1995). A nonparametric approach for mapping quantitative trait loci. *Genetics* **139**, 1421–1428.

Lander, E. S. & Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185–199.

Li, W. & Chen, Z. (2009). Multiple interval mapping for quantitative trait loci with a spike in the trait distribution. *Genetics* **182**, 337–342.

Loudet, O., Chaillou, E. S., Camilleri, C., Bouchez, D. & Daniel-Vedele, F. (2002). Bay-0 × Shahdara recombinant inbred line population: a powerful tool for the genetic dissection of complex traits in *Arabidopsis*. *Theoretical and Applied Genetics* **104**, 1172–1184.

Martinez, O. & Curnow, R. N. (1992). Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. *Theoretical and Applied Genetics* **85**, 480–488.

Moulton, L. H. & Halsey, N. A. (1995). A mixture model with detection limits for regression analyses of antibody response to vaccine. *Biometrics* **51**, 1570–1578.

Moulton, L. H., Curriero, F. C. & Barroso, P. F. (2002). Mixture models for quantitative HIV RNA data. *Statistical Methods in Medical Research* **11**, 317–325.

R Development Core Team. (2008). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rowe, H. C., Hansen, B. G., Halkier, B. A. & Kliebenstein, D. J. (2008). Biochemical networks and epistasis shape in the *Arabidopsis thaliana* metabolome. *Plant Cell* **20**, 1199–1216.

Satagopan, J. M., Yandell, B. S., Newton, M. A. & Osborn, T. C. (1996). A Bayesian approach to detect quantitative trait loci using Markov Chain Monte Carlo. *Genetics* **144**, 805–816.

Thomson, P. C. (2003). A generalized estimating equations approach to quantitative trait locus detection of non-normal traits. *Genetics Selection Evolution* **35**, 257–280.

Wang, S., Basten, C. J. & Zeng, Z.-B. (2007). Windows QTL Cartographer Version 2.5. Statistical Genetics, North Carolina State University.

Xu, S. & Atchley, W. R. (1996). Mapping quantitative trait loci for complex binary diseases using line crosses. *Genetics* **143**, 1417–1424.

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

Zou, F., Fine, J. P. & Yandell, B. S. (2002). On empirical likelihood for a semiparametric mixture model. *Biometrika* **89**, 61–75.