Theor Appl Genet (2006) 113:809–819
DOI 10.1007/s00122-006-0333-z

ORIGINAL PAPER

Joint modeling of additive and non-additive genetic line effects in single field trials

Helena Oakey · Arūnas Verbyla · Wayne Pitchford · Brian Cullis · Haydn Kuchel

Received: 11 January 2006 / Accepted: 3 June 2006 / Published online: 2 August 2006

Abstract A statistical approach is presented for selection of best performing lines for commercial release and best parents for future breeding programs from standard agronomic trials. The method involves the partitioning of the genetic effect of a line into additive and non-additive effects using pedigree based inter-line relationships, in a similar manner to that used in animal breeding. A difference is the ability to estimate non-additive effects. Line performance can be assessed by an overall genetic line effect with greater accuracy than when ignoring pedigree information and the additive effects are predicted breeding values. A generalized definition of heritability is developed to account for the complex models presented.

Communicated by M. Cooper

H. Oakey (✉) · A. Verbyla
BiometricsSA, School of Agriculture and Wine, University of Adelaide, PMB 1, Glen Osmond, SA 5064, Australia
e-mail: Helena.Oakey@adelaide.edu.au

W. Pitchford
Animal Science, School of Agriculture and Wine, University of Adelaide, Roseworthy, SA 5371, Australia

B. Cullis
Biometrics, NSW Department of Primary Industries,
Wagga Wagga Agricultural Institute,
Private Mail Bag Wagga, Wagga,
NSW 2650, Australia

H. Kuchel
Plant Breeding Unit, Australian Grain Technologies,
Perkins Building, Roseworthy, SA 5371, Australia

Introduction

In most crop breeding programs there are two fundamental goals. The main goal is to identify the best performing lines for commercial release, and the second goal is to identify lines that can be used as parents in future crosses.

The selection of best performing lines for traits of interest is undertaken through well-designed breeding trials conducted across multiple environments and analyzed appropriately. Good trial design and subsequent statistical analysis enable efficient separation of genetic and environmental effects. Suitable designs may include classical designs such as incomplete block, row–column, α latinized row–column (John et al. 2002) to the more recently devised designs which are efficient for a prespecified correlation structure (Martin et al. 2004). The method of analysis should support the design used and the aim of the trial. Although most approaches for single trials (for example, Besag and Kempton 1986; Cullis and Gleeson 1991; Gilmour et al. 1997) consider the genetic effects of the lines as fixed effects, if the ultimate aim of the analysis is selection line effects should be treated as random (Smith et al. 2005).

The suitability of lines as parents can be conducted through specialized mating designs such as the diallel cross (see Topal et al. 2004 for a recent example). These designs allow the partitioning of the genetic effect of a line into additive and non-additive effects. The additive effects or breeding values obtained for each line measure the potential of a line as a parent (Falconer and Mackay 1996) and for a diallel cross are termed “general combining ability”. The non-additive effects obtained for each line are associated with dominance and epistatic effects and for a diallel cross are termed ‘‘specific combining ability’’. However, there are several disadvantages of formal mating designs. Firstly, only small numbers of lines can be examined at once. Secondly, they are necessarily conducted in addition to any breeding trials and usually performed after or near the commercial release of a line therefore restricting their usefulness. Because of these disadvantages, the suitability of lines as parents is often assessed in the same way as their potential for commercial release, that is, by examining their overall genetic effect. However, if the attributes of a released line are a result of interactions between genes (epistasis), then this approach is less than ideal. In this case, the performance of the line is greater than the sum of alleles leading to an inflated assessment of breeding potential.

The additive genetic effect or breeding value is widely determined in animal breeding trials and is used to assess the potential of an animal as a parent (see Brown et al. 2000 for a recent example in sheep), since it is not simple to replicate genotypes. The approach involves the incorporation of the pedigree information of animals into the trial analysis in the form of the additive relationship matrix **A** (Henderson 1976). The animal approach incorporating the pedigree information has however, only recently been advocated for use in breeding programs for plants, see for example Durel et al. (1998) and Dutkowski et al. (2002). Davik and Honne (2005) also incorporate the pedigree information, but in a diallel setting.

In this paper, an approach is developed where the pedigree information is incorporated into the analysis of single trials for field crops. This involves the use of an additive relationship matrix (allowing for inbreeding) and hence the estimation of breeding values. In addition, as lines can be replicated in plant breeding trials, the analysis can also estimate non-additive genetic effects. In self-pollinated or inbred lines, these non-additive effects will reflect epistatic interactions because inbreeding will largely eliminate dominance effects. However, in hybrid crops both dominance and epistatic effects may be reflected in non-additive effects. Thus, a single analysis will allow *both* the selection of potential parents for future breeding programs using additive effects and promising commercial lines combining both additive and non-additive effects, i.e. the overall or total genetic effect. Genetic effects are treated as random effects which is consistent with the classical quantitative genetics approach and with the underlying aim which is selection of superior parents and commercial lines.

Information on genetic variance parameters normally only available from formal mating designs is produced as a bi-product of such an approach. A generalized definition of heritability is developed as the classical definition which arises from simple quantitative genetic models will not be appropriate for the mixed models considered in this paper.

The approach presented here is a mixed model form of an ‘‘extended’’ classical quantitative genetics model. It follows a long and ongoing tradition to attempt to model the gene to phenotype relationship (see Cooper and Hammer 2005 for a recent review).

This paper is structured as follows. The motivating example used to illustrate the approach is presented and an overview of current single trial analysis is outlined. Incorporation of pedigree information is discussed. This leads to defining a generalized heritability for the resulting mixed model, which is used to analyze the data and forms the basis of discussion that concludes the paper.

Materials and methods

Description of motivating example

The data considered in this paper was produced as part of the national Australian Grain Technologies’ (AGT) network of advanced trials. A total of 253 advanced wheat lines were tested as part of the 2004 Stage 3 trialling system. The pedigree of 129 of these lines was known, while the other lines consisted of lines with unknown pedigrees or filler lines. The genetic information of these latter lines is not normally relevant. However, the inclusion of these lines is important as they provide information about environmental variation. Data was collected and is presented on 14 trials, grown in locations around Australia. Most trials were laid out in rectangular arrays of plots, comprising 12 columns by 42 rows. One trial had 18 columns by 28 rows. Plots were sown 1.32 m × 5 m and reduced to 1.32 m × 3.2 m before anthesis by herbicide application. Seed was sown on a volume basis aiming for an average 200 seeds per square meter. Most lines were sown at all trials. Trials were designed using the nearest neighbour option within Agrobase II (Agronomix, Canada) with two replicates per line. Yield was recorded in grams per plot and converted to kilograms per hectare for presentation.

Standard statistical approach for single trial analysis

In this paper, the performance of lines is analyzed following the approach of Eckermann et al. (2001) for single trials. Genetic line effects are included as random which supports the ultimate aim of the analysis which is selection (Smith et al. 2005). The environmental variation often present in field trials is modeled according to Gilmour et al. (1997), who allow for the three possible sources of environmental variation, namely global, extraneous, and local. Here, in addition, design and randomization based terms are included (Cullis et al. 2006).

Thus, the statistical model fitted for a single trial, referred to as the *Standard* model is

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

The $(n \times 1)$ vector of phenotypic yield observations $\mathbf{y}$ is arranged as trial rows within columns, where $\boldsymbol{\tau}$ is a $(t \times 1)$ vector consisting of t fixed terms, and includes an overall or population mean for the lines with pedigrees and similarly one for the filler lines. Global field variation such as linear row or linear column effects and extraneous field variation such as that introduced through management practices (for example, harvest order and varying plot-size) or gradient effects are also included if required. $\mathbf{X}$ is the corresponding $(n \times t)$ design matrix.

The random vector of (overall) genetic line effects $\mathbf{g}^{(m \times 1)}$ of m lines with pedigree information is assumed normally distributed with mean zero and variance σ_g^2 $\mathbf{I}_m$, where $\mathbf{I}_m$ represents an $(m \times m)$ identity matrix. The corresponding design matrix $\mathbf{Z}_g$ is $(n \times m)$ and relates observations to lines.

The vector $\mathbf{u}^{(b \times 1)}$ consists of subvectors $\mathbf{u}_i^{(b_i \times 1)}$ where the subvector $\mathbf{u}_i$ corresponds to the ith random term. The corresponding design matrix $\mathbf{Z}_u^{(n \times b)}$ is partitioned conformably as $[\mathbf{Z}_{u_1} \cdots \mathbf{Z}_{u_b}]$. The subvectors are assumed mutually independent with variance $\sigma_i^2\mathbf{I}_{b_i}$. The subvectors include random terms for extraneous field variation such as random row or column variation and also design and randomization based blocking factors. A subvector $\mathbf{u}_g$ for filler line effects, that is lines included in the trial which do not have pedigree information, is also included.

The $(n \times 1)$ residual vector $\boldsymbol{\eta}$ represents local stationary variation. It is the sum of two independent vectors, $\xi^{(n\times1)}$ representing a spatially dependent mean zero random stationary process and $\zeta^{(n\times1)}$ a zero mean process representing measurement error. The measurement error term ζ has variance $\sigma_n^2\mathbf{I}_n$ and the spatial dependent term ξ has variance $\sigma_e^2\boldsymbol{\Sigma}^{(n\times n)}$, where the matrix $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_c \otimes \boldsymbol{\Sigma}_r)$, represents the Kronecker product between auto-regressive processes of order one (AR1) in the column and row directions, respectively. Thus, the residual vector $\boldsymbol{\eta}$ has distribution $\boldsymbol{\eta}\sim N(\mathbf{0}, \mathbf{R})$, where $\mathbf{R} = \sigma_e^2\boldsymbol{\Sigma} + \sigma_n^2\mathbf{I}_n$.

Thus, the line term $\mathbf{g}$ reflects the genetic variation and the fixed $\boldsymbol{\tau}$, random $\mathbf{u}$ and residual $\boldsymbol{\eta}$ terms reflect the design and conduct of the trial, and as such provide the underlying structure for non-genetic variation.

Extending the *Standard* statistical approach

The example data set contains lines with pedigree information and replication. Therefore the $(m \times 1)$ vector of (overall) genetic line effects $\mathbf{g}$ can be partitioned into a vector of additive line effects $\mathbf{a}$ and a vector of non-additive effects $\mathbf{i}$, such that $\mathbf{g} = \mathbf{a} + \mathbf{i}$. In general, the components of non-additive effects, dominance and epistasis cannot be distinguished by this method. However, because lines in this data set have been inbred for at least five generations they are assumed homozygous due to inbreeding, and therefore the dominance effect of a line is generally assumed to be zero. As a result the *non-additive* effects are referred to here as *epistatic* effects.

Incorporating this partitioned vector of genetic line effects into the *Standard* model 1, the extended model or *Pedigree* model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}_g\mathbf{a} + \mathbf{Z}_g\mathbf{i} + \mathbf{Z}_u\mathbf{u} + \boldsymbol{\eta} \tag{2}$$

where terms $\mathbf{X}\boldsymbol{\tau}$, $\mathbf{Z}_u$ $\mathbf{u}$, $\boldsymbol{\eta}$, and $\mathbf{Z}_g$ are defined as in the *Standard* model 1. The vector of epistatic effects $\mathbf{i}^{(m \times 1)}$ for the m lines with pedigree information has distribution $\mathbf{i} \sim N(\mathbf{0},\sigma_i^2\ \mathbf{I}_m)$.

The vector of additive effects $\mathbf{a}^{(m \times 1)}$ of the m lines with pedigree information has distribution, $\mathbf{a} \sim N(\mathbf{0},\sigma_a^2\mathbf{A})$, where $\mathbf{A}^{(m \times m)}$ is the known additive relationship matrix. If the additive relationship matrix is denoted as $\mathbf{A} = \{a_{jt}\}$, then the diagonal terms are $a_{jj} = 1 + F_j$, where F_j is the inbreeding coefficient of line j and the off-diagonal terms are $a_{jt} = 2f_{jt}$ where $2f_{jt}$ is the numerator of the coefficient of relationship (Wright 1922), and f_{jt} is the coefficient of parentage between lines j and t.

The additive relationship matrix and its inverse can be calculated in two ways. Firstly, $\mathbf{A}$ and its inverse $\mathbf{A}^{-1}$ can be calculated following the approach of Henderson (1976) but with modification for lines that have been selfed (see The additive relationship matrix-adjustment for self-fertilization). The package ASReml (Gilmour et al. 2005) provides speedy calculation of $\mathbf{A}^{-1}$ based on the efficient algorithm of Meuwissen and Luo (1992) and includes the modification for lines that have been selfed (see Appendix). Alternatively, $\mathbf{A}$ can be found by initially establishing a coefficient of parentage matrix, where the diagonal elements of the matrix are $f_{jj} = 0.5(1 + F_j)$ and the off-diagonal elements of the matrix are f_{jt}. The algorithm of Sneller (1994) can be used with an appropriate adjustment for selfing (see The coefficient of parentage matrix-adjustment for self-fertilization). To obtain the additive relationship matrix **A** it is sufficient to multiply all the elements of the coefficient of parentage matrix by two.

The vector of additive effects **a** and epistatic effect **i** are assumed to be mutually independent, so that the vector of overall or total genetic effects $\mathbf{g} = \mathbf{a} + \mathbf{i}$ has distribution $\mathbf{g} \sim N(0,\sigma_i^2\,\mathbf{I}_m + \sigma_a^2\,\mathbf{A})$.

Generally, whether a test line is taken to the next stage of the breeding and commercialization process will depend on how its performance compares to a control line. Therefore, the conditional probability that the genetic effect of the jth line is greater than that of the control line given the data [expressed as $P(g_j - g_{\text{cntl}} > 0\ |\mathbf{y})$] is used to rank lines. The control line should be included in the trial and have pedigree information.

Diagnostics for the models fitted include plotting a sample variogram for examining spatial covariance structure and residual plots (see Gilmour et al. 1997 for details). These models are fitted using the software ASReml (Gilmour et al. 2005). Estimation of variance components is by residual or restricted maximum likelihood (REML, Patterson and Thompson 1971), using the average information REML algorithm (Gilmour et al. 2005). The ASReml code to fit the *Pedigree* model Eq. 2 at one of the trials is included in ASReml code for fitting the *Pedigree* model Eq. 2.

Heritability generalized

Heritability is a measure used to quantify the percentage of total variation that can be explained by the genotypic component. Although the definition arises in a number of ways, it is based on a simple quantitative genetics model (Falconer and Mackay 1996) for a randomly mating population. Broad sense means line heritability is given by

$$H^2 = \sigma_g^2/(\sigma_g^2 + \sigma^2/r) \tag{3}$$

while narrow sense means line heritability requires a pedigree and a relationship matrix and is given by

$$h^2 = \sigma_a^2/(\sigma_a^2 + \sigma^2/r) \tag{4}$$

The model presented for analysis of trial data, Eq. 2, does not adhere to the standard assumptions. Thus, Eqs. 3 and 4 may not be appropriate.

Cullis et al. (2006) consider the problem of defining heritability in more complex settings. Their definition is based on average pairwise prediction error variance that is appropriate for general error covariance matrices and diagonal genetic covariance matrices.

To develop a general approach, heritability is defined as the squared correlation between the realized (or predicted) and the true genetic effect (Falconer and Mackay 1996). This definition implicitly assumes a single genetic effect, whereas in general we have a vector of genetic effects. In the standard quantitative model this is not an issue, because genetic effects have a ‘‘common heritability’’. In more complex models this no longer holds.

To reduce the genetic effect to a scalar quantity, consider a linear combination of the true genetic effects, namely $\mathbf{c}^T\ \mathbf{g}$, and the corresponding predicted genetic effects, namely $\mathbf{c}^T\tilde{\mathbf{g}}$. There are many choices for **c** and the derivation of generalized heritability results in a canonical set of vectors **c**.

A generic mixed model is used to present the approach. Thus, suppose

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}_g\mathbf{g} + \mathbf{Z}_u\mathbf{u} + \boldsymbol{\eta} \tag{5}$$

where $\mathbf{g} \sim N(\mathbf{0},\mathbf{G})$, $\mathbf{u} \sim N(\mathbf{0},\mathbf{U})$, and $\boldsymbol{\eta} \sim N(0,R)$. The models *Standard* Eq. 1 and *Pedigree* Eq. 2 are specific cases. Note that

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\tau},\mathbf{V})$$

where $\mathbf{V} = \mathbf{R} + \mathbf{Z}_g\ \mathbf{G}\mathbf{Z}_g^T + \mathbf{Z}_u\ \mathbf{U}\mathbf{Z}_u^T$.

For the genetic effect $\mathbf{c}^T\tilde{\mathbf{g}}$, the heritability is defined as

$$H_c^2 = \frac{\operatorname{cov}(\mathbf{c}^T\mathbf{g},\mathbf{c}^T\tilde{\mathbf{g}})^2}{\operatorname{var}(\mathbf{c}^T\mathbf{g})\operatorname{var}(\mathbf{c}^T\tilde{\mathbf{g}})} = \frac{\mathbf{c}^T\mathbf{G}\mathbf{Z}_g^T\mathbf{P}_v\mathbf{Z}_g\mathbf{G}\mathbf{c}}{\mathbf{c}^T\mathbf{G}\mathbf{c}}$$

where $\mathbf{P}_v = \mathbf{V}^{-1} - \mathbf{V}^{-1}\ \mathbf{X}\ (\mathbf{X}^T\ \mathbf{V}^{-1}\ \mathbf{X})^{-1}\ \mathbf{X}^T\ \mathbf{V}^{-1}$. We begin by choosing **c** to maximize the heritability subject to $\mathbf{c}^T\ \mathbf{G}\mathbf{c} = 1$ (normalization with respect to **G**).

We consider the Lagrangian $\text{Ł}_{\mathbf{c}}$, where,

$$\text{Ł}_{\mathbf{c}} = \mathbf{c}^T\mathbf{G}\mathbf{Z}_g^T\mathbf{P}_v\mathbf{Z}_g\mathbf{G}\mathbf{c} - \lambda(\mathbf{c}^T\mathbf{G}\mathbf{c} - 1) \tag{6}$$

and choose **c** to maximize $\text{Ł}_{\mathbf{c}}$. The detailed solution to this problem is presented in Generalized definition of heritability. The vector **c** that maximizes H_c^2 is an eigenvector of the matrix $\mathbf{Z}_g^T\ \mathbf{P}_v\ \mathbf{Z}_g\ \mathbf{G}$ with associated eigenvalue λ. In fact

$$\max_{\mathbf{c}} H_c^2 = \lambda$$

so that this eigenvalue is a component of the full heritability.

The full set of eigenvalues of $\mathbf{Z}_g^T$ $\mathbf{P}_v$ $\mathbf{Z}_g$ $\mathbf{G}$ will characterize the full heritability. Let $\lambda_1,\lambda_2,\ldots,\lambda_m$ be the full set of eigenvalues. Some of these eigenvalues will be zero because of constraints on $\tilde{\mathbf{g}}$. Suppose the first s are zero. The generalized heritability is defined as

$$H^2 = \frac{\sum_{i=1}^{m} \lambda_i}{m-s} = \frac{\sum_{i=s+1}^{m} \lambda_i}{m-s} \tag{7}$$

The classical quantitative genetics model, with m test lines each with r replicates (and total number of observations $n = mr$) is

$$\mathbf{y} = \mathbf{1}_n\mu + \mathbf{Z}_g\mathbf{g} + \boldsymbol{\eta} \tag{8}$$

where $\mathbf{Z}_g = \mathbf{I}_m \otimes \mathbf{1}_r$, $\mathbf{g} \sim N(\mathbf{0},\sigma_g^2\ \mathbf{I}_m)$ is the vector of genetic effects and $\boldsymbol{\eta} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$. In this case the matrix $\mathbf{Z}_g^T$ $\mathbf{P}_v$ $\mathbf{Z}_g$ $\mathbf{G}$ has one zero eigenvalue and $m - 1$ repeated eigenvalues that equal $H^2 = \sigma_g^2/\sigma_g^2 + \sigma^2/r$, the mean line heritability. Thus, Eq. 7 reduces to the mean line heritability.

In the classical quantitative genetic model Eq. 8, $\mathbf{g} \sim N(\mathbf{0},\ \sigma_a^2\mathbf{A})$ and the number of replicates $r = 1$, so that lines are related with additive relationship matrix $\mathbf{A}$, the generalized heritability is a narrow sense heritability and can be shown to be

$$h^2 = \frac{1}{m-1}\sum_{i=2}^{m} \frac{\varsigma_i\sigma_a^2}{\varsigma_i\sigma_a^2 + \sigma^2}$$

Again, $s = 1$. The ς_i are the eigenvalues of $(\mathbf{I}_m - \mathbf{P}_{\mathbf{1}_m})\mathbf{A}$; where $\mathbf{P}_{\mathbf{1}_m}$ is the projection matrix onto $\mathbf{1}_m$. This differs from the usual narrow sense heritability given by Eq. 4. The generalized definition takes into consideration the pedigree structure rather than implicitly assuming independence of lines.

In the *Pedigree* model we can obtain a broad sense heritability (H^2) by considering $\mathbf{G} = \sigma_i^2\ \mathbf{I}_m + \sigma_a^2\ \mathbf{A}$ and a narrow sense heritability (h^2) by considering $\mathbf{G} = \sigma_a^2\ \mathbf{A}$. However, it is not possible to present analytical solutions and numerical methods must be used to calculate the heritability.

Importantly, we can write

$$\mathbf{Z}_g^T\mathbf{P}_v\mathbf{Z}_g\mathbf{G} = \mathbf{I}_m - \mathbf{G}^{-1}\mathbf{C}^{ZZ}$$

where $\mathbf{C}^{ZZ}$ is the prediction error variance matrix for $\mathbf{g}$, so that eigenvalue calculations can be based on $\mathbf{I}_m - \mathbf{G}^{-1}\ \mathbf{C}^{ZZ}$.

For large problems an approximation to the generalized heritability may be very useful. Using the property that the trace of a matrix is the sum of the eigenvalues of that matrix, an approximate heritability is

$$H^2 = \left(1 - \frac{\mathrm{tr}(\mathbf{G}^{-1}\mathbf{C}^{ZZ})}{m}\right)$$

and the trace term can be found by summing element by element product of the two matrices. This ignores the possibility of zero eigenvalues.

Results

For the AGT data, the coefficient of parentage matrix was calculated using International Crop Information System (ICIS), which uses the algorithm of Sneller (1994). A modification due to selfing was implemented (see The coefficient of parentage matrix-adjustment for self-fertilization). This coefficient of parentage matrix was then used to calculate the additive relationship matrix $\mathbf{A}$.

For each trial, the *Standard* model and the *Pedigree* model were fitted. A summary of non-genetic or environmental variation is presented for each trial (Table 1). The column and row correlations of the stationary spatial variation from the *Pedigree* and *Standard* model were very similar, so that only those from the *Pedigree* model are presented (Table 1). The column correlation parameter was not significant for four trials. Notice that the row AR1 correlation is very large indicating strong smooth spatial variation at all trials. A measurement error term was significant ($P < 0.05$) at 13 trials; at one trial (Robinvale) it was not significant.

The *Standard* and *Pedigree* models each had the same environmental terms fitted, so that the *Standard* model was a sub-model of the *Pedigree* model. A residual or restricted maximum likelihood ratio test (REMLRT) is used to compare these models and test the significance of the additive component, but as the null hypothesis H_0 was on the boundary ($\sigma_a^2 = 0$), the reference distribution was nonstandard. The P value was approximated using a mixture of half χ_0^2 and half χ_1^2 (Self and Liang 1987; Stram and Lee 1994, but see Crianiceanu and Ruppert 2004 for a discussion on this approximation).

The *Pedigree* model was better than the *Standard* model at all trials indicating that the additive proportion of the overall genetic variation was (highly) significant (Table 2). The variance of the difference between a random term $\mathbf{g}$ and its Best Linear Unbiased Predictor (BLUP) $\tilde{\mathbf{g}}$ is known as the prediction error variance or var($\tilde{\mathbf{g}} - \mathbf{g}$). For all trials, the average estimated prediction error variance was lower under the *Pedigree* model, which was expected under a model

Table 1 Environmental terms fitted in the analysis of yield (tonne/ha) for each of the trials. All trials had a random block term added to account for the randomization of the trial design

Trial	Location	Environmental terms		Column[a] AR1	Row[a] AR1
		Random	Fixed		
1	Coomalbidgup[b]	Spl(row) column[c]	Linear row, harvest order, row:(linear column)	0.54	0.83
2	Coonalpyn[b]	Column		0	0.84
3	Kapunda[b]		Linear column	0.38	0.79
4	Merredin[b]	Column		0	0.91
5	Mingenew[b]		Linear column	0.21	0.81
6	Minnipa[b]	Spl(row)[c]	Linear row	0.43	0.87
7	Narrabri[b]	Column, row		0.40	0.81
8	Narrandera[b]		Linear column	0.35	0.84
9	Pinnaroo[b]	Spl(column), column[c]	Linear column, plot size, linear row	0.32	0.47
10	Robinvale	Row		0.18	0.71
11	Roseworthy[b]			0.48	0.92
12	Scaddon[b]	Column	Linear row	0	0.92
13	Temora[b]	Column	Linear row	0	0.79
14	Wongan Hills[b]	Row	Linear row	0.64	0.93

[a] Column and row correlations presented were from the *Pedigree* model

[b] A measurement error term was fitted at these trials

[c] Spl(*term*) indicates a smoothing spline (Verbyla et al. 1999) of *term* was fitted

which describes the underlying distribution of **g** more accurately. Note that the prediction error variance estimated under both models is approximate because the variance components in the prediction error variance are replaced by their estimated REML values. This is also true of the BLUPs and hence these are empirical BLUPs or E-BLUPs.

The total genetic variation found at the 14 trials, varied enormously. Merredin, Wongan Hills, and Robinvale had comparably small total genetic variation and Narrabri by far the greatest total genetic variation. At a particular trial, the overall genetic variation of **g** being predicted by the *Pedigree* model was higher than under the *Standard* model (Table 3). In some trials the difference was substantial. For the *Pedigree* model, the proportion of the total genetic variation represented by the additive component varied across trials. At five trials all genetic variation was found to be additive. The REML estimate of the epistatic variance at these trials was on the boundary.

The broad sense heritability of the *Standard* model was higher than the *Pedigree* model (Table 3). This higher heritability is likely to be the result of an upward bias as a result of an incorrect model (Costa e Silva et al. 1994). In particular, the *Standard* model assumes independence of line when in fact correlation (in the form of the **A** matrix) exists between lines. The narrow sense heritability which is able to be determined under the *Pedigree* model is a more appropriate indicator of heritability (Viana 2005) and as such is the preferable indicator.

There were high correlations between the overall total genetic E-BLUPs of the *Standard* ($\tilde{\mathbf{g}}$) and the *Pedigree* model ($\tilde{\mathbf{g}} = \tilde{\mathbf{a}} + \tilde{\mathbf{i}}$) (Table 3). This agreement

Table 2 Tests of significance for improvement in the prediction of yield (tonne/ha) resulting from the *Standard* versus *Pedigree* model and the average prediction error variance of the total genetic effect (**g**) for the *Standard* and the *Pedigree* model

Trial	Location	REMLRT[a]	*P* value of additive component	Average prediction error variance	
				Standard	Pedigree
1	Coomalbidgup	8.32	0.0020	234	226
2	Coonalpyn	29.6	< 0.0001	184	168
3	Kapunda	15.7	< 0.0001	171	160
4	Merredin	12.2	0.0002	48.9	46.9
5	Mingenew	12.8	0.0002	164	157
6	Minnipa	5.90	0.0076	57.8	56.5
7	Narrabri	19.7	< 0.0001	360	349
8	Narrandera	20.9	< 0.0001	95.8	89.9
9	Pinnaroo	18.8	< 0.0001	130	114
10	Robinvale	19.6	< 0.0001	59.2	53.2
11	Roseworthy	18.7	< 0.0001	178	168
12	Scaddon	3.24	0.0359	160	155
13	Temora	14.2	< 0.0001	152	140
14	Wongan Hills	15.9	< 0.0001	52.3	48.6

[a] Residual or restricted maximum likelihood ratio test of H_o, $\sigma_a^2 = 0$

Table 3 Total or overall genetic variance σ_g^2 of yield (tonne/ha) at each of the trials from the *Standard* and *Pedigree* models and broad (H^2) and narrow sense (h^2) heritability (calculated using the generalized heritability formula 7)

Trial	Location	Standard		Pedigree				Correlation[b]	
		σ_g^2	H^2	σ_g^2	Percent additive[a]	H^2	h^2	$(\tilde{\mathbf{g}}, \tilde{\mathbf{a}} + \tilde{\mathbf{i}})$	$(\tilde{\mathbf{g}}, \tilde{\mathbf{a}})$
1	Coomalbidgup	90.26	0.69	110.29	77.08	0.64	0.42	0.986	0.934
2	Coonalpyn	59.91	0.71	68.20	100.00	0.60[c]	0.60	0.971	0.971
3	Kapunda	19.06	0.24	26.15	100.00	0.22[c]	0.22	0.813	0.813
4	Merredin	2.27	0.47	2.38	52.66	0.43	0.18	0.961	0.767
5	Mingenew	45.67	0.70	55.05	81.30	0.64	0.45	0.984	0.940
6	Minnipa	8.89	0.81	9.96	63.12	0.77	0.38	0.996	0.911
7	Narrabri	375.34	0.82	441.82	81.36	0.77	0.54	0.993	0.958
8	Narrandera	17.33	0.73	23.82	100.00	0.66[c]	0.66	0.982	0.982
9	Pinnaroo	14.12	0.40	16.12	100.00	0.32[c]	0.32	0.872	0.872
10	Robinvale	4.23	0.58	4.68	92.65	0.48	0.42	0.944	0.921
11	Roseworthy	55.45	0.71	60.31	76.10	0.64	0.41	0.982	0.918
12	Scaddon	29.32	0.56	37.88	77.95	0.53	0.35	0.977	0.924
13	Temora	22.18	0.47	29.58	100.00	0.41[c]	0.41	0.930	0.930
14	Wongan Hills	3.81	0.64	5.17	100.00	0.56[c]	0.56	0.966	0.966

[a] Additive variation as a percent of the total or overall genetic variation (σ_g^2) of the *Pedigree* model

[b] $\tilde{\mathbf{g}}$ is the E-BLUP of $\boldsymbol{g}$ from Eq. 1 and $\tilde{\mathbf{a}}$ and $\tilde{\mathbf{i}}$ are the E-BLUPs of $\boldsymbol{a}$ and $\boldsymbol{i}$, respectively, from Eq. 2

[c] The epistatic variance component was on the boundary at these trials, therefore H^2 and h^2 are equivalent

was reflected in terms of the top 20 ranking lines. Across all trials, an average of 80% of the top 20 ranking lines were the same under both models.

In trials, where the epistatic component of the genetic variation was significant, the correlations between the genetic E-BLUPs of the *Standard* ($\tilde{\mathbf{g}}$) and the *additive* genetic E-BLUPs of the *Pedigree* model ($\tilde{\mathbf{a}}$) were lower (Table 3), than in the comparison of the correlation between the overall total genetic E-BLUPs of the *Standard* and *Pedigree* model. However, the lower correlations do not reflect the differences in the top 20 ranking lines. If decisions on the best potential parents were based on the predicted yield under the *Standard* model rather than on the *additive* predicted yield of the *Pedigree* model then 30% of these decisions would be incorrect (Fig. 1).

Discussion

This paper develops a statistical approach that can be used in crop breeding trials with pedigree information and replication of lines. It involves fitting a model referred to here as the *Pedigree* model that predicts additive and non-additive genetic effects of test lines and simultaneously models spatial variation. In inbred lines the non-additive effects will represent epistatic effects. However, with hybrid crops, non-additive effects will also include dominance effects. The approach offers advantages over current methods in that it enables the selection of the best performing line for commercial release and the best parents for further crosses in a single analysis and from a standard crop breeding trial.

The overall genetic effect or value of a line is obtained from the sum of additive and non-additive effects (epistatic effects). Both terms are deemed important in determining the commercial worth of a line as it is the overall performance of a line and therefore overall (or total) genetic value that is important. The epistatic component is clearly important at the majority of trials. In some cases it makes a line suitable for a particular environment or conversely it is what deems it unsuitable. The total genetic variation explained is higher under the *Pedigree* model. The *Pedigree* model has also been shown to predict a genetic effect that has a lower prediction error variance on average than that determined under the *Standard* model and therefore is a preferable means of obtaining the genetic value of a line.

The additive effects of the *Pedigree* model are breeding values and as such are the preferable means of determining potential parents for breeding programs. The breeding value of every line (with pedigree information) can be obtained without resorting to specialized trial designs such as diallel crosses which require extra resources and are limited in the number of lines that can be included. Using the *Standard* model to determine potential parents may lead to the selection of lines that do not have the highest breeding value and therefore may lead to subsequent breeding programs that are not optimal.

A concern with this method is that the relationship matrix **A** is based on expected (average) relationships

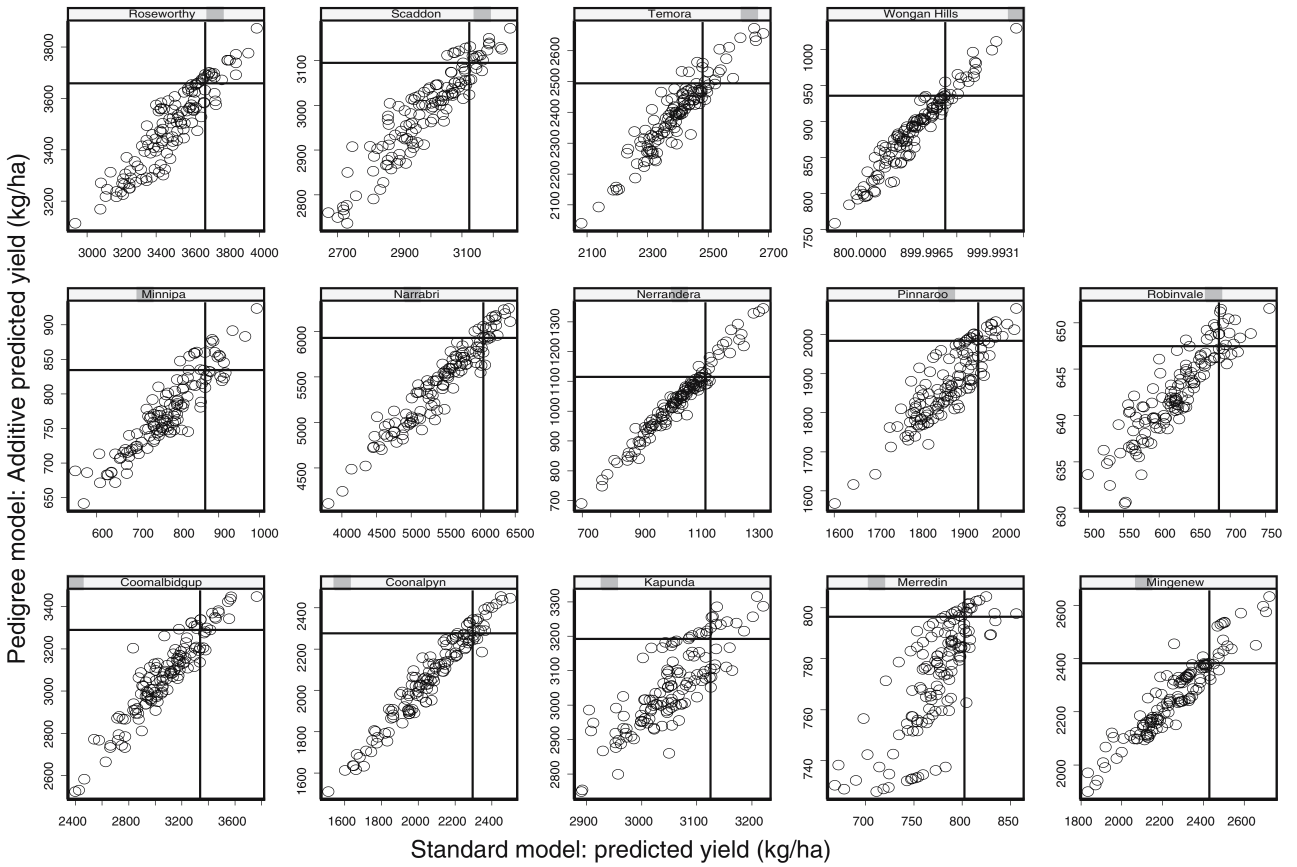

Fig. 1 The *additive* predicted (breeding value) yield (kg/ha) for the *Pedigree* model plotted against the predicted yield (kg/ha) of the *Standard* model. *Horizontal and vertical lines* show the cut off for the top 20 ranking varieties under the *Pedigree* and *Standard* model, respectively. Each trial has been plotted on an individual scale to enhance the presentation

between individuals. For instance full-siblings will have identical coefficients of parentage with other individuals, even though it is likely they do not share identical genotypes. In particular, in plant populations where selection of lines over many generations is undertaken, the relationship between full siblings may be much greater than expected and could be much higher with one parent than the other. If genotypic information was available (in the form of marker data for instance) then a more accurate estimation of the relationship between individuals could be determined (see Crepieux et al. 2004). The selection of lines that occurs in plant breeding trials may also result in a biased estimate of the additive variance. Van der Werf and de Boer (1990) suggest bias is eliminated when relationship information of all selected ancestors is included. In the example presented here, every attempt was made to do this with lines of known pedigree, so that in most cases ancestry was traced back several generations and used in the formation of the relationship matrix. Van der Werf and de Boer (1990) also found that ''bias was smaller in a small population and (or) when selection had been practised for just a few generations''. This phenomena is discussed by Walsh (2005), and may help counteract bias introduced by selection.

The development of a generalized definition of heritability enables pedigree and environmental information to be taken into consideration in models which do not conform to the simple quantitative model which assumes independence of lines.

The approach presented requires a separate fit for each trial, so that genotypic effects vary across environments. This means that these effects are confounded with the genotype by environment interaction. Thus, it is important to extend the ideas of this paper to multienvironment trials which allow the genotype by environment effects to be quantified, and this is a current topic of research.

Acknowledgements Thanks to the Grain Research Development Corporation (GRDC) who fund the research of Helena Oakey through a Grains Industry Research Scholarship. Arūnas Verbyla and Brian Cullis also thank GRDC. We are grateful to the breeders and technical staff of Australian Grain Technologies for access to the grain yield and pedigree data used in this study.

Appendix

The additive relationship matrix-adjustment for self-fertilization

In plant breeding, the test lines that are included in trials are often the result of five or six generations of self-fertilization. The method of Henderson (1976) was developed for use in animal pedigrees, and as such requires for any particular line that is a result of n generations of self-fertilization that all the previous $n - 1$ generations of lines involved in its development are included in the pedigree. Clearly, in plant breeding trials where each test line has undergone the self-fertilization process up to n times, this would require an (unnecessarily) large pedigree to be recorded in order to obtain accurate estimates of a_{jt}. A modification in the calculation of the inbreeding coefficient F_j and therefore in the a_{jj} value, can be incorporated into the algorithm, so that it is unnecessary to include the $n - 1$ generation of lines in the pedigree, just the number of generations n of self-fertilization need be recorded for each line.

If both parents, s and d of individual j are known then, the adjustment under n generations of self-fertilization is given by

$$a_{jj} = 2 - 0.5^{n-1} + 0.5^n a_{sd} \qquad (9)$$

which reduces to Henderson's equation under no self-fertilization, i.e. $n = 1$, also note that a_{jj} tends to 2 as n tends to infinity.

Under n generations of self-fertilization, when one parent is known or when no parents are known the value of a_{jj} can be shown to be

$$a_{jj} = 2 - 0.5^{n-1}.$$

The coefficient of parentage matrix-adjustment for self-fertilization

The method of Sneller (1994) does not take into consideration self-fertilization. A modification in the calculation of the inbreeding coefficient F_j and therefore f_{jj} is necessary when dealing with individuals that have been self-fertilized for n generations.

Under self-fertilization, the coefficient of parentage f_{jj} of j in the nth generation is given by half Eq. 9 as follows:

$$f_{jj} = 1 - 0.5^n + 0.5^n f_{sd} \qquad (10)$$

When one parent is known or when no parents are known the value of f_{jj} is $f_{jj} = 1 - 0.5^n$

ASReml code for fitting the *Pedigree* model 2

The following is the code for the *.as* ASReml file used for fitting the *Pedigree* model to a trial.

```
single trial model
 block 2       #block term -factor with 2 levels
 column 12     #column term -factor with 12 levels
 row 42        #row term -factor with 42 levels
 yield         #response variable
 lrow          #linear row term -variable centred at mean row
 lcol          #linear column term -variable centred at mean column
 line 253      #factor with 253 levels
 ped 2         #factor with two levels: 1=lines with pedigree, 2=filler lines
 knownped 129  #factor with levels 1:129, for lines with pedigrees
               #and "NA"s for filler lines
 filler 124    #factor with levels 1:124 for filler lines
               #and "NA"s for lines with pedigrees
stage3.giv     #the A inverse file
stage3rba.asd !skip1 !mvinclude!slow !maxit 20 #the data file
yield ˜  mu ped, #the fixed terms of the model
!r block  filler knownped  ide(knownped) units #the random terms of the model
!f mv            #estimate missing values
1 2 1            #number of sites, number of R-structure components, number of G-structures
12 column AR1    #number of columns with AR1 structure
42 row AR1       #number of rows with AR1 structure
knownped 1       #G structure for lines with pedigree
knownped 0 GIV1 !GP  #specifies the file stage3.giv as the corresponding G structure
predict ped 1 knownped !vpv !onlyuse knownped ide(knownped) #pev matrix (total)
predict ped 1 knownped !vpv !onlyuse knownped #pev matrix (additive)
```

The *stage3.giv* is a file containing the inverse of the additive relationship matrix. ASReml requires a file which is just the lower triangle of this matrix. It is important to ensure that the numbering of lines in *knownped* factor corresponds directly to the ordering of rows and columns in the "*.giv*" file, so that row one and column one of the **A** inverse matrix contain the additive relationships of individual 1, which should correspondingly be labeled as 1 in the *knownped* factor. The "*.giv*" file can be created in ASReml if a pedigree file is supplied, and ASReml now implements the adjustment for inbred lines.

The *stage3rba.asd* is a text file containing the data. The *knownped* and *filler* columns have been created from the *line* column in which the lines are numbered from 1:253. In particular, the *knownped* is a column which has been defined as a factor with 129 levels. The levels correspond to the lines with known pedigree. It has "NA"s in the positions which correspond to filler lines. The *filler* is a column which has been defined as a factor with 124 levels, filler lines are defined as 1:124 and lines which have pedigree have "NA"s. The *ped* column is a factor which has two levels so that separate overall means can be fitted for filler lines and lines with known pedigree.

The additive genetic effect for each line is fitted by including the term *knownped* in the random part of the model specification and the epistatic genetic effect for each line is fitted by including the term *ide(knownped)* in the random part of the model specification, the *units* term is the measurement error term.

The last two lines are the predict statements to obtain the elements of the estimated prediction error variance matrix, so that the generalized heritability can be calculated. ASReml places these in the "*.pvs*" file. The estimated prediction error variance of the total genetic effects is used for calculating a broad sense heritability and those of the additive effects for calculating a narrow sense heritability. Calculation of generalized heritability was carried out using R (R Development Core Team 2005). The R code is available from the corresponding author.

Generalized definition of heritability

The Lagrangian given by Eq. 6 is to be optimized with respect to $\mathbf{c}$. Thus, differentiating $\text{Ł}_{\mathbf{c}}$ with respect to $\mathbf{c}$ and setting to zero, we find

$$\mathbf{Z}_g^T\mathbf{P}_v\mathbf{Z}_g\mathbf{G}\mathbf{c} = \lambda\mathbf{c}. \tag{11}$$

Thus, $\mathbf{c}$ is an eigenvector of the matrix $\mathbf{Z}_g^T\ \mathbf{P}_v\ \mathbf{Z}_g\ \mathbf{G}$ with eigenvalue λ. Not only can the $\mathbf{c}$ that maximizes the squared correlation be found, but a complete set of eigenvectors $\mathbf{c}$ for $\mathbf{Z}_g^T\ \mathbf{P}_v\ \mathbf{Z}_g\ \mathbf{G}$ with associated eigenvalues. Notice that from Eq. 11

$$\begin{aligned}\mathbf{c}^T\mathbf{G}\mathbf{Z}_g^T\mathbf{P}_v\mathbf{Z}_g\mathbf{G}\mathbf{c} &= \lambda\mathbf{c}^T\mathbf{G}\mathbf{c}\\ &= \lambda\end{aligned}$$

using the constraint. Thus, the eigenvalues provide a set of heritability components that can be used to provide an overall measure of heritability.

From results on mixed models, $\mathbf{G}\mathbf{Z}_g^T\mathbf{P}_v\ \mathbf{Z}_g\ \mathbf{G} = \mathbf{G} - (\mathbf{Z}_g^T\ \mathbf{S}\mathbf{Z}_g + \mathbf{G}^{-1})^{-1}$ where $\mathbf{S} = \mathbf{R}^{-1} - \mathbf{R}^{-1}\ \mathbf{X}(\mathbf{X}^T\ \mathbf{R}^{-1}\mathbf{X})^{-1}\ \mathbf{X}^T\ \mathbf{R}^{-1}$. Now $(\mathbf{Z}_g^T\ \mathbf{S}\mathbf{Z}_g + \mathbf{G}^{-1})^{-1} = \mathbf{C}^{ZZ}$ is the partition of the inverse of the mixed model coefficient matrix corresponding to $\mathbf{g}$. This latter term $\mathbf{C}^{ZZ}$ is also equivalent to the prediction error variance matrix (i.e. $\mathrm{var}(\tilde{\mathbf{g}} - \mathbf{g})$), an estimate of which is available in the software ASReml (Gilmour et al. 2005) via the predict statement. So

$$\mathbf{Z}_g^T\mathbf{P}_v\mathbf{Z}_g\mathbf{G} = \mathbf{I}_m - \mathbf{G}^{-1}\mathbf{C}^{ZZ} \tag{12}$$

and eigenvalues of this matrix are required to determine the generalized heritability.

References

Besag J, Kempton R (1986) Statistical analysis of field experiments using neighbouring plots. Biometrics 42:231–251

Brown D, Tier B, Reverter A, Banks R, Graser H (2000) OVIS: a multiple trait breeding value estimation program for genetic evaluation of sheep. Wool Technol Sheep Breed 48

Cooper M, Hammer GL (2005) Preface to special issue: complex traits and plant breeding—can we understand the complexities of gene-to-phenotype relationships and use such knowledge to enhance plant breeding outcomes? Aust J Agric Res 56:869–872

Costa e Silva J, Borralho NMG, Potts BM (1994) Additive and non-additive genetic parameters from clonally replicated and seedling progenies of *Eucalyptus globulus*. Genetics 138:963–971

Crepieux S, Lebreton C, Servin B, Charmet G (2004) Quantitative trait loci QTL detection in multicross inbred designs: recovering QTL identical-by-descent status information from marker data. Genetics 168:1737–1749

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

Cullis BR, Gleeson AC (1991) Spatial analysis of field experiments—an extension to two dimensions. Biometrics 47:1449–1460

Cullis BR, Smith A, Coombes N (2006) On the design of early generation variety trials with correlated data. J Agric Biol Environ Stat (in press)

Davik J, Honne B (2005) Genetic variance and breeding values for resistance to wind-borne disease [*Sphaeotheca macularis* (wallr. exfr.)] in strawberry (*Fragaria x ananassa* duch.) estimated by exploring mixed models and spatial models and pedigree information. Theor Appl Genet 111:256–264

Durel CE, Laurens F, Fouillet A, Lespinasse Y (1998) Utilization of pedigree information to estimate genetic parameters from large unbalanced data sets in apple. Theor Appl Genet 96:1077–1085

Dutkowski GW, Costa e Silva J, Gilmour AR, Lopez GA (2002) Spatial analysis methods for forest genetic trials. Can J For Res 32:2201–2214

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

Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics, 4th edn. Longman Group Ltd

Gilmour AR, Cullis B, Verbyla AP (1997) Accounting for natural and extraneous variation in the analysis of field experiments. J Agric Biol Environ Stat 2:269–293

Gilmour AR, Cullis BR, Gogel B, Welham SJ, Thompson R (2005) ASReml, user guide. Release 2.0. VSN International Ltd, Hemel Hempstead

Henderson CR (1976) A simple method for computing the inverse of a numerator relationship matrix used in the prediction of breeding values. Biometrics 32:69–83

John J, Ruggiero K, Williams E (2002) ALPHA(n)-designs. Aust NZ J Stat 44:457–465

Martin R, Eccleston J, Chan B (2004) Efficient factorial experiments when the data are spatially correlated. J Stat Plan Inference 126:377–395

Meuwissen THE, Luo Z (1992) Computing inbreeding coefficients in large populations. Genet Sel Evol 24:305–313

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

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

Self SG, Liang K-Y (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J Am Stat Assoc 82:605–610

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 143:1–14

Sneller CH (1994) SAS programs for calculating coefficients of parentage. Crop Sci 34:1679–1680

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

Topal A, Aydin C, Akgun N, Babaoglu M (2004) Diallel cross analysis in durum wheat (*Triticum durum* Desf.) identification of best parents for some kernel physical features. Field Crops Res 87:1–12

Verbyla AP, Cullis BR, Kenward M, Welham S (1999) The analysis of designed experiments and longitudinal data using smoothing splines (with discussion). Appl Stat 48:269–311

Viana JMS (2005) Dominance, epistasis, heritabilities and expected genetic gain. Genet Mol Biol 28:67–74

Walsh B (2005) The struggle to exploit non-additive variation. Aust J Agric Res 56:873–881

van der Werf J, de Boer IJM (1990) Estimation of additive genetic variances when base populations are selected. J Anim Sci 68:3124–3132

Wright S (1922) Coefficients of inbreeding and relationship. Am Nat 56:330–338