2012

Multivariate whole genome average interval mapping: QTL analysis for multiple traits and/or environments

Arunas P. Verbyla
University of Adelaide

Brian R. Cullis
University of Wollongong, bcullis@uow.edu.au

Publication Details

Verbyla, A. P. & Cullis, B. R. (2012). Multivariate whole genome average interval mapping: QTL analysis for multiple traits and/or environments. Theoretical and Applied Genetics: international journal of plant breeding research, 125 (5), 933-953.

Multivariate whole genome average interval mapping: QTL analysis for multiple traits and/or environments

Abstract
A major aim in some plant-based studies is the determination of quantitative trait loci (QTL) for multiple traits or across multiple environments. Understanding these QTL by trait or QTL by environment interactions can be of great value to the plant breeder. A whole genome approach for the analysis of QTL is presented for such multivariate applications. The approach is an extension of whole genome average interval mapping in which all intervals on a linkage map are included in the analysis simultaneously. A random effects working model is proposed for the multivariate (trait or environment) QTL effects for each interval, with a variance-covariance matrix linking the variates in a particular interval. The significance of the variance-covariance matrix for the QTL effects is tested and if significant, an outlier detection technique is used to select a putative QTL. This QTL by variate interaction is transferred to the fixed effects. The process is repeated until the variance-covariance matrix for QTL random effects is not significant; at this point all putative QTL have been selected. Unlinked markers can also be included in the analysis. A simulation study was conducted to examine the performance of the approach and demonstrated the multivariate approach results in increased power for detecting QTL in comparison to univariate methods. The approach is illustrated for data arising from experiments involving two doubled haploid populations. The first involves analysis of two wheat traits, a-amylase activity and height, while the second is concerned with a multienvironment trial for extensibility of flour dough. The method provides an approach for multi-trait and multienvironment QTL analysis in the presence of non-genetic sources of variation.

Keywords
whole, environments, traits, multiple, analysis, qtl, mapping, interval, average, genome, multivariate

Disciplines
Physical Sciences and Mathematics

Publication Details
Verbyla, A. P. & Cullis, B. R. (2012). Multivariate whole genome average interval mapping: QTL analysis for multiple traits and/or environments. Theoretical and Applied Genetics: international journal of plant breeding research, 125 (5), 933-953.

Noname manuscript No.
(will be inserted by the editor)

Multivariate whole genome average interval mapping: QTL analysis for multiple traits and/or environments

Arūnas P. Verbyla · Brian R. Cullis

the date of receipt and acceptance should be inserted later

Abstract A whole genome approach for the analysis of quantitative trait loci (QTL) is presented for multivariate situations. These situations include QTL by trait and QTL by environment interactions. The focus is on plant-based studies but the methods can be used in other contexts. The approach is built on an extension of whole genome average interval mapping in which all intervals on a linkage map are included in the analysis simultaneously. A random effects working model is proposed for the multivariate QTL effects for each interval, with a variance-covariance matrix linking the variates in a particular interval. The significance of the variance-covariance matrix for the QTL effects is tested and if significant, an outlier detection technique is used to select a putative QTL. This QTL by variate interaction is transferred to the fixed effects. The process is repeated until the variance-covariance matrix for QTL random effects is not significant. Unlinked markers can also be included in the analysis. The approach is illustrated using a QTL analysis for two wheat traits, α-amylase and height, and also for a multi-environment wheat trial. Both experiments involve a doubled haploid population.

Keywords Factor analytic structure · Multi-environment · Multi-trait · Multivariate · QTL analysis · WGAIM

Introduction

Most research studies in plants involve measurement or scoring of several variables or of a single variable under different conditions or treatments. For example, several traits may be measured or observed in an experiment and QTL by trait interactions may be of interest. Multi-environment trials are common in plant-based studies and QTL by environment interaction is then of interest. The genetic control of these multivariate measurements may involve both common and separate origins, and understanding their nature is important in making genetic progress. This is particularly true for marker assisted selection where co-location or closely linked QTL for several

Arūnas P. Verbyla
School of Agriculture, Food and Wine, The University of Adelaide, PMB 1, Glen Osmond, SA 5064, Australia
E-mail: ari.verbyla@adelaide.edu.au

Arūnas P. Verbyla
Mathematics, Informatics and Statistics and Food Futures National Research Flagship, CSIRO, Urrbrae, SA 5064, Australia

Brian R. Cullis
Biometrics, Industry and Innovation New South Wales, Wagga Wagga, NSW 2650, Australia

traits may inhibit the ability to pyramid such QTL. While the focus in this paper is on plant-based studies, the ideas and methods can be applied in other areas.

In the plant area, past researchers have investigated multivariate methods. For example, Tinker and Mather (1995) consider an approach to multi-environment analysis while Hackett et al (2001) present a review and an interval mapping method based on multivariate regression. Verbyla et al (2003) use a factor analytic model in an interval mapping setting and Vargas et al (2006) consider factorial regression and partial least squares methods, also for multi-environment analysis. More recently, Boer et al (2007) consider QTL analysis for multi-environment trials, using a genome scan approach within a mixed models setting, and move to using environmental variable by QTL interaction terms in an attempt to explain the QTL by environment interactions found. Malosetti et al (2008) consider multi-trait multi-environment analysis. These authors use mixed models, a preliminary genome scan and a backward elimination approach for putative QTL selected using the preliminary genome scan.

Most of the methods proposed both in the univariate and multivariate setting, involve some type of genome scan, often at various levels. Thus, usually a large number of analyses is required. For example, composite interval mapping (Zeng 1994; Jansen 1994) might be utilized in an attempt to allow for background genetic variation. A preliminary scan is required. Subsequently, it may not be clear how many co-factors to use. In addition, these methods also suffer from multiple testing issues and hence the need to use LOD scores.

Verbyla et al (2007) presented a whole genome average interval mapping approach for single trial QTL analysis. This method uses all the intervals on a linkage map simultaneously and avoids the difficult issues regarding repeated genome scans. The approach involves a working model in which every interval is allowed a QTL size that is initially assumed a random effect. The working model provides a mechanism for determining if QTL are present and a stopping rule for the selection process. An outlier detection technique is used in a forward selection process to select the QTL. The method was shown to be much more powerful than composite interval mapping (Zeng 1994; Jansen 1994), although there is a small increase in selecting false positives.

In this paper, an approach is presented for multivariate QTL analysis using a whole genome interval mapping approach. Thus all interval by variate QTL effects are included in the model simultaneously. These multivariate genetic QTL sizes are modelled as random effects with an associated variance-covariance matrix allowing correlation between the variates, be they traits or environments. A likelihood ratio test is described for significance of the QTL variance-covariance matrix. If significant, a multivariate outlier detection method based on a Cholesky decomposition (Golub and van Loan 1996) of the QTL variance-covariance matrix is used to select the most likely interval for a QTL. Multivariate QTL are chosen in a forward selection process and progressively moved to the fixed effects model. The approach allows both genetic and non-genetic effects to be included in the model simultaneously.

Materials

Late Maturity α-amylase in wheat

Late maturity α-amylase (LMA) in wheat is a defect where potentially high levels of the enzyme pl α-amylase accumulates in the ripening grain. The expression of the enzyme and its accumulation in wheat grain has detrimental consequences for processing by end-users to produce value-added wheat products and usually results in downgrading of the grain quality and loss of premiums to farmers. LMA is a difficult trait to phenotype because it is induced by temperature changes. Experiments to investigate LMA are therefore complex and involve multiple phases.

Experimental details

A total of 194 doubled haploid (DH) lines derived from a cross between an advanced breeding line, WW1842, and the line Whistler, were used in the experiment. The phases of the experiment were growth, temperature induction, further growth and assaying seeds using using ELISA plates. A complete account of the experiment can be found in Tan et al (2010) while the design of the experiment is discussed in Butler et al (2009).

The first growth phase was conducted at Cobbity, NSW. Two micro-climate rooms were used, with 220 pots in each room, subdivided into two blocks and two sides within each block of 55 pots, arranged in an 11 $\times$ 5 rectangular array. Randomization of parents and DH lines was restricted; 130 lines had two replicates while 60 lines had three replicates. Lines were assigned to pots so that each room contained either one (160) or two (30) pots of each line, randomized so that each side within each block contained only one pot of each of 55 lines. Four plants were grown in each pot and at anthesis spikes from healthy plants were tagged.

The induction phase was carried out 26-28 days after anthesis. Pots were assigned to induction cohorts (pots within many of the DH lines were induced on different days) for exposure to cool temperatures and were transferred to a cool temperature room. After 8-10 days the plants were returned to their original position in the micro-climate rooms until the plants reached harvest ripeness. This is the second growth phase, and at the commencement of this phase, the height of the plants in a pot was measured.

In the assay phase of the process, the aim was to assay approximately 5 grains from each primary tiller from each of the 4 plants per pot. Tillers from a total of 425 pots were deemed sufficiently healthy to produce a reliable result so that only 1375 tillers (out of a potential 1700) were used. Grain numbers per tiller varied from 1 to 62, with a median of 13. This presented challenges for the allocation of grains to the ELISA plates.

Each ELISA plate had 96 wells arranged in a 12 column $\times$ 8 row array to which seeds were assigned. In the design, the number of seeds per pot varied from 1 to 22, with 25% of pots having less than 19 seeds. With the additional requirement of at least one blank (negative control) per plate, the seeds for assay were ultimately distributed over 75 plates. Plates were grouped into 15 sets of 5, each group of plates being a near complete duplicate of each line, pot and plant.

Using the design, supernatant extract using a single seed (100 ml) was measured into the appropriate position of antibody-coated ELISA plates. The optical densities (OD) were measured at 450 nm with the micro-plate photometer (Multiskan Ascent, Thermo Scientific) and thus OD and height constitute the two traits of interest.

Genetic information

A total of 697 DArT markers (http://www.triticarte.com.au/) and 101 polymorphic microsatellite (SSR) markers were genotyped on the WW1842 x Whistler DH population. A linkage map was constructed using MapManager QTXb20 (Manly et al 2001).

For QTL analysis redundant markers (coincident on the linkage map) were removed and the resultant map had a total of 437 markers (101 SSR and 336 DArT markers). The reduced linkage map was then checked using the qtl package (Broman et al 2009) in the R environment (R Development Core Team 2009) using a combination of double cross-over statistics and likelihood based methods. This resulted in substantial changes to both the order of markers within linkage groups and to the groups themselves. The final map had 31 linkage groups with an additional 14 unlinked markers that were included for analysis. The total length of the linkage map was 3160cM, with an average spacing of 7.5cM between markers.

Dough Rheology

Mann et al (2009) provide details of experiments that were conducted using the Kukri × Janz doubled haploid population in Queensland, Australia. Kukri has unique high dough strength while Janz is considered to have genes for "wide adaptation" and high yield in Australia. This cross is therefore of interest to study the genetic basis of dough rheology, in particular dough strength and extensibility. The aim of the experiments was to examine quality of wheat, from milling to final loaf characteristics.

Experimental details

Field trials were conducted at a number of sites. Two sites, Biloela and Lundavra, in the years 2001 and 2002 form the basis of the analysis presented here. At those two sites two replicates of 156 doubled haploid lines (and the parents and other standard lines) were planted in two replicates of a Latinised row-column design, laid out in a two-dimensional array of up to 10 columns by 36 rows (32 rows at Biloela in 2001). A number of traits were measured in a series of multi-phase experiments. Grain samples were milled using a Buhler mill in an incomplete block design with milling days forming the blocks. For Biloela 2001, the milling consisted of 106 days by 9 samples per day, with each site processed as a separate block of days. For 2002, limited grain yield for the Lundavra site restricted laboratory duplication, so the Biloela and Lundavra plots were randomized within the milling design, with the majority of laboratory duplication coming from the Biloela site. The milling design was again an incomplete block design of 92 milling days by 10 samples per day. The milled grain was processed to obtain dough and the trait examined in this paper, maximum extensibility, were determined from two dough pieces of 150 g each from one mix, using the Extensograph (Brabender Duisburg, Germany).

Genetic information

A genetic linkage map consisting of 246 segregating loci spread over 21 linkage groups (chromosomes) and scored over 172 genetic lines was used in the analysis. The markers were mainly microsatellites analyzed by Syngenta Toulouse (France) and CSIRO Plant Industry (Australia) laboratories. The map was checked in the statistical software package R (R Development Core Team 2009) using the qtl library (Broman et al 2009). The map length was 3400cM with an average spacing of 14cM between markers.

Methods

Overview

Whole genome average interval mapping (WGAIM) was presented by Verbyla et al (2007) for univariate QTL analysis. Linear mixed models formed the underlying basis of analysis. A working model was used in which QTL effects for each interval were assumed to be random effects. Intervals not markers formed the basis of analysis through averaging of the location of putative QTL in each interval. If sufficient variance was associated with these random interval QTL effects (as judged by a likelihood ratio test), at least one putative QTL was assumed to be present. An outlier detection technique was used to select the most likely chromosome and subsequently the most likely interval for a putative QTL. This interval was moved to the fixed effects in a forward selection approach. The process was repeated until the variance attributable to the random QTL size effects was no longer significant.

The multivariate whole genome average interval mapping (MVWGAIM) approach presented in this paper mirrors the univariate method. A working model is proposed for random QTL effects for intervals and incorporates differing variances for the multivariate response, and differing correlations between pairs of variates. An approximate likelihood ratio test is used to determine if the QTL random effects provide sufficient variance-covariance structure to warrant selection of a putative QTL, and an outlier model based on a Cholesky decomposition or a factor analytic approximation allows selection to proceed. As for WGAIM, the forward selection process firstly involves determining the chromosome most likely to contain a QTL and then selecting the interval on that chromosome as most likely to contain the putative QTL. One difference is that the multivariate outlier detection leads to fixed QTL effects for each of the variates. Some of the individual variate QTL effects may not be significant as a multivariate outlier can exist in a dimension lower than the full multivariate dimension. This reflects the possible varying level of common (possibly pleiotropic) effects in the multi-trait situation, and QTL by environment interactions in the multi-environment situation.

Linear mixed model

Mixed models form the basis for analysis. These models are of the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}_0\mathbf{u}_0 + \mathbf{Z}\mathbf{g} + \mathbf{e} \tag{1}$$

where $\mathbf{X}$, $\mathbf{Z}_0$ and $\mathbf{Z}$ are known design matrices for the fixed terms, random terms and genetic effects respectively, $\boldsymbol{\tau}$ is the vector of fixed term parameters, $\mathbf{u}_0$ is a vector of random terms, $\mathbf{g}$ is the vector of genetic effects, and $\mathbf{e}$ is a vector of residual random terms. These latter two terms are assumed independent, mean zero with covariance matrices $\mathbf{G}_0$ and $\mathbf{R}$ respectively. The form of these matrices will depend on the application.

The vector $\mathbf{y}$ consists of data that has multivariate structure. This might simply be multiple traits, responses on multiple treatments or multiple environments, or combinations of these; the term variates is used for the generic multivariate vector. We denote by t the dimension of the underlying multivariate response for a single experimental unit.

The form that $\mathbf{R}$ takes will reflect the nature of the multivariate analysis. Thus for a multi-trait analysis, it will be appropriate to provide for different variances for different traits, and for correlation between traits. In addition, the nature of the experiment will dictate what additional sources of variation are required. For example spatial variation in the field may need to be accounted for in field measurements, and laboratory variation for quality measurements. Both of these components may be required for traits at differing phases in the measurement process as presented by Smith et al (2006).

Estimation of the fixed effects and variance parameters in (1) is based on residual maximum likelihood (REML) as original proposed by Patterson and Thompson (1971). Best linear unbiased prediction (BLUP) is used for the random effects; see Robinson (1991).

Multivariate whole genome average interval mapping

The whole genome model for the genetic effects $\mathbf{g}$ in the multivariate case is given by

$$g_{ij} = \sum_{k=1}^{c} \sum_{l=1}^{r_k - 1} q_{i;kl} a_{j;kl} + p_{ij} \tag{2}$$

where g_{ij} is the genetic effect for line i for variate j, there are c chromosomes, and r_k markers on chromosome k and hence $r_k - 1$ intervals. The total number of markers is $r_\cdot = \sum_k r_k$ and hence the number of intervals is $r_\cdot - c$. The terms $q_{i;kl}$ are the unknown QTL indicators for line i, either -1 or 1 for doubled haploid (DH) or recombinant inbred lines (RIL), depending on the parental origin, for the lth interval on the kth chromosome, while $a_{j;kl}$ is the QTL size for variate j in that interval. The term p_{ij} is a polygenic effect that provides a genetic residual and reflects the possible small contribution of a large number of genes that may impact on the genetic expression of variate j in line i.

We use the regression approach (Haley and Knott 1992; Martinez and Curnow 1992) for QTL mapping and hence for each interval, $q_{i;kl}$ is replaced by its expected value given the two markers defining the interval. If $\mathbf{M}$ is the matrix of marker scores ($n_g \times r_\cdot$) with element $m_{i;kl}$ being the marker score for marker l on chromosome k for line i, using the results of Whittaker et al (1996) we find in a similar manner to Verbyla et al (2007) that

$$g_{ij} = \sum_{k=1}^{c} \sum_{l=1}^{r_k - 1} (m_{i;kl} \lambda_{k;l,l} + m_{i;k,l+1} \lambda_{k;l+1,l}) a_{j;kl} + p_{ij} \tag{3}$$

The terms $\lambda_{k;l,l}$ and $\lambda_{k;l+1,l}$ are functions of the recombination frequency for interval l on chromosome k, denoted by $\theta_{k;l,l+1}$ and the recombination frequency between the left marker defining the interval and the putative QTL, denoted by $\theta_{k;l}$. As in Verbyla et al (2007) there are too many parameters ($\theta_{k;l}$) to estimate and $\lambda_{k;l,l}$ and $\lambda_{k;l+1,l}$ are replaced by their expected value (assuming the distance from the left flanking marker and the putative QTL is uniformly distributed). Thus both are replaced by

$$\lambda_{kl;E} = \frac{\theta_{k;l,l+1}}{2d_{k;l,l+1}(1-\theta_{k;l,l+1})} \tag{4}$$

where

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

is Haldane's distance between markers l and $l+1$ on chromosome k. If we form the matrix of genetic effects $\mathbf{G}$ we can write (3) as

$$\mathbf{G} = \mathbf{M}\mathbf{\Lambda}_E\mathbf{A} + \mathbf{P} \tag{5}$$

where the matrix $\mathbf{\Lambda}_E$ is a block diagonal matrix (with blocks corresponding to chromosomes or linkage blocks), with kth block $\mathbf{\Lambda}_k$ being $r_k \times r_k - 1$. The only non-zero elements of $\mathbf{\Lambda}_k$ are the two central diagonals, and the two values in each column are identical and given by (4). If $\mathbf{M}_E = \mathbf{M}\mathbf{\Lambda}_E$, in vector form (5) is given by

$$\mathbf{g} = (\mathbf{I}_t \otimes \mathbf{M}_E)\mathbf{a} + \mathbf{p} \tag{6}$$

The working model for $\mathbf{a}$ is discussed below.

Unlinked markers

There are situations where it is desirable to include unlinked markers in an analysis. Thus consider the term for a single marker that occupies chromosome or linkage group $c+1$. The QTL on linkage group $c+1$ contributes a term

$$q_{ij;c+1}a_{j;c+1}$$

Given the marker scores $m_{i;c+1}$ on the marker for genotypes i, the regression approach for single marker regression replaces $q_{ij;c+1}$ by

$$\mathrm{E}\left(q_{ij;c+1}|m_{i;c+1}\right) = (1-2\theta_{c+1})m_{i;c+1} \tag{7}$$

where θ_{c+1} is he recombination fraction between the putative QTL and the marker. It is not possible with a single marker to estimate both the size and location (recombination fraction) of a QTL, and in the spirit of Verbyla et al (2007) we integrate out the location or distance using a uniform distribution for the distance between the QTL and the marker. Thus if we use Haldane's distance, we have

$$\theta_{c+1} = \frac{1}{2}(1-e^{-2d_{c+1}})$$

where d_{c+1} is the distance between the marker and the QTL. Notice that

$$1 - 2\theta_{c+1} = e^{-2d_{c+1}} \tag{8}$$

and we assume that d_{c+1} is uniform distributed over the range $(0, \infty)$. This uniform distribution is improper in the Bayesian sense (Gelman et al 2004, page 61). We integrate out d_{c+1} in (7) using (8), namely

$$\int_0^\infty e^{-2x} dx = \frac{1}{2}$$

so that our regression model (7) becomes

$$\mathrm{E}\left(q_{ij;c+1} | m_{i;c+1}\right) = \frac{1}{2} m_{i;c+1}$$

Thus the uncertainty in the position of the QTL down-weights the marker by a value of 2. Thus WGAIM with unlinked markers proceeds using the marker scores divided by two.

Polygenic effects

A natural model for the polygenic effects $\mathbf{p}$ is $\mathbf{p} \sim N(\mathbf{0}, \mathbf{G}_p \otimes \mathbf{I}_{n_g})$. The matrix $\mathbf{G}_p$ may be an unstructured (and hence fully parameterized) $t \times t$ variance-covariance matrix, or it may take on another form. This model also assumes that the genotypes are uncorrelated and therefore unrelated. If a pedigree is available, relationship matrices can also be included as in Oakey et al (2006) and Oakey et al (2007).

Factor analytic models have been used for the analysis of multi-environment trials (Smith et al 2005) and can provide a very good and numerically stable approximation to the unstructured model. Thus we may consider the model

$$\mathbf{p} = (\mathbf{\Lambda}_p \otimes \mathbf{I}_{n_g})\mathbf{f}_p + \boldsymbol{\xi}_p \tag{9}$$

where $\mathbf{f}_p$ is the $n_f n_g \times 1$ vector of n_f factor effects for each of the n_g genetic lines, with distribution $\mathbf{f}_p \sim N(\mathbf{0}, \mathbf{I}_{n_f n_g})$, $\boldsymbol{\xi}_p$ is a residual random vector for genetic variation not explained by the factors and it is assumed $\boldsymbol{\xi}_p \sim N(\mathbf{0}, \mathbf{\Psi}_p \otimes \mathbf{I}_{n_g})$; $\mathbf{\Psi}_p$ is a $t \times t$ diagonal matrix. In addition $\mathbf{f}_p$ and $\boldsymbol{\xi}_p$ are assumed to be independent. The matrix $\mathbf{\Lambda}_p$ consists of loadings for each environment for each factor and indicate the relative expression of the underlying factor for each environment.

Under the FA model the variance matrix $\mathbf{G}_p$ takes on the form

$$\mathbf{G}_p = \mathbf{\Lambda}_p \mathbf{\Lambda}_p^T + \mathbf{\Psi}_p$$

Working model for $\mathbf{a}$

The working model is a natural extension of the univariate specification given by Verbyla et al (2007). Thus the QTL sizes are assumed $\mathbf{a} \sim N(\mathbf{0},\ \mathbf{G}_a \otimes I_{r.-c})$, where $\mathbf{G}_a$ is a $t \times t$ variance-covariance matrix, allowing for the t variates.

Stopping rule

A multivariate QTL exists if $\mathbf{G}_a \neq \mathbf{0}$. Thus we would like test the hypothesis $H_0 : \mathbf{G}_a = \mathbf{0}$ to establish if a QTL exists and if the test is retained, the selection process concludes. If the test is rejected, there is evidence that at least one putative QTL exists and a process is used to select the most likely interval for a putative QTL.

The test of $H_0 : \mathbf{G}_a = \mathbf{0}$ is non-standard, just as in the univariate case. From a practical point of view there are major difficulties with such a test. If a variance of a variate is zero, covariances or correlations with other variates are not defined. To overcome this problem, the approach taken is to initially fit a model with only variances for the multivariate sizes and test the significance of this so-called diagonal variance matrix. This establishes the presence of variation that is necessary for a QTL to exist. Once established, a correlated model is fitted.

There is one additional complication. Fitting a correlated polygenic effect with a diagonal working model for the random effect sizes distorts the null distribution of the test statistic. Thus at the stage of testing for putative QTL, the polygenic effects are also fitted using a diagonal variance model. In this case, if $\hat{\ell}$ is the maximized residual log-likelihood including the diagonal variance model for putative QTL and $\hat{\ell}_0$ is the maximized residual log-likelihood omitting the diagonal variance model, the likelihood ratio test statistic is found by

$$X^2_{LR} = 2(\hat{\ell} - \hat{\ell}_0) \tag{10}$$

and this statistic has an approximate distribution under the null hypothesis (zero diagonal variance matrix) that is a mixture of chi-squared distributions. In fact the mixture consists of chi-squared distributions from zero to t degrees of freedom with approximate null distribution given by

$$X^2_{LR} \sim \left(\frac{1}{2}\right)^t \sum_{k=0}^{t} \binom{t}{k} \chi^2_k \tag{11}$$

where χ^2_k represents a chi-square distribution on k degrees of freedom. This choice is investigated in a small simulation study.

Outlier detection and selection of QTL

As in Verbyla et al (2007) we use an outlier detection approach to select putative QTL. The alternative outlier model (AOM) is used again, but on a transformed scale. Thus we consider the Cholesky decomposition of $\mathbf{G}_a$, namely $\mathbf{G}_a = \mathbf{L}_a\mathbf{L}_a^T$ where $\mathbf{L}_a$ is a lower triangular matrix (that is having all elements above the diagonal equal to zero). Then we can write

$$\mathbf{a} = (\mathbf{L}_a \otimes \mathbf{I}_{r.-c})\mathbf{f}_a$$

where $\mathbf{f}_a \sim N(\mathbf{0},\ \mathbf{I}_t \otimes \mathbf{I}_{r.-c})$ are independent standard normal variates. The outlier model is based on $\mathbf{f}_a$ and hence matches the strategy used in the univariate case by Verbyla et al (2007).

The AOM is used for chromosomes in the first instance, and then for intervals within the selected chromosome. Thus the effects $\mathbf{f}_a$ are modified for chromosome k by (if intervals are nested within variates)

$$\mathbf{f}_{ak} = \mathbf{f}_a + (\mathbf{I}_t \otimes \mathbf{D}_k)\boldsymbol{\delta}_k \tag{12}$$

where $\boldsymbol{\delta}_k \sim N(\mathbf{0}, \sigma^2_{ak}\mathbf{I}_t \otimes \mathbf{I}_{r_k-1})$ is a vector of departures for chromosome k. This outlier model provides for variance inflation for chromosome k via the variance component σ^2_{ak}. The matrix $\mathbf{D}_k$ is a $(r_\cdot - c) \times (r_k - 1)$ matrix with an identity matrix for the $r_k - 1$ intervals on chromosome k and zeros elsewhere.

The AOM given by (12) results in QTL sizes

$$\mathbf{a}_k = (\mathbf{L}_a \otimes \mathbf{I}_{r_\cdot - c})\mathbf{f}_{ak}$$

with a modified variance-covariance matrix of $(1 + \sigma^2_{ak})(\mathbf{G}_a \otimes \mathbf{I}_{r_k-1})$ for chromosome k. Thus the AOM allows for a rescaling of the underlying variate QTL size variance-covariance matrix and chromosomes indicating such an inflation are flagged as possibly containing a QTL.

Denoting the elements of $\mathbf{f}_a$ by $f_{a,jkl}$, and their best linear unbiased predictions by $\tilde{f}_{a,jkl}$, it is shown in APPENDIX A that the score statistic is a function of

$$t^2_k = \frac{\sum_{j=1}^{t}\sum_{l=1}^{r_k-1}\tilde{f}^2_{a,jkl}}{\sum_{j=1}^{t}\sum_{l=1}^{r_k-1}\mathrm{var}\left(\tilde{f}_{a,jkl}\right)} \tag{13}$$

This statistic is used to rank the chromosomes in terms of their outlier status. The largest statistic indicates the biggest departure from the null hypothesis ($\sigma^2_{ak} = 0$) and hence this chromosome is selected as being most likely to contain a putative QTL.

The same argument within the selected chromosome can be used to select the most likely interval using the statistic

$$t^2_{kl} = \frac{\sum_{j=1}^{t}\tilde{f}^2_{a,jkl}}{\sum_{j=1}^{t}\mathrm{var}\left(\tilde{f}_{a,jkl}\right)}$$

The selected interval is placed in the fixed effects part of the model as an interaction between the interval and the factor defining the variates. The process of selection continues until the stopping rule is invoked.

Lastly, it is possible to use a factor analytic approximation in deriving the statistics for QTL selection for larger problems. Details are given in APPENDIX B.

Final assessment of significance of QTL

Single multivariate QTL are determined using the above process using forward selection. Selected QTL are fitted as fixed effects as they are chosen. The model fitted depends on the context. For example, in a multi-trait situation, the trait by QTL interval is fitted in the fixed effects. In a multi-environment or multi-treatment setting, a main effect for the QTL interval is fitted together with the interaction between the environment or treatment and the QTL interval. This is because the measurement is the same across environments and treatments and a

common QTL is a sensible outcome. The final output of an analysis will therefore depend on the situation but will consist of individual Wald statistics of the appropriate effects, be they main effects or interactions.

Computation

All computations were performed in R (R Development Core Team 2009) using the asreml (Butler et al 2007), with multivariate QTL analysis using components of the wgaim package (Taylor et al 2009).

Results

Simulation study

A small simulation study was conducted to examine the null distribution for the stopping rule (the test of the diagonal variance matrix being zero) using (10) and (11). data was simulated for population sizes (n_p) of 100, 200 and 400 with 2 replicates and 4 environments. The null model (no QTL) was given by ($i = 1, 2, \ldots, n_p$; $j = 1, 2, 3, 4$; $k = 1, 2$)

$$y_{ijk} = \mu_j + u_{pij} + e_{ijk}$$

where $\mu_j = 10$, the polygenic effects u_{pj} were simulated using the covariance matrix

$$\mathbf{G}_p = \begin{bmatrix} 1.0 & 0.9 & 0.7 & 0.5 \\ 0.9 & 1.0 & 0.7 & 0.3 \\ 0.7 & 0.7 & 1.0 & 0.5 \\ 0.5 & 0.3 & 0.5 & 1.0 \end{bmatrix}$$

for the 4 traits for each line and independent standard normal errors e_{ijk}. Note however, that only a diagonal variance matrix is fitted in examining the distribution of the residual likelihood ratio statistic. Two thousand (2000) simulations were run for each population size.

The linkage map for each population size was generated as outlined in Verbyla et al (2007) and consisted of 9 linkage groups of 11 markers, with an original spacing of 10 cM (the distances used in the simulation were estimated using the simulated marker data on the individuals).

Percentage points for the three population sizes are given in Table 1. For all population sizes the estimated percentage points are less than or close to the corresponding nominal points, suggesting the mixture of chi-squared distributions is a good approximation to the distribution under the null hypothesis.

Late Maturity α-amylase

The optical density (OD) and height data were firstly analyzed separately and subsequently together in a bivariate analysis. In all analyses the non-doubled haploid lines (the parental lines, Spica and the negative control) were

Table 1 Estimated proportion of values out of 2000 simulations of the residual likelihood ratio statistic exceeding the nominal critical value for levels 0.10, 0.05 and 0.01 for population sizes 100, 200 and 400. The simulation involved 4 variates.

	Critical	Population Size		
Probability	Value	100	200	400
0.10	4.96	0.079	0.082	0.070
0.05	6.50	0.046	0.048	0.034
0.01	10.02	0.019	0.011	0.009

omitted from the analysis. These lines are not of direct interest and in particular showed extremes in OD that would have biased the results.

Height

For height, the baseline mixed model without QTL effects was given symbolically by

$$\mathsf{ht} = 1 + \mathbf{id} + \mathbf{Room} + \mathbf{Room.Block} + \mathbf{Room.Block.Side} + \mathbf{error} \tag{14}$$

reflecting the nested structure of the glasshouse experiment. This is a simple variance components model in which random effects are presented in bold. The terms in the model have their obvious meaning. The 1 represents the constant or mean height. The design or blocking factors in the glasshouse were **Room**, **Block** which is nested in room and **Side** which is nested in block within a room; all factors have 2 levels. The nesting of effects introduces the terms like **Room.Block**. The genetic effects are given by the factor **id** and had 194 levels. The **error** was assumed independent and identically distributed.

Using (14) as the baseline model, QTL effects were found using asreml in R and are presented in Table 2. Four significant QTL were found for height; the interval on chromosome 4D is close to the height gene *Rht-D1*, while the interval on 4B is consistent with the *Rht-B1* gene in wheat.

Table 2 Height QTL for the LMA glasshouse experiment

	Left		Right			Wald	
Chromosome	Marker	dist(cM)	Marker	dist(cM)	Size	Statistic	P-value
3D	gwm3	105.0	wPt-3863	118.4	-1.888	-3.44	$< .01$
4B	wPt-3908	90.7	wPt-6149	94.1	-3.803	-7.18	$< .01$
4D	wPt-0472	36.5	wPt-0710	38.7	6.761	12.59	$< .01$
6D	wPt-4830	0.0	barc146	3.4	1.435	2.66	0.01

The estimates of non-genetic effects were very similar under both the base and QTL models. The polygenic component was considerably reduced after QTL were fitted and it was found that 64% of the original polygenic variance was explained by the selected QTL.

Optical density or LMA

For LMA, the baseline model was given by

$$\begin{aligned} \mathsf{tod} = 1 + \mathbf{id} + \mathbf{Room} + \mathbf{Room.Block} + \mathbf{Room.Block.Side} + \\ \mathbf{Pot} + \mathbf{Pot.Tiller} + \mathbf{Induction} + \mathbf{GoSlide} + \\ \mathbf{Slide} + \mathbf{Slide.SlideRow} + \mathbf{Slide.SlideCol} + \mathbf{error} \end{aligned} \tag{15}$$

where tod is the transformed optical density ($-1/od^3$) which was used because of the highly skewed nature of the optical density *od*. All but the constant term 1 are random effects. Determination of OD is a multi-phase process. Thus in addition to the variation through room, block and side, there is possible variation between **Pot**s and tillers within pots, **Pot.Tiller**, through **Induction** group, and finally slide variation through groups of slides (**GoSlide**), **Slide**s and variation as specified by rows and columns within slide (**Slide.SlideRow** and **Slide.SlideCol**). The **error** was assumed independent and identically distributed.

Twelve putative QTL were found using (15) as the baseline model. These are given in Table 3. Interestingly, the height QTL on 4B (in an adjacent interval) and 4D are also found for LMA, but there are many QTL specific to LMA.

Table 3 LMA QTL for the LMA glasshouse experiment on the optical density (OD) transformed scale

	Left		Right			Wald	
Chromosome	Marker	dist(cM)	Marker	dist(cM)	Size	Statistic	P-value
2D	Barc159	0.0	gwm320	8.7	-0.026	-2.04	0.04
3A	rPt-9057	30.5	wPt-4077	44.6	0.05	3.85	$< .01$
3B	barc147	124.3	rPt-7228	131.1	-0.046	-3.71	$< .01$
3D	wPt-0732	2.3	wPt-6262	20.5	0.055	4.09	$< .01$
4B	barc020	99.9	gwm113	106.0	-0.084	-6.84	$< .01$
4D	wPt-0472	36.5	wPt-0710	38.7	0.084	4.07	$< .01$
4D	wPt-0710	38.7	wPt-4572	98.1	0.077	2.72	$< .01$
5BL	wPt-4791	55.0	barc232	56.9	0.027	2.18	0.03
5D	barc143	69.1	wPt-6225	70.6	-0.035	-2.85	$< .01$
6A	gwm427	0.0	wPt-2880	34.5	-0.023	-1.59	0.11
Unlinked	gwm301				-0.077	-3.12	$< .01$
Unlinked	wPt-0877				-0.066	-2.31	0.02

For optical density, 63% of the polygenic variance was explained by the selected QTL.

Joint analysis of Height and LMA

The joint analysis of height and transformed optical density is conducted using the models (14) and (15). In addition, correlation between height and transformed optical density is included in the model for genetic effects and pot effects. The QTL explained 67% of the polygenic variance for both traits.

Sixteen QTL were found in the joint QTL analysis and are presented in Table 4. Two were not significant when examined as fixed effects, even though they were selected. There were 5 QTL common to both height and LMA, and in addition 2 QTL were specific to height and 5 QTL were specific to optical density or LMA.

Table 4 Multivariate (height (ht) and transformed LMA (od)) QTL results for the LMA glasshouse experiment. Terms in bold are significant using a Wald test.

Chromosome	Left	dist(cM)	Right	dist(cM)	Trait	Size	Wald	P-value
1D	wPt-1799	0.0	wPt-0459	13.8	ht	-0.36	-0.65	0.52
					od	2.94	2.34	**0.02**
3B	barc147	124.3	rPt-7228	131.1	**ht**	-1.14	-2.08	**0.04**
					od	5.07	4.13	< .01
3A	wPt-2910	15.2	wPt-4692	15.8	ht	-0.68	-1.32	0.19
					od	-4.59	-3.96	< .01
3D	gwm341	0.0	wPt-0732	2.3	ht	0.02	0.04	0.97
					od	-5.11	-4.39	< .01
3D	gwm3	105.0	wPt-3863	118.4	**ht**	-2.07	-3.74	< .01
					od	0.97	0.78	0.44
4B	wPt-6149	94.1	barc020	99.9	**ht**	-3.84	-7.20	< .01
					od	7.69	6.39	< .01
4D	wPt-0472	36.5	wPt-0710	38.7	**ht**	6.67	11.40	< .01
					od	-12.24	-9.24	< .01
5B	wPt-8604	7.1	wPt-9724	8.2	ht	-0.26	-0.47	0.64
					od	-4.55	-3.72	< .01
5B	wPt-1250	76.8	wPt-1548	77.4	**ht**	-1.13	-2.03	**0.04**
					od	4.43	3.56	< .01
5BL	wPt-4791	55.0	barc232	56.9	ht	0.38	0.70	0.48
					od	-3.60	-2.95	< .01
5D	barc143	69.1	wPt-6225	70.6	ht	0.31	0.57	0.57
					od	4.22	3.51	< .01
6D	wPt-4830	0.0	barc146	3.4	**ht**	1.33	2.46	**0.01**
					od	-1.94	-1.61	0.11
7A	wPt-0556	121.0	rPt-6430	124.1	**ht**	-2.11	-3.81	< .01
					od	3.64	2.93	< .01
Unlinked	gwm301				ht	-0.88	-0.83	0.40
					od	6.80	2.87	< .01
Unlinked	wPt-0877				ht	0.60	0.48	0.63
					od	1.85	0.66	0.51
Unlinked	stm5tcacI				ht	1.72	1.64	0.10
					od	-2.08	-0.88	0.38

The QTL selected using separate analyses for height and optical density and those selected using a joint analysis are presented in Table 5. Tick marks indicate selected QTL. There is some consistency between the two sets of analyses, with 9 of the 18 QTL being (much) the same, with an additional optical density QTL being consistent. The selected intervals are not always the same; often adjacent intervals are selected in the individual analyses while the joint analysis provides a selection that takes into account both traits simultaneously.

Dough Rheology

Individual site QTL analyses

Individual site analyses are conducted before the joint multi-environment analysis that allows QTL × environment interactions to be assessed.

For the single site analyses the baseline model was of the form

Table 5 Summary of the selected QTL intervals for the models for each of height and transformed optical density and the joint analysis. The left hand marker for the interval selected in the joint analysis is given as an indicator. One interval was selected (S) in a univariate analysis but was not selected in the joint analysis. The label † indicates the interval selected in the univariate did not match the selected interval from the joint analysis but was on the same chromosome. The last unlinked effect was selected (S) in the joint analysis but none of the effects was significant using a Wald test.

QTL		Univariate		Joint	
Chromosome	Left Marker	ht	od	ht	od
1D	wPt-1799				✔
2D	barc159		✔		
3A	wPt-2910		✔†		✔
3B	barc147		✔	✔	✔
3D	gwm341		✔†		✔
3D	gwm3	✔		✔	
4B	wPt-6149	✔†	✔†	✔	✔
4D	wPt-0472	✔	✔†	✔	✔
5B	wPt-8604				✔
5B	wPt-1250			✔	✔
5B	wPt-4791		✔		✔
5D	barc143		✔		✔
6A	gwm427		S		
6D	wPt-4830	✔		✔	
7A	wPt-0556			✔	✔
Unlinked	gwm301		✔		✔
Unlinked	wPt-0877		✔		
Unlinked	stm5cacI			S	S

$$\mathsf{ext} = \mathsf{Type} + \mathbf{id} + \mathbf{Rep} + \mathbf{Rep.Block} + \mathbf{Column.Row} + \mathbf{MeasDay} + \mathbf{Labno} + \mathbf{error}$$

where ext the maximum extensibility, Type is a factor with level DH for doubled haploid lines, and other lines having their own level, Rep and Block are design factors reflecting the blocking structure at the field level, Column.Row is a field plot effect, and MeasDay and Labno are factors for the measurement day and laboratory samples; in the latter case there were duplicate measures of extensibility for some samples. The factor id is the genetic effect due to doubled haploids. The error was adequately modelled using a constant variance.

Having fitted a baseline model for each of the 4 sites, QTL analysis was conducted using wgaim in the R environment. Details are omitted, but the percentage of polygenic variance explained by the QTL found for each site was around 45%.

Multi-environment QTL analysis

The model for the multi-environment analysis of extensibility was given by

$$\begin{aligned}\mathsf{ext} = \mathsf{Expt}*\mathsf{Type} &+ \mathbf{fa(Expt,1).id} + \mathbf{diag(Expt).Rep} + \mathbf{diag(Expt).Rep.Block} + \\ &\mathbf{diag(Expt).Column.Row} + \mathbf{diag(Labsection).MeasDay} + \\ &\mathbf{diag(Labsection).Labno} + \mathbf{diag(Labsection):error}\end{aligned}$$

where the additional components in the joint analysis involve the site or environment factor Expt. The model is very similar to the single site form, with interactions or crossing with Expt. The genetic doubled haploid lines are correlated across environments using a first order Factor analytic model (fa(Expt,1).id) as discussed by Smith et al (2001). Other effects allow for separate (using a diagonal variance structure) variance components for field

and laboratory effects; the material was tested in three sections which constitute the levels of the Labsection factor.

The percentage of variance explained by the QTL by environment effects ranged from 40% to 51%.

The selected QTL are given in Table 6. For this QTL by environment analysis it is possible to test for a common effect across environments for each QTL. These Wald tests are presented in Table 7 where the interval effects are represented symbolically; for example, X.11.4 is interval 4 on chromosome 1A, as listed in Table 6. There are two QTL that were significant across all sites and provide a common contribution at those sites; one involves the glutenin *GluD1*. The other QTL selected have varying levels of expression across sites, from 1 to 3 sites showing significant association. This highlights the QTL by environment interaction for extensibility.

Table 6 Multivariate QTL for the extensibility: Two common QTL across environments and 5 multi-environment by QTL interactions. Terms in bold are significant using a Wald test.

	Left		Right					Wald	
Chr	Marker	dist(cM)	Marker	dist(cM)	Year	Location	Size	Statistic	P-value
1A	cfd021a	8.4	NW2343	11.4	**2001**	**Biloela**	-0.549	-4.16	< .01
					2001	Lundavra	-0.149	-1.17	0.24
					2002	Biloela	-0.175	-1.58	0.11
					2002	Lundavra	-0.145	-0.96	0.33
1B	NW1355	68.6	Bx7	73.9	**2001**	**Biloela**	0.710	4.89	< .01
					2001	Lundavra	0.129	0.92	0.36
					2002	Biloela	0.236	1.88	0.06
					2002	Lundavra	0.370	2.31	0.02
1B	gwm191.1	86.9	NW1103	227.5	**all**	**both**	0.541	2.29	**0.02**
1D	GluD1	83.9	cfd048	85.9	**2001**	**Biloela**	-0.383	-2.97	< .01
					2001	**Lundavra**	-1.051	-8.33	< .01
					2002	**Biloela**	-0.719	-6.58	< .01
					2002	**Lundavra**	-0.969	-6.55	< .01
4D	Rht2	0.0	cfd071	21.2	2001	Biloela	0.029	0.15	0.84
					2001	**Lundavra**	-0.470	-3.33	< .01
					2002	**Biloela**	-0.552	-4.51	< .01
					2002	Lundavra	-0.163	-0.99	0.32
4D	NW2208	132.1	gwm609	142.6	**all**	**both**	-0.328	-3.46	< .01
7A	NW2062	232.5	FC1	250.9	**2001**	**Biloela**	0.659	4.47	< .01
					2001	Lundavra	0.166	1.16	0.25
					2002	**Biloela**	0.271	2.17	**0.03**
					2002	Lundavra	0.304	1.84	0.07

Table 7 Wald Statistics for Environment by QTL interactions. Conditional F statistics were calculated (labelled F) with denominator degrees of freedom (den df) estimated using Kenward and Roger (1997). The P-values suggest two QTL do not interact with the environment.

Term	df	den df	F	P-value
Expt:X.1D.6	3.00	177.3	7.19	< .01
Expt:X.4D.1	3.00	178.8	5.37	< .01
Expt:X.4D.7	3.00	176.4	0.55	0.65
Expt:X.1B.16	3.00	178.9	1.69	0.17
Expt:X.7A.9	3.00	176.8	2.74	0.05
Expt:X.1A.4	3.00	180.6	3.09	0.03
Expt:X.1B.13	3.00	179.3	2.58	0.06

The consistency of QTL between individual and multi-site analyses is presented in Table 8. Five QTL are consistent whereas 7 are not. The multi-environment analysis leads to different QTL being selected from the univariate analyses because evidence for QTL is accumulated across environments.

Table 8 Summary of the selected QTL intervals for the models for each site both from individual site analyses and the multi-environment analysis. The left hand marker for the interval selected in the joint analysis is given as an indicator. Two intervals were selected in univariate analyses but were not selected in the joint analysis. The label † indicates the interval selected did not match the selected interval from the joint analysis.

QTL		Single Site Analysis				Multi-site analysis			
Chromosome	Left Marker	01B	01L	02B	02L	01B	01L	02B	02L
1A	cfd021a	✔				✔			
1B	NW1355	✔†			✔	✔			✔
1B	gwm191.1			✔	✔	✔	✔	✔	
1D	GluD1	✔†	✔†	✔	✔	✔	✔	✔	✔
4A	NW2277				✔				
4B	Rht1	✔							
4D	Rht2						✔	✔	
4D	NW2208	✔†	✔†	✔		✔	✔	✔	✔
5D	cfd018	✔	✔						
5D	barc110	✔							
7A	NW2062		✔†			✔		✔	✔
7B	barc065				✔				

Discussion

The approach outlined in this paper is a natural extension of whole genome average interval mapping to the multivariate situation, and hence carries forward the advantages outlined in Verbyla et al (2007). A working model, stopping rule and outlier detection technique characterize the forward selection procedure. Where relevant, common QTL can be determined across the multivariate specification at the final stage of analysis.

Multivariate methods offer increased power for detection of QTL through correlation, and also allow common QTL to be determined directly (for example in QTL by environment or QTL by trait settings) rather than indirectly using univariate analyses. The forward selection nature of the process provides a simple method for detection but can result in models with non-significant QTL (in the conventional sense of a test of fixed effects). Methods to overcome this problem are the subject of current research.

A key aspect of the approach is the ability to include and hence allow for non-genetic sources of variation in the analysis, thereby reducing the likelihood of false positives due to omission of such effects. The examples illustrate the incorporation of such effects. Thus the methods discussed in this paper provide a comprehensive approach to multivariate QTL analysis.

Acknowledgements The authors gratefully acknowledge the financial support of the Grains Research and Development Corporation (GRDC) through the Statistics for the Australian Grains Industry (SAGI) project.

Appendix A: The score statistic under the alternative outlier model

The full (marginal) model for $\mathbf{y}$ under the AOM (12) can be written as

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\tau},\ \mathbf{H}) \tag{16}$$

where if $\mathbf{M}_{E,k} = \mathbf{M}_E\mathbf{D}_k$, the selected columns of $\mathbf{M}_E$ corresponding to the kth chromosome, the variance matrix $\mathbf{H}$ is given by

$$\mathbf{H} = \mathbf{R} + \mathbf{Z}_0\mathbf{G}_0\mathbf{Z}_0^T + \mathbf{Z}\left\{(\mathbf{L}_a\mathbf{L}_a^T) \otimes \mathbf{M}_E\mathbf{M}_E^T\right\}\mathbf{Z}^T + \mathbf{Z}\left\{(\sigma^2_{ak}\mathbf{L}_a\mathbf{L}_a^T) \otimes \mathbf{M}_{E,k}\mathbf{M}_{E,k}^T\right\}\mathbf{Z}^T$$

As in Verbyla et al (2007), an outlier statistic is developed using the score for σ^2_{ak} under the null hypothesis $H_0:\ \sigma^2_{ak} = 0$. If $\mathbf{P} = \mathbf{H}^{-1} - \mathbf{H}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{H}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{H}^{-1}$, the REML score for σ^2_{ak} evaluated at zero is

$$U_k(0) = -\frac{1}{2}\left\{\mathrm{tr}\left(\mathbf{PZ}(\mathbf{L}_a\mathbf{L}_a^T \otimes \mathbf{M}_{E,k}\mathbf{M}_{E,k}^T)\mathbf{Z}^T\right) - \mathbf{y}^T\mathbf{PZ}(\mathbf{L}_a\mathbf{L}_a^T \otimes \mathbf{M}_{E,k}\mathbf{M}_{E,k}^T)\mathbf{Z}^T\mathbf{Py}\right\} \tag{17}$$

Noting the the BLUP for $\mathbf{f}_a$ is

$$\tilde{\mathbf{f}}_a = (\mathbf{L}_a \otimes \mathbf{M}_{E,k})^T\mathbf{Z}^T\mathbf{Py}$$

and denoting the elements of $\tilde{\mathbf{f}}_a$ by $\tilde{f}_{a,jkl}$ for the BLUP for interval l on chromosome k for variate j, the score can be written as

$$\begin{aligned} U_k(0) &= -\frac{1}{2}\sum_{j=1}^{t}\sum_{l=1}^{r_k-1}\left\{\mathrm{var}\left(\tilde{f}_{a,jkl}\right) - \tilde{f}^2_{a,jkl}\right\} \\ &= \frac{1}{2}\left(\sum_{j=1}^{t}\sum_{l=1}^{r_k-1}\mathrm{var}\left(\tilde{f}_{a,jkl}\right)\right)(t_k^2 - 1) \end{aligned}$$

where t_k^2 is given by (13). Thus t_k^2 indicates the departure from $U_k(0) = 0$. This statistic therefore provides evidence that σ^2_{ak} departs from zero for chromosome k. The chromosome most likely to contain a QTL is the one with largest t_k^2.

Appendix B: Using a Factor Analytic approximation in the decomposition of the QTL genetic covariance matrix

The Cholesky decomposition of $\mathbf{G}_a$ requires many parameters for larger multivariate problems and an approximation becomes both sensible and necessary. It is possible to use a factor analytic approximation for the full covariance model $\mathbf{G}_a$ which mirrors the use of FA models in the analysis of multi-environment trials. Thus we suppose

$$\mathbf{G}_a = \mathbf{\Lambda}_a\mathbf{\Lambda}_a^T + \mathbf{\Psi}_a$$

which arises from a model for $\mathbf{a}$ of the form

$$\mathbf{a} = (\mathbf{\Lambda}_a \otimes I_{r.-c})\mathbf{f} + \boldsymbol{\xi}$$

where $\mathbf{f} \sim N(\mathbf{0},\ \mathbf{I}_{n_f(r.-c)})$, n_f is the number of factors, and residual term is such that $\boldsymbol{\xi} \sim N(\mathbf{0},\ \mathbf{\Psi} \otimes \mathbf{I}_{r.-c})$. The matrix $\mathbf{\Psi}$ is a $t \times t$ diagonal matrix with elements ψ_j.

Now $\mathbf{G}_a$ can be written as

$$\mathbf{G}_a = \mathbf{L}_a \mathbf{L}_a^T \qquad \mathbf{L}_a = \left[\mathbf{\Lambda}_a \ \mathbf{\Psi}_a^{1/2} \right]$$

where $\mathbf{\Psi}_a^{1/2}$ is the diagonal matrix consisted of square roots of the elements of $\mathbf{\Psi}_a$. Thus $\mathbf{a} = (\mathbf{L}_a \otimes I_{r.-c})\mathbf{f}_a$ where

$$\mathbf{f}_a = \begin{bmatrix} \mathbf{f} \\ \mathbf{\Psi}^{-1/2}\boldsymbol{\xi} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \boldsymbol{\xi}_a \end{bmatrix}$$

Here $\boldsymbol{\xi}_a$ is a scaled version of $\boldsymbol{\xi}$ and $\mathbf{f}_a$ provides an approximation to the corresponding vector under the Cholesky decomposition. The AOM model carries over naturally to this approximation.

References

Boer MP, Wright D, Feng L, Podlich DW, Luo L, Cooper M, van Eeuwijk FA (2007) A mixed-model quantitative trait loci (QTL) analysis for multiple-environment trial data using environmental covariables for QTL-by-environment interactions, with an example in maize. Genetics 177:1801–1813

Broman KW, Wu H, Churchill G, Sen S, Yandell B (2009) qtl: Tools for analyzing QTL experiments. URL http://www.biostat.jhsph.edu/ kbroman/qtl, r package version 1.11-12

Butler DB, Tan MK, Cullis BR (2009) Improving the accuracy of selection for late maturity α-amylase in wheat using multi-phase designs. Crop and Pasture Science 60:1202–1208

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

Gelman A, Carlin JB, Stern HS, Rubin DB (2004) Bayesian Data Analysis, 2nd edn. Chapman and Hall/CRC

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

Hackett CA, Meyer RC, Thomas WTB (2001) Multi-trait QTL mapping in barley using multivariate regression. Genetical Research 77:95–106

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

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

Kenward MJ, Roger JH (1997) Small sample inference for fixed effects from restricted maximum likelihood. Biometrics 53:983–997

Malosetti M, Ribaut JM, Vargas M, Crossa J, van Eeuwijk FA (2008) A multi-trait multi-environment QTL mixed model with an application to drought and nitrogen stress trials in maize (*Zea mays* l.). Euphytica 161:241–257

Manly KF, Cudmore RH, Meer JM (2001) Map manager QTX, cross-platform software for genetic mapping. Mammalian Genome 12:930–932

Mann G, Diffey S, Cullis BR, Azanza F, Martin D, Kelly A, McIntyre L, Schmidt A, Ma W, Nath A, Kutty I, Leyne PE, Rampling L, Quail KJ, Morell MK (2009) Genetic control of wheat quality: interactions between chromosomal regions determining protein content and composition, dough rheology, and sponge and dough baking properties. Theoratical and Applied Genetics DOI: 10.1007/s00122-009-1000-y

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

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

Oakey H, Verbyla AP, Cullis BR, Wei X, Pitchford WS (2007) Joint modelling of additive and non-additive (genetic line) effects in multi-environment trials. Theoretical and Applied Genetics 114:1319–1332

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

R Development Core Team (2009) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org, ISBN 3-900051-07-0

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

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 AB, Cullis BR, Thompson R (2001) Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. Biometrics 57:1138–1147

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

Tan MK, Verbyla AP, Cullis BR, Martin P, Milgate AW, Oliver JR (2010) Genetics of late maturity α-amylase in a doubled haploid wheat population. Crop and Pasture Science 61:153–161

Taylor JD, Diffey S, Verbyla AP, Cullis BR (2009) wgaim: Whole Genome Average Interval Mapping for QTL detection using mixed models. R package version 0.02-1

Tinker NA, Mather DE (1995) Methods for qtl analysis with progeny replicated in multiple environments. Journal of Quantitative Trait Loci 1:http://probe.nalusda.gov:8000/otherdocs/jqtl/index.html

Vargas M, van Eeuwijk FA, Crossa J, Ribaut JM (2006) Mapping QTLs and QTL $\times$ environment interaction for CYMMYT maize drought stress program using factorial regression and partial least squares methods. Theoretical and Applied Genetics 112:1009–1023

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

Verbyla AP, Cullis BR, Thompson R (2007) The analysis of QTL by simultaneous use of the full linkage map. Theoretical and Applied Genetics 116:95–111

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

Zeng ZB (1994) Precision mapping of quantitative trait loci. Genetics 136:1457–1468