RWGAIM: an efficient high-dimensional random whole genome average (QTL) interval mapping approach

ARŪNAS P. VERBYLA[1,2]*, JULIAN D. TAYLOR[1,2] AND KLARA L. VERBYLA[3]

[1] *School of Agriculture, Food and Wine, The University of Adelaide, PMB 1, Glen Osmond, SA 5064, Australia*
[2] *Mathematics, Informatics and Statistics and Food Futures National Research Flagship, CSIRO, Glen Osmond SA 5064, Australia*
[3] *Mathematics, Informatics and Statistics and Food Futures National Research Flagship, CSIRO, Acton ACT 2601, Australia*

Summary

Mapping of quantitative trait loci (QTLs) underlying variation in quantitative traits continues to be a powerful tool in genetic study of plants and other organisms. Whole genome average interval mapping (WGAIM), a mixed model QTL mapping approach using all intervals or markers simultaneously, has been demonstrated to outperform composite interval mapping, a common approach for QTL analysis. However, the advent of high-throughput high-dimensional marker platforms provides a challenge. To overcome this, a dimension reduction technique is proposed for WGAIM for efficient analysis of a large number of markers. This approach results in reduced computing time as it is dependent on the number of genetic lines (or individuals) rather than the number of intervals (or markers). The approach allows for the full set of potential QTL effects to be recovered. A proposed random effects version of WGAIM aims to reduce bias in the estimated size of QTL effects. Lastly, the two-stage outlier procedure used in WGAIM is replaced by a single stage approach to reduce possible bias in the selection of putative QTL in both WGAIM and the random effects version. Simulation is used to demonstrate the efficiency of the dimension reduction approach as well as demonstrate that while the approaches are very similar, the random WGAIM performs better than the original and modified fixed WGAIM by reducing bias and in terms of mean square error of prediction of estimated QTL effects. Finally, an analysis of a doubled haploid population is used to illustrate the three approaches.

1. Introduction

High-throughput high-dimensional marker platforms are now available for many species. The Affymetrix Genome-Wide Human SNP Array 6.0 (http://www.affymetrix.com) has been available for some time and has almost 1 million SNPs in a 1·8 million marker array. New generation sequencing techniques (Varshney *et al.*, 2009) promise an explosion of data that will be challenging for association studies, quantitative trait locus (QTL) analysis and genomic prediction. There is a need to provide statistical methods that handle high-dimensional data. The focus of this paper is QTL analysis in plants, however, the methods presented could be applied to any species.

Estimated sizes of QTLs are often used by researchers to further their understanding of the impact of individual QTL. However, it has been understood for some time that QTL analysis leads to biased estimates of putative QTL effect sizes. Beavis (1994, 1998) discusses the inflated effects that result from QTL analyses, Xu (2003) provides a theoretical justification, while Melchinger *et al.* (1998) also conclude that sizes are inflated. Developing methods that reduce bias is desirable.

There are many methods available for the analysis of QTL; it is not possible to review all the methods so a selection of key papers are discussed. Initial methods of analysis used individual markers (Weller, 1986) and interval mapping using mixture models (Lander & Botstein, 1989). Haley & Knott (1992) and Martinez & Curnow (1992) introduced the regression approach for interval mapping. The bias in simple interval mapping led to composite interval mapping (CIM; Zeng, 1994; Jansen, 1994) in which co-factors are used to control for other QTLs when searching for a single QTL. This, however, requires the selection of co-factors. Interval

* Corresponding author: School of Agriculture, Food and Wine, The University of Adelaide, PMB 1, Glen Osmond, SA 5064, Australia. Tel: +618 8303 8769. e-mail: ari.verbyla@adelaide.edu.au

mapping and CIM usually involve fitting models at regular intervals along the genome and permutation for determining genome-wide thresholds for significance which can be very time-consuming. To overcome this latter problem, Piepho (2001) presents a fast approximate method for determining the threshold for both interval mapping and CIM. Multiple interval mapping (MIM; Kao *et al.*, 1999) was introduced to try and improve detection of QTL by allowing multiple rather than single QTL to be fitted at the same time. This approach requires some knowledge of how many QTLs are present and there are model selection issues. The aim of CIM and MIM is to allow for multiple QTLs in the analysis because the presence of other QTL impacts on the estimation of each QTL.

Bayesian methods have been used extensively in QTL analysis, in particular for mapping multiple QTLs (Satagopan *et al.*, 1996; Heath, 1997; Sillanpää & Arjas, 1998, 1999; Xu, 2003; Yi, 2004; Wang *et al.*, 2005); for an excellent review, see Yi & Shriner (2008). Che & Xu (2010) present a fast method for determining thresholds in a Markov chain Monte Carlo approach. There are methods (Xu, 2003; Yi *et al.*, 2003) that incorporate all the markers simultaneously, thereby removing the need to select co-factors or the specific number of QTLs in MIM.

Verbyla *et al.* (2007) presented a non-Bayesian whole genome average interval mapping (WGAIM) approach for QTL analysis of a single trait in a single trial in which all the intervals on a linkage map are used simultaneously. While this adds some complexity, it overcomes the need for repeated genome scans and a threshold for QTL detection is readily available. In addition, because mixed models are used for analysis, non-genetic effects, such as terms for experimental design, are easily included. WGAIM uses forward selection hence the number of analyses are greatly reduced. A simple random effects working model is used in which all the intervals are allowed to contain a possible QTL and a likelihood ratio test of significance of the random effects working model is conducted to decide if selection of a putative QTL is warranted. If selection is warranted, a two-stage outlier procedure (firstly linkage groups then intervals within a selected linkage group) is used to select the most likely putative QTL, that is then moved to the fixed effects part of the model. The approach shown by Verbyla *et al.* (2007) was much more powerful than CIM, although there is a small increase in selecting false positives. The approach has been implemented in the R environment (R Development Core Team, 2012) in the package wgaim; see Taylor & Verbyla (2011). Note that although the original formulation of WGAIM used intervals, the wgaim package also allows analysis using markers rather than intervals.

While WGAIM performs well, there are aspects of the approach that require refinement. Since WGAIM uses all the intervals (or markers) simultaneously, an efficient method of analysis is required when the number of intervals (or markers) is very large, typically much larger than the number of observations on the trait of interest. The (selection) bias of WGAIM was not examined by Verbyla *et al.* (2007). The outlier procedure uses two stages and this can influence the results. This paper addresses the high-dimensional problem, the issue of bias and the outlier procedure. The high-dimensionality question is addressed by reformulating the random effects model for QTL sizes in WGAIM in such a way that the size of a random term is equal to the number of lines with marker data when the number of markers is greater than the number of lines. The question of selection bias is addressed by assigning each putative QTL to a random effect, rather than moving them to the fixed effects component of the model, in the forward selection process. Thus, the resulting model has multiple random components, a separate component for each putative QTL and one component for non-QTL. This procedure is called RWGAIM in this paper. The other modification of WGAIM is the reduction of the outlier procedure from two stages, linkage groups then intervals, to a one-stage procedure based only on intervals. This is applicable to WGAIM, yielding a method called WGAIMI, and to RWGAIM.

The paper is organized as follows. Firstly, the WGAIM approach is outlined. The extension that allows for more markers than genetic lines on a trait is presented. Then a modification of WGAIM is described in which each putative QTL is placed in a random effect rather than a fixed effect as they are selected. Thirdly, an intuitive argument is used to justify using a single stage outlier procedure for selection of putative QTL. Three simulations are conducted to examine the three aspects. In the first, the number of markers was varied for a fixed number of lines to compare timings using a conventional and the dimension reduction approach for analysis. This simulation involves low to moderate numbers of markers to facilitate the comparison. The second simulation considers genuine high-dimensional situations, with both population sizes and marker numbers varying. Lastly, a simulation study using the setup along the lines of Verbyla *et al.* (2007) is used to assess the impact of using random effects for the QTL selected and the single stage outlier procedure using WGAIM, WGAIMI and RWGAIM. Detection of QTL, false positive rates and importantly the estimated sizes of QTL are summarized. A wheat-doubled haploid population is used to conduct a QTL analysis using WGAIM, WGAIMI and RWGAIM to provide a further comparison of results across the methods. The paper finishes with a brief discussion.
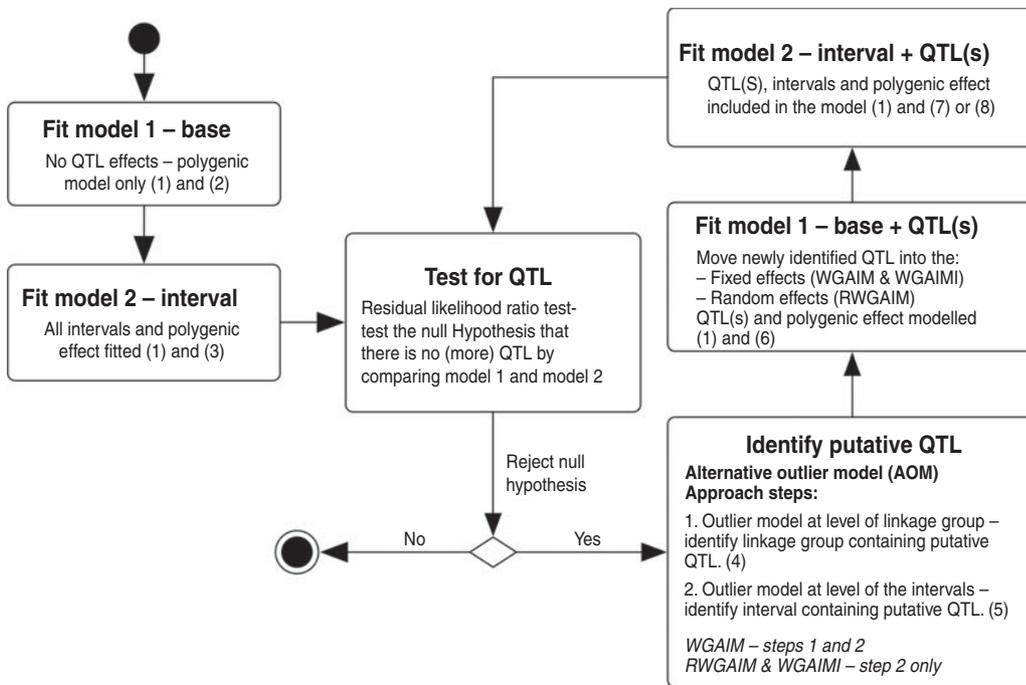
Fig. 1. Scheme for WGAIM, WGAIMI and RWGAIM. Difference between the approaches is indicated in italics and the equations in the text corresponding to each step are indicated in brackets where relevant.

2. Methods

The general scheme and differences between the WGAIM, WGAIMI and RWGAIM approaches are outlined in Fig. 1 and are detailed below. Equation numbers in Fig. 1 refer to those in the text.

(i) *WGAIM*

Verbyla *et al.* (2007) used mixed models as the basis for QTL analysis because the experiments to be analysed involved design effects and possibly other non-genetic effects that should be accounted for in the analysis. The models fitted to the data are of the form

$$y = X\tau + Z_0 u_0 + Z_g u_g + e, \tag{1}$$

where the vector y ($n \times 1$) consists of trait data. The components of (1) reflect the nature of the study. Thus, X ($n \times t$) and Z_0 ($n \times b$) are known design matrices for fixed and random effects, respectively, that arise from the design of the study and non-genetic sources of variation (Smith *et al.*, 2005, 2006), τ is the vector of fixed effects parameters, and u_0 is a vector of random effects. The vector of residual effects is denoted by e and this term and u_0 are assumed independent, mean zero with variance–covariance matrices R and G_0, respectively. The forms for R and G_0 will reflect the nature of the analysis.

If there are n_g lines of interest, the vector of genetic effects, u_g, is $n_g \times 1$. The design matrix Z_g in (1) assigns the appropriate genetic effect to each observation and thus consists of zeros and ones.

The WGAIM approach requires a baseline model to be fitted. This means specifying the genetic effects u_g. The standard model is the so-called infinitesimal model or polygenic model for which

$$u_g = u_p, \tag{2}$$

where $u_p \sim N(\mathbf{0}, \sigma_p^2 I_{n_g})$; if pedigree information is available, more complex models can be fitted (Oakey *et al.*, 2006).

The next step in WGAIM is to include the interval (or marker) regression. This is called the working model in Verbyla *et al.* (2007). Firstly, note if there are c linkage groups (or chromosomes) with linkage group k having r_k markers, hence we assume

$$u_g = \sum_{k=1}^{c} \sum_{j=1}^{r_k - 1} q_{kj} a_{kj} + u_p, \tag{3}$$

where q_{kj} is a vector of scores for a potential QTL in interval j of linkage group k; a_{kj} is the corresponding size of the QTL. The QTL scores reflect the two possible genotypes, AA and BB for doubled haploid and recombinant inbred lines, and AB and BB for back-cross populations; the scores are 1 and -1 for the two genotypes. As q_{kj} is unknown, by using the regression approach (Haley & Knott, 1992; Martinez & Curnow, 1992), it is replaced by its conditional expectation given flanking markers m_{kj} and $m_{k,j+1}$. The expression is given in Verbyla *et al.* (2007) and involves $\theta_{k;j,j+1}$, the recombination frequency between markers j and $j+1$ on linkage group k, and θ_{kj}, the recombination fraction between marker j and the

possible QTL in that interval. Verbyla *et al.* (2007) further average out the location by assuming the distance from the marker j to the QTL is uniformly distributed on the interval, thereby eliminating θ_{kj}. This results in the model

$$
\begin{aligned}
\boldsymbol{u}_g &= \sum_{k=1}^{c} \sum_{j=1}^{r_k-1} (\boldsymbol{m}_{kj} + \boldsymbol{m}_{k,j+1}) \, \psi_{kj} \, a_{kj} + \boldsymbol{u}_p, \\
&= \boldsymbol{M}_E \boldsymbol{a} + \boldsymbol{u}_p,
\end{aligned}
\tag{4}
$$

where $\psi_{kj} = \theta_{k;j,j+1}/2d_{k;j,j+1}(1-\theta_{k;j,j+1})$ and $d_{k;j,j+1}$ is Haldane's distance between markers j and $j+1$ on linkage group k. Note that $\boldsymbol{M}_E$ is a known matrix of pseudo-markers spanning the whole genome. To complete the working model, it is assumed that $\boldsymbol{a} \sim N(\boldsymbol{0}, \sigma_a^2 \boldsymbol{I}_{r-c})$ where $r = \sum_{k=1}^{c} r_k$. This is a very simple model and aims to provide a mechanism for determining if a putative QTL can be selected.

The WGAIM process starts by fitting (1) with (2) and (1) with (4) and conducting a residual likelihood ratio test of the hypothesis $H_0 \colon \sigma_a^2 = 0$. Under H_0, the statistic has a distribution that is a mixture of chi-squared distributions, namely $0 \cdot 5\chi_0^2 + 0 \cdot 5\chi_1^2$ (Stram & Lee, 1994). If the hypothesis is rejected, this suggests that there is at least one putative QTL.

To select the most likely interval for the putative QTL, an outlier detection approach is used. Verbyla *et al.* (2007) use the alternative outlier model (AOM). In the original formulation, the process occurred in two steps. Firstly, an outlier model was proposed at the level of the linkage group. Thus, the model proposed was

$$
\boldsymbol{u}_g = \boldsymbol{M}_E \, (\boldsymbol{a} + \boldsymbol{D}_k \boldsymbol{\delta}_k) + \boldsymbol{u}_p,
$$

where $\boldsymbol{D}_k$ is an $(r-c) \times (r_k-1)$ matrix with a one in position corresponding to each interval j on linkage group k and zero elsewhere. It is assumed $\boldsymbol{\delta}_k \sim N(0, \sigma_{a,k}^2 \boldsymbol{I}_{r_k-1})$. The idea behind this formulation is that if a putative QTL exists on linkage group k, the size of the effects of the putative QTL will be inflated by $\boldsymbol{\delta}_k$ through an increase in variance. Although it appears that many models have to be fitted, in fact a procedure based on a score statistic for the hypothesis $H_0 \colon \sigma_{a,k}^2 = 0$ was used by Verbyla *et al.* (2007), and this only relies on the fit of (1) with (4). The selection of the linkage group that contains the putative QTL is based on the outlier statistics given by

$$
t_k^2 = \frac{\sum_{j=1}^{r_k-1} \tilde{a}_{kj}^2}{\sum_{j=1}^{r_k-1} \mathrm{var}(\tilde{a}_{kj})},
\tag{5}
$$

where $\tilde{a}_{kj}$ is the best linear unbiased predictor (BLUP) of the size a_{kj} of a QTL in interval j on linkage group k, and $\mathrm{var}(\tilde{a}_{kj})$ is its variance. The linkage group with the largest statistic t_k^2 is selected as containing the putative QTL. Once the linkage group is selected, the

interval on that linkage group most likely to contain a putative QTL is selected. This involves a similar AOM, for intervals on linkage group k, namely

$$
\boldsymbol{u}_g = \boldsymbol{M}_E(\boldsymbol{a} + \boldsymbol{d}_{kj}\delta_{kj}) + \boldsymbol{u}_p,
$$

where $\boldsymbol{d}_{kj}$ is a vector with a one in position corresponding to interval j on linkage group k and $\delta_{kj} \sim N(0, \sigma_{a,kj}^2)$. A score statistic for the hypothesis $H_0 \colon \sigma_{a,kj}^2 = 0$ is used again and the outlier statistic is

$$
t_{kj}^2 = \frac{\tilde{a}_{kj}^2}{\mathrm{var}(\tilde{a}_{kj})}.
\tag{6}
$$

The interval with the largest t_{kj}^2 is selected as containing the putative QTL. This QTL interval is moved to the fixed effects. The forward selection continues with (2) and (4) replaced by

$$
\boldsymbol{u}_g = \boldsymbol{m}_{E,1} a_1 + \boldsymbol{u}_p,
\tag{7}
$$

$$
\boldsymbol{u}_g = \boldsymbol{m}_{E,1} a_1 + \boldsymbol{M}_{E,-1} \boldsymbol{a}_{-1} + \boldsymbol{u}_p,
\tag{8}
$$

where $\boldsymbol{m}_{E,1}$ is the vector of pseudo-marker scores for the first putative QTL and a_1 is the size of that QTL. $\boldsymbol{M}_{E,-1}$ is the matrix of scores for those pseudo-markers that have not been selected as putative QTL and $\boldsymbol{a}_{-1}$ is the vector of effect sizes for those pseudo-markers.

The process is repeated. Thus both (1) with (7) and (1) with (8) are fitted and the likelihood ratio test for $H_0 \colon \sigma_a^2 = 0$ is carried out again. If H_0 is rejected, another putative QTL is selected using the approach based on the AOM. After s such selections the working model (4) has become

$$
\boldsymbol{u}_g = \sum_{l=1}^{s} \boldsymbol{m}_{E,l} a_l + \boldsymbol{M}_{E,-s} \boldsymbol{a}_{-s} + \boldsymbol{u}_p,
\tag{9}
$$

so that we have two terms involving the pseudo-marker effects. The first term involves the putative QTL; a_l is the size of the lth QTL and $\boldsymbol{m}_{E,l}$ are the corresponding pseudo-marker scores for the interval defining l. The second term contains the interval that has not been selected so that $\boldsymbol{a}_{-s}$ is the vector of sizes of QTL that may yet be chosen (most of them will not be selected) with corresponding scores $\boldsymbol{M}_{E,-s}$. After all putative QTL have been selected, the QTL are summarized by listing the putative QTL, the size of QTL found, and a P-value indicating the level of significance for each putative interval.

(ii) *High-dimensional analysis*

A major issue concerns the possible high dimension of $\boldsymbol{M}_E$ in (4) or of $\boldsymbol{M}_{E,-s}$ in (9) at stage s of selection. For ease of presentation, we consider the first step of selection. The matrix $\boldsymbol{M}_E$ is of size $n_g \times r - c$, and $r - c$ may be very large, larger than n_g. If $n_g > r - c$

there is no need to invoke dimension reduction. If $n_g < r - c$, it is possible to carry out efficient computations not in dimension $r - c$, but in dimension n_g which may be much smaller, and subsequently recover the $r - c$ effects necessary for selection.

Our approach is similar to that proposed by Stranden & Garrick (2009) in the context of genomic selection. For generality, suppose that $a \sim N(\mathbf{0}, \sigma_a^2 G_a)$. In WGAIM, $G_a = I_{r-c}$. Taylor *et al.* (2012) consider an alternative approach based on penalized likelihoods and in that case at stage k of the iterative process involved, $G_a = W_k$ where W_k is a diagonal matrix of known weights. In these cases, G_a is a known matrix and the results of this section will hold for any method in which this is the case or can be constructed to be so. Note that while WGAIM involved pseudo-markers for intervals, these can be replaced by marker scores (and would be in the genomic prediction situation).

Thus, suppose for the moment that no putative QTLs have been selected ($s = 0$ in the discussion above). The variance model for the random effects $M_E a$ is of the form

$$\sigma_a^2 M_E G_a M_E^T, \tag{10}$$

which is of size $n_g \times n_g$. The matrix $M_E G_a M_E^T$ is known. Then, we can re-formulate the working mixed model (4) as follows. Let

$$Z_E = (M_E G_a M_E^T)^{1/2}$$

denote the matrix square root; this can be found by using the singular value decomposition (Golub & van Loan, 1996). If we define an $n_g \times 1$ vector a^* as

$$a^* \sim N(\mathbf{0}, \sigma_a^2 I_{n_g}),$$

the working model term for interval effects $M_E a$ can be replaced by

$$Z_E a^*$$

and the same variance model (10) is obtained. Thus, we can estimate n_g effects rather than $r - c$.

Most importantly for QTL analysis, however, the $r - c$ effects can be recovered. Note firstly that the BLUP for a^* is given by

$$\tilde{a}^* = \sigma_a^2 (M_E G_a M_E^T)^{1/2} Z_g^T P y,$$

where $\quad P = H^{-1} - H^{-1} X (X^T H^{-1} X)^{-1} X^T H^{-1} \quad$ and $H = R + Z_o G_o Z_o^T + \sigma_g^2 Z_g Z_g^T + \sigma_a^2 Z_g M_E G_a M_E^T Z_g^T$. The BLUP of a is given by

$$\tilde{a} = \sigma_a^2 G_a M_E^T Z_g^T P y,$$

so that we can write

$$\tilde{a} = G_a M_E^T (M_E G_a M_E^T)^{-1/2} \tilde{a}^*, \tag{11}$$

thereby recovering the estimated size of potential QTLs in each interval from an analysis using the modified model.

To calculate the outlier statistics given by (5) and (6), we need the variance of $\tilde{a}$. This can be found using

$$\begin{aligned} \mathrm{var}(\tilde{a}) = {} & G_a M_E^T (M_E G_a M_E^T)^{-1/2} \, \mathrm{var}(\tilde{a}^*) \\ & \times (M_E G_a M_E^T)^{-1/2} M_E G_a \end{aligned} \tag{12}$$

and so the variance matrix of the BLUPs of a^* is required, that is $\mathrm{var}(\tilde{a}^*)$, again a lower-dimensional calculation after fitting the reduced dimensional model. Only the diagonal elements of $\mathrm{var}(\tilde{a})$ are required together with (11) to calculate the outlier statistics (5) and (6). The full matrix need not be calculated in practice.

(iii) *Random WGAIM*

The details presented in Verbyla *et al.* (2007) differ in only one aspect for a random version of WGAIM (RWGAIM). In WGAIM (Verbyla *et al.* 2007) the unselected effects are random effects (the working model in that paper) such that in (9), $a_{-s} \sim N(\mathbf{0}, \sigma_a^2 I)$ and the selected effects a_s are fixed effects. The proposal for RWGAIM, is to assume $a_l \sim N(0, \sigma_{al}^2)$, so that the sizes are random effects, have their own variance (resulting in a random regression on the pseudo-marker scores $m_{E,l}$) and all have a different variance to the unselected effects. This makes sense as putative QTL effects exhibit variation from zero because they are QTL. Thus, putative QTL are assumed to come from their own distribution and non-QTL come from another distribution. The distributions differ in their variances and not their means. This approach has the flavour of hierarchical models proposed by Griffin & Brown (2010) and others in high-dimensional variable selection.

(iv) *QTL selection using interval statistics only*

In Verbyla *et al.* (2007) and as outlined above, the selection of a QTL takes place in two stages, with both stages using outlier statistics. In the first stage, an outlier statistic was calculated for each linkage group using (5), and the linkage group with the largest statistic was selected as containing a putative QTL. In the second stage, the interval with the largest outlier statistic on the selected linkage group was selected as containing the putative QTL using (6). This process works effectively in many situations. However, the size of the linkage group has an impact on the size of the outlier statistic in the first stage. This can be seen from (5). On small linkage groups, the size of effects relative to their variance will be larger than for large linkage groups. This is because while there will generally be a small number of large sizes where a

putative QTL is present, the variances will generally be of comparable size across the whole linkage group. Thus, using the outlier statistic for linkage groups favours the selection of small linkage groups at the first stage. This problem is overcome if selection is based only on the outlier statistics for the intervals, that is (6), and the first stage is omitted. This is the approach adopted in WGAIMI and RWGAIM.

(v) *Significance of QTL*

Once all putative QTLs have been selected, they each constitute a random effect term. The size of the QTL effect is then a BLUP. It is no longer appropriate to test the hypothesis that the effect is zero in order to assess the significance of the putative QTL. Tests of hypotheses pertain to unknown parameters, and random effects involve distributions of effects.

To provide a measure of the strength of a QTL, the conditional distribution of the true (random) QTL effect a_l say, given the data are used. That is, under the normality assumptions for a linear mixed model,

$$a_l | \boldsymbol{y_2} \sim N(\tilde{a}_l, \sigma^2_{\mathrm{PEV}, l}),$$

where $\boldsymbol{y_2}$ is the component of the data free of fixed effects (Verbyla, 1990). The mean of this conditional distribution is the BLUP of a_l, that is the estimated size of the QTL $\tilde{a}_l$, and the variance is the prediction error variance (PEV) of a_l. Thus, the proper assessment of the impact of the QTL involves determining how far $\tilde{a}_l$ is from zero relative to the variation as measured by $\sigma_{\mathrm{PEV}, l}$. Clearly, distributions for which the mean is far from zero relative to the variation indicate a strong QTL effect, whereas distributions for which the mean is close to zero will indicate a weaker QTL.

To quantify this measure of strength of each putative QTL, it is possible to calculate a probability somewhat like a P-value, for which values close to 0 indicate that the QTL is strong. Consider the statistic

$$X_l^2 = \left(\frac{a_l - \tilde{a}_l}{\sigma_{\mathrm{PEV}, l}} \right)^2,$$

which has a chi-squared distribution on 1 degree of freedom. Zero on the original scale is $c_l^2 = \tilde{a}_l^2 / \sigma^2_{\mathrm{PEV}, l}$ on the chi-squared scale and so

$$\mathrm{Pr}(X_l^2 > c_l^2)$$

provides a measure of the strength of the putative QTL through how far $\tilde{a}_l$ is from zero relative to $\sigma_{\mathrm{PEV}, l}$. In RWGAIM, this probability can be used to indicate the strength of the putative QTL.

Note also that while LOD scores are not necessary for WGAIM or RWGAIM (because the multiple testing problem is largely avoided), a LOD score for RWGAIM is given by

$$\mathrm{LOD}_l = \frac{1}{2} \log_{10} \left(e^{c_l^2} \right).$$

Lastly, the percentage of genetic variance accounted for by each putative QTL is of interest. This can be determined approximately for both WGAIM and RWGAIM as follows.

The genetic effects are specified by (9); once all putative QTL have been found s is the number of QTLs. If $\boldsymbol{q}_l$ is the vector of QTL scores (-1 or 1) for QTL l, to facilitate an expression for the variance, we use a combination of (3) and (9) namely,

$$\boldsymbol{u}_g = \sum_{l=1}^{s} \boldsymbol{q}_l a_l + \boldsymbol{M}_{E, -s} \boldsymbol{a}_{-s} + \boldsymbol{u}_p. \tag{13}$$

Consider a single line i. The variance of the genetic effect for line i, u_{gi}, is then given by

$$\mathrm{var}(u_{gi}) = \left\{ \sum_{l=1}^{s} a_l^2 + \sum \sum_{l \neq l'} (1 - 2\theta_{ll'}) a_l a_{l'} \right\} + \sigma_a^2 \boldsymbol{m}_{Ei, -s}^T \boldsymbol{m}_{Ei, -s} + \sigma_p^2,$$

where we have allowed for linked QTL, $\theta_{ll'}$ is the recombination fraction for QTL l and l', and $\boldsymbol{m}_{Ei, -s}^T$ is the ith row of $\boldsymbol{M}_{E, -s}$. Thus, each line has its own (different) variance. To obtain an approximate measure of the total genetic variance, we define an 'average' line effect, u_g^*, by averaging the multiplier of σ_a^2 over lines. Thus, if

$$m_{E, -ss} = \frac{\sum_{i=1}^{n_g} \boldsymbol{m}_{Ei, -s}^T \boldsymbol{m}_{Ei, -s}}{n_g},$$

the variance of the 'average' line effect is taken as

$$\mathrm{var}\left(u_g^* \right) = \sum_{l=1}^{s} a_l^2 + \sum \sum_{l \neq l'} (1 - 2\theta_{ll'}) a_l a_{l'} + \sigma_{al}^2 m_{E, -ss} + \sigma_p^2. \tag{14}$$

Equation (14) specifies the approximate total genetic variance. A difficulty with using (14) to find percentage of variance is that the individual contributions of QTL can add up to more than 100% because of the covariance of linked QTL in repulsion. Given that the calculation is not precise, the covariances are ignored in the calculation so that

$$\mathrm{var}\left(u_g^* \right) = \sum_{l=1}^{s} a_l^2 + \sigma_{al}^2 m_{E, -ss} + \sigma_p^2 \tag{15}$$

is used for the total variance. Using (15), the percentage variance attributed to the lth QTL is then

$$PV_l = 100 \frac{a_l^2}{\mathrm{var}\left(u_g^* \right)}. \tag{16}$$

To calculate these values, a_l and other unknown parameters are replaced in (15) and (16) by their estimates from either the WGAIM or RWGAIM approach.

(vi) *Computation*

Both the high-dimensional approach and the RWGAIM approach have been implemented in the R environment (R Development Core Team, 2012) in the package `wgaim` (Taylor *et al.*, 2011) which is available on CRAN (http://cran.r-project.org). This package requires the `asreml` (Butler *et al.*, 2011) and `qtl` (Broman *et al.*, 2003, 2012) packages in R. In particular, it is the power of `asreml` that allows the complex models to be fitted. The `wgaim` package includes diagnostic plots and linkage map plots that allow the QTL to be displayed.

3. Materials

(i) *Simulation studies*

Three simulation studies were conducted.

(a) *Comparison of conventional and dimension reduction analysis*

To demonstrate the impact of the dimension reduction approach, a small simulation study was carried out. The population size was $n_g = 200$ lines with two replicates. The phenotypic data were generated using the model ($i = 1, 2, …, n_g, k = 1, 2$)

$$y_{ik} = \mu + \sum_{l=1}^{n_q} q_{il} a_l + u_{pi} + e_{ik}, \tag{17}$$

where the mean μ was 10, the polygenic effects u_{pi} are assumed $N(0, 0·5)$ and e_{ik} were assumed independent and identically distributed $N(0, 1)$. There were $n_q = 5$ QTLs, with q_{il} in (17) having a simulated allelic value (-1 or 1) for line i for QTL l. The size of this QTL is a_l and this is 0·5 for all QTLs. Thus, the mean line heritability (allowing for the two replicates) is approximately 0·85. The aim of having such a high heritability was to ensure that all 5 QTL would be detected to enable the timings to be compared.

The genetic data consisted of 10 linkage groups with the number of markers per linkage group varying from 21 to 101 in steps of ten markers; this allows the impact to be gauged from the case where the number of lines is close to the number of markers (210) to the case where the number of markers (1010) greatly exceeds the number of lines. Each linkage group was 250 cM long so that the density of markers changed from 12·5 cM intervals to 2·5 cM intervals. The five QTL were placed in the middle of each of five linkage groups. The bi-allelic marker scores were generated for each of the 200 lines for each scenario and the linkage map was then estimated as it would be in practice.

Two methods of analysis were carried out using RWGAIM using the one-stage outlier approach. One involved using the dimension reduction approach, while the other used the original method that involves using all the intervals explicitly. The aim was to provide a comparison of the time taken for each scenario using each of the methods.

(b) *High-dimensional simulation study*

To examine the effectiveness of the dimension reduction approach in genuinely high-dimensional situations, a second simulation was conducted. Using the same model as the first simulation, data were generated for population sizes 500, 1000 and 2000 lines in combination with 10 010, 20 010 and 50 010 markers (10 linkage groups with 101, 201 and 501 markers respectively) with QTLs at markers in the middle of 5 of the 10 linkage groups. As given above the linkage groups were 250 cm in length so that the density of the markers changed from 2·5 to 0·5 cM. The analysis of the generated data used markers rather than intervals. The computer processor timings of the analysis for each simulation from the nine combinations (population size by marker number) were determined.

(c) *Power, false discovery rate and bias*

The third simulation is similar to that conducted by Verbyla *et al.* (2007). The aim is to investigate three methods of QTL analysis, namely the original WGAIM approach using the two-stage outlier procedure, WGAIM using a one-stage outlier procedure (WGAIMI) and RWGAIM using a one-stage outlier procedure. These methods are expected to give similar results for QTL analysis apart from two key areas, short linkage groups and bias in estimated QTL sizes.

Nine linkage groups were simulated. On all linkage groups the markers were spaced at 10 cM. Linkage groups 5 and 9 had only two markers while the remaining linkage groups had 11 markers. Seven QTL ($n_q = 7$) were placed at the midpoints of C1·4 (linkage group 1, interval 4), C1·8, C2·4, C2·8, C3·6, C4·4 and C5·1 as in Verbyla *et al.* (2007). C1·4 and C1·8 were in repulsion while C2·4 and C2·8 were in coupling; these QTL are 40 cM apart. Population sizes were 100, 200 and 400 lines each with two replicates.

Introducing two very short linkage groups was aimed at investigating the impact of the two-stage outlier procedure using linkage groups and the intervals as opposed to using a single-stage outlier procedure based on the intervals. This highlights the impact in an extreme situation.

Five hundred (500) simulations were carried out for each population size using (17) with $a_l = 0.378$, apart from C1·8 for which $a_l = -0.378$. The polygenic effect had variance 0·5 and the residual variance was 1. The mean line heritability (allowing for the two replicates) is therefore 0·5.

For each population size, the three methods of analysis were carried out for each simulated dataset. Three sets of summaries were calculated for each method of QTL selection for each set of 500 simulations, namely

1. the proportion of correct detection of each QTL, including proportions for the QTL in coupling and repulsion,
2. the proportion of false positives, both linked (on the same chromosome as actual QTL) and un-linked (on chromosomes not containing QTL), and
3. means of estimated sizes of QTL effects for simulations where the QTL was selected. Empirical mean square error of prediction of the estimates were also found using all 500 simulations (not just those simulations where a true QTL was found).

As in Verbyla *et al.* (2007), a QTL was considered detected, if the correct interval or intervals on either side were selected.

(ii) *Salinity trial*

To illustrate the use of the random version of WGAIM, we analysed data from the experiment presented in Genc *et al.* (2010). The aim of the experiment was to examine QTL for various traits under a salinity treatment and a non-saline environment in a hydroponic experiment.

A doubled-haploid bread wheat population (Berkut × Krichauff) was grown in supported hydroponics to identify QTLs associated with salinity tolerance traits. The experiment consisted of two runs in a growth cabinet. In each run, there were four trolleys on each side of the growth cabinet (blocks), each controlling two tubs. Each tub consisted of 9 rows and 5 columns with 3 positions being allocated to the input and output of the supported hydroponics and hence not having a line assigned. In each replication, the 152 DH lines were randomly allocated across two trolleys (and hence four tubs) with the two parents (Berkut and Krichauff) and two other standard lines (Baart and Kharchia) being also randomly allocated to each tub; thus, there were four replicates of each of the DH lines. Some traits were only measured on the saline treatment including shoot magnesium concentration (in parts per million per kg of dry matter) which is the trait analysed in this paper. Saline conditions tend to decrease magnesium concentration and searching for QTL that mitigate this effect is of interest. We only consider the first run because
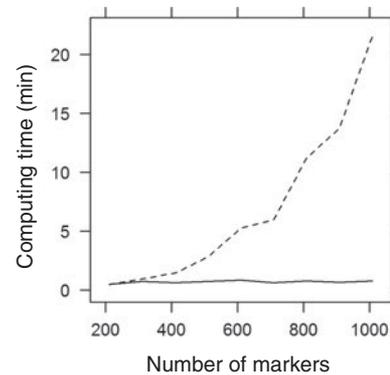

Fig. 2. Computation time (in minutes) for analysis using the dimension reduction approach (solid line) and the approach using markers explicitly (broken line) for numbers of markers from 210 to 1010.

analysing two runs potentially involves a multi-environment QTL analysis which is beyond the scope of this paper.

Genotyping using 216 SSR markers was carried out using standard PCR conditions and subsequent gel electrophoretic separation on 8 % polyacrylamide gels. In addition to the 216 SSRs and *Vrn* genes, 311 DArT markers were scored by Triticarte Pty. Ltd (http://www.triticarte.com.au). The linkage map was constructed using Map Manager QTX version QTXb20 (Manly *et al.*, 2001). Marker order was verified using RECORD (Van Os *et al.*, 2005). Segregation ratios of two genotype classes (Berkut allele and Krichauff allele) at each locus were tested and markers deviating significantly from the expected ratio of 1:1 were excluded from QTL analysis. Re-estimation of genetic distance was performed using the Lander & Green (1987) algorithm in the R (R Development Core Team, 2012) package `qtl` (Broman *et al.*, 2012). Coincident markers were removed resulting in 379 distinct markers on 23 linkage groups for QTL analysis. The total length of the map was 3770 cM with an average spacing of 10·6 cM between the markers.

4. Results

(i) *Simulation study: comparison of two methods of analysis*

The simulations were conducted on a 64 Bit PC using Windows 7, with an Intel® Core™ i7-2860QM CPU at 2·50 GHz, with 16 Gb RAM at 2·39 GHz. The two methods of analysis are identical apart from dimension reduction and hence both methods resulted in the same five putative QTL as they should.

Figure 2 presents the CPU time required for each analysis from 210 to 1010 markers for the two methods of analysis. The dimension reduction approach resulted in times represented by the solid line, while the original method using all the markers explicitly

Table 1. *High-dimensional simulations for varying population sizes and number of markers. Non-redundant markers varied with population size. Both the processor time and processor time per step of the forward selection process are presented*

Population size	Number of markers	Non-coincident markers	Time (min)	Time per step (min)
500	10 010	7263	4·29	0·71
	20 010	9311	4·22	0·70
	50 010	11 176	4·78	0·80
1000	10 010	9142	37·01	6·18
	20 010	14 185	38·50	6·41
	50 010	19 841	40·03	6·67
2000	10 010	9947	338·87	56·48
	20 010	18 316	345·11	57·52
	50 010	31 491	363·91	60·65

resulted in the timings given by the broken line. The dimension reduction approach resulted in times that remained generally constant despite an increasing number of markers compared with the original approach where times increased exponentially. These trends for the two approaches would be expected to continue should the number of markers increase causing the original formulation to quickly become prohibitive, while the dimension reduction approach is expected to be able to deal with considerably large datasets.

(ii) *Simulation study: high-dimensional analysis*

The dimension reduction approach was used for high-dimensional situations (10 010, 20 010 and 50 010 markers) for 500, 1000 and 2000 lines. In all the cases, 2000 lines were used in generation of marker data. Then, a random sample of lines was taken to obtain the data for 500 and 1000 lines for the simulations. Table 1 details the results in terms of the actual number of non-redundant markers and the computer processor time used in the analyses. The markers will be redundant because of very low recombination fractions and limited number of lines. Thus, for 500 lines, it was very difficult to retain large numbers of markers, with improvement as the number of lines increased. Note that with five QTLs, there are six steps in the forward selection process. Hence the time required per step can be calculated. This provides a base time line to gauge how long an analysis with a particular number of QTLs is likely to take. This is included in Table 1.

Firstly, note that for population sizes 500 and 1000 the analyses for the simulated data for all marker sizes were conducted using markers rather than intervals on the 64 bit PC outlined in the previous simulation. For population size 2000, the simulations were run on a high-performance computer requesting 16 Gb of memory. Only a single node was used.

For population size 500, the number of redundant markers restricted the size of the problem that could be investigated. The number of non-redundant markers increases with population size and subsequently the processor time goes up by a factor of 9-fold for each step as we move from 500 to 1000 to 2000. Thus, for 2000 lines and approximately 30 000 markers, the analysis took about 6 h.

The conclusion from this simulation study is that QTL analysis can be performed even for very high-dimensional data, possibly using high-performance computing facilities. The limitations are memory and storage usage in the R environment and the ability of the asreml software to handle such large dense datasets. These aspects increase the required time for analysis of very high dimensions.

(iii) *Simulation study: power, false discovery rate and bias*

The results of the simulations with seven QTLs as described above are presented in Tables 2–6. Each aspect is now discussed in turn.

Table 2 presents the proportion of simulations in which each QTL was successfully found using the original WGAIM (applying the two-stage outlier procedure), WGAIMI (applying the one-stage outlier procedure) and RWGAIM (applying the one-stage outlier procedure). The results are very similar for all three methods although WGAIM finds marginally more QTLs in total for all population sizes. This is mainly because original WGAIM tends to find the QTL at C5·1 more often than the other methods; this is a linkage group with only one interval. The WGAIM approach also tends to find more of the QTL in repulsion (C1·4 and C1·8) than the other two methods but less QTLs in coupling (C2·4 and C2·8)

Table 2. *Proportion of 500 simulations in which the QTL was detected for WGAIM, WGAIMI and random WGAIM (RWGAIM) analyses. The standard error for a proportion can be based on binomial distribution, that is $\sqrt{\hat{p}(1-\hat{p})/500}$. For $\hat{p}=0{\cdot}1$ (and hence $\hat{p}=0{\cdot}9$), the standard error is $0{\cdot}013$, for $\hat{p}=0.3$ the standard error is $0{\cdot}020$ and for $\hat{p}=0.5$ the standard error is $0{\cdot}022$*

Population	Method	C1·4	C1·8	C2·4	C2·8	C3·6	C4·4	C5·1	Total
100	WGAIM	0·416	0·512	0·710	0·410	0·704	0·520	0·948	4·220
	WGAIMI	0·368	0·474	0·680	0·420	0·756	0·604	0·730	4·032
	RWGAIM	0·352	0·474	0·676	0·482	0·768	0·624	0·742	4·118
200	WGAIM	0·962	0·950	0·734	0·626	0·750	0·804	0·996	5·822
	WGAIMI	0·956	0·934	0·740	0·630	0·792	0·832	0·908	5·792
	RWGAIM	0·952	0·930	0·736	0·640	0·788	0·822	0·904	5·772
400	WGAIM	0·976	0·974	0·942	0·942	0·984	1·000	1·000	6·818
	WGAIMI	0·972	0·970	0·934	0·944	0·982	1·000	1·000	6·802
	RWGAIM	0·962	0·972	0·934	0·954	0·986	1·000	1·000	6·808

Table 3. *Two way tables for the QTL in repulsion (C1·4 and C1·8) with proportions of the 500 simulations for each population size for the combinations of non-detected $\overline{D}$ and detected D QTL. C1·4 is on the left and C1·8 on the top of each 2×2 table. The standard error for a proportion can be based on binomial distribution, that is $\sqrt{\hat{p}(1-\hat{p})/500}$. For $\hat{p}=0{\cdot}1$ (and hence $\hat{p}=0{\cdot}9$) the standard error is $0{\cdot}013$, for $\hat{p}=0{\cdot}3$ the standard error is $0{\cdot}020$ and for $\hat{p}=0{\cdot}5$ the standard error is $0{\cdot}022$*

Method		Population size					
		100		200		400	
		$\overline{D}$	D	$\overline{D}$	D	$\overline{D}$	D
WGAIM	$\overline{D}$	0·390	0·194	0·010	0·028	0·010	0·014
	D	0·098	0·318	0·040	0·922	0·016	0·960
WGAIMI	$\overline{D}$	0·424	0·208	0·014	0·030	0·014	0·014
	D	0·102	0·266	0·052	0·904	0·016	0·956
RWGAIM	$\overline{D}$	0·430	0·218	0·014	0·034	0·014	0·024
	D	0·096	0·256	0·056	0·896	0·014	0·948

than RWGAIM. Apart from C5·1, however, it is RWGAIM that finds more QTLs although the differences are small. All three methods perform very well. Approximate standard errors can be found using the standard results for binomial proportion and are indicated in all the tables. These standard errors provide an idea of the scale of differences between methods. For example, the difference between the three methods for the QTL at C5·1 is clearly highly significant.

Two-way tables for the proportions of QTLs detected for the two QTLs in repulsion are given in Table 3. Results for WGAIM, WGAIMI and RWGAIM are generally very similar, though the WGAIM method detects both QTLs more often than the other methods for all population sizes. For the two QTLs in coupling, the corresponding results are given in Table 4. Although the detection rates compared with repulsion were lower, RWGAIM shows a small improvement over WGAIM and WGAIMI in detecting both QTLs.

Table 5 gives the proportion of false positives for the three methods. False positives can be linked (on the same chromosome as true QTL, C1–C4, false positives cannot be attributed to C5 because it has only one interval) or unlinked (on chromosomes where no QTL exist, C6–C9). The rates of false positives are the number found across all 500 simulations divided by 500. For linkage group C9 which has one interval, the WGAIM method leads to false positives at a much larger rate than WGAIMI and RWGAIM. This reflects the problem of favouring short linkage groups when using WGAIM with a two-stage outlier procedure. This increase in false positives also leads to a decrease in false positives on the other linkage groups that do not contain QTLs (in comparison with WGAIMI and RWGAIM). The rates of linked false

Table 4. *Two way tables for the QTL in coupling (C2·4 and C2·8) with proportions of 500 simulations for each population size for the combinations of non-detected $\overline{D}$ and detected D QTL. C2·4 is on the left and C2·8 on the top of each 2×2 table. The standard error for a proportion can be based on binomial distribution, that is $\sqrt{\hat{p}(1-\hat{p})/500}$. For $\hat{p}=0\cdot1$ (and hence $\hat{p}=0\cdot9$) the standard error is $0\cdot013$, for $\hat{p}=0\cdot3$ the standard error is $0\cdot020$ and for $\hat{p}=0\cdot5$ the standard error is $0\cdot022$*

		Population size					
		100		200		400	
Method		$\overline{D}$	D	$\overline{D}$	D	$\overline{D}$	D
WGAIM	$\overline{D}$	0·080	0·210	0·054	0·212	0·010	0·048
	D	0·510	0·200	0·320	0·414	0·048	0·894
WGAIMI	$\overline{D}$	0·090	0·230	0·046	0·214	0·012	0·054
	D	0·490	0·190	0·324	0·416	0·044	0·890
RWGAIM	$\overline{D}$	0·078	0·246	0·048	0·216	0·012	0·054
	D	0·440	0·236	0·312	0·424	0·034	0·900

Table 5. *Proportion of false QTLs in 500 simulations detected for the three methods of analysis. Both linked (putative QTLs are on chromosomes with QTL, C1–C4, C5 has only one interval and cannot lead to false positives) and unlinked (putative QTLs are on chromosomes without QTLs, C6–C9) are presented. The standard error for a proportion can be based on binomial distribution, that is $\sqrt{\hat{p}(1-\hat{p})/500}$. For $\hat{p}=0\cdot1$ (and hence $\hat{p}=0\cdot9$) the standard error is $0\cdot013$, for $\hat{p}=0\cdot3$ the standard error is $0\cdot020$ and for $\hat{p}=0\cdot5$ the standard error is $0\cdot022$*

Population	Method	C1	C2	C3	C4	C6	C7	C8	C9	Total Linked	Unlinked
100	WGAIM	0·238	0·232	0·154	0·124	0·048	0·076	0·042	0·136	0·748	0·302
	WGAIMI	0·242	0·192	0·166	0·166	0·084	0·146	0·074	0·020	0·766	0·324
	RWGAIM	0·198	0·194	0·156	0·118	0·062	0·118	0·060	0·014	0·666	0·254
200	WGAIM	0·080	0·130	0·112	0·122	0·018	0·024	0·016	0·124	0·444	0·182
	WGAIMI	0·072	0·110	0·124	0·122	0·054	0·038	0·030	0·010	0·428	0·132
	RWGAIM	0·034	0·102	0·118	0·112	0·052	0·036	0·026	0·008	0·366	0·122
400	WGAIM	0·030	0·054	0·052	0·016	0·016	0·008	0·006	0·040	0·152	0·070
	WGAIMI	0·032	0·048	0·044	0·014	0·012	0·014	0·008	0·004	0·138	0·038
	RWGAIM	0·024	0·036	0·038	0·014	0·012	0·014	0·008	0·004	0·112	0·038

positives are similar for WGAIM and WGAIMI, but are reduced for the RWGAIM method. Lastly, note that RWGAIM shows reduced rates of false positives overall when compared with WGAIM and WGAIMI for all population sizes.

As a final comment we note that the results of the simulations for QTLs in repulsion and coupling of the two-way tables might be expected to be symmetric. Furthermore, for single QTL, we might expect similar proportions of detection if position is taken into account (for example C3·6 and C4·4). This is not evident from the results and it is not immediately clear why this is the case. It can only be conjectured that the possible cause might be the forward selection process.

One aspect of WGAIM that was not examined by Verbyla *et al.* (2007) was bias in the estimated QTL effect sizes. This was examined in the current simulation study and the results are given in Table 6. The estimated sizes of effects were averaged over the simulations in which the effects were detected. Firstly, both WGAIM and WGAIMI show similar bias, and both show higher bias (all true QTL sizes are ±0.378) than RWGAIM. Thus, RWGAIM reduces the bias particularly for the smaller population sizes, though bias still persists. The bias at population size 200 is smaller for RWGAIM while all methods exhibit little bias at population size 400. Empirical standard errors are provided to provide an indication of significance

Table 6. *Mean estimated size of QTL effects* (*true size is 0·378 for all but C1·8 which has a size of* −0·378) *with empirical standard error* (SE) *and mean square error of prediction* (MSEP) *for each method across 500 simulations for each population size*

		Population size								
		100			200			400		
Method	Interval	Size	SE	MSEP	Size	SE	MSEP	Size	SE	MSEP
WGAIM	C1·4	0·508	0·0093	0·0897	0·441	0·0062	0·0283	0·392	0·0036	0·0101
	C1·8	−0·461	0·0085	0·0925	−0·455	0·0058	0·0270	−0·396	0·0033	0·0089
	C2·4	0·515	0·0082	0·0948	0·444	0·0071	0·0659	0·377	0·0035	0·0137
	C2·8	0·556	0·0080	0·0803	0·443	0·0053	0·0488	0·379	0·0040	0·0155
	C3·6	0·491	0·0052	0·0580	0·385	0·0037	0·0396	0·390	0·0028	0·0062
	C4·4	0·509	0·0088	0·0879	0·390	0·0035	0·0321	0·372	0·0023	0·0028
	C5·1	0·391	0·0055	0·0193	0·347	0·0035	0·0076	0·407	0·0024	0·0038
WGAIMI	C1·4	0·518	0·0117	0·0997	0·435	0·0064	0·0304	0·389	0·0034	0·0098
	C1·8	−0·465	0·0106	0·1006	−0·459	0·0072	0·0363	−0·395	0·0034	0·0096
	C2·4	0·538	0·0083	0·0961	0·440	0·0073	0·0658	0·377	0·0035	0·0134
	C2·8	0·558	0·0078	0·0825	0·445	0·0054	0·0483	0·379	0·0039	0·0159
	C3·6	0·482	0·0079	0·0609	0·379	0·0054	0·0383	0·391	0·0028	0·0064
	C4·4	0·496	0·0102	0·0838	0·387	0·0042	0·0300	0·372	0·0023	0·0027
	C5·1	0·430	0·0049	0·0468	0·361	0·0033	0·0180	0·407	0·0024	0·0038
RWGAIM	C1·4	0·452	0·0071	0·0833	0·410	0·0047	0·0204	0·369	0·0033	0·0092
	C1·8	−0·419	0·0079	0·0970	−0·441	0·0042	0·0184	−0·376	0·0031	0·0098
	C2·4	0·503	0·0076	0·0883	0·412	0·0061	0·0599	0·369	0·0033	0·0116
	C2·8	0·512	0·0070	0·0696	0·425	0·0054	0·0473	0·378	0·0033	0·0143
	C3·6	0·450	0·0054	0·0456	0·362	0·0037	0·0348	0·380	0·0027	0·0055
	C4·4	0·451	0·0060	0·0641	0·361	0·0035	0·0300	0·363	0·0023	0·0030
	C5·1	0·398	0·0052	0·0445	0·341	0·0035	0·0199	0·400	0·0025	0·0036

between methods for each of the QTL for each population size.

The empirical mean square error of prediction (MSEP) for each QTL (indexed by l) and each population size was found using

$$\text{MSEP}_l = \frac{\sum_{s=1}^{500}(a_l - \hat{a}_{sl})^2}{500},$$

where a_l is the true size of the QTL and $\hat{a}_{sl}$ is the estimated size in simulation s. These are given in Table 6; note that in this calculation all simulations were used so that if a true QTL was not detected, its estimated size was zero. The MSEPs are almost always smaller for RWGAIM suggesting this method is better than both WGAIM and WGAIMI in terms of estimation (or prediction) of putative QTL effect sizes. Notice that the mean effect sizes for C5.1 across the three methods are similar for the three population sizes, but that the MSEPs is smaller for WGAIM for population sizes 100 and 200 where this method detects more of this true QTL than the other two methods.

(iv) *Salinity trial analysis*

Magnesium concentration was divided by 100 to facilitate presentation of the results. The analysis begins with consideration of the experimental design.

The experiment involves a nested blocking structure in which the factor Tubs (2 levels) are nested in Trolley (2 levels) which is nested in the factor Side (2 levels) of the growth cabinet. Thus, without marker data, the initial model is given symbolically by

conc = Type + id + Side + Side.Trolley
        + Side.Trolley.Tub + e,

where conc is the magnesium concentration, the factor Type distinguishes doubled haploid lines from other lines that do not have marker information, id is the genotype factor for lines in the experiment, and Side, Trolley and Tub are factors each with two levels that describe the physical layout (blocking) of the experiment. Terms like Side.Trolley indicate nesting of Trolley in Side. All but Type are random effects. The residual error e was modelled to take into account possible correlation between tubs that were controlled by a single trolley.

In the course of modelling the data, row and column effects in some Tubs were observed and included in the model. This model is called the baseline model. Table 7 provides estimated variance components for the baseline model, and the final WGAIM, WGAIMI and RWGAIM models. For the baseline model, it was found that only Tubs in Trolley 1 on Side 1 were substantially correlated at the residual level; the other

Table 7. *Estimated variance components for the baseline model and the final WGAIM, WGAIMI and RWGAIM models. ints refers to the model term $M_{E-s}.a_{-s}$ qtl followed by a linkage group refers to the QTL random effect using RWGAIM*

	Model			
Term	Baseline	WGAIM	WGAIMI	RWGAIM
ints	—	0·0004	0·0002	0·0003
qtl: 3A(1)	—	—	—	0·09
qtl: 3A(2)	—	—	—	0·38
qtl: 3D	—	—	—	0·11
qtl: 4B	—	—	—	0·15
qtl: 5A	—	—	—	0·35
qtl: 5B	—	—	—	0·44
qtl: 5D	—	—	—	0·09
qtl: 6B	—	—	—	0·12
id	2·08	0·72	0·72	0·73
Side	0·43	0·44	0·40	0·41
Side.Trolley	0·37	0·37	0·41	0·40
Side.Trolley.Tub	0	0	0	0
Side 1 Trolley 1 Tub 2 Row	0·98	0·76	0·80	0·74
Side 1 Trolley 2 Tub 2 Row	0·37	0·37	0·44	0·46
Side 1 Trolley 1 Tub 1 Column	0·05	0·08	0·18	0·12
Side 1 Trolley 2 Tub 1 Column	0·23	0·22	0·18	0·25
Side 1 Trolley 2 Tub 2 Column	0·54	0·59	0·51	0·44
Side 2 Trolley 1 Tub 1 Column	0·38	0·41	0·34	0·33
Side 2 Trolley 1 Tub 2 Column	0·28	0·29	0·39	0·36
Side 2 Trolley 2 Tub 2 Column	0·50	0·41	0·35	0·37
Side 1 Trolley 1 correlation	0·62	0·56	0·73	0·69
Side 1 Trolley 1 Tub 1 variance	0·89	0·45	0·68	0·58
Side 1 Trolley 1 Tub 2 variance	0·35	0·22	0·23	0·24
Side 1 Trolley 2 correlation	0	0	0	0
Side 1 Trolley 2 Tub 1 variance	1·02	1·09	1·10	1·04
Side 1 Trolley 2 Tub 2 variance	0·34	0·40	0·44	0·34
Side 2 Trolley 1 correlation	0	0	0	0
Side 2 Trolley 1 Tub 1 variance	1·27	1·75	1·37	1·51
Side 2 Trolley 1 Tub 2 variance	0·69	0·62	0·64	0·70
Side 2 Trolley 2 correlation	0	0	0	0
Side 2 Trolley 2 Tub 1 variance	1·46	1·53	1·44	1·54
Side 2 Trolley 2 Tub 2 variance	0·38	0·57	0·58	0·56

trolleys were then fitted with only separate error variance for the two Tubs. Notice that there is a systematic effect observed in terms of error variance. In all trolleys, Tub 1 has a higher estimated variance than Tub 2. Modelling this is therefore potentially important in QTL analysis.

The analysis of magnesium concentration was carried out using WGAIM, WGAIMI and RWGAIM. The three methods resulted in 9, 8 and 8 putative QTLs respectively. The results are given in Table 8. Six QTLs were essentially the same or in closely linked intervals for WGAIM and RWGAIM, while WGAIMI and RWGAIM had seven putative QTLs in common. The sizes of effects under RWGAIM are generally smaller than those of the other two methods. The percentage of the total genetic variance attributed to each QTL is given in Table 8, using (16), while LOD scores for each QTL are also provided.

The approach based on WGAIM resulted in the selection of three putative QTLs that were on smaller linkage groups. In fact the linkage group 1D1 has two markers. This reflects the issue raised regarding WGAIM, namely the tendency to favour small linkage groups in selection of putative QTLs. These three putative QTLs were not found using the WGAIMI or RWGAIM approach. On the other hand, putative QTLs were found on 5D and 6B using RWGAIM (4A and 6B using WGAIMI) that were not found using WGAIM.

The estimated variance components for the final WGAIM, WGAIMI and RWGAIM models are given in Table 7 and show the similarity of the three approaches. The reduction in the id variance component is similar in all methods. Using the percentages in Table 7, approximately 67, 72 and 67 % of the genetic variance was explained by the putative QTL for each of the three methods. Large individual contributions come from the *Vrn* genes and the second interval on 3A. These effects are also reflected in the variance components for QTL that are estimated in

Table 8. *Putative QTL found using WGAIM, WGAIMI and RWGAIM*

Method	Linkage group	Marker Left	dist(cM)	Marker Right	dist(cM)	Size	P-value or Prob	% var	LOD
WGAIM	1D1	wPt-4647	0	wmc147	3·89	0·221	0·016	1·9	1·25
	2A	wPt-5647	11·02	gwm071	63·44	0·427	0·001	7·1	2·32
	3A	gwm732a	2·18	barc1121	8·95	−0·292	0·004	3·3	1·80
	3A	wmc343	97·33	cfa2262	101·48	0·539	0·000	11·3	5·92
	3B	wPt-2757	7·15	wPt-7015	78·87	0·351	0·016	4·8	1·27
	3D	cfd223	43·91	cfd152	61·76	0·346	0·001	4·6	2·33
	4B	gwm495	51·74	gwm149	55·21	0·411	0·000	6·5	2·99
	5A	wPt-1370	235·31	Vrn1A	239·06	0·605	0·000	14·2	8·82
	5B	wmc289	82·60	VrnBR3/R4	91·16	−0·588	0·000	13·4	6·03
WGAIMI	3A	barc1121	8·95	wPt-2866	16·77	−0·394	0·000	5·6	3·20
	3A	gwm002	80·59	gwm005	83·91	0·713	0·000	18·3	10·98
	4A	wPt-4660	55·56	wmc468	56·94	−0·308	0·001	3·4	2·42
	4B	wPt-1505	55·87	gwm251	55·87	0·403	0·000	5·9	3·78
	5A	wPt-1370	235·31	Vrn1A	239·06	0·641	0·000	14·8	10·76
	5B	wmc289	82·60	VrnBR3/R4	91·16	−0·650	0·000	15·2	8·56
	5D	cfd19a	84·61	gwm292a	99·29	−0·308	0·002	3·4	2·14
	6B	wPt-3118	65·32	wPt-4858	87·02	−0·392	0·000	5·5	2·65
RWGAIM	3A	gwm732a	2·18	barc1121	8·95	−0·287	0·002	3·3	2·02
	3A	gwm002	80·59	gwm005	83·91	0·607	0·000	14·9	8·44
	3D	wPt-4237	19·05	cfd223	43·91	0·321	0·002	4·2	1·99
	4B	wPt-1505	55·87	gwm251	55·87	0·372	0·000	5·6	3·18
	5A	wPt-1370	235·31	Vrn1A	239·06	0·588	0·000	14·0	9·16
	5B	wmc289	82·60	VrnBR3/R4	91·16	−0·658	0·000	17·5	8·60
	5D	cfd19a	84·61	gwm292a	99·29	−0·287	0·002	3·3	1·99
	6B	cfd001a	89·02	wPt-5333	89·63	−0·339	0·001	4·6	2·57

the RWGAIM method; see the values in the column labelled RWGAIM in Table 7.

5. Discussion

The need for methods that allow high-dimensional analysis is becoming more apparent and the approach presented in this paper provides a way to do so in QTL analysis and could potentially be extended for use in Genomic Selection (Meuwissen *et al.*, 2001). The effective dimension is reduced to the number of genetic lines in the data that have marker information. This dimension reduction allows faster model fitting and the marker or interval effects can be recovered. Thus, QTL selection can proceed in an efficient manner. In addition, the approach of this paper can accommodate non-genetic sources of variation.

Bias of the sizes of putative QTLs is always a problem. By placing putative QTLs in a random effects term, the bias can be reduced because of shrinkage of effects towards zero (Robinson, 1991). While this does not fully remove the bias due to selection, it does reduce the impact for small to moderate population sizes. For large populations, bias in estimated effects using WGAIM appears to be small. The simulation study showed the improvement in bias

using RWGAIM. Apart from bias, the two methods performed similarly in the simulation study but did differ in their putative QTLs in the example. Experience suggests that RWGAIM may be more robust in terms of LOD scores for putative QTLs and has reduced mean square error of prediction.

The original WGAIM approach used two stages of outlier statistics for selection of putative QTLs. The simulation study showed that using WGAIM, false positive rates for short linkage groups are likely to be higher. On the other hand, if genuine QTL exist for such linkage groups, their detection is enhanced by using WGAIM. Thus, there is a trade-off when using WGAIM and the recommendation is to use RWGAIM with a single stage outlier procedure based only on the intervals on the linkage map. This balances finding QTLs with better false positive rates.

The methods of this paper are available in software that includes the power of mixed model analysis, allowing inclusion of effects that pertain to the design of the study and to other non-genetic sources of variation that may need to be included in the analysis. Markers rather than intervals can be used in the analysis if that is desirable. In this case the matrix M_E is replaced by M in the derivations.

In conclusion, this paper presents an approach for QTL analysis with a large number of markers,

provides an approach with reduced bias and reduced mean square error of prediction, and places all intervals on the linkage map on the same footing for selection of putative QTLs.

References

Beavis, W. D. (1994). The power and deceit of QTL experiments: lessons from comparative QTL studies. In *Proceedings of the 49th Annual Corn and Sorghum Industry Research Conference*, pp. 250–266. American Seed Trade Association, Washington, DC.

Beavis, W. D. (1998). QTL analyses: power, precision and accuracy. In *Molecular Dissection of Complex Traits*, (ed. A. H. Patterson), pp. 145–162. New York: CRC Press.

Broman, K. W., Wu, H., Churchill, G., Sen, S., Arends, D., Jansen, R., Prins, P., Yandell, B. (2012). *qtl: Tools for Analyzing QTL Experiments*. R package version 1.23–16. http://cran.r-project.org/web/packages/qtl/index.html.

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

Butler, D. G., Cullis, B. R., Gilmour, A. R. & Gogel, B. J. (2011). Mixed models for S language environments: ASReml-R reference manual. Technical report, Queensland Department of Primary Industries.

Che, X. & Xu, S. (2010). Significance test and genome selection in Bayesian shrinkage analysis. *International Journal of Plant Genomics* **2010**, Article ID 893206, 11 pages.

Genc, Y., Oldach, K., Verbyla, A. P., Lott, G., Hassan, M., Tester, M., Wallwork, H. & McDonald, G. K. (2010). Sodium exclusion QTL associated with improved seedling growth in bread wheat under salinity stress. *Theoretical and Applied Genetics* **121**, 877–894.

Golub, G. & van Loan, C. (1996). *Matrix Computations*, 3rd edn. Baltimore, MD: The Johns Hopkins University Press.

Griffin, J. E. & Brown, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis* **5**, 171–188.

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.

Heath, S. C. (1997). Markov chain Monte Carlo segregation and linkage analysis of oligogenic models. *American Journal of Human Genetics* **61**, 748–760.

Jansen, R. C. (1994). Controlling the Type I and Type II errors in mapping quantitative trait loci. *Genetics* **138**, 871–881.

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

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

Lander, E. S. & Green, P. (1987). Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences, USA* **84**, 2363–2367.

Manly, K. F., Cudmore, Jr R. H. & Meer, J. M. (2001). Map Manager QTX, cross platform software for genetic mapping. *Mammalian Genome* **12**, 930–932.

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.

Melchinger, A. E., Utz, H. F. & Schön, C. C. (1998). Quantitative trait loci (QTL) mapping using different testers and independent population samples in maize reveals low power of QTL detection and large bias in estimates of QTL effects. *Genetics* **149**, 383–403.

Meuwissen, T. H., Hayes, B. J. & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819V1829.

Oakey, H., Verbyla, A., Pitchford, W., Cullis, B. & Kuchel, H. (2006). Joint modelling of additive and non-additive genetic line effects in single field trials. *Theoretical and Applied Genetics* **113**, 809–819.

Piepho, H.-P. (2001). A quick method for computing approximate thresholds for quantitative trait loci detection. *Genetics* **157**, 425–432.

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science* **6**, 15–51.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Satagopan, J., 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.

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

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

Smith, A., Cullis, B. & Thompson, R. (2005). The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. *Journal of Agricultral Science* (*Cambridge*) **143**, 449–462.

Smith, A. B., Lim, P. & Cullis, B. R. (2006). The design and analysis of multi-phase quality trait experiments. *Journal of Agicultural Science* (*Cambridge*) **144**, 393–409.

Stram, D. O. & Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics* **50**, 1171–1177.

Stranden, I. & Garrick, D. J. (2009). Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. *Journal of Dairy Science* **92**, 2971–2975.

Taylor, J. D. & Verbyla, A. P. (2011). R package wgaim: QTL analysis in bi-parental populations using linear mixed models. *Journal of Statistical Software* **40**, 1–18.

Taylor, J. D., Diffey, S., Verbyla, A. P. & Cullis, B. R. (2011). *wgaim: Whole Genome Average Interval Mapping for QTL detection using mixed models*. R package version 1.1. http://cran.r-project.org/web/packages/wgaim/index.html.

Taylor, J. D., Verbyla, A. P., Cavanagh, C. C. & Newberry, M. (2012). Variable selection in linear mixed models using an extended class of penalities. *Australian and New Zealand Journal of Statistics* in press.

Van Os, H., Stam, P., Visser, R. & Van Eck, H. J. (2005). RECORD: a novel method for ordering loci on a genetic linkage map. *Theoretical and Applied Genetics* **112**, 30–40.

Varshney, R. K., Nayak, S. N., May, G. D. & Jackson, S. A. (2009). Next generation sequencing technologies and their implications for crop genetics and breeding. *Trends in Biotechnology* **27**, 522–530.

Verbyla, A. P. (1990). A conditional derivation of residual maximum likelihood. *Australian Journal of Statistics* **32**, 227–230.

Verbyla, A. P., Cullis, B. R. & Thompson, R. (2007). The analysis of QTL by simultaneous use of the full linkage map. *Theoretical and Applied Genetics* **116**, 95–111.

Wang, H., Zhang, Y., Li, X., Masinde, G. L., Mohan, S., Baylink, D. J. & Xu, S. (2005). Bayesian shrinkage estimation of quantitative trait loci parameters. *Genetics* **170**, 465–480.

Weller, J. I. (1986). Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. *Biometrics* **42**, 627–640.

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

Yi, N. (2004). A unified Markov chain Monte Carlo framework for mapping multiple quantitative trait loci. *Genetics* **167**, 967–975.

Yi, N. & Shriner, D. (2008). Advances in Bayesian multiple QTL mapping in experimental designs. *Heredity* **100**, 240–252.

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

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