Theor Appl Genet (2007) 116:95–111
DOI 10.1007/s00122-007-0650-x

ORIGINAL PAPER

The analysis of QTL by simultaneous use of the full linkage map

Arūnas P. Verbyla · Brian R. Cullis · Robin Thompson

Received: 5 July 2006 / Accepted: 17 September 2007 / Published online: 20 October 2007

Abstract An extension of interval mapping is presented that incorporates all intervals on the linkage map simultaneously. The approach uses a working model in which the sizes of putative QTL for all intervals across the genome are random effects. An outlier detection method is used to screen for possible QTL. Selected QTL are subsequently fitted as fixed effects. This screening and selection approach is repeated until the variance component for QTL sizes is not statistically significant. A comprehensive simulation study is conducted in which map uncertainty is included. The proposed method is shown to be superior to composite interval mapping in terms of power of detection of QTL. There is an increase in the rate of false positive QTL detected when using the new approach, but this rate decreases as the population size increases. The new approach is much simpler computationally. The analysis of flour milling yield in a doubled haploid population illustrates the improved power of detection of QTL using the approach, and also shows how vital it is to allow for sources of non-genetic variation in the analysis.

Communicated by J.-L. Jannink.

A. P. Verbyla
School of Agriculture, Food and Wine,
The University of Adelaide, PMB 1, Glen Osmond,
SA 5064, Australia

A. P. Verbyla (✉)
Statistical Bioinformatics—Agribusiness,
Mathematical and Information Sciences,
CSIRO, PMB 2, Glen Osmond, SA 5064, Australia
e-mail: Ari.Verbyla@adelaide.edu.au

B. R. Cullis
Biometrics, Wagga Wagga Research Institute,
New South Wales Department of Primary Industries, Wagga
Wagga, NSW 2650, Australia

R. Thompson
School of Mathematical Sciences, Queen Mary College,
University of London, Mile End Rd., London E1 4NS, UK

R. Thompson
Centre for Mathematical and Computational Biology,
Department of Biomathematics and Bioinformatics,
Rothamsted Research, Harpenden, Hertfordshire AL5 2JQ, UK

Introduction

The determination of genomic regions or genes that influence the expression of a quantitative trait is important in the plant and animal breeding context. For plants, this can lead to rapid improvements in agronomic and quality traits, in disease resistance, and tolerance to both biotic and abiotic stresses. In the livestock industry, this can lead to improvements in quality traits, such as marbling in beef, and wool characteristics in sheep. In this paper the focus is on plant breeding, but the methods extend naturally to the livestock situation.

Interval mapping using maximum likelihood (Lander and Botstein 1989) or regression methods (Haley and Knott 1992; Martinez and Curnow 1992) has been the standard approach for the analysis of QTL and is available in most software packages. In the plant breeding context, these approaches require additional sources of variation to be accommodated (Moreau et al. 1999; Eckermann et al. 2001; Smith et al. 2001), and these analyses are outside the capabilities of standard software for QTL analysis. In addition, if multi-environment trials are conducted, as is

often the case, combining information and allowing for QTL by environment interaction in the presence of other sources of variation becomes necessary (Piepho 2000; Verbyla et al. 2003).

A major problem with interval mapping is the piecemeal nature of the analysis. The analysis proceeds by either examining association at various (usually regularly spaced) distances along the linkage map or by direct analysis using flanking markers (Whittaker et al. 1996). This piecemeal nature introduces the issue of multiple testing, and these tests are correlated if the intervals are on the same chromosome. In statistical terms we have a very difficult model selection problem (see Broman and Speed 2002 and the related discussion). There are other more technical issues associated with a genome scan, namely, estimates of parameters that are related to non-genetic effects can change as the analysis proceeds along the genome. This refitting of incidental parameters makes the process time-consuming, particularly in multi-environment analysis.

An extension of interval mapping that is widely used is composite interval mapping (CIM, Zeng 1994; Jansen 1994). CIM attempts to include a summary of background genetic variation in the analysis. Initial interval mapping is used to select fixed co-factors to be fitted in the subsequent interval mapping analysis. A co-factor is removed when the interval mapping scan is within 10 cM of the co-factor. A parametric bootstrap approach is used to establish a genome-wide threshold for determination of QTL.

A natural approach to allow for background genetic variation is to include all markers simultaneously in the QTL analysis. Thus Whittaker et al. (2000) use ridge regression in the context of marker-assisted selection, while Gianola et al. (2003) propose treating the markers (on the same chromosome) as correlated random effects. As ridge regression can be interpreted in terms of a random effects model, these two approaches are similar. Yi et al. (2003) extend this idea to allow for a complex mean-variance relationship for marker effects and use Markov chain Monte Carlo methods (analytical progress is not possible because of the model specification). Recent work by Foster et al. (2007) involves hierarchical modelling of the variances of marker effects and the use of the least absolute shrinkage and selection operator (Tibshirani 1996); see also Kiiveri (2004) for an extension. Most of the above approaches treat the markers as the genetic information (which they are) but in a manner that suggests the putative QTL are at the markers.

In this paper, an extension of interval mapping is proposed in which all intervals on the linkage map required for analysis are included in the model simultaneously. The approach is investigated in a two-part simulation study. The case of a single chromosome is investigated to establish simple properties of the approach and is followed by a broader and more realistic study based on the simulation study carried out by Broman and Speed (2002). In both cases, the genome-wide type I error rate is shown to be less than the nominal level. In the larger simulation study, the type I error rate is shown to approach the nominal level, for example 5%, as the population size increases. The power of the approach is demonstrated in comparison to the composite interval mapping method (Zeng 1994; Jansen 1994). The analysis of a flour milling yield experiment illustrates how the method fits naturally into a complex setting where additional components of non-genetic variation need to be included. The importance of incorporating non-genetic variation in the model for QTL detection is also illustrated. The paper concludes with discussion.

Materials

To motivate our development we use data from two field experiments conducted in 1999 and 2000 which involved 175 from a total of 180 doubled haploid (DH) lines from the Sunco × Tasman mapping population. This population was developed as part of the Grains Research and Development Corporation National Wheat Molecular Marker Program in Australia.

Both field trials were designed as randomized complete block designs with two replicates of each DH line and additional plots of parental and commercial lines. Each trial was laid out in the field as a rectangular array of 38 rows and 12 columns. Grain samples from most of the field plots were then milled using a Buhler mill. For the 1999 field trial, none of the field plots were replicated in the milling process but an additional 47 so-called milling control samples were included at regular intervals during milling of the field samples. The field plots were randomly assigned to mill days and mill order within mill days. The laboratory measurement phase took a total of 38 mill days with 11 samples milled per day. In 2000, 23% of the field samples were replicated in the milling process. Thus a total of 456 samples were milled over 38 mill days with 12 samples per mill day. Field plot samples were randomly assigned to mill days and mill order within mill days.

The trait that is examined in this paper is milling yield, which is one of the most commercially important quality traits in wheat breeding in Australia. Smith et al. (2006, 2001) have shown that milling yield is subject to large amounts of non-genetic sources of variation, hindering both genetic progress using traditional breeding approaches and efficient and accurate identification of QTL. Preliminary analysis of the phenotypic data on the

Sunco × Tasman population is presented in Smith et al. (2006) and the models discussed in that paper are the basis of further analysis presented in this paper.

Lehmensiek et al. (2005) describe in detail the construction of the linkage map for the Sunco × Tasman population. Of their original 345 markers (a mixture of AFLP, RFLP and microsatellite markers and protein analysis) we discarded 58 which were co-located with one or more other markers. Re-estimation of genetic distance was performed using the Lander and Green (1987) algorithm in the R (R Development Core Team 2006) package qtl (Broman et al. 2005).

Lehmensiek et al. (2006) carried out a QTL analysis of these data with the improved linkage map and using the methods of Verbyla et al. (2003).

Methods

The regression approach for QTL analysis (Haley and Knott 1992; Martinez and Curnow 1992) is used throughout this paper. A fundamental reason for using regression methods is the ability to easily include additional sources of variation in the model, both fixed and random effects, and hence the models discussed are linear mixed models. The approach presented below builds on the interval mapping method of Whittaker et al. (1996), and extended by Moreau et al. (1999) and Eckermann et al. (2001), for doubled haploid or recombinant inbred populations in field crops. However, the ideas and methods are generally applicable for other population structures.

Intervals rather than the markers themselves are the basis of our approach because we believe QTL are rarely at a marker. In addition, QTL will induce a correlated association between surrounding markers and the trait of interest. Thus the sizes of marker association near a QTL should be correlated in the model, something that is difficult to formulate; this correlation arises from our interval based approach in a natural way.

As the details of our approach are lengthy, an overview of the full process is presented.

Overview

A key step forward in reformulating QTL analysis is to recognise that determining QTL is a selection process. Each interval on the linkage map may contain a QTL, and our aim is to assemble the evidence for each interval, rank this evidence and ultimately select the intervals on the basis of the strength of the evidence. Our approach involves a number of components:

1. A working model based on interval mapping is formulated, in which the QTL sizes are assumed to be random effects, one for each interval. These effects are assumed to be independent normally distributed with common variance.
   To eliminate the large number of parameters, a model is proposed in which the location of a putative QTL is an interval rather than at a specific point. While this may seem to be a limitation, estimation of the location is often very imprecise and hence finding the most likely interval for a QTL would seem to be sufficient. In the context of haplotype analysis and fine mapping of QTL, the full parameterisation discussed in the paper can be used and the location estimated.
2. We use forward selection. Our full selection process consists of multiple stages but in comparison to other interval mapping approaches, the number of stages is greatly reduced; the number of models fitted at the QTL stage in our approach is usually twice the number of QTL plus 2. Firstly, a baseline model is fitted that specifies all genetic effects that do not involve markers, and all relevant non-genetic effects. For example, a polygenic effect for genetic lines would be appropriate and effects for experimental design and management of the experiment would be included. This baseline model may be the result of a model building process, as in the analysis of the Sunco × Tasman milling yield data presented below.
   The baseline model is augmented by random regression genetic effects for every interval in the linkage map. The significance of this genome-wide random regression term is examined using a residual likelihood ratio test (REMLRT). If the random regression term is significant, this indicates a putative QTL exists and a selection procedure for a putative QTL is conducted; otherwise the process concludes.
3. The selection of a putative QTL uses an outlier detection method. The alternative outlier model (Cook et al. 1982; Thompson 1985) is used to formulate a score based statistic for QTL screening and selection. Using the score based statistic, the chromosome with the largest score based statistic for a putative QTL is selected. The rationale behind this first step is that not only will the specific interval that contains the QTL be inflated, so will surrounding intervals. A component of the score statistic is then used to select the most likely interval for a QTL on the selected chromosome. This putative QTL is moved to the fixed effects part of the model. The process of selection is repeated until the test that the random regression term has zero variance is not rejected; this indicates no further QTL are likely to be present.

The details of the approach are presented below. Interval mapping is discussed and leads to a description of the working model and the REMLRT. Finally, the outlier detection method is presented.

Preliminaries

There are three sources of data in QTL analysis that are available on the genetic lines of interest. These sources are the trait or phenotypic data for the genetic lines, together with design, management and other variables that impact on the expression of the phenotype. The second source of data are markers and their scored values for each of the genetic lines. The third source of data is derived from the marker scores, namely a genetic linkage map that specifies the linkage groups (which may correspond to chromosomes) together with marker order and genetic distance as determined by recombination events.

The phenotypic data is represented by ($\boldsymbol{y}$, $\boldsymbol{X}$, $\boldsymbol{Z}$, $\boldsymbol{Z}_g$) where $\boldsymbol{y}$ is the $n \times 1$ vector of trait values on the n observational units and $\boldsymbol{X}$ and $\boldsymbol{Z}$ are $n \times t$ and $n \times b$ matrices of fixed effect and random effect variates and factors reflecting design, management and other sources of variation. The matrix $\boldsymbol{Z}_g$ is $n \times l$ relates the observed data to each of the l genetic lines and is a binary matrix.

The marker scores are arranged in an $l \times r$ matrix $\boldsymbol{M}$ of r marker scores on the l genetic lines. This is the second source of data and provides genetic information on each of the lines.

Interval mapping methods rely on the availability of a linkage map for the population, that is, an ordered set of markers usually arranged in the chromosomal structure for the organism (for example wheat or barley), together with estimated recombination fractions between adjacent markers. This is the third source of data. Let c denote the number of chromosomes and r_k denote the number of markers on chromosome k; the number of markers generally varies across chromosomes. Let $\boldsymbol{M}_k$ denote the matrix of scored markers on chromosome k. Then the matrix of markers scores can be ordered as $\boldsymbol{M} = [\boldsymbol{M}_1 \ \boldsymbol{M}_2 \ \ldots \ \boldsymbol{M}_c]$ and $r = r_1 + \cdots + r_c$. For genetic line i, $m_{i,k;j}$ and $m_{i,k;j+1}$ will denote the jth and $j+1$th marker scores for a pair of adjacent markers on chromosome k; this is the jth interval on chromosome k. The vector of these scores across all lines will be denoted by $\boldsymbol{m}_{k;j}$ and $\boldsymbol{m}_{k;j+1}$ respectively, and are columns j and $j+1$ of $\boldsymbol{M}_k$.

Let $\theta_{k;j,j+1}$ denote the recombination fraction between markers j and $j+1$ (the jth interval) on chromosome k. The genetic distance (in Morgans) for interval j, $d_{k;j,j+1}$, will be based on Haldane's distance,

$$d_{k;j,j+1} = -\frac{1}{2}\log(1 - 2\theta_{k;j,j+1})$$

although other distance measures could be used. The abbreviation cM will be used for centi-Morgan distance based on 100 $d_{k;j,j+1}$. The calculations used in interval mapping assume that recombination events occur at random along the genome, and this corresponds to this distance measure.

Missing trait data is usually estimated in the analysis, together with the effects of interest. Missing marker scores are handled as in Martinez and Curnow (1994).

The total genetic effect for genetic line $i = 1,2,\ldots,l$ will be denoted by g_i and the vector of these effects by $\boldsymbol{g}$. We begin with the general model (see Verbyla et al. 2003 for details)

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}_g\boldsymbol{g} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{\tau}$ is a $t \times 1$ vector of fixed effects, $\boldsymbol{u}$ is a $b \times 1$ vector of random effects assumed $N(\boldsymbol{0}, \sigma^2\boldsymbol{G}(\boldsymbol{\gamma}))$, and $\boldsymbol{\epsilon}$ is the residual vector, assumed $N(\boldsymbol{0}, \sigma^2\boldsymbol{R}(\boldsymbol{\phi}))$. The latter two effect vectors are assumed mutually independent. The fixed, random and residual terms reflect the design and conduct of the trial, and as such provide the underlying structure for non-genetic variation through the associated structures, namely $\boldsymbol{\tau}$ and the parameters of the covariance matrices $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$.

The vector of genotypic effects is of prime interest and is decomposed as follows. If we have a single QTL,

$$g_i = q_i a + p_i \tag{2}$$

where a represents the size of the QTL (on the scale of the trait), q_i is unknown, but is either -1 or 1 for doubled haploid lines depending on the parental allele at the QTL (Whittaker et al. 1996), and p_i is the residual or polygenic effect, assumed to be distributed $N(0,\sigma^2\gamma_g)$. In vector form (2) is given by

$$\boldsymbol{g} = \boldsymbol{q}a + \boldsymbol{p} \tag{3}$$

with $\boldsymbol{p} \sim N(\boldsymbol{0}, \sigma^2\gamma_g \boldsymbol{I}_l)$.

Interval mapping: the regression approach

Our approach is based on interval mapping using the regression approach of Whittaker et al. (1996). The regression method for interval mapping involves replacing q_i by its expected value given the flanking markers that define the interval being examined.

Consider chromosome k. Let $\theta_{k;j}$ denote the recombination fraction between marker j and the putative QTL in the jth interval on chromosome k, and $\theta^*_{k;j}$ denote the recombination fraction between the putative QTL in the jth interval and marker $j+1$. Thus $0 \le \theta_{k;j}, \theta^*_{k;j} \le \theta_{k;j,j+1}$ and a form of Trow's formula (1913), $(1 - 2\theta_{k;j})(1 - 2\theta^*_{k;j}) = 1 - 2\theta_{k;j,j+1}$ connects these recombination rates. Using

Haldane's distance measure, the distance from marker j to the QTL is

$$d_{k;j} = -\frac{1}{2}\log(1 - 2\theta_{k;j}) \tag{4}$$

Figure 1 presents the physical representation for chromosome k with recombination frequencies, distances and a putative QTL in interval j (the notation for the size is introduced below).

Whittaker et al. (1996) show that the conditional expectation of the QTL genotype given two flanking markers is

$$E(q_i|m_{i,k;j}, m_{i,k;j+1}, \theta_{k;j}, \theta_{k;j,j+1}) = m_{i,k;j}\lambda_{k;j,j} + m_{i,k;j+1}\lambda_{k;j+1,j} \tag{5}$$

where

$$\lambda_{k;j,j} = \lambda_{k;j,j}(\theta_{k;j};\, \theta_{k;j,j+1}) = \frac{(1 - \theta_{k;j,j+1} - \theta_{k;j})(\theta_{k;j,j+1} - \theta_{k;j})}{\theta_{k;j,j+1}(1 - \theta_{k;j,j+1})(1 - 2\theta_{k;j})} \tag{6}$$

$$\lambda_{k;j+1,j} = \lambda_{k;j+1,j}(\theta_{k;j};\, \theta_{k;j,j+1}) = \frac{\theta_{k;j}(1 - \theta_{k;j})(1 - 2\theta_{k;j,j+1})}{\theta_{k;j,j+1}(1 - \theta_{k;j,j+1})(1 - 2\theta_{k;j})} \tag{7}$$

Note that $0 \le \lambda_{k;j,j} \le 1$ and $0 \le \lambda_{k;j+1,j} \le 1$ and as these two variables are functions of the unknown $\theta_{k;j}$, they are also unknown.

At this point, a change in notation for the size of the QTL effect is appropriate. Interval analysis implicitly assumes there may be a QTL in every interval. Thus we let $a_{k;j}$ denote the size of a putative QTL in the jth interval on chromosome k and replace a by $a_{k;j}$. Applying (5) in (2) and hence in (3), we have in vector form

$$\begin{aligned} \boldsymbol{g} &= (\boldsymbol{m}_{k;j}\lambda_{k;j,j} + \boldsymbol{m}_{k;j+1}\lambda_{k;j+1,j})a_{k;j} + \boldsymbol{p} \\ &= \boldsymbol{m}_{k;j}\alpha_{k;j,j} + \boldsymbol{m}_{k;j+1}\alpha_{k;j+1,j} + \boldsymbol{p} \end{aligned} \tag{8}$$

where $\alpha_{k;j,j} = \lambda_{k;j,j}\, a_{k;j}$ and $\alpha_{k;j+1,j} = \lambda_{k;j+1,j}\, a_{k;j}$. In (8), the subscripts $k;j$ and $k;j + 1$ have been used to indicate the interval being examined and hence there are only two regression parameters in each fit of flanking markers across the genome.

The full model for analysis of interval j on chromosome k is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}_g\boldsymbol{M}_{k;j}\boldsymbol{\alpha}_{k;j} + \boldsymbol{Z}_g\boldsymbol{p} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon}$$

where $\boldsymbol{M}_{k;j} = [\boldsymbol{m}_{k;j} \quad \boldsymbol{m}_{k;j+1}]$ and $\boldsymbol{\alpha}_{k;j} = [\alpha_{k;j,j}\, \alpha_{k;j+1,j}]^{\mathrm{T}}$.

This model is fitted for each k and appropriate j, so that $\boldsymbol{\tau}$ is re-estimated, as are parameters associated with $\boldsymbol{p}$, $\boldsymbol{u}$ and $\boldsymbol{\epsilon}$, namely $\sigma^2, \gamma_g, \boldsymbol{\gamma}$ and $\boldsymbol{\phi}$. For a QTL to be in an interval requires the sign of the fitted regression coefficients $[\alpha_{k;j,j}, \alpha_{k;j+1,j}]^{\mathrm{T}}$ to be the same (Whittaker et al. 1996).

Working statistical model

As in conventional interval mapping we assume every interval may contain a QTL. Unlike interval mapping however, we include all intervals in a single analysis, and assume a working model of a simple random effect for the size of QTL across the genome.

The genetic model we use is an extension of (2), namely

$$g_i = \sum_{k=1}^{c}\sum_{j=1}^{m_k - 1} q_{i,k;j}a_{k;j} + p_i \tag{9}$$

where $q_{i,k;j}$ is the indicator of parental type (or allele number) of a putative QTL in the jth interval on chromosome k and $a_{k;j}$ is the size of a putative QTL in that interval. As a working model we assume $a_{k;j} \sim N(0, \sigma^2\, \gamma_a)$ and that the sizes are mutually independent. Thus we assume a QTL may exist in every interval and that the presence of a significant QTL variance ratio γ_a suggests at least one QTL may be present.

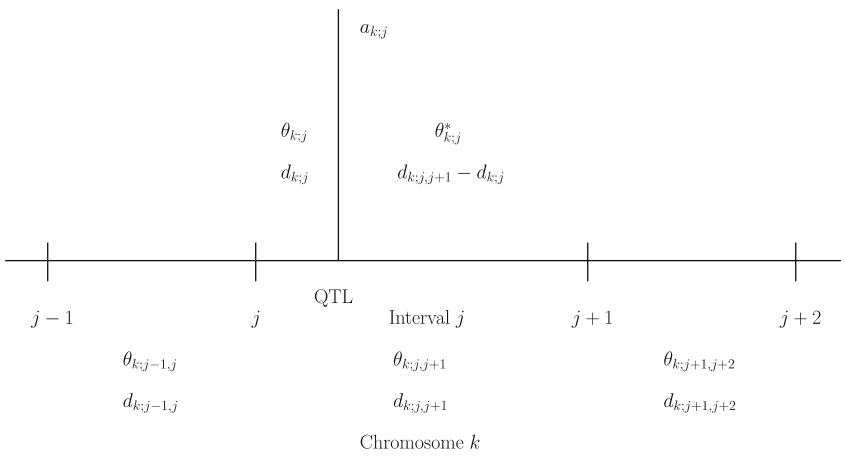

Fig. 1 Notation for interval mapping: recombination frequencies, distances and QTL for intervals on chromosome k

We do not believe the working model reflects reality as far as QTL effects is concerned. It is a vehicle that will allow the detection of putative QTL that behaves like a penalty as in ridge regression (Whittaker et al. 2000).

As for interval mapping we replace $q_{i,k;j}$ in (9) by its expected value given markers j and $j+1$ on chromosome k. Thus using (8) we can write the vector of genetic effects $\boldsymbol{g}$ as

$$\boldsymbol{g} = \sum_{k=1}^{c} \sum_{j=1}^{m_k-1} (\boldsymbol{m}_{k;j}\lambda_{k;j,j} + \boldsymbol{m}_{k;j+1}\lambda_{k;j+1,j})a_{k;j} + \boldsymbol{p}$$

or succinctly as

$$\boldsymbol{g} = \boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{a} + \boldsymbol{p} \tag{10}$$

where $\boldsymbol{a}$ is the vector of sizes of effects, and $\boldsymbol{\Lambda}$ is a block diagonal matrix of size $r \times (r-c)$, with kth block

$$\boldsymbol{\Lambda}_k = \begin{bmatrix} \lambda_{k;1,1} & 0 & 0 & \dots & 0 \\ \lambda_{k;2,1} & \lambda_{k;2,2} & 0 & \dots & 0 \\ 0 & \lambda_{k;3,2} & \lambda_{k;3,3} & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & \lambda_{k;m_k-1,m_k-1} \\ 0 & 0 & 0 & \dots & \lambda_{k;m_k,m_k-1} \end{bmatrix}$$

so that the blocks correspond to chromosomes or linkage groups. Note that $\boldsymbol{\Lambda}$ contains unknown recombination frequencies $\theta_{k;j}$.

Equation (10) can be written as:

$$\boldsymbol{g} = \boldsymbol{M}\boldsymbol{a}_M + \boldsymbol{p}$$

where $\boldsymbol{a}_M = \boldsymbol{\Lambda}\boldsymbol{a}$ are marker effects, rather than interval effects. Under the independent normality assumption for $\boldsymbol{a}$, we see that

$$\boldsymbol{a}_M \sim N(\boldsymbol{0},\ \sigma_a^2\boldsymbol{\Lambda}\boldsymbol{\Lambda}^T)$$

The matrix $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}}$ is tri-diagonal and shows that the QTL sizes *induce* a variance-covariance structure between the size of marker effects, something that would be expected. This suggests that any analysis based on markers should incorporate some form of variance–covariance structure. Model (10) results in a variance-covariance matrix which has the form of a generalized moving average process (Diggle 1990), although the the matrix $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}}$ is not of full rank; it is a so-called reduced rank form (Thompson et al. 2003). In contrast, a full rank autoregressive structure was proposed for the marker variance–covariance matrix by Gianola et al. (2003).

Returning to (10), $\boldsymbol{g}$ given $\boldsymbol{\Lambda}$ follows a normal distribution with mean vector $\boldsymbol{0}$ and variance matrix given by

$$\mathrm{var}(\boldsymbol{g}|\boldsymbol{\Lambda}) = \sigma^2\left(\gamma_a\boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{M}^{\mathrm{T}} + \gamma_g\boldsymbol{I}_l\right)$$

Apart from σ^2, there are $r-c+2$ distinct parameters in this variance matrix, namely the $r-c$ parameters $\theta_{k;j}$, and γ_a and γ_g. The full model for the trait is based on (1) and (10). The distribution of $\boldsymbol{y}$ given $\boldsymbol{\Lambda}$ is

$$\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\tau},\ \sigma^2\boldsymbol{H}) \tag{11}$$

where

$$\boldsymbol{H} = \boldsymbol{R} + \gamma_a\boldsymbol{Z}_g\boldsymbol{M}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{M}^{\mathrm{T}}\boldsymbol{Z}_g^T + \gamma_g\boldsymbol{Z}_g\boldsymbol{Z}_g^{\mathrm{T}} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}$$

If a QTL is at a marker, the matrix $\boldsymbol{\Lambda}$ will contain two identical columns with a single unit entry. Thus the model will be singular in terms of the sizes of the effects in the two intervals. They in fact coincide, and obviously both cannot be estimated. This issue is resolved in the following section.

For large r relative to the sample size n, problems may arise in estimation of $\lambda_{k;j,j}$ and hence $\theta_{k;j}$. Even if estimation is possible, the estimates may be very poor as information on the precise location of a QTL is likely to be small. An approach is presented in the next section that overcomes this difficulty.

Reducing the parameterisation

To eliminate the parameters $\lambda_{k;j,j}$, we assign a prior distribution regarding the location of the QTL and our aim is to integrate (or average out) these parameters over each interval. Without prior information, a QTL in any interval can occur at any location within that interval. We therefore assume that the distance from the left hand marker to a putative QTL, $d_{k;j}$, is uniformly distributed. Notice that it must be distance and not the recombination fraction that takes on the uniform distribution, because distance is additive whereas recombination fractions are not. Thus using Haldane's distance measure (4), it is assumed $d_{k;j} \sim U[0,\ d_{k;j,j+1}]$, where U denotes the uniform distribution defined on the range specified. With this specification, the $d_{k;j}$ or $\theta_{k;j}$ need to be integrated out to form a marginal distribution. Unfortunately, this is analytically intractable.

To make progress, we replace $\boldsymbol{\Lambda}$ in (10) by its expected value, namely $\boldsymbol{\Lambda}_E = \mathrm{E}(\boldsymbol{\Lambda})$, in the manner of the regression approach to interval mapping. The non-zero elements of $\boldsymbol{\Lambda}_E$ are shown in the Appendix to be

$$\mathrm{E}\left(\lambda_{k;j,j}\right) = \mathrm{E}\left(\lambda_{k;j+1,j}\right) = \frac{\theta_{k;j,j+1}}{2d_{k;j,j+1}(1-\theta_{k;j,j+1})} \tag{12}$$

Note that this implies that we have a regression on "pseudo-markers"

$$\frac{\theta_{k;j,j+1}}{d_{k;j,j+1}(1-\theta_{k;j,j+1})}\,\frac{1}{2}\left(\boldsymbol{m}_{k;j} + \boldsymbol{m}_{k;j+1}\right) \tag{13}$$

for each interval. Thus the average of the markers scores for the markers defining the interval are scaled according to

their separation in terms of recombination fraction and genetic distance. Figure 2 is a graph of the factor that multiplies the average marker scores for each genetic line as given in (13) plotted against recombination fraction. Notice that for most reasonable intervals, the recombination fraction will be of the order of 0.1 (with distance approximately 10 cM), and the multiplying factor is close to 1. At large recombination fractions and hence large distances between markers, the factor in (13) will depress the marker average. Thus wider intervals will have smaller "pseudo-markers" and this will impact on the size of effect predicted for that interval and on the associated prediction error variance.

Lastly, using such "pseudo-markers" (13) is not intuitively obvious and arises through the use of interval mapping.

Our genetic model is based on these moment calculations and is given by

$$\boldsymbol{g} = \boldsymbol{M}\boldsymbol{\Lambda}_E\boldsymbol{a} + \boldsymbol{p}$$

Under this model, if $\boldsymbol{M}_E = \boldsymbol{Z}_g\boldsymbol{M}\boldsymbol{\Lambda}_E$, the model specified by (11) now has variance matrix $\sigma^2\,\boldsymbol{H}$ where

$$\boldsymbol{H} = \boldsymbol{R} + \gamma_a\boldsymbol{M}_E\boldsymbol{M}_E^{\mathrm{T}} + \gamma_g\boldsymbol{Z}_g\boldsymbol{Z}_g^{\mathrm{T}} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}$$

and $\boldsymbol{M}_E$ is a *known* matrix.

Estimation

The full model for QTL analysis can be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{M}_E\boldsymbol{a} + \boldsymbol{Z}_g\boldsymbol{p} + \boldsymbol{Z}\boldsymbol{u} + \epsilon \qquad (14)$$

which is a linear mixed model. Thus residual maximum likelihood or REML (Patterson and Thompson 1971) and best linear unbiased prediction or BLUP (Robinson 1991) are appropriate methods for analysis.

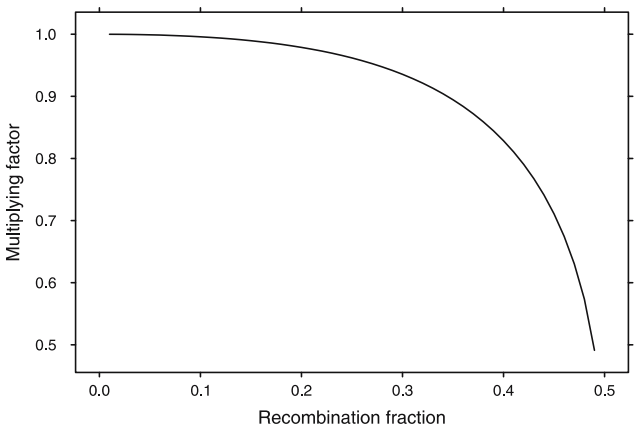

Fig. 2 Multiplying factor for "pseudo-markers" defined by (13), against recombination fraction

The main advantage of (14) is that all intervals are used simultaneously in one analysis. Thus the benefits of composite interval mapping are built into the approach through the inclusion of all intervals in the analysis, and polygenic and non-genetic effects are estimated allowing for all possible QTL.

Test for the presence of a QTL

The process of finding a QTL and incorporating the QTL into a final model involves three basic steps.

If the interval marker random regression has a variance ratio γ_a that is significantly greater than zero, this suggests that at least one QTL is present. Thus the first step involves fitting models both with and without random regression effects for the size of QTL, and testing if $H_0{:}\gamma_a = 0$. If not significant the process is terminated.

Thus we fit (14) obtaining a maximized residual log-likelihood $\log(L)$, say, and fit the model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}_g\boldsymbol{p} + \boldsymbol{Z}\boldsymbol{u} + \epsilon$$

obtaining a maximised residual log-likelihood $\log(L_0)$. The REMLRT for testing $H_0{:}\gamma_a = 0$ is based on the statistic

$$-2\log\Lambda = 2\{\log(L) - \log(L_0)\}$$

which under H_0 has a distribution which is a mixture of chi-squared distributions, namely $0.5\chi_0^2 + 0.5\chi_1^2$ (Stram and Lee 1994).

If the hypothesis is rejected, the process proceeds to the second step, the outlier detection stage.

Detection of QTL: an outlier model

All intervals on the linkage map can be classified into two groups. The first group consists of the intervals not containing a QTL and is large in number. The size of estimated QTL effects for these intervals will be small, because these intervals do not contain a QTL. The second group is small in number and consists of the intervals that contain a QTL. The size of QTL effect for these intervals will reflect the presence of a QTL and hence will be "large". Thus the QTL size effects represent outliers in comparison to the majority of intervals as given by the first group. An approach to detect outliers may therefore be used to select intervals for putative QTL.

We use the alternative outlier model (AOM) introduced for linear models (Cook et al. 1982; Thompson 1985) and developed for linear mixed models in an unpublished PhD thesis of Gogel (1997); see also Gogel et al. (2001).

The AOM for the size vector for all intervals on chromosome k is given by

$$\boldsymbol{a}^{ko} = \boldsymbol{a} + \boldsymbol{E}_k\boldsymbol{\delta}_k \tag{15}$$

where $\boldsymbol{\delta}_k$ is a vector of random effects and $\boldsymbol{E}_k^{\mathrm{T}} = [\boldsymbol{0}\ \ \boldsymbol{I}_{rk-1}\boldsymbol{0}]$ maps these effects to chromosome k. We assume that $\boldsymbol{\delta}_k \sim N(\boldsymbol{0}, \sigma^2\gamma_{a,k}\boldsymbol{I}_{r_k-1})$. This model modifies the size of the QTL effects for intervals $j = 1,2,...,r_k - 1$ by inflating the variance on that chromosome and hence allowing larger predicted random size effects. Under (15), the full mixed model is given by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{M}_E\boldsymbol{a} + \boldsymbol{M}_E\boldsymbol{E}_k\boldsymbol{\delta}_k + \boldsymbol{Z}_g\boldsymbol{p} + \boldsymbol{Z}\boldsymbol{u} + \epsilon \tag{16}$$

In order to develop a statistic that can be used to locate QTL, we consider the score (first derivative of the residual log-likelihood) for $\gamma_{a,k}$ evaluated at zero. If there is no QTL on the chromosome, the score will have mean zero. We firstly seek the chromosomes that have a score that is inflated as this will suggest a QTL may be present on that chromosome.

If $\boldsymbol{P} = \boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}$, the REML score for $\gamma_{a,k}$ under (16) evaluated at $\gamma_{a,k} = 0$ is given by

$$U_k(0) = -\frac{1}{2}\left\{\mathrm{tr}\left(\boldsymbol{P}\boldsymbol{M}_E\boldsymbol{E}_k\boldsymbol{E}_k^{\mathrm{T}}\boldsymbol{M}_E^{\mathrm{T}}\right) - \frac{1}{\sigma^2}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{M}_E\boldsymbol{E}_k\boldsymbol{E}_k^{\mathrm{T}}\boldsymbol{M}_E^{\mathrm{T}}\boldsymbol{P}\boldsymbol{y}\right\} \tag{17}$$

The score can be simplified as follows. If $\tilde{\boldsymbol{a}}$ is the BLUP for the sizes $\boldsymbol{a}$ under (14) the vector $\tilde{\boldsymbol{w}}_k = \boldsymbol{E}_k^{\mathrm{T}}\boldsymbol{M}_E^{\mathrm{T}}\boldsymbol{P}\boldsymbol{y}$ which appears in (17) can be written as:

$$\begin{aligned}\tilde{\boldsymbol{w}}_k &= \frac{1}{\gamma_a}\boldsymbol{E}_k^{\mathrm{T}}\tilde{\boldsymbol{a}} \\ &= \frac{\tilde{\boldsymbol{a}}_k}{\gamma_a}\end{aligned} \tag{18}$$

If

$$\boldsymbol{C}_{k,k} = \boldsymbol{E}_k^{\mathrm{T}}\boldsymbol{M}_E^{\mathrm{T}}\boldsymbol{P}\boldsymbol{M}_E\boldsymbol{E}_k$$

it can be shown $\mathrm{E}\left(\tilde{\boldsymbol{w}}_k^{\mathrm{T}}\tilde{\boldsymbol{w}}_k\right) = \sigma^2\mathrm{tr}\left(\boldsymbol{C}_{k,k}\right)$ and $\mathrm{E}\left(\tilde{\boldsymbol{a}}_k^{\mathrm{T}}\tilde{\boldsymbol{a}}_k\right) = \sigma^2\gamma_a^2\mathrm{tr}\left(\mathrm{C}_{k,k}\right)$. With these definitions and using properties of the trace (17) can be written as:

$$\begin{aligned}U_k(0) &= \frac{1}{2}\left(\tilde{\boldsymbol{w}}_k^{\mathrm{T}}\tilde{\boldsymbol{w}}_k\sigma^2 - \mathrm{tr}\left(\boldsymbol{C}_{k,k}\right)\right) \\ &= \frac{\mathrm{tr}\left(\boldsymbol{C}_{k,k}\right)}{2}\left(t_k^2 - 1\right)\end{aligned} \tag{19}$$

where

$$t_k^2 = \frac{\tilde{\boldsymbol{w}}_k^{\mathrm{T}}\tilde{\boldsymbol{w}}_k}{\sigma^2\mathrm{tr}\left(\boldsymbol{C}_{k,k}\right)} = \frac{\tilde{\boldsymbol{a}}_k^{\mathrm{T}}\tilde{\boldsymbol{a}}_k}{\sigma^2\gamma_a^2\mathrm{tr}\left(\boldsymbol{C}_{k,k}\right)} = \frac{\sum_{j=1}^{r_k-1}\tilde{a}_{k;j}^2}{\sum_{j=1}^{r_k-1}\mathrm{var}\left(\tilde{a}_{k;j}\right)}$$

using (18). Thus the BLUP of the QTL sizes and their estimated variances under (14) are required to calculate the statistic t_k^2 and naturally provide the basis of outlier detection. If $\boldsymbol{C}$ is the coefficient matrix of the mixed model equations (see for example Henderson 1950; Robinson 1991) and $\boldsymbol{C}_{k;j,k;j}^{M_EM_E}$ is the component of $\boldsymbol{C}^{-1}$ relating to $\tilde{a}_{k;j}$, then $\sigma^2\boldsymbol{C}_{k;j,k;j}^{M_EM_E}$ is the prediction error variance of $\tilde{a}_{k;j}$ that is generally available in mixed model software. The estimated variances of $\tilde{a}_{k;j}$ can then be found using

$$\mathrm{var}\left(\tilde{a}_{k;j}\right) = \sigma^2\gamma_a - \sigma^2\boldsymbol{C}_{k;j,k;j}^{M_EM_E}$$

If there is no QTL on the chromosome, the score statistic has mean zero and hence a "large" deviation of the observed score from zero indicates a QTL may be present. Thus large t_k^2 suggest a QTL is present and the chromosome with the largest t_k^2 is selected as being most likely to contain a QTL.

Having determined the chromosome most likely to contain a QTL, the intervals within that chromosome are examined. If we replace $\boldsymbol{E}_k$ in the derivation of the score by a single column $\boldsymbol{e}_{k;j}$ that selects interval j on chromosome k (the selected chromosome), the score for each interval on chromosome k can be determined. In fact for a single interval j on chromosome k, (19) can be written as:

$$U_{k;j}(0) = \frac{c_{k;j,j}}{2}\left(t_{k;j}^2 - 1\right)$$

where

$$t_{k;j}^2 = \frac{\tilde{w}_{k;j}^2}{\sigma^2 c_{k;j,j}} = \frac{\tilde{a}_{k;j}^2}{\sigma^2\gamma_a^2 c_{k;j,j}} = \frac{\tilde{a}_{k;j}^2}{\mathrm{var}\left(\tilde{a}_{k;j}\right)}$$

and the $t_{k;j}^2$ reflect the importance of an interval with respect to a putative QTL. Again if $t_{k;j}^2$ is large this suggests a QTL may be present in the interval and hence the interval with largest $t_{k;j}^2$ is chosen as the likely position for the QTL.

The selected QTL interval is now moved to the fixed effects and the process repeated until the random effects QTL component is not significant.

Final model

When the selection process concludes, all putative QTL will appear as fixed effects. Thus if S putative QTL are selected, the final model is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \sum_{s=1}^{S}\boldsymbol{m}_{E,s}a_s + \boldsymbol{Z}_g\boldsymbol{p} + \boldsymbol{Z}\boldsymbol{u} + \epsilon \tag{20}$$

where $\boldsymbol{m}_{E,s} = \boldsymbol{Z}_g\boldsymbol{M}_s\boldsymbol{\lambda}_{E,s}$ is the appropriate vector for the sth putative QTL, $\boldsymbol{M}_s$ is a $n \times 2$ matrix of the marker scores defining the interval and $\boldsymbol{\lambda}_{E,s}$ is the appropriate column of $\boldsymbol{\Lambda}_E$.

The final step in the process is to fit (20) and to summarise the QTL information. We call the approach presented whole genome average interval mapping (WGAIM).

Implementation

The WGAIM approach presented above has been implemented by Simon Diffey of the New South Wales Department of Primary Industries. The implementation is in R (R Development Core Team 2006) and is built on the libraryqtl (Broman et al. 2005). The model fitting involves using theasreml package (Butler et al. 2007) in R. The implementation was used in the simulation study and the analysis of the Sunco × Tasman flour yield trials presented below.

Simulation study: details

To examine the performance of WGAIM, a simulation study was conducted. The new method was compared to composite interval mapping (CIM), because CIM is a widely used approach. The simulation study was conducted in two parts. The first part involves examining the performance of the method for a single chromosome while the second part is similar to the simulation study of Broman and Speed (2002) and involves multiple chromosomes.

The number of genetic lines and hence the population size considered is l = 100, 200 and 400. These values were chosen to reflect the size of population common in crop trials (and hence differ from Broman and Speed 2002). In all simulations, two replicates of each line were generated. This differs from Broman and Speed (2002) who consider a single replicate for each line. However, in the case of crops and other plants, replication of genetic lines in trials is common, and as described below, our simulations are constructed to ensure that our results are comparable to those of Broman and Speed (2002).

The simple model for the simulated y_{ij}, $i = 1,\ldots,l; j = 1,2$ is given by

$$y_{ij} = \mu + g_i + e_{ij}$$

where $e_{ij} \sim N(0, \sigma^2)$ are independent random variables. The genetic line effects g_i are given by

$$g_i = \sum_{s=1}^{S} q_{is} a_s + p_i$$

where there are S QTL, and q_{is} for QTL s is either -1 or 1 depending on the parental allele that line i carries. The polygenic effects $p_i \sim N(0, \sigma_g^2)$ are independent and also independent of e_{ij}. The leading term represents the QTL present in the simulations, with the number S and their configuration being presented below for the two parts of the simulation study. For each simulation scenario, 2000 simulations were carried out.

In all simulations, $\sigma^2 = 1$, $\sigma_g^2 = 0.5$ and $a_s = 0.378$. Thus genetic line means will have a variance of $\sigma^2/2 + \sigma_g^2 = 1$ and we call this the line mean variance below as it relates to the stratum of variation due to the genetic lines. This is equivalent to the phenotypic variance of 1 in the simulation study of Broman and Speed (2002). The total line mean variance will depend on the number of QTL in the simulation and also on whether they are linked. For example, for a single chromosome with two QTL in coupling of size a_1 and a_2, respectively, and a distance of d Morgans apart, the total mean line variance is

$$a_1^2 + 2e^{-2d} a_1 a_2 + a_2^2 + 1$$

where we have assumed Haldane's distance (4).

The underlying chromosomal structures are 100 cM in length, with 11 equally spaced markers. Broman and Speed (2002) locate the QTL at markers, but in our study, the QTL are located at midpoints of intervals in configurations to be discussed below. This is the more likely situation in practice. The genetic lines are simulated on the basis of this underlying structure, and hence each line in the simulation has scores on the 11 markers for each chromosome and on the QTL.

For each population size (l = 100,200,400), a linkage map was constructed on the basis of the simulated marker scores, and also including the simulated QTL scores to accurately place the QTL on the estimated linkage map. This was required to confirm the detection or non-detection of a QTL using WGAIM and CIM. Note that while the underlying structure has chromosomes of length 100 cM and equally spaced markers, the estimated linkage map for a chromosome varied in length from 80 to 120 cM, with marker spacing varying from 4 to 25 cM. The qtl (Broman et al. 2005) package in R (R Development Core Team 2006) was used to develop the linkage map and then to check the order of the markers and QTL. The final linkage map was used for all 2000 simulations for the particular scenario and population size. Thus map uncertainty was included in the simulation study.

Analysis for each simulation using WGAIM was carried out as outlined in the Methods section. For composite interval mapping, an initial interval mapping scan was conducted every 5 cM at what we call imputed markers. As the length of a chromosome varied, the number of locations for the interval mapping scan varied. A number of cofactors (with the number depending on the scenario) were determined using forward selection and were included in the model for analysis. A parametric bootstrap was used to set the genome-wide threshold for the F test for significant QTL; see Broman and Speed (2002) for further details regarding the parametric bootstrap. Either 25,000 (Part 1) or 20,000 (Part 2) bootstrap samples were used (and found

to yield a stable threshold). This was a very time-consuming process.

QTL under the CIM model were selected based on an F-statistic, with peaks in a 5 cM scan above the threshold being selected. For scan positions within 10 cM of a co-factor, the co-factor was removed from the augmented model. Two peaks were deemed to be separate QTL if the minimum F-statistic between the peaks was less than half the larger of the two peaks.

Comparison of WGAIM and CIM is complicated by the fact that one method is based on intervals while the other is based on marker or specific position. In the context of markers rather than intervals, Broman and Speed (2002) defined successful detection of a QTL if the correct marker or a marker 10 cM either side of the correct marker is selected. However, those authors assumed a fixed map and in reality the map distances will vary depending on the simulated population of lines. For WGAIM we define correct detection of a QTL if the interval containing the QTL or either of two neighbouring intervals are selected. In effect, WGAIM uses the midpoints of the intervals as pseudo-markers and hence in this sense our definition of correct detection provides for an average interval of 20 cM and approximately matches the approach of Broman and Speed (2002). For CIM, an equivalent approach is to allow the five imputed markers about the QTL to be included in the definition of correct detection. On average this will also mean a width of 20 cM. However for both WGAIM and CIM, there will be variation in this width of 20 cM across the genome. With these definitions for WGAIM and CIM, the comparisons should be fair, if not providing exactly equivalent results.

All methods can falsely discover QTL. The type I error rate for WGAIM in the case of no QTL is based on the (REML) likelihood ratio test (REMLRT) that the variance component for the distribution of size of QTL is zero. At each stage of testing, the REMLRT is compared to the 95th percentile of the asymptotic null distribution, namely the mixture distribution $0.5\chi^2_0 + 0.5\chi^2_1$ (Stram and Lee 1994). For the case of no QTL, a type I error occurs if the first such test is rejected.

For simulations where QTL are present, falsely discovered QTL are called extraneous. These extraneous QTL may be linked, so that a falsely selected QTL is on the same chromosome as a true QTL, or unlinked in the case of a chromosome that does not contain a QTL. This latter situation results in a false discovery rate (FDR). WGAIM and CIM were examined for both types extraneous QTL. In the simulations the rate of detection of extraneous QTL is measured by the mean number of extraneous QTL per simulation; this is defined as the sum of the number of extraneous QTL found by the number of simulations for that number of extraneous QTL divided by the total number of simulations. By definition, the FDR will be at least the Type I error rate. FDR is often viewed as an important measure of the performance of an approach, as discovering extraneous QTL can result in poor decision making for marker assisted selection.

Part 1

The first simulation study is based on a single chromosome. Four scenarios were considered, namely

1. No QTL, to establish the Type I error rate and false discovery rate (FDR) for WGAIM.
2. $S = 1$ QTL at 45 cM in the underlying chromosome, comprising 12.5% of the total line mean variance.
3. $S = 2$ QTL in coupling, at 45 and 85 cM, respectively, jointly 29.3% of the total line mean variance because of linkage.
4. $S = 2$ QTL in repulsion, again at 45 cM and 85 cM respectively, jointly 13.6% of the total line mean variance because of linkage.

The methods of QTL analysis are WGAIM, and CIM with 1 co-factor (CIM1) for $S = 1$ and 2 co-factors (CIM2) for $S = 2$. The aim was to compare WGAIM with the best equivalent CIM approach.

Part 2

The second simulation is based on the comprehensive simulation study of Broman and Speed (2002). Genetic lines are assumed to have a total of 9 chromosomes, each with 11 equally spaced markers at a spacing of 10 cM. There are two scenarios, namely

1. No QTL, to establish the Type I error rate and FDR for more than a single chromosome.
2. $S = 7$ QTL. The locations of the QTL are denoted by (Ck, Ij) where Ck is chromosome k and Ij is interval j on that chromosome. QTL are located at the midpoints of the intervals (C1, I4), (C1, I8), (C2, I4), (C2, I8), (C3, I6), (C4, I4), (C5, I1), unlike Broman and Speed (2002) who assumed that each QTL was located at the left-hand marker of each interval. The favourable QTL allele is coded as 1 for all but the second QTL (C1, I8), for which it is –1. Thus the two QTL on chromosome 1 are in repulsion while the two QTL on chromosome 2 are in coupling. The proportion of the total line mean variance due to the 7 QTL was $7 \times 0.378^2/(7 \times 0.378^2 + 1) = 1/(1 + 1) = 0.50$, again matching Broman and Speed 2002). There are additional terms in this calculation for the QTL in coupling and

repulsion due to linkage, but these additional terms cancel.

The methods of QTL analysis are WGAIM, and CIM with 7 co-factors (CIM7), the latter chosen by forward selection. Broman and Speed (2002) found that CIM7 performed well in their simulation study, and as the correct number of cofactors are included our comparison is again with the best CIM approach.

Simulation study: results

Part 1: Type I error rate and FDR

For case of a single chromosome with no QTL, the Type I error rate and FDR for WGAIM are presented in Table 1. The Type I error rate of the our test is generally less than the set nominal 5% level, as can be seen in Table 1. This result is consistent with the results reported by Crainiceanu and Ruppert (2004); see also an unpublished PhD thesis by Welham (2006). The second measure presented in Table 1 is the mean number of extraneous QTL or FDR. The rates are consistent across population sizes and are very low.

We did not investigate Type I error rates for CIM as these are well established and found using parametric bootstrap methods.

Part 1: Single QTL

Turning to the simulation results for a single QTL, the proportion of correct identifications of the QTL for WGAIM and CIM1 are given in Table 2. The trends are very clear. For both WGAIM and CIM1 the proportion of simulations in which the QTL is correctly identified increases as the population size increases. WGAIM is best overall.

WGAIM has a higher mean number of extraneous (and of course in this case linked) QTL. CIM1 has a lower rate of false positives than WGAIM at the larger population sizes. For a small population size, detection of a QTL is difficult. In this case, it may be considered an advantage to at least find a QTL, which while not located in the correct interval, is linked to the correct QTL. WGAIM does this for a population size of 100.

Part 1: Two QTL in coupling or repulsion

For two QTL in coupling or repulsion on a single chromosome, the results for each QTL are presented in Table 3. Two-way tables of non-detection (labelled not *D*) and detection (labelled *D*) of the intervals I4 (left of the table) and I8 (top of the table) are presented for the two QTL in coupling or repulsion. The marginal Totals are also presented in this table. Examining the two-way tables allows us to investigate the ability of WGAIM and CIM2 to detect the two QTL simultaneously. The mean number of linked extraneous QTL found in the simulations is also reported in Table 3.

We begin with two QTL in coupling. Firstly, the rate of correctly identifying either QTL for a population size of 100 is low for both methods. WGAIM correctly detects the two individual QTL with rates of 0.709 and 0.587, while CIM2 has corresponding rates of 0.528 and 0.489. WGAIM is able to correctly detect both with a rate of 0.337, while CIM2 is at a low rate of 0.118. Both methods improve as the population size increases, with WGAIM being uniformly better than CIM2.

The rate of extraneous QTL for WGAIM is higher than CIM2. For small population sizes where it is hard to find QTL, it may be helpful to find a QTL for which the interval is not correct, but which is linked to the correct QTL. This component of the simulation study shows that increased correct detection of coupled QTL by WGAIM comes with an increase in extraneous QTL.

For two QTL in repulsion, a population size of 100 is again too small for reliable QTL detection. WGAIM has rates of correct detection at 0.528 and 0.550 for the two individual QTL, compared to 0.489 and 0.553 for CIM2. Note however that WGAIM correctly identifies both QTL at a rate of 0.433 compared to 0.317 for CIM2. Again the rate of detection of the two QTL for all methods improves as the population size increases. WGAIM is again best overall in terms of correctly finding the QTL.

Table 1 Simulation study Part 1: single chromosome

l	Type I error	FDR
100	0.031	0.037
200	0.026	0.037
400	0.029	0.036

Type I error rates and mean number of extraneous QTL (FDR) for WGAIM

Table 2 Simulation study Part 1: single chromosome and single QTL

Population size	Correct		Linked extraneous	
	WGAIM	CIM1	WGAIM	CIM1
100	0.720	0.701	0.165	0.180
200	0.962	0.958	0.097	0.057
400	0.991	0.995	0.112	0.026

Rate of correct identification of the single QTL and the mean number of linked extraneous QTL found

Table 3 Simulation study Part 1: single chromosome and two QTL in coupling or repulsion

Population Size	Coupling						Repulsion					
	WGAIM			CIM2			WGAIM			CIM2		
100	Not *D*	*D*	Total	Not *D*	*D*	Total	Not *D*	*D*	Total	Not *D*	*D*	Total
Not *D*	0.041	0.250	0.291	0.201	0.271	0.472	0.307	0.117	0.424	0.275	0.236	0.511
D	0.372	0.337	0.709	0.410	0.118	0.528	0.143	0.433	0.576	0.172	0.317	0.489
Total	0.413	0.587	*E*: 0.214	0.611	0.389	*E*: 0.129	0.450	0.550	*E*: 0.237	0.447	0.553	*E*: 0.145
200	Not *D*	*D*	Total	Not *D*	*D*	Total	Not *D*	*D*	Total	Not *D*	*D*	Total
Not *D*	0.002	0.067	0.069	0.028	0.134	0.162	0.017	0.064	0.081	0.034	0.095	0.129
D	0.043	0.888	0.931	0.090	0.748	0.838	0.053	0.866	0.919	0.088	0.783	0.871
Total	0.045	0.955	*E*: 0.153	0.118	0.882	*E*: 0.071	0.070	0.930	*E*: 0.195	0.122	0.878	*E*: 0.112
400	Not *D*	*D*	Total	Not *D*	*D*	Total	Not *D*	*D*	Total	Not *D*	*D*	Total
Not *D*	0.001	0.005	0.006	0.017	0.017	0.034	0.000	0.014	0.014	0.002	0.013	0.015
D	0.028	0.966	0.994	0.061	0.905	0.966	0.024	0.962	0.986	0.025	0.960	0.985
Total	0.029	0.971	*E*: 0.146	0.078	0.922	*E*: 0.102	0.024	0.976	*E*: 0.188	0.027	0.973	*E*: 0.058

Rate of correct identification of the two QTL is presented as a series of two by two tables. Interval I4 is represented on the left and interval I8 at the top of each two by two table. *D* denotes correctly detected or identified and not *D* means not identified. Rate of extraneous linked QTL is presented in the lower right of each two by two table and is prefaced by *E*: for extraneous

The mean number of extraneous QTL selected with the two QTL in repulsion is higher for WGAIM than for CIM2, as for two QTL in coupling.

Overall WGAIM is more powerful than CIM in detecting true QTL, but the method also detects a higher number of extraneous QTL.

Part 2: Type I error rate and FDR

We turn to the more realistic setting of multiple chromosomes that mirrors a typical QTL mapping situation. For the simulation scenario of no QTL, the type I error rates and false discovery rates (FDR) for WGAIM are presented in Table 4. As for the single chromosome situation, the type I error rates are below the nominal 5% level but with the increased number of intervals (and hence levels of the working model random effects) the rate is larger than the single chromosome case given in Table 1 which is very positive. In addition, as the population size increases, the type I error rate tends to the nominal 5% level, as would be expected.

The FDR in Table 4 are very stable across population sizes and again are low.

Part 2: Identification rates for each QTL

The correct detection rates for individual QTL are presented in Table 5. CIM7 performs quite badly for population size 100. This is particularly true for the QTL in coupling. The performance of both methods improves as the population size increases. WGAIM again has the best overall detection rates. One interesting feature is the decrease in the rate of correct detection in comparison with the single chromosome simulations, particularly when compared with CIM2 in the case of QTL in coupling or repulsion.

There are seven QTL in the simulations. The total number of correctly identified QTL for the three methods and three population sizes are also given in Table 5. The results found in our study with regard to CIM are consistent with those found by Broman and Speed (2002). All three methods improve their power as the population size increases. The mean number of correctly identified QTL using WGAIM is always higher than for CIM7, often considerably so. This shows that WGAIM is much more powerful than CIM at finding true QTL.

The rates of detection of extraneous QTL are presented in Table 6. The mean numbers of linked extraneous QTL are presented for each chromosome, together with an overall total. The mean number of unlinked extraneous QTL are presented as a total for chromosomes 6–9.

Table 4 Simulation study Part 2: nine chromosomes

l	Type I error	FDR
100	0.039	0.054
200	0.041	0.052
400	0.046	0.055

Type I error rates and mean number of extraneous QTL (FDR)

Table 5 Simulation study Part 2: nine chromosomes

Rate of correct identification of each QTL and the total mean number out of 7

Size	Method	Interval							Total
		(C1, I4)	(C1, I8)	(C2, I4)	(C2, I8)	(C3, I6)	(C4, I4)	(C5, I1)	
100	WGAIM	0.470	0.445	0.468	0.722	0.679	0.603	0.564	3.951
	CIM7	0.132	0.056	0.208	0.223	0.335	0.241	0.181	1.376
200	WGAIM	0.940	0.952	0.623	0.733	0.790	0.787	0.773	5.598
	CIM7	0.706	0.810	0.234	0.330	0.584	0.555	0.614	3.833
400	WGAIM	0.961	0.968	0.937	0.918	0.980	0.994	0.997	6.755
	CIM7	0.900	0.946	0.919	0.896	0.990	0.989	0.998	6.638

The rates for linked extraneous QTL show that for population size 100, where correct detection is difficult, WGAIM has higher rates than CIM7. For $l = 200$, WGAIM has higher rates, except for chromosome 2 where the two QTL are in coupling. Here CIM7 has much higher rates. The total rates are very similar for both methods.

For a population size $l = 400$, the rates for both methods are very good and are very similar.

The detection of unlinked extraneous QTL is always higher for WGAIM, but decreases as the population size increases. Thus the increased power of detecting true QTL that WGAIM affords, is accompanied by an increase (albeit small) in detection of false QTL.

Part 2: Rates for QTL in repulsion and coupling

Table 7 gives the rates for correct identification of QTL in repulsion and coupling presented as two-way tables.

For the two QTL in repulsion, CIM7 is very poor for population sizes of 100 and 200, where results are much worse than in Table 3. There is considerable improvement for a populations size of 400. In contrast, WGAIM gradually improves detection of both QTL, and has much better rates of detection than CIM across the board, with a marked improvement in performance with a population size of 200.

The patterns are similar for two QTL in coupling. Again CIM7 performs poorly for population sizes of 100 and 200, improving considerably for a size of 400. WGAIM gradually improves detection of both QTL, and has much better rates of detection than CIM7 across the board. This study highlights the difficulty in detecting QTL in coupling for small and moderate populations sizes, even with a more powerful method, namely WGAIM.

Analysis of Sunco–Tasman flour yield experiment

The analysis commences with determining the appropriate non-genetic model for the flour yield data. Smith et al. (2006) present an approach to the analysis of quality trait data from so-called multi-phase experiments. Flour yield is measured in a two-phase (field and milling phases) experiment. Their modelling includes both blocking factors to respect the randomisation processes in the design phases as well as accounting for other sources of non-genetic variation and spatial and/or temporal correlation from the field and laboratory processes. They demonstrate that such modelling significantly increases the response to selection in traditional breeding analysis and suggest that this modelling should also enhance accurate detection of QTL for quality traits. We exclude the QTL terms from the initial modelling to establish an appropriate phenotypic model that can be used as the baseline model in the outlier QTL detection methodology, WGAIM.

Smith et al. (2006) provide details of the analysis of the phenotypic data for the 1999 data. The baseline model included terms for spatial correlation, temporal correlation within mill days, an overall regression on mill order and a random regression for mill days on mill order. A similar approach was used for the 2000 data, where the final model included temporal correlation within mill days, mill days and regressions on mill order and field row. Both models include a polygenic effect for the doubled haploid lines.

Tables 8 and 9 present a summary of the QTL identified for 1999 and 2000, respectively using WGAIM and a 5% threshold for the likelihood ratio test. Using the full non-genetic modelling approach together with the outlier method resulted in 13 and 9 QTL being identified for 1999 and 2000, respectively. Of these a total of 8 could be realistically regarded as the same QTL. The Z statistic reflects the importance of the selected QTL (it is the estimate of the size of the QTL divided by it's standard error) and a value larger than 2 in absolute magnitude is associated with a significant effect in standard statistical analysis.

The estimated polygenic variance for the final models including the QTL were greatly reduced when compared to the baseline models; in fact the detected QTL represent 70

Table 6 Simulation study Part 2: nine chromosomes

Mean number of linked extraneous QTL on chromosomes 1–5 and unlinked extraneous QTL on chromosomes 6–9

Size	Method	Linked						Unlinked C6–C9
		C1	C2	C3	C4	C5	Total	
100	WGAIM	0.227	0.203	0.150	0.100	0.069	0.749	0.199
	CIM7	0.038	0.092	0.040	0.023	0.005	0.198	0.005
200	WGAIM	0.098	0.145	0.105	0.144	0.049	0.541	0.090
	CIM7	0.043	0.258	0.044	0.105	0.018	0.468	0.001
400	WGAIM	0.032	0.029	0.029	0.009	0.008	0.107	0.025
	CIM7	0.038	0.025	0.009	0.009	0.001	0.082	0.003

Table 7 Simulation study Part 2: nine chromosomes and two QTL in repulsion at (C1, I4) and (C1, I8) and coupling at (C2, I4) and (C2, I8)

Rate of identification of the QTL is presented by two by two tables, with I4 on the left and I8 on the top of each two by two table. *D* denotes correctly detected or identified and not *D* means not identified

Population Size	Repulsion				Coupling			
	WGAIM		CIM7		WGAIM		CIM7	
100	Not *D*	*D*	Not *D*	*D*	Not *D*	*D*	Not *D*	*D*
Not *D*	0.418	0.112	0.830	0.038	0.057	0.475	0.574	0.218
D	0.137	0.333	0.114	0.018	0.221	0.247	0.203	0.005
200	Not *D*	*D*	Not *D*	*D*	Not *D*	*D*	Not *D*	*D*
Not *D*	0.017	0.043	0.095	0.199	0.060	0.317	0.468	0.298
D	0.031	0.909	0.095	0.611	0.207	0.416	0.202	0.032
400	Not *D*	*D*	Not *D*	*D*	Not *D*	*D*	Not *D*	*D*
Not *D*	0.015	0.024	0.018	0.082	0.006	0.057	0.007	0.074
D	0.017	0.944	0.036	0.864	0.076	0.861	0.097	0.822

and 85% of the original total genetic component of flour yield found under the baseline model in the 1999 and 2000 trials, respectively.

The impact in the QTL analysis of incorporating the variation due to the environment, and also due to the experimental design and conduct of the experiment was also examined. The outlier QTL procedure was repeated using the simplistic model of independently normally distributed genotype or line effects and phenotypic error, respectively. Thus the field variation and laboratory variation are condensed to a single error variance parameter. The QTL found under this simplistic model are presented in Tables 8 and 9 under the heading "No modelling". There were only 5 and 6 QTL identified at each of the two sites with three in common. These results demonstrate the potential benefits of using the most efficient non-genetic models for these types of data.

The results using the outlier method can also be compared with the results of Lehmensiek et al. (2006). These authors only identified 3 and 4 QTL for 1999 and 2000, respectively, with 3 QTL in common. Importantly, the outlier method identified the same QTL that were identified by Lehmensiek et al. (2006), but in addition detected further QTL, most of which were common across the two years. This is a remarkable result, given that Lehmensiek et al. (2006) used efficient non-genetic models in their analyses. Thus WGAIM exhibits increased power of detection of QTL in comparison to the random effects multi-environment interval mapping method due to Verbyla et al. (2003).

Discussion

The outlier approach presented in this paper provides a mechanism to simultaneously incorporate all intervals on a linkage map into a model and hence into a method for QTL analysis. The model is based on an extension of the simple interval mapping approach of Whittaker et al. (1996) that results in a regression on flanking markers. Our extension of interval mapping introduces the notion of a working model that is used to determine QTL using forward selection. The exact location of a QTL in an interval is not estimated in order to overcome the problem of having a very large number of parameters that would need to be estimated. At the final stage of analysis, when all selected QTL are fitted as fixed effects, the location of a QTL in an interval can be estimated. We have avoided doing so because the point location of a QTL is typically very poorly estimated.

Table 8 QTL summary for 1999 Sunco × Tasman data for QTL using a full non-genetic model (full modelling) and QTL analysis using the outlier method but without modelling non-genetic variation (No modelling)

QTL	Full modelling				No modelling			
	Ci	Ij	Position	*Z* statistic	Ci	Ij	Position	*Z* statistic
1	2B	I5	(51.2, 54.8)	4.63	2B	I5	(51.2, 54.8)	6.68
2	1B	I16	(247.3, 256.7)	–3.46	1B	I14	(90.9, 239.5)	–2.89
3	5A	I13	(90.6, 95.1)	–3.48	5A	I13	(90.6, 95.1)	– 3.13
4	3D	I13	(161.5, 175.3)	3.46				
5	6B	I4	(5.1, 7.2)	–3.37				
6	1D	I5	(36.6, 68.9)	–2.70	1D	I5	(36.6, 68.9)	–2.49
7	6D	I6	(50, 51.2)	–2.71				
8	7D	I4	(94, 98.2)	2.07	7D	I4	(94, 98.2)	1.86
9	1B	I4	(10.7, 11)	2.39				
10	2B	I10	(61.8, 95.9)	2.39				
11	4D	I4	(18.9, 30)	2.46				
12	4A	I14	(154.7, 157.6)	–2.31				
13	5D	I8	(117.9, 132.1)	–2.62				

Ci Ij gives the chromosome and the interval on the chromosome of the QTL, while Position is the distance in cM along the chromosome to the left and right flanking markers for each QTL. The Z-statistic is the estimate of the size of QTL divided by the standard error of the estimate

The effectiveness of the random effects formulation for selecting putative QTL reflects the stability offered by this model, and thus while QTL sizes are shrunk toward zero, these sizes together with the outlier approach presented are able to isolate genuinely important effects very effectively. Rather than requiring a LOD score, our approach uses a sequential test procedure based on the residual likelihood ratio statistic allows a genome-wide assessment of significance. The simulation studies presented show that while a theoretical justification for a critical value for a specified type I error rate is not available, a simple approach of using a 5% test at each stage leads to Type I error rates that are conservative and tend to the nominal 5% as the population sizes increases. The false discovery rate in the case of no QTL is also shown to be very low.

The simulations show that the WGAIM outlier approach detects a much higher number of genuine QTL than CIM and hence is a much more powerful at detecting genuine QTL than CIM. The performance of WGAIM is particularly striking in situations where QTL are in coupling or repulsion, where CIM fails quite badly, even for quite large populations sizes.

A consequence of the increased rate of correct detection of QTL using WGAIM, is the increase in the rate of false positives or FDR compared to CIM. The FDR for WGAIM decreases as the population size increases. However, a higher FDR might be expected for a method that has much

Table 9 QTL summary for 2000 Sunco × Tasman data for QTL using a full non-genetic model (Full modelling) and QTL analysis using the outlier method but without modelling non-genetic variation (No modelling)

QTL	Full modelling				No modelling			
	Ci	Ij	Position	*Z* statistic	Ci	Ij	Position	*Z* statistic
1	2B	I6	(54.8, 59.6)	9.31	2B	I6	(54.8, 59.6)	7.34
2	7D	I3	(86.6, 94)	4.01	7D	I4	(94, 98.2)	4.42
3	4B	I2	(0, 12)	–4.82	4B	I2	(0, 12)	–4.11
4	4D	I2	(0, 1.8)	3.90	4D	I2	(0, 1.8)	3.27
5	1B	I14	(90.9, 239.5)	–4.50				
6	6B	I6	(8.9, 9.4)	–5.45	6B	I6	(8.9, 9.4)	–4.73
7	5A	I14	(95.1, 102.1)	–3.79	5A	I14	(95.1, 102.1)	–3.82
8	1B	I2	(0, 9)	2.57				
9	4A	I3	(10.3, 11.4)	2.49				

Ci Ij gives the chromosome and the interval on the chromosome of the QTL, while Position is the distance in cM along the chromosome to the left and right flanking markers for each QTL. The Z-statistic is the estimate of the size of QTL divided by the standard error of the estimate

greater power of detection of genuine QTL. Thus WGAIM results in a trade-off of detection of real and false QTL, with the trade-off greatly favoring correct detection. Thus we believe the discovery rate of true QTL out-weighs the increase in the rate of false positives when compared to CIM.

A very important aspect of WGAIM, is the reduction in computational complexity. The selection of co-factors, both in terms of number and location is not required as all intervals are included. Obtaining thresholds for CIM is very time-consuming and is not required for WGAIM. The thresholds used in WGAIM are based on asymptotic theory. This might be seen as a negative, but with the complex nature of analysis of crop trials, the fact that this approach works is in fact a bonus. Lastly, the model fitting can be carried out in available software, namely ASReml (Gilmour et al. 2007; Butler et al. 2007).

The regression approach for QTL analysis also allows non-genetic effects to be included in the analysis in a routine fashion. Thus the outlier method for QTL analysis is available in complex experimental situations. The Sunco × Tasman doubled haploid wheat population data from trials in 1999 and 2000 provide an example of the ability of the method to incorporate important sources of non-genetic variation. Flour yield is a complex trait and the outlier approach enabled the determination of many putative QTL. The number both for individual trials and those in common across trials greatly surpasses the results from previous QTL analyses.

The impact of non-genetic variation on the detection of QTL can be substantial. The analysis of flour yield for the Sunco × Tasman population highlighted the improvement possible in incorporating the multi-phase nature of the data generation process. More genuine QTL and fewer extraneous QTL should be the result. Thus the manner in which phenotypic data or trait data are generated must be understood and incorporated in the analysis.

Our approach uses a forward selection strategy. Forward selection does not always produce optimal solutions and we are currently investigating an alternative approach. In addition, methods for more complex situations, such as multi-environment analysis that allows for QTL × environment interactions, multi-trait QTL analysis and the impact of treatments on the size of QTL are being developed.

Acknowledgments We gratefully acknowledge the Grains Research and Development Corporation (GRDC) for support through Key Programme 3 of their National Statistics Project. We thank the Australian Winter Cereals Molecular Marker Program and it's predecessor the National Wheat Molecular Marker Program, both funded by GRDC, for the flour milling yield data analysed in this paper. We are grateful to Simon Diffey, New South Wales Department of Primary Industries, for his excellent implementation of the approach using R and the qtl package. Lastly, we thank the Associate Editor and the referees whose comments have led to substantial improvements and clarifications being incorporated into the paper.

Appendix

The expectation result (12) relies on Haldane's mapping function (4). In terms of recombination frequencies,

$$\theta_{k;j} = \frac{1}{2}(1 - e^{-2d_{k;j}}), \quad 1 - \theta_{k;j} = \frac{1}{2}(1 + e^{-2d_{k;j}}),$$
$$1 - 2\theta_{k;j} = e^{-2d_{k;j}}$$

Thus for example, on substituting for $\theta_{k;j}$ in (7)

$$\lambda_{k;j+1,j}(d_{k;j}) = \frac{1 - 2\theta_{k;j,j+1}}{\theta_{k;j,j+1}(1 - \theta_{k;j,j+1})} \frac{e^{2d_{k;j}} - e^{-2d_{k;j}}}{4}$$

and hence assuming $d_{k;j} \sim U[0,\ d_{k;j,j+1}]$ we find (using x as the dummy variable for integration of the distance)

$$\begin{aligned}
\mathrm{E}(\lambda_{k;j+1,j}) &= \int_0^{d_{k;j,j+1}} \lambda_{k;j+1,j}(x) \frac{1}{d_{k;j,j+1}} dx \\
&= \frac{1 - 2\theta_{k;j,j+1}}{4 d_{k;j,j+1} \theta_{k;j,j+1}(1 - \theta_{k;j,j+1})} \\
&\quad \times \left(\frac{e^{2d_{k;j,j+1}}}{2} + \frac{e^{-2d_{k;j,j+1}}}{2} - 1 \right) \\
&= \frac{1 - 2\theta_{k;j,j+1}}{4 d_{k;j,j+1} \theta_{k;j,j+1}(1 - \theta_{k;j,j+1})} \\
&\quad \times \left(\frac{1}{2(1 - 2\theta_{k;j,j+1})} + \frac{1 - 2\theta_{k;j,j+1}}{2} - 1 \right) \\
&= \frac{\theta_{k;j,j+1}}{2 d_{k;j,j+1}(1 - \theta_{k;j,j+1})}
\end{aligned}$$

as given in (12). The result for $\lambda_{k;j,j}$ follows by symmetry or by repeating the integration process explicitly using (6).

References

Broman KW, Speed TP (2002) A model selection approach for the identification of quantitative trait loci in experimental crosses. J R Stat Soc B64:641–656

Broman KW, Wu H, with ideas from Gary Churchill, Sen S, & contributions from Brian Yandell (2005) qtl: Tools for analyzing QTL experiments. R package version 1.01-9

Butler DG, Cullis BR, Gilmour AR, Gogel BJ (2007) ASReml-R, reference manual. Technical report, Queensland Department of Primary Industries

Cook RD, Holschuh N, Weisberg S (1982) A note on an alternative outlier model. J R Stat Soc B44:370–376

Crainiceanu C, Ruppert D (2004) Likelihood ratio tests in linear mixed models with one variance component. J R Stat Soc B66:165–185

Diggle PJ (1990) Time series analysis: a biostatistical approach. Oxford University Press, Oxford

Eckermann PJ, Verbyla AP, Cullis BR, Thompson R (2001) The analysis of quantitative traits in wheat mapping populations. Aust J Agric Res 52:1195–1206

Foster SD, Verbyla AP, Pitchford WS (2007) Incorporating LASSO effects the linear mixed model for the detection of QTL. J Agric Biol Environ Stat 12:300–314

Gianola D, Perez-Enciso M, Toro MA (2003) On marker-assisted prediction of genetic value: beyond the ridge. Genetics 163:347–365

Gilmour AR, Gogel BJ, Cullis BR, Thompson R (2007) ASReml Users Guide. VSN International Ltd., Release 2.0

Gogel BJ (1997) Spatial analysis of multi-environment variety trials. PhD thesis, Department of Statistics, The University of Adelaide

Gogel BJ, Welham SJ, Verbyla AP, Cullis BR (2001) Outlier detection in linear mixed effects: summary of research. Technical Report P106, The University of Adelaide, Biometrics

Haley CS, Knott SA (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity 69:315–324

Henderson CR (1950) Estimation of genetic parameters (abstract). Ann Math Stat 21:309–310

Jansen RC (1994) Controlling the type I and type II errors in mapping quantitative trait loci. Genetics 138:871–881

Kiiveri HT (2004) A Bayesian approach to variable selection when the number of variables is very large. In: Science and statistics: a Festchrift for Terry speed. Lecture Notes. Institute of Mathematical Statistics, pp 127–144

Lander ES, Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199

Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci USA 84:2363–2367

Lehmensiek A, Eckermann PJ, Verbyla AP, Appels R, Sutherland MW, Daggard GE (2005) Curation of wheat maps to improve map accuracy and QTL detection. Aust J Agric Res 56:1347–1354

Lehmensiek A, Eckermann PJ, Verbyla AP, Appels R, Sutherland MW, Martin D, Daggard GE (2006) Flour yield QTLs in three Australian doubled haploid wheat populations. Aust J Agric Res 57:1115–1122

Martinez O, Curnow RN (1992) Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. Theor Appl Genetics 85:480–488

Martinez O, Curnow RN (1994) Missing markers when estimating quantitative trait loci using regression mapping. Heredity 73:198–206

Moreau L, Monod H, Charcosset A, Gallais A (1999) Marker-assisted selection with spatial analysis of unreplicated field trials. Theor Appl Genetics 98:234–242

Patterson HD, Thompson R (1971) Recovery of interblock information when block sizes are unequal. Biometrika 58:545–554

Piepho H-P (2000) A mixed-model approach to mapping quantitative trait loci in barley on the basis of multiple environment data. Genetics 156:2043–2050

R Development Core Team (2006) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0

Robinson GK (1991) That BLUP is a good thing: the estimation of random effects. Stat Sci 6:15–51

Smith AB, Cullis BR, Appels R, Campbell AW, Cornish GB, Martin D, Allen HM (2001) The statistical analysis of quality traits in plant improvement programs with application to the mapping of milling yield in wheat. Aust J Agric Res 52:1207–1219

Smith AB, Lim P, Cullis BR (2006) The design and analysis of multi-phase quality trait experiments. J Agric Sci (Cambridge) 144:393–409

Stram DO, Lee JW (1994) Variance components testing in the longitudinal mixed effects model. Biometrics 50:1171–1177

Thompson R (1985) A note on restricted maximum likelihood estimation with an alternative outlier model. J R Stat Soc B47:53–55

Thompson R, Cullis B, Smith A, Gilmour A (2003) A sparse implementation of the average information algorithm for factor analytic and reduced rank variance models. Aust N Z J Stat 45:445–459

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc B58:267–288

Trow A (1913) Forms of reproduction: primary and secondary. J Genetics 2:313–324

Verbyla AP, Eckermann PJ, Thompson R, Cullis BR (2003) The analysis of quantitative trait loci in multi-environment trials using a multiplicative mixed model. Aust J Agric Res 54:1395–1408

Welham SJ (2006) Smoothing spline methods within the mixed model framework. PhD thesis, London School of Hygiene and Tropical Medicine, The University of London

Whittaker JC, Thompson R, Visscher PM (1996) On the mapping of QTL by regression of phenotype on marker-type. Heredity 77:23–32

Whittaker JC, Thompson R, Denham MC (2000) Marker-assisted selection using ridge regression. Genet Res Camb 75:249–252

Yi N, George V, Allison DB (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