

# A multi-trait multi-environment QTL mixed model with an application to drought and nitrogen stress trials in maize (*Zea mays* L.)

Marcos Malosetti · Jean Marcel Ribaut ·  
Mateo Vargas · José Crossa · Fred A. van Eeuwijk

Received: 4 May 2007 / Accepted: 4 October 2007 / Published online: 7 December 2007  
© The Author(s) 2007

**Abstract** Despite QTL mapping being a routine procedure in plant breeding, approaches that fully exploit data from multi-trait multi-environment (MTME) trials are limited. Mixed models have been proposed both for multi-trait QTL analysis and multi-environment QTL analysis, but these approaches break down when the number of traits and environments increases. We present models for an efficient QTL analysis of MTME data with mixed models by reducing the dimensionality of the genetic variance–covariance matrix by structuring this matrix using direct products of relatively simple matrices representing variation in the trait and environmental dimension. In the context of MTME data, we address how to model QTL by environment interactions and the genetic basis of heterogeneity of variance and correlations between traits and environments. We illustrate our approach with an example including five traits across eight stress trials in CIMMYT maize. We detected 36 QTLs affecting yield, anthesis-silking interval, male

flowering, ear number, and plant height in maize. Our approach does not require specialised software as it can be implemented in any statistical package with mixed model facilities.

**Key words** Genetic correlation · Mixed model · Multi-environment trial · Multi-trait analysis · QTL by environment interaction · Pleiotropy

## Introduction

Plant breeders have an interest in multiple trait evaluations of germplasm rather than single trait evaluations, because good varieties combine optimal values for several traits to maximise productivity and quality. Multiple traits are of further interest to breeders when searching for indirect traits in selection schemes. By exploiting genetic correlations between traits, secondary traits can be used to improve primary ones that have low heritability or are difficult to measure. Genotype by environment interaction (GEI) complicates the analysis of single traits when breeders evaluate their genotypes across a range of environments. The occurrence of GEI in multi-trait data provides an even larger challenge to the breeder. Data in plant breeding programmes often have a multi-trait multi-environment (MTME) structure, but limited statistical methodology is available to correctly represent genetic and phenotypic variation in MTME data.

---

M. Malosetti (✉) · F. A. van Eeuwijk  
Biometris, Wageningen University, PO Box 100,  
Wageningen 6700 AC, The Netherlands  
e-mail: marcos.malosetti@wur.nl

J. M. Ribaut · M. Vargas · J. Crossa  
International Maize and Wheat Improvement Center  
(CIMMYT), Apdo postal 6-6641, Mexico DF 06600,  
Mexico

The simplest approach to analyse MTME data is to perform a series of single-trait single-environment analyses and then combine the results in some kind of meta-analysis. Methods of analysis for single traits in single environments need not be simple, but nevertheless often have the form of analysis of variance and regression models with single error terms combined with least squares procedures for parameter estimation, although mixed model analyses allowing multiple random terms would be more appropriate.

A combined MTME analysis is more powerful than a collection of single-trait single-environment analyses and allows a more realistic analysis of the data, as GEI and genetic correlations between traits can be directly modelled. Conversely, a MTME analysis requires more elaborate models. For the analysis of GEI in single traits, multiplicative fixed models have become popular (van Eeuwijk 1995; Crossa and Cornelius 2002). Well known examples of such models are the regression on the mean model (Finlay and Wilkinson 1963) and the Additive Main effects and Multiplicative Interaction effects model, or, AMMI model, (Gollob 1968; Gabriel 1978; Gauch 1988). Fixed multiplicative models can be generalised to multi-trait GEI analysis, for example in the form of three-mode principal components (Kroonenberg and Basford 1989; Basford et al. 1991; Crossa et al. 1995). By determining low dimensional approximations (principal components) to the structure present in the three classification modes of genotypes, environments, and traits, GEI patterns across traits can be studied (van Eeuwijk and Kroonenberg 1995; Varela et al. 2006). Fixed multiplicative models are useful tools for a first exploration of MTME data. For inferential purposes, however, we prefer a mixed model approach.

The modelling of the genetic (co)variances between traits and environments in combination with the modelling of the heterogeneous residuals is a condition to be fulfilled to arrive at reliable conclusions about genotypic differences. Mixed models are a natural framework for the analysis of such complex data sets, especially when the data are unbalanced (see a recent review in Smith et al. 2005). By modelling the response as the result of random and fixed factors and covariables, a wide range of possible (co)variance structures can be used to model the data, improving tests and estimates of treatment effects. The literature shows many examples of the

use of mixed models for complicated genotype by environment data (Denis et al. 1997; Piepho 1997; Cullis et al. 1998; Smith et al. 2001), but is less prolific with respect to the analysis of genotype by trait by environment data (Smith et al. 2007). Estimates for parameters in mixed models can be obtained by residual maximum likelihood (Patterson and Thompson 1971), which nowadays is implemented in statistical packages such as Genstat (Payne et al. 2006), ASReml (Gilmour et al. 2006), SAS (SAS Institute 1999), and R (R Development Core Team 2005), among others. The mixed model methodology is therefore a very suitable framework for plant breeders to analyse their complex data sets.

A less recognised application of mixed models is in the detection of quantitative trait loci (QTLs). By following the principles of regression-based QTL mapping (Haley and Knott 1992; Martínez and Curnow 1992), molecular marker information can be integrated into mixed models to test for the effect of DNA polymorphisms on phenotypic traits. Within the context of MTME data, the integration of molecular marker data into mixed models can help to identify regions (QTLs) with effects on multiple traits in multiple environments. This type of QTL analysis provides a valuable tool for investigating issues such as: (1) the occurrence of QTL by environment interaction (QEI), which is caused by changing expression of QTLs across environments; (2) the causes of genetic correlations between traits, which result from either linked QTLs or pleiotropic QTLs; and (3) the changes in genetic correlations between traits across environments, which are caused by linked or pleiotropic QTLs showing QEI. Multi-trait or multi-environment QTL mapping approaches have been presented in the literature (Jiang and Zeng 1995; Knott and Haley 2000; Piepho 2000; Verbyla et al. 2003; Malosetti et al. 2004; Emrich et al. 2007; Boer et al. 2007). In all those examples, the problem is either reduced to the multi-trait (MT), or to the multi-environment (ME) dimension. Recently, Malosetti et al. (2006) extended the QTL model to the MTME level using mixed model methodology.

In this paper, we further elaborate the mixed model methodology for MTME QTL mapping. The approach consists in first identifying an efficient model for genetic correlations by imposing some structure on the (co)variance matrix. Once a suitable and parsimonious model is identified, molecular

marker information is included to extend the phenotypic model into a MTME QTL model. To illustrate this method, we re-analyse a maize F<sub>2</sub> reference population from CIMMYT, in which five traits were evaluated over a range of several stress/non-stress environments. Part of the data used here was previously analysed by a single-trait single-environment approach (Ribaut et al. 1996, 1997) and another part by a single-trait multi-environment approach (Vargas et al. 2006).

## Materials and methods

### Plant material, phenotypic and molecular data

The data used correspond to an F<sub>2</sub> maize reference population from CIMMYT maize drought breeding program, which was derived from the cross of a drought-tolerant line (P<sub>1</sub>) with a drought susceptible line (P<sub>2</sub>). Here we provide a brief description of genotypes, trials and molecular marker procedures, more details are given in Ribaut et al. (1996, 1997). DNA from 211 F<sub>2</sub> plants was extracted to produce information for 132 co-dominant markers on 10 linkage groups. Phenotypic evaluations were performed on 211 F<sub>2:3</sub> families, each one derived from an original F<sub>2</sub> plant. The families were evaluated under different water and nitrogen regimes during 1992, 1994 and 1996. In the winter of 1992 three water regimes were imposed on the trials: well watered (WW), intermediate stress (IS) and severe stress (SS). In the winter of 1994, only the IS and SS trials were available. Nitrogen availability varied in the 1996 trials, with two low nitrogen treatments (LN, in winter and summer) and one high-nitrogen treatment (HN in summer). In each of the trials, five traits were evaluated: grain yield (YLD), the time gap between male and female flowering, that is, the anthesis-silking interval (ASI), days to male flowering (MFLW), the number of ears per plant (ENO) and plant height (PH).

### Mixed model for the MTME data

Consider a MTME data set consisting of  $I$  genotypes, evaluated in  $J$  environments with measurements on  $K$  traits (in our example,  $I = 211$ ,  $J = 8$ , and  $K = 5$ ).

Define an  $N \times 1$  vector  $\underline{y}$ , with  $N = IJK$ , that contains all the observations sorted by trait within environment and within genotype. Random variables will be underlined. The typical element of the observation vector  $\underline{y}$  is  $\underline{y}_{ijk}$ , so that within  $\underline{y}$  the trait index  $k$  runs fastest and the genotype index  $i$  runs slowest. We will now develop a mixed model to describe the observations (Smith et al. 2005). Given that the interest is in the genetic variation within the population rather than the genotypes themselves, we assume genotypes to be random. Trait-environment combinations are taken as fixed. A general formulation of a mixed model for the MTME data is:

$$\underline{y} = X\beta + Z\underline{u} + \underline{e}. \quad (1)$$

The response vector  $\underline{y}$ , is modelled by a set of fixed effects collected in vector  $\beta$  and random effects collected in vectors  $\underline{u}$  and  $\underline{e}$ .  $X$  and  $Z$  are design matrices assigning the fixed and random effects to the observations. Vector  $\beta$  contains the trait means within environments across genotypes,  $\mu_{(j,k)}$ , so that  $\beta = (\mu_{(1,1)}, \mu_{(1,2)}, \dots, \mu_{(1,k)}, \mu_{(2,1)}, \dots, \mu_{(J,K)})'$ . Vector  $\underline{u}$  collects the random genotypic effects per trait by environment combination,  $u_{(i,j,k)}$ , so that  $\underline{u} = (u_{(1,1,1)}, u_{(1,1,2)}, \dots, u_{(1,1,K)}, u_{(1,2,1)}, \dots, u_{(1,J,K)}, u_{(2,1,1)}, \dots, u_{(I,J,K)})'$ . Random genetic effects are assumed normally distributed,  $\underline{u} \sim N(0, G)$ , with  $G$  the genetic (co)variance matrix ( $\text{vcov}^G$ ). Finally,  $\underline{e}$  is a vector of non-genetic residuals associated with each observation and normally distributed,  $\underline{e} \sim N(0, R)$ . The phenotypic (co)variance is given by:  $V(\underline{y}) = ZGZ' + R$ .

From a breeder's point of view, the  $\text{vcov}^G$  is of special interest as it reflects the magnitude and pattern of relationships between genetic effects. Random genetic effects across a set of environments will not be independent if there are genes/QTLs with effects across those environments. In addition, genetic effects for different traits are not independent if genes/QTLs for different traits are linked or pleiotropic. The effect of genes/QTLs across environments will often not be equal in size, and sometimes not even in sign, leading to heterogeneous genetic variances. The model for  $\text{vcov}^G$  should reflect these relationships and the heterogeneities in genetic variation. Under the very unrealistic assumption of complete independence between genetic effects across environments and traits,  $G$  has a simple form, with non-zero values on the diagonal (the genetic variances per trait per environment) and zeroes

**Table 1** Different models for the genetic (co)variance matrix (G)

	G matrix formation	Models for $G_{m \times m}^{\text{trait-env}} / G_{j \times j}^{\text{env}} \otimes G_{k \times k}^{\text{trait}}$	Number of parameters in $G^a$	BIC	CPU time <sup>b</sup>
Model 1	$G = G_{m \times m}^{\text{trait-env}}$	DIAG	40	23,070	1.3 s
Model 2		UNIF	41	22,535	22.7 s
Model 3		FA1	80	21,836	2 m 3.6 s
Model 4		UN	820	20,263	3 m 38.1 s
Model 5	$G = G_{j \times j}^{\text{env}} \otimes G_{k \times k}^{\text{trait}}$	DIAG $\otimes$ DIAG	$8 + 5 - 1 = 12$	20,589	1.2 s
Model 6		UNIF $\otimes$ DIAG	$9 + 5 - 1 = 13$	23,690	3.4 s
Model 7		FA1 $\otimes$ DIAG	$16 + 5 - 1 = 20$	20,910	3.5 s
Model 8		UNIF $\otimes$ UN	$9 + 15 - 1 = 13$	20,522	15.5 s
Model 9		FA1 $\otimes$ UN	$16 + 15 - 1 = 30$	19,709	15.0 s

Models 1–4 use the factorial combination of traits and environments as different traits. Models used are diagonal (DIAG), uniform (UNIF), factor analytic order 1 (FA1) and unstructured (UN). Models 5–9 use the direct product of covariance matrices for environments and traits. Bayesian Information Criterion is presented to compare the goodness of fit of the different models. The computing time required to fit the model is shown in the last column

<sup>a</sup> The number of parameters for the models 5–9 follows from the sum of the parameters for the component matrices minus the number of identification constraints. Note that the total number of (co)variance parameters in the model is equal to the number presented for the G matrix in this column plus 40 from the number of parameters in the R matrix

<sup>b</sup> Pentium<sup>®</sup> 4, CPU 3.61 GHz, 3.12 GB RAM

off-diagonal. This model implies neither common QTLs between environments nor linked or pleiotropic QTLs for traits. A more realistic model would allow for covariances caused by common QTLs between environments and by linked or pleiotropic QTLs for traits. The most general model, allows the G matrix to contain unique genetic variances and covariances. This is the so-called unstructured model, which, in practice can be difficult to fit due to the high number of parameters to be estimated. Between the unrealistic independence model and the fully unstructured  $\text{vcov}^G$  model there are a number of more parsimonious models that approximate the unstructured  $\text{vcov}^G$  by imposing some structure on it.

Table 1 shows a non-exhaustive list of different models that can be used to model the  $\text{vcov}^G$ . Model choice is an iterative process and will depend on the particular data set, so predefined ‘good’ models are hard, if not impossible, to provide. The models in Table 1 can be grouped in two. The first group (including models 1–4) considers the factorial combination of traits and environments, interpreting each trait–environment combination as a trait by itself. With  $J$  environments and  $K$  traits, a total of  $M$  ‘new traits’ ( $M = JK$ ) can be defined. Models 5–9 form a group of models that exploits the direct product of (co)variance matrices for environments and traits. Note that model 1 represents the unrealistic model of

complete independence between genetic effects, and model 4 is the unstructured model, with the full G matrix. Models 2, 3, and 5–9 provide approximations to the full G matrix. Model 2 adds one parameter to model 1, which imposes a uniform genetic correlation between traits and environments. Model 3 uses a multiplicative model called factor analytic model of order 1, to approximate a fully unstructured (co)variance matrix (Oman 1991; Gogel et al. 1995). The rest of the models combine in different ways the diagonal (DIAG), the uniform (UNIF), the factor analytic of order 1 (FA1), and the unstructured (UN) models. The choice of the best model for the data can be based on a goodness of fit criterion such as the Bayesian Information Criterion, or BIC (Schwarz 1978). The BIC is calculated as  $\text{BIC} = -2 \log L + \log(N) \times p$ , with  $\log L$  the residual loglikelihood,  $N$  the sample size, and  $p$  the number of variance and covariance parameters in the model. The smaller the value of BIC, the better the model is. It must be noted that the effective sample size to use in the calculation is not clearly defined within the mixed model framework (Pauker 1998). An upper limit would be the total number of observations, while a lower limit would correspond to the total number of individuals (genotypes in this case). Staying on the conservative side, we used the number of genotypes as an estimate for the sample size in the expression for BIC.

### A QTL mixed model for the MTME data

The phenotypic model discussed in the preceding section, serves as the basis for a more elaborate model in which the effect of a particular genomic region on the phenotype is tested. For individual genotypes, molecular markers offer information at the DNA level. QTLs can then be identified by testing the association between polymorphisms at the DNA level with variation at the phenotypic level. A QTL model arises from Eq. 1 by including the effect of a putative QTL as follows:

$$\underline{y} = X\beta + X^{\text{QTL}}\alpha + Z\underline{u}^* + \underline{e}. \quad (2)$$

The extra term in the model is composed of a design matrix  $X^{\text{QTL}}$ , which is derived from molecular marker information (a further description of this key matrix will follow), and a vector of fixed QTL effects ( $\alpha$ ). In an MTME model, vector  $\alpha$  has dimensions  $M \times 1$  and contains the additive genetic QTL effects for all the traits in each of the environments:  $\alpha = (\alpha_{(1,1)}, \alpha_{(1,2)} \dots \alpha_{(1,K)}, \alpha_{(2,1)} \dots \alpha_{(J,K)})'$ . The random genetic effects, now collected in a vector  $\underline{u}^*$ , result from the effects of QTLs outside the tested region, that is, the genetic background. Genetic background effects are assumed normally distributed:  $\underline{u}^* \sim N(0, G^*)$ . Note that  $G^*$  represents the part of the genetic (co)variance that is not explained by the QTL.

A key element in the QTL model is the design matrix  $X^{\text{QTL}}$ , which contains the so-called genetic predictors. Genetic predictors are a function of the conditional probabilities of the QTL genotype given flanking marker information (Lynch and Walsh 1998). In an  $F_2$  population, at a marker position, the genetic predictor corresponding to an additive genetic QTL effect of an individual will take the value  $-1$ ,  $0$ , or  $1$  depending on whether the individual's marker genotype is  $aa$ ,  $Aa$ , or  $AA$  (interpreted as the QTL genotypes  $qq$ ,  $Qq$ , and  $QQ$  respectively). In between markers, QTL genotypes are not directly observable, but conditional probabilities of genotypes  $qq$ ,  $Qq$ , and  $QQ$  can be calculated from flanking markers (Jiang and Zeng 1997). In between marker positions, the value of the additive genetic predictor is equal to the difference of the conditional probabilities for QTL genotypes  $QQ$  and  $qq$ :  $\text{Pr}(QQ|\text{flanking markers}) - \text{Pr}(qq|\text{flanking markers})$ . The values of the genetic predictors for each of the  $I$  genotypes can be estimated and

collected in a vector  $p = (x_1, x_2 \dots x_I)'$ . The design matrix of genetic predictors ( $X^{\text{QTL}}$ ) can then be expressed as:  $X^{\text{QTL}} = p \otimes I_M$ , that is, the direct product (Kronecker product) of vector  $p$  and the identity matrix of dimension  $M$ . Note that the present configuration of matrix  $X^{\text{QTL}}$  assumes a pleiotropic effect of the QTL on all traits. However, the design matrix  $X^{\text{QTL}}$  can be modified to exclude the effect of a particular trait–environment combination by removing the corresponding columns (and therefore also reducing the vector of estimated effects).

Note that although dominance has not been considered in the model here, it can be fitted by including an extra term containing dominance genetic predictors. At marker positions, dominance genetic predictors will take the values  $0$ ,  $1$ ,  $0$  for  $aa$ ,  $Aa$ , and  $AA$  respectively, and in between markers will be equal to  $\text{Pr}(Qq|\text{flanking markers})$  (Lynch and Walsh 1998). In a preliminary analysis, we scanned the genome and we did not find any significant dominance effect (data not shown). The observation of dominance not being important in this population is consistent with previous results on the same population (Vargas et al. 2006). We therefore excluded dominance from the model.

The extension from a single QTL model to a multi-QTL model is straightforward and is given by Eq. 3.

$$\underline{y} = X\beta + \sum_q X_q^{\text{QTL}}\alpha^q + Z\underline{u}^* + \underline{e}. \quad (3)$$

The QTL section includes the additive effects of all detected QTLs in the genome. Note that the form of the design matrix  $X_q^{\text{QTL}}$  defines whether that position is pleiotropic or not. Linked QTLs are determined by consecutive design matrices  $X_q^{\text{QTL}}$  and  $X_{q+1}^{\text{QTL}}$  whose positions are close on the same chromosome. The degree with which the QTL model explains the genetic variance can be assessed by comparing the diagonal estimates of matrix  $G^*$  from fitting Eq. 3, with the diagonals of the  $G$  matrix obtained from fitting Eq. 1.

### QTL mapping: scanning and testing procedure

With the modelling framework determined, we still need a strategy to develop a multi-QTL model (Eq. 3) from a phenotypic model (Eq. 1). Model search is not a simple task and there is often no unique solution to

this problem. Here we followed a procedure which can be split into two main steps: (1) genome-wide scan with tests for single QTLs, (2) multi-QTL model refinement by backward selection from a model containing all the significant, but still putative, QTLs detected in step 1. All models were fitted with Genstat<sup>®</sup>, version 9 (Payne et al. 2006), see Appendix II for source code.

In step 1 we performed a genome-wide scan using a one-QTL model, which consisted of fitting Eq. 2 at regular intervals along the genome. This strategy required the calculation of genetic predictors on a regular grid across the genome. We chose 5 cM as the maximum distance between consecutive predictors. At each evaluation point within the genome, we used a Wald test (Verbeke and Molenberghs 2000; Payne et al. 2006) to find out whether one or more of the trait–environment QTL effects were significantly different from zero ( $H_0: \alpha_{(1,1)} = \alpha_{(1,2)} = \dots = \alpha_{(1,K)} = \alpha_{(2,1)} = \dots = \alpha_{(J,K)} = 0$ ). The test statistic for a Wald test,  $W$ , for the effect of a QTL in a single environment is calculated as:  $W = \frac{\hat{\alpha}_{(j,k)}^2}{SE^2}$ , with  $\hat{\alpha}_{(j,k)}$  the estimated QTL effect in environment  $j$  for trait  $k$ , and  $SE$  the associated standard error. Under the null hypothesis  $W$  is chi-square distributed with 1 degree of freedom. Instead of testing QTL effects for all traits simultaneously, a more informative sub-test can be carried out to test QTL effects for specific traits. For example, for trait 1 the null hypothesis then becomes:

$$H_0 : \alpha_{(1,1)} = \alpha_{(1,2)} = \dots = \alpha_{(1,J)} = 0.$$

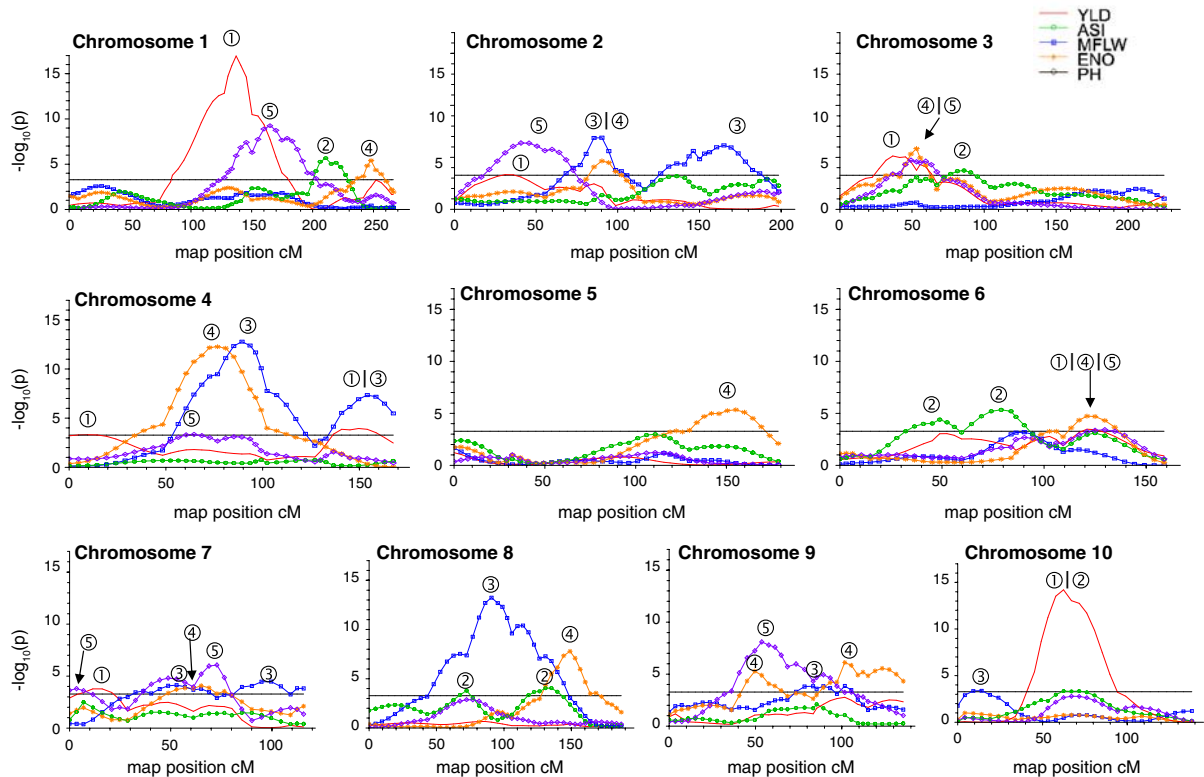
The values of the Wald statistics or the associated tail probabilities,  $P$ , expressed as  $-\log_{10}(P)$ , serve to produce plots analogous to the usual LOD score profiles in QTL mapping. By plotting the  $-\log_{10}(P)$  along the chromosomes, we identified putative QTLs at those positions for which peaks in the profile exceeded a threshold value. We used a Bonferroni-based multiple-test control threshold, using the estimation of the effective number of tests along the genome proposed by Li and Ji (2005). We control the genome-wide alpha level at 0.05, which corresponded to a point-wise alpha level of 0.05 divided by the effective number of tests along the genome. All QTLs identified in this way, constituted the starting set of QTLs (predictors) for the backward selection procedure of the second step in the procedure.

QTLs identified in step 1 showed a significant effect for one or more traits, so the design matrix needed to be adapted as described in the previous section (see Appendix I). Step 2 started by fitting Eq. 3 including all putative QTLs in the model. Backward selection consisted in removing putative QTLs from the model when the associated Wald test conditional on all the other putative QTLs being in the model was not significant ( $P > 0.05$ ). In each step the position showing the largest  $P$  value (above 0.05) was excluded from the model and the process repeated until no position had an associated  $p$  value larger than 0.05. From the fit of the final multi-QTL model we estimated the individual QTL effects and standard errors.

## Results

The results of fitting a phenotypic model (Eq. 1) assuming different models for the  $vcov^G$  are given in Table 1. As mentioned before, models 1–4 (group 1) are based on the idea of treating trait–environment combinations as traits. In contrast, models 5–9 (group 2) use direct products of covariance matrices for traits and environments. While models in group 1 are more flexible, they require the estimation of a higher number of parameters than those in group 2. Note that the simplest model in group 1 (model 1) has a higher number of parameters than the most complex one in group 2 (model 9). The higher number of parameters of models in group 1 resulted in longer computing time as is shown in Table 1. A higher number of parameters creates also more convergence difficulties. For example, although we eventually could fit an unstructured model, convergence was only achieved after supplying appropriate initial values. The BIC showed that all models in group 2 (except for model 5) performed better than those in group 1, with model 9 being the best. Based on these results we decided to use model 9 in the QTL mapping stage.

The result of the first step of our QTL mapping approach, that is, the fit of Eq. 2 across the chromosomes, is summarised in Fig. 1. A total of 36 regions, revealed by peak values in the profiles, were identified as harbouring putative QTLs. In some cases, profiles related to different traits showed peaks at the same position, which we considered as an indication of pleiotropy. However, we recognise that strictly speaking this might not be true, as very close linkage



**Fig. 1** Result of a multi-trait multi-environment QTL scan. The profiles correspond to yield (YLD), ASI, male flowering (MFLW), ear number (ENO), and plant height (PH). The marks:

①, ②, ③, ④ and ⑤ indicate QTL positions for YLD, ASI, MFLW, ENO and PH, respectively. When the same QTL has an effect on more than one trait the marks are separated by ‘|’

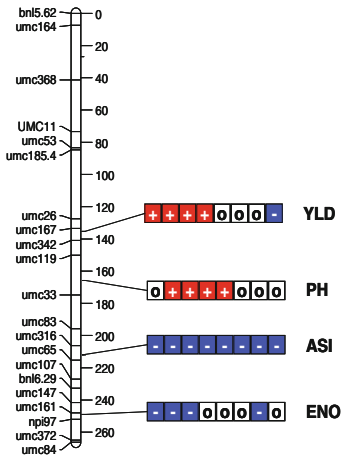
can not be excluded as a possibility. Four chromosome regions had exactly coinciding peaks for two traits, and at chromosome 6 three traits had a peak in the same region. All 36 positions were regarded as candidate QTLs and constituted the starting point for the backwards selection procedure. None of the candidates were eliminated in the backward selection stage. Therefore our final QTL model consisted of 36 QTLs of which 31 related to a single trait, four affected two traits, and the remaining QTL affected three traits (Fig. 2 shows the chromosome position and the affected trait(s) for each of the 36 QTLs).

In addition to the affected trait(s) and chromosome position, Fig. 2 shows the results of the test for the specific effect of the QTLs in each of the eight environments (represented by blocks inside the QTL bars). Significant effects are indicated by either a ‘+’ or ‘-’ sign, depending on whether the allele from the drought-susceptible parent (P2) increased or decreased the trait value. Non-significant effects are indicated by ‘o’. Inspection of Fig. 2 raises the impression that most

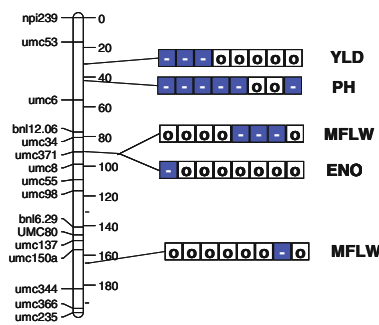
of the QTLs showed inconsistent effects across environments. Inconsistent QTL effects underlie GEI. The occurrence of GEI was expected for trials involving contrasting environmental conditions. Relatively consistent QTL effects were identified for ASI on chromosomes 1, 6 and 10, for MFLW on chromosomes 4, and 9, and for PH on chromosomes 3, 6 and 9. No QTL showed consistent effects on YLD and ENO, which could be the consequence of a more complex genetic basis of these traits (especially YLD).

The positions and effects of the different QTLs can help to understand the causes of genetic correlations between traits. Neighbouring or pleiotropic QTLs with consistent effects on different traits will induce consistent genetic correlations. Consistent genetic correlations were not observed in general, although some consistent positive correlations were suggested by the two linked QTLs for PH and MFLW around the middle of chromosome 9. The majority of the genetic correlations induced by linked or pleiotropic QTLs were inconsistent across environments. For example,

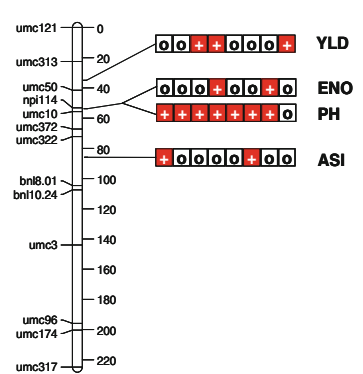
Chromosome 1



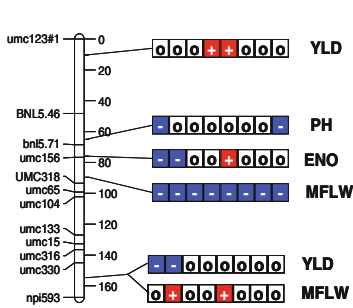
Chromosome 2



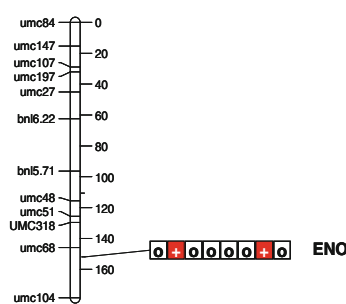
Chromosome 3



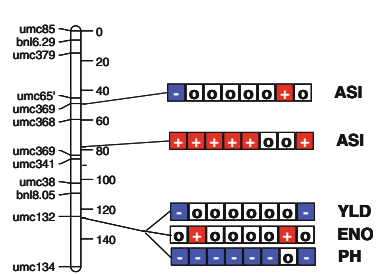
Chromosome 4



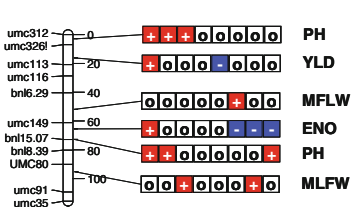
Chromosome 5



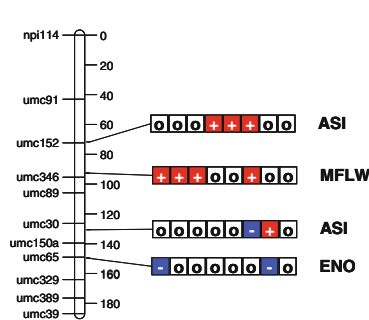
Chromosome 6



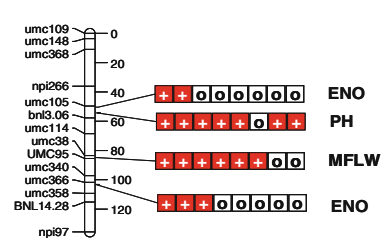
Chromosome 7



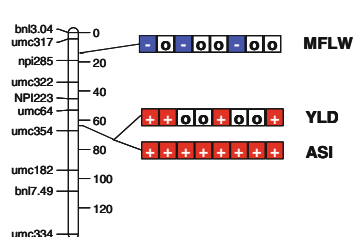
Chromosome 8



Chromosome 9



Chromosome 10



◀ **Fig. 2** Distribution of QTLs detected for yield (YLD), ASI, male flowering (MFLW), ear number (ENO), and plant height (PH). Each QTL is represented by a bar with a connector to the corresponding position on the chromosome. Pleiotropic QTLs are indicated by connectors pointing to more than one trait. The eight blocks constituting each bar show the effect of the QTL in each environment (from left to right: NS92a, IS92a, SS92a, IS94a, SS94a, LN96a, LN96b, HN96b). Significant effects are indicated by either '+' (red background) or '-' (blue background) depending on whether the allele from the drought-susceptible parent (P2) increases or decreases the trait's value. Non-significant effects are indicated by 'o' (white background)

the two linked QTLs on top of chromosome 2 (30–40 cM) induced a positive correlation between YLD and PH in the first three environments, but that correlation disappeared in the rest of the environments.

As YLD is the most important trait, it is interesting to point out where the QTLs for YLD were located and what their relationships were with QTLs for other traits. YLD QTLs were present on all chromosomes except at chromosomes 5, 8 and 9. Yield QTLs were in many cases linked or pleiotropic to PH and inducing positive correlations between the traits. Linked QTLs for YLD and PH were observed at chromosomes 1, 2, 3, 4 and 7. Pleiotropy was observed for chromosome 6. The YLD QTL on chromosome 3 was also linked to an ENO QTL, and pleiotropy was observed on chromosome 6. In all cases the correlation between YLD and ENO was positive. Correlations between YLD and ASI were mostly caused by a pleiotropic QTLs on chromosome 10 and some linkage on chromosomes 1, 3 and 6, but no clear direction of the correlation was observed. Finally, a MFLW QTL was pleiotropic to YLD on chromosome 4 (inducing a negative correlation), and some weak linkage was observed on chromosomes 2, 7 and 10 without a clear direction of the sign of the correlation.

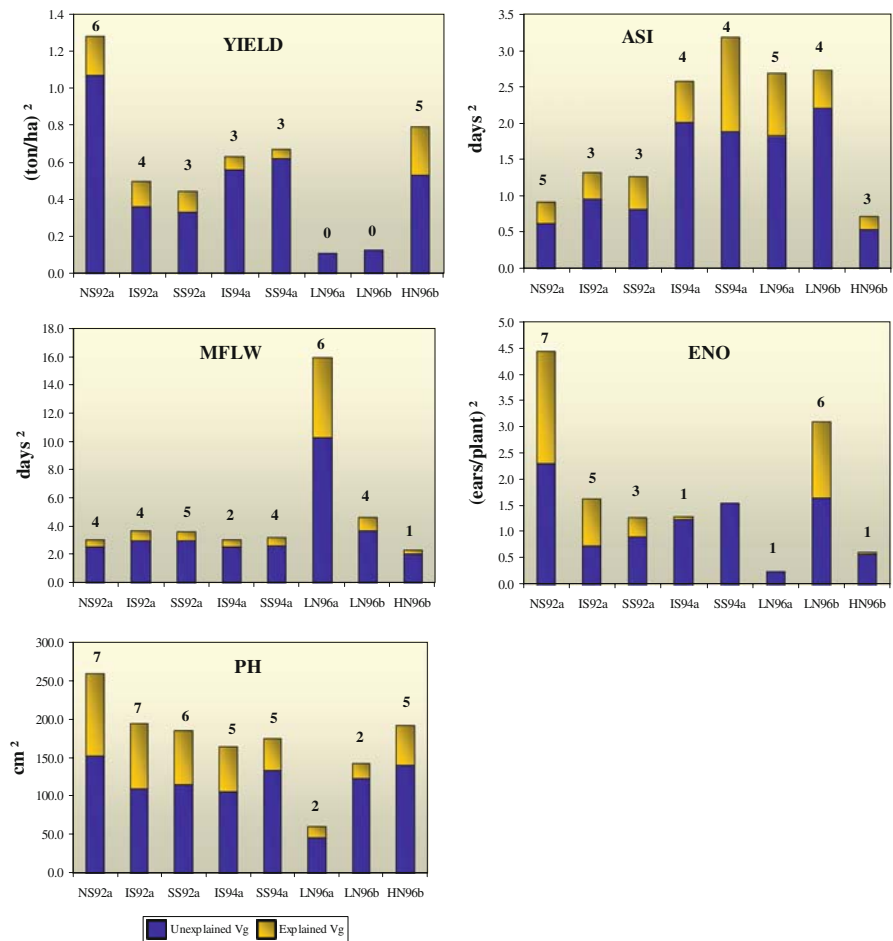
Figure 3 shows the total genetic variance, the number of detected QTLs and the proportion of explained variation by the detected QTLs for each trait and environment. The total genetic variance was given by the diagonal of the G matrix of the phenotypic model (Eq. 1), and the unexplained variance was given by the diagonal of the G\* matrix after fitting the full QTL model (Eq. 3 including 36 QTLs). Estimates were produced by fitting Eqs. 1 and 3 based on model 3, which allowed better estimates of the variance components due to its higher flexibility in comparison to model 9. In terms of the total

genetic variance, Fig. 3 shows that genotypes gave less consistent responses across environments for YLD, ASI and ENO than for MFLW and PH. The heterogeneity of genetic variance across environments underlines the existence of GEI, which is in agreement with our findings of inconsistent QTL effects across environments. In the case of MFLW and PH, less GEI occurred as the genetic variance was more homogenous across environments. The only exception was LN96a (low nitrogen trial in winter season), which seemed to produce a very distinct reaction of the genotypes with a longer male flowering period but more homogenous plant height. The proportion of explained variation ranged from zero to 55% (Fig. 3). On average ASI and PH were the traits that had the highest proportion of explained genetic variance (29% and 31%, respectively), followed by MFLW and ENO with 19% and 23%, respectively, and finally YLD with 15%. As expected, the number of detected QTLs was in general related to the amount of genetic variation observed.

## Discussion

Plant breeders routinely deal with data involving collections of genotypes evaluated for multiple traits across multiple environments. Mixed models offer a suitable framework to jointly analyse such data without imposing unrealistic assumptions, like zero genetic correlations between environments and traits, and constant variance across environments. In MTME QTL analysis, the flexibility of mixed models can be fully exploited. However, for several traits across multiple environments, the number of (co)variance parameters increases rapidly, causing serious difficulties in fitting the models and requiring substantial computing time. In our approach to QTL mapping, model fitting at individual genomic evaluation points should be fast as many evaluations are required along the genome. One way out of the above problem is to impose some structure on the genetic (co)variance matrix, which leads to more parsimonious models. In this paper, we compared different models to structure the genetic (co)variances between five traits and eight environments, making it possible to perform a MTME QTL analysis involving 40 trait–environment combinations.

**Fig. 3** Bar plot of the total genetic variance for yield (YLD), ASI, male flowering (MFLW), ear number (ENO), and plant height (PH) in eight environments in Mexico. The proportion of explained variance by the final QTL model is indicated in yellow. The total number of detected QTLs is indicated on top of each bar



Two major strategies were used to model the (co)variance matrix  $G$ . The first strategy considers each trait–environment combination as a new ‘trait’. The second approach uses the direct product of covariance matrices for traits and environments. We observed that the second strategy was a better option than the first one, at least for our data. The use of direct products of matrices to construct (co)variance matrices gave a good fit to the data, while considerably reducing the number of parameters. In terms of computing time, the second approach produced models that converged much faster than those based on the first strategy. Fast fitting of models is desirable as we needed to fit a similarly structured mixed model at more than 400 chromosome positions. To give some idea about the required computing time for a QTL analysis, our data consisted of 8,440 observations and needed almost 2 h of a Pentium<sup>®</sup> 4, 3.6 GHz processor and 3.1 GB of RAM memory, to run a

genome-wide QTL scan (based on model 9, Table 1). Fitting the final QTL model, which consisted of 36 different QTLs, took slightly over 3 min.

Based on a mixed model, which combined both efficiency in terms of number of parameters and goodness of fit, we were able to perform a MTME QTL analysis on five traits in eight environments. The first key aspect of this mixed model approach is that QTL effects were tested taking into account the genetic correlation structure in the data. In a simulation study, Piepho (2005) showed that ignoring genetic correlations in multi-environment data leads to a substantial increase of the type I error rate when testing for QTL effects. We therefore expect that our model approach will reduce the risk of over-optimistic conclusions. The second important aspect is that by using a parsimonious model for  $G$ , a larger number of trait–environment combinations can be included in comparison to other multi-trait QTL mapping

approaches. Most of the multi-trait QTL mapping approaches model the genetic (co)variances by an unstructured model (Jiang and Zeng 1995; Korol et al. 1998; Knott and Haley 2000; Hackett et al. 2001; Lund et al. 2003), which in practice limits the number of traits that can be handled. It is symptomatic that the applications of multi-trait QTL mapping under the use of unstructured (co)variance matrices never included more than just a few traits, i.e., mostly 2–5 (Calinski et al. 2000; Hackett et al. 2001; Szyda et al. 2003; Mercadé et al. 2005; Olsen et al. 2005). With few traits and/or environments, the unstructured model is a possible option, although not necessarily the optimal one. However, if the number of environments and traits increases, as is the case in this paper, the unstructured model will eventually fail and more efficient modelling approaches need to be used.

The MTME QTL analysis detected 36 QTLs affecting grain yield and other important traits in maize across a wide range of stress conditions. In addition to the improvement in power to detect QTLs which was demonstrated using simulation by Jiang and Zeng (1995) and Knott and Haley (2000), an integrated QTL analysis produces useful information concerning the genetic determination and relation between traits. The set of QTLs could be categorised as consisting of two classes: QTLs that affected only one trait (31 in total) versus pleiotropic QTLs that affected two or more traits (the remaining five QTLs). The basis of genetic correlations can be understood in terms of pleiotropic QTLs and linked QTLs for multiple traits. QTL locations informed us about which of the two mechanisms was more plausible.

Our QTL analysis provided also insight in the causes of GEI. QTLs with consistent effects across environments were distinguished from those whose effects were highly influenced by the environment, the latter being responsible for GEI. Size and sign of QTL effects across environments give an indication of the importance of the particular QTL as cause of observed GEI. Large differences in QTL effects between environments underlie strong GEI, the extreme being a reversal of sign implying cross-over GEI. Environment-specific QTL effects are a valuable piece of information for the breeder at the moment of pyramiding favourable alleles for broad or specific adaptation.

A desirable feature of our approach was that all information was produced within the same model class, thereby avoiding the burden of having to

combine results from different analyses outside a formal framework. For example, three of the eight environments (NS92a, IS94a and SS94a) were previously used in a conventional single-trait single-environment QTL analysis (Ribaut et al. 1996, 1997). For yield four, five and four QTLs were reported in NS92a, IS94a and SS94a, respectively, which fairly well agrees with the number of detected QTLs in those three environments in the analysis here (six, four and three QTLs, see Fig. 3). However, the single-trait analysis, failed to give an integrated answer with respect to QTL locations and effects, so QTLs detected in one environment were not strictly comparable to those found in a second environment. In a multi-environment strategy, Vargas et al. (2006) integrated the information across the eight environments to model QEI, but kept the traits separately (YLD and ASI). Their main conclusions coincided with the ones found here, with the major QTLs for YLD identified on chromosome 1 and 10, and for ASI found on chromosomes 1, 6 and 10. However, some power seemed to have been gained in the MTME approach, as we detected eight and seven QTLs for YLD and ASI, instead of the six and five YLD and ASI QTLs reported by Vargas et al. (2006).

In conclusion, this paper shows how high dimensional data sets (traits  $\times$  environments) can be used in the identification of genetic factors underlying trait variation and covariation. The approach exploits the flexibility of mixed modelling, which has the added advantage of being readily available to the breeding community. The approach does not require any specific software other than a package with mixed model facilities, although it does require some extra intervention from the user. That last requirement is largely compensated by the improvement and reliability of the results that is expected to follow from the use of a more realistic model for the genetic (co)variances.

**Acknowledgments** This research was done in the framework of the project, ‘An eco-physiological–statistical framework for the analysis of G $\times$ E and QTL $\times$ E as occurring in abiotic-stress trials, with applications to the CIMMYT drought-stress programs in tropical maize and bread wheat’, Generation Challenge Program, Competitive Grant #4.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.







Genstat codes to fit the models in Table 1

### Model 1:

```
VCOMPONENTS [fixed = te; constant = omit;
experiment = te]random = geno.te
VSTRUCTURE [term = geno.te] factor = te;
model = diagonal
REML Y = Y
```

Short code explanation

The VCOMPONENTS statement defines the fixed and random effects (fixed= fixed model terms, random= random model terms). The ‘fixed = te’ specification fits an intercept term for each trait – environment combination. Note that this is because the overall constant is excluded by the command ‘constant = omit’. Finally, by including the ‘experiment=’ option, a different residual term is fitted for each trait–environment combination.

The VSTRUCTURE statement is used to impose a model on the random terms. The option term= indicates which term to model, the option factor= over which factor levels to form the direct product, and the option model= indicates which model to use for the (co)variance matrix.

The REML statement fits the model by REsidual Maximum Likelihood and has as parameter the response vector ‘Y’.

### Model 2:

```
VCOMPONENTS [fixed = te; constant = omit;
experiment = te]random = geno.te
VSTRUCTURE [term = geno.te] factor =
te; model = uniform; heterogeneity =
outside
REML Y = Y
```

### Model 3:

```
VCOMPONENTS [fixed = te; constant = omit;
experiment = te]random = geno.te
VSTRUCTURE [term = geno.te] factor = te;
model = fa
REML Y = Y
```

### Model 4:

```
VCOMPONENTS [fixed = te; constant = omit;
experiment = te]random = geno.te
VSTRUCTURE [term = geno.te] factor =
te; model = unstructured
REML Y = Y
```

### Model 5:

```
VCOMPONENTS [fixed = te; constant = omit;
experiment = te]random = geno.env.trait
VSTRUCTURE [term = geno.env.trait] factor =
env, trait; model = diagonal, diagonal
REML Y = Y
```

### Model 6:

```
VCOMPONENTS [fixed = te; constant = omit;
experiment = te]random = geno.env.trait
VSTRUCTURE [term = geno.env.trait] factor =
env, trait; model = uniform, diagonal;
heterogeneity = outside,*
REML Y = Y
```

### Model 7:

```
VCOMPONENTS [fixed = te; constant = omit;
experiment = te]random = geno.env.trait
VSTRUCTURE [term = geno.env.trait] factor =
env, trait; model = fa, diagonal
REML Y = Y
```

### Model 8:

```
VCOMPONENTS [fixed = te; constant = omit;
experiment = te]random = geno.env.trait
VSTRUCTURE [term = geno.env.trait] factor =
env, trait; model = uniform, unstructured;
heterogeneity = outside,*
REML Y = Y
```

### Model 9:

```
VCOMPONENTS [fixed = te; constant = omit;
experiment = te]random = geno.env.trait
VSTRUCTURE [term = geno.env.trait] factor =
env, trait; model = fa, unstructured
REML Y = Y
```

Genstat code to fit a one-QTL model based on the best model (model 9):

```

VCOMPONENTS [fixed = te + mk.te;
constant = omit; experiment = te]
random = geno.env.trait
VSTRUCTURE [term = geno.env.trait] factor =
env, trait; model = fa, unstructured
REML Y

```

## References

- Basford KE, Kronenberg PM, DeLacy IH (1991) Three-way methods for attribute genotype  $\times$  environment data: an illustrated partial survey. *Field Crops Res* 27:131–157
- 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
- Calinski T, Kaczmarek Z, Krajewski P, Frova C, Sari-Gorla M (2000) A multivariate approach to the problem of QTL localization. *Heredity* 84:303–310
- Crossa J, Basford K, Taba S, DeLacy IH, Silva E (1995) Three-mode analyses of maize using morphological and agronomical attributes measured in multilocation trials. *Crop Sci* 35:1483–1491
- Crossa J, Cornelius P (2002) Linear-bilinear models for the analysis of quantitative-environment interaction. In: Kang MS (ed), *Quantitative genetics, genomics, and plant breeding*. CAB International, Wallingford, UK, pp 305–322
- Cullis B, Gogel B, Verbyla A, Thompson R (1998) Spatial analysis of multi-environment early generation variety trials. *Biometrics* 54:1–18
- Denis JB, Piepho HP, van Eeuwijk FA (1997) Modelling expectation and variance for genotype by environment data. *Heredity* 79:162–171
- Emrich K, Price A, Piepho HP (2007) Assessing the importance of genotype  $\times$  environment interaction for root traits in rice using a mapping population III: QTL analysis by mixed models. *Euphytica* (this issue)
- Finlay KW, Wilkinson GN (1963) The analysis of adaptation in a plant breeding programme. *Aust J Agric Res* 14:742–754
- Gabriel KR (1978) Least squares approximation of matrices by additive and multiplicative models. *J R Stat Soc Ser B* 40:186–196
- Gauch HG (1988) Model selection and validation for yield trials with interaction. *Biometrics* 44:705–715
- Gollob HF (1968) A statistical model which combines features of factor analysis and analysis of variance techniques. *Psychometrika* 33:73–115
- Gilmour AR, Gogel BJ, Cullis BR, Thompson R (2006) *AS-Reml User Guide Release 2.0* VSN International Ltd., Hemel Hempstead, HP1 1ES, UK
- Gogel BJ, Cullis BR, Verbyla AP (1995) REML estimation of multiplicative effects in multi-environment variety trials. *Biometrics* 51:744–749
- Hackett CA, Meyer RC, Thomas WTB (2001) Multi-trait QTL mapping in barley using multivariate regression. *Genet Res* 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
- Jiang C, Zeng Z-B (1995) Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics* 140:1111–1127
- Jiang CJ, Zeng ZB (1997) Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines. *Genetica* 101:47–58
- Knott SA, Haley CS (2000) Multitrait least squares for quantitative trait loci detection. *Genetics* 156:899–911
- Korol AB, Ronin YI, Nevo E, Hayes PM (1998) Multi-interval mapping of correlated trait complexes. *Heredity* 80:273–284
- Kroonenberg PM, Basford KE (1989) An investigation of multi-attribute genotype response across environments using three-mode principal component analysis. *Euphytica* 44:109–123
- Li J, Ji L (2005) Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity* 95:221–227
- Lund MS, Sorensen P, Gulbrandtsen B, Sorensen DA (2003) Multitrait fine mapping of quantitative trait loci using combined linkage disequilibria and linkage analysis. *Genetics* 163:405–410
- Lynch M, Walsh B (1998) *Genetics and analysis of quantitative traits*. Sinauer Associates, Sunderland MA
- Malosetti M, Boer MP, Bink MCAM, van Eeuwijk FA (2006) Multi-trait QTL analysis based on mixed models with parsimonious covariance matrices. In: *Proceedings of the 8th World Congress on Genetics Applied to Livestock Production*, August 13–18, Belo Horizonte, MG, Brasil. <http://www.wcgalp8.org.br/wcgalp8>. Article 25-04
- Malosetti M, Voltas J, Romagosa I, Ullrich SE, van Eeuwijk FA (2004) Mixed models including environmental covariables for studying QTL by environment interaction. *Euphytica* 137:139–145
- Martínez O, Curnow RN (1992) Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. *Theor Appl Genet* 85:480–488
- Mercadé A, Estellé J, Noguera JL, Folch JM, Varona L, Silió L, Sánchez A, Pérez-Enciso M (2005) On growth, fatness, and form: a further look at porcine Chromosome 4 in an Iberian  $\times$  Landrace cross. *Mamm Genome* 16:374–382
- Olsen HG, Lien S, Gautier M, Nilsen H, Roseth A, Berg PR, Sundsaasen KK, Svendsen M, Meuwissen THE (2005) Mapping of a milk production quantitative trait locus to a 420-kb region on bovine chromosome 6. *Genetics* 169:275–283
- Oman SD (1991) Multiplicative effects in mixed model analysis of variance. *Biometrika* 78:729–739
- Patterson HD, Thompson R (1971) Recovery of inter-block information when block sizes are unequal. *Biometrika* 58:545–554
- Pauler DK (1998) The Schwarz criterion and related methods for normal linear models. *Biometrika* 85:13–27
- Payne RW, Harding SA, Murray DA, Soutar DM, Baird DB, Welham SJ, Kane AF, Gilmour AR, Thompson R,

- Webster R, Tunnicliffe Wilson G (2006) The guide to GenStat release 9, Part 2: statistics. VSN International, Hemel Hempstead
- Piepho HP (1997) Analyzing genotype-environment data by mixed models with multiplicative terms. *Biometrics* 53: 761–766
- Piepho HP (2000) A mixed-model approach to mapping quantitative trait loci in barley on the basis of multiple environment data. *Genetics* 156:2043–2050
- Piepho HP (2005) Statistical tests for QTL and QTL-by-environment effects in segregating populations derived from line crosses. *Theor Appl Genet* 110:561–566
- R Development Core Team (2005) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>
- Ribaut JM, Hoisington DA, Deutsch JA, Jiang C, Gonzalez de Leon D (1996) Identification of quantitative trait loci under drought conditions in tropical maize.1. Flowering parameters and the anthesis-silking interval. *Theor Appl Genet* 92:905–914
- Ribaut JM, Jiang C, Gonzalez de Leon D, Edmeades GO, Hoisington DA (1997) Identification of quantitative trait loci under drought conditions in tropical maize.2. Yield components and marker-assisted selection strategies. *Theor Appl Genet* 94:887–896
- SAS Institute (1999) SAS/Stat User's guide, Version 8. SAS Institute, Cary, NC
- Schwarz G (1978) Estimating the dimension of a model. *Ann Stat* 6:461–464
- Smith A, Cullis B, Thompson R (2001) analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. *Biometrics* 57:1138–1147
- Smith A, Stringer J, Wei X, Cullis B (2007) Varietal selection for perennial crops where data relate to multiple harvests from a series of field trials. *Euphytica* 157:253–266
- Smith AB, Cullis BR, Thompson R (2005) The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. *J Agric Sci* 143:449–462
- Szyda J, Grindflek E, Liu Z, Lien S (2003) Multivariate mixed inheritance models for QTL detection on porcine chromosome 6. *Genet Res* 81:65–73
- van Eeuwijk FA (1995) Linear and bilinear models for the analysis of multi-environment trials: I. An inventory of models. *Euphytica* 84:1–7
- van Eeuwijk FA, Kroonenberg PM (1995) The simultaneous analysis of genotype by environment interaction for a number of traits using three-way multiplicative modelling. *Biuletyn Oceny Odmian (Cultivar Testing Bulletin)* 26–27: 83–96
- Varela M, Crossa J, Rane J, Joshi AK, Trethowan R (2006) Analysis of a three-way interaction including multi-attributes. *Aust J Agric Res* 57:1185–1193
- Vargas M, van Eeuwijk F, Crossa J, Ribaut J-M (2006) Mapping QTLs and QTL  $\times$  environment interaction for CIMMYT maize drought stress program using factorial regression and partial least squares methods. *Theor Appl Genet* 112:1009–1023
- Verbeke G, Molenberghs (2000) Linear mixed models for longitudinal data. Springer Verlag, New York
- Verbyla A, Eckermann PJ, Thompson R, Cullis B (2003) The analysis of quantitative trait loci in multi-environment trials using a multiplicative mixed model. *Aust J Agric Res* 54:1395–1408