Theor Appl Genet (2014) 127:2193–2210
DOI 10.1007/s00122-014-2373-0

ORIGINAL PAPER

Factor analytic and reduced animal models for the investigation of additive genotype-by-environment interaction in outcrossing plant species with application to a *Pinus radiata* breeding programme

Brian R. Cullis · Paul Jefferson · Robin Thompson · Alison B. Smith

Received: 26 February 2014 / Accepted: 25 July 2014 / Published online: 22 August 2014

Abstract

Key message **Modelling additive genotype-by-environment interaction is best achieved with the use of factor analytic models. With numerous environments and for outcrossing plant species, computation is facilitated using reduced animal models**.

Abstract The development of efficient plant breeding strategies requires a knowledge of the magnitude and structure of genotype-by-environment interaction. This information can be obtained from appropriate linear mixed model analyses of phenotypic data from multi-environment trials. The use of factor analytic models for genotype-by-environment effects is known to provide a reliable, parsimonious and holistic approach for obtaining estimates of genetic correlations between all pairs of trials. When breeding for outcrossing species the focus is on estimating additive genetic correlations and effects which is achieved by including pedigree information in the analysis. The use of factor analytic models in this setting may be computationally prohibitive when the number of environments is moderate to large. In this paper, we present an approach that uses an approximate reduced animal model to overcome the computational issues associated with factor analytic models for additive genotype-by-environment effects. The approach is illustrated using a *Pinus radiata* breeding dataset involving 77 trials, located in environments across New Zealand and south eastern Australia, and with pedigree information on 315,581 trees. Using this approach we demonstrate the existence of substantial additive genotype-by-environment interaction for the trait of stem diameter measured at breast height. This finding has potentially significant implications for both breeding and deployment strategies. Although our approach has been developed for forest tree breeding programmes, it is directly applicable for other outcrossing plant species, including sugarcane, maize and numerous horticultural crops.

Introduction

One of the key determinants of genetic gain in plant improvement programmes is the accurate prediction of the genetic value of an individual. The definition of genetic value in the plant breeding context, requires careful consideration, as it is necessarily not only dependant on the traits of interest, but also the target set of environments. In forest tree breeding programmes, deployment of elite germplasm will occur across a well-defined, but heterogeneous, set of target environments. The Radiata Pine Breeding Company (RPBC) programme, for example, aims to breed and provide germplasm for deployment across New Zealand, the central and southern tablelands of New South Wales and Tasmania. Informed breeding strategies and selection of elite germplasm require a thorough examination of the extent and nature of genotype-by-environment (G × E) interaction.

Communicated by Jose Crossa.

B. R. Cullis · A. B. Smith (✉)
National Institute for Applied Statistics Research Australia,
University of Wollongong, Wollongong, Australia
e-mail: alismith@uow.edu.au

B. R. Cullis
Computational Informatics, CSIRO, Canberra, Australia

P. Jefferson
Radiata Pine Breeding Company, Rotorua, New Zealand

R. Thompson
Rothamsted Research, Harpenden and Queen Mary College,
London, UK

The phenotypic panel for assessing G × E is a so-called multi-environment trial (MET). METs consist of (genetic) trials grown over many years and locations which are placed within the target set of environments. Fully efficient (that is, one-stage) mixed model approaches for analysing plant improvement MET datasets are in widespread use in Australia (Cullis et al. 2010). These approaches are usually based on the methods of Smith et al. (2001) which accommodate all aspects of trial design and allow for complex modelling of the variance structure of the residuals. The key aspect of the Smith et al. (2001) approach is the use of factor analytic (FA) models for modelling the variance structure of the G × E effects. Smith et al. (2001) originally only considered modelling the total genotype-by-environment effects where information on pedigrees is not included in the analysis. Burgueno et al. (2011) demonstrated the utility of the FA model for modelling G × E using six METs in potatoes, maize and wheat. Each of these MET datasets had small numbers of genotypes and relatively small numbers of environments, but they were able to demonstrate an advantage of the FA model over other models when there was complex G × E. Oakey et al. (2006, 2007) extended the approach of Smith et al. (2001) to incorporate modelling both additive and non-additive effects in a single and multi-environment trial setting. Cullis et al. (2010) applied these extensions to the analysis of yield and oil data from a series of canola breeding trials.

One of the many advantages of using FA models in the analysis of large MET datasets which possess a high degree of imbalance, is the ability to adequately capture the often complex variance structure without the use of an excessive number of variance parameters (Kelly et al. 2007). As the number of environments, t, becomes large, the number of variance parameters for the unstructured variance model (that is, $t(t+1)/2$) increases to unacceptably high levels and hence it is impractical and inefficient to consider fitting the unstructured variance model to MET datasets with moderate to large t and poor connectivity. Baltunis et al. (2010), Raymond (2011), Apiolaza (2012) have recently considered the analysis of tree breeding MET datasets which had moderate to poor connectivity and moderate $t = 8$, 26 and 18, respectively. The variance structure of the additive genotype-by-environment effects was modelled using $t(t-1)/2$ pairwise bivariate unstructured variance models. The approach is inherently inefficient, is largely uninformative and often results in estimated genetic correlations between environments which are inadmissable (that is, greater than 1 or less than −1).

Another advantage of the use of FA models is that the (not uncommon) case of a less than full rank variance structure for the G × E effects can be accommodated, provided that an appropriate estimation algorithm is implemented. In the case of FA models for G × E effects in the absence of pedigree information or residual G × E effects when pedigree information is included, the approach of Thompson et al. (2003) may be used. This approach was extended by Kelly et al. (2009) for the case of FA models for additive G × E effects.

There has been limited adoption of FA models in the analysis of tree breeding MET datasets. Costa e Silva et al. (2006) used FA models for the analysis of stem diameter measured at breast height (DBH) from a MET dataset involving 15 *Eucalyptus globulus* progeny trials established in four states of Australia. They fitted FA models of order 1 and 2 to the sub-race and family within sub-race by environment effects, respectively. These models provided a good fit to the variance structures and demonstrated the presence of substantial G × E for both sets of effects with estimated genetic correlations ranging from −0.23 to 1.00 and 0.25 to 1.00, respectively.

Hardner et al. (2010) also used FA models for the analysis of DBH in a MET dataset involving 841 hybrid clones of eucalypts sown across 21 saline environments in Australia. They fitted FA models to the family and genotype within family by environment effects, respectively. Due to the small number of families (8), attempts to fit high-order FA variance models failed. However, FA models for the genotype within family effects revealed substantial G × E, with the estimated genetic correlations ranging from −0.53 to 0.99.

Zapata-Valenzuela (2012) used an FA1 model for the variance structure of the additive genetic by environment effects in the analysis of a MET dataset on loblolly pines with 16 environments and 463 clones. They investigated two other variance models, namely the compound symmetric and heterogenous variance common correlation model. Although there was no formal approach to model selection the FA model provided an improvement in fit compared to the other two models.

One of the key advantages of the FA model is its links with multiple regression and principal component analysis. The FA model can be formulated as a random genetic regression of genetic effects on (unknown) environmental covariates (that is, factor loadings), with a different slope (factor score) for each genotype. This regression is termed a latent regression model and simple assumptions regarding the distribution of the slopes and the residual term leads to the FA variance structure for the between environment genetic covariance matrix. The environmental covariates can be rotated to be orthogonal, resulting in the so-called principal component solution. This then allows for a meaningful examination of the nature and extent of G × E using graphical tools such as biplots (Kempton 1984), latent regression plots (Thompson et al. 2003) and heatmaps of the estimated between environment genetic correlation matrix with rows and columns ordered using clustering or mixture model approaches (Cullis et al. 2010). Hardner et al. (2010) used the biplot to explore the G × E for the genotype within family by environment effects, demonstrating marked rank changes between environments.

The aim of this paper is to present an approach for investigating additive genotype-by-environment interaction in outcrossing plant species using MET datasets with moderate to large numbers of trials. We illustrate the method with an example from the RPBC breeding programme comprising $t = 77$ trials with pedigree information on 315,581 trees. The approach of Cullis et al. (2010) would be computationally prohibitive for a dataset of this size, since it would involve a set of mixed model equations (MME) of the order of 21e6. We present an approach that is based on the reduced animal model of Quass and Pollack (1980). Our so-called approximate reduced animal model results in a significant reduction in the dimension of the MME and is used both for the estimation of variance parameters, including those associated with the FA model, and the prediction of random effects, including breeding values for both backward and forward selections.

Motivating example

The MET dataset comprised 77 trials grown in a range of environments across New Zealand and New South Wales (Australia) with planting dates spanning the period 1968 to 2005. The trait of interest in this paper is stem diameter (cm) measured at breast height (DBH). Trial information is summarised in Table 1. A total of 34 trials comprised trees predominantly from open pollinated (OP) families with the remaining trials comprising trees from closed pollinated (CP) families. Trees derived from an OP family are those for which only one parent is known. Progeny derived from a CP family are those in which both parents are known. The CP families were produced by crossing a small number of male trees with a large number of female trees. Clonal material was used in 7 trials (E36, E37, E38, E39, E42, E43, E44). Many of the trials contained progeny from so-called controls which were mixed parent seedlots in which the individual parents could not beidentified. Excluding controls, the number of families per trial ranged from 19 to 942 (see Table 1). Note that in the OP and CP trials there is no replication of individuals since all trees are $\mathbf{F}_1$ progeny of OP or CP families, so there is a one-to-one correspondence between plots and trees.

To describe the experimental designs for each trial used in this study we utilise many of the basic concepts and nomenclature found in Bailey (2008), which include the definition of so-called plot (or blocking) structures, treatment structures, the observational unit and experimental unit for comparative experiments. Here the treatment structure is the genetic material (i.e. trees) which is grouped (for the majority of trials) into so-called genetic sets. For each trial, the genetic sets contained roughly the same number of families and each family contained the same number of individuals. The plot structure (hereafter referred to as blocking structure) for most trials was blocks, main plots within blocks and plots (ie. single trees) within main plots within blocks. For trials E17, E42, E43 and E44 the blocking structure was blocks, main plots within blocks, incomplete blocks within main plots within blocks and plots within incompleteblocks within main plots within blocks. For trials E3, E55, E65, E66, E71, E72, E73 and E77 the blocking structure was blocks, main plots within blocks, multi-tree plots in main plots within blocks and plots within multi-tree plots within main plots within blocks. The randomisation of genetic material to plots obeyed the blocking structure used for each trial, in that genetic sets were randomly assigned to main plots within blocks, and individuals within genetic sets were randomly assigned to either plots within main plots within blocks or groups of individuals from the same family were randomly assigned to multi-tree plots within main plots within blocks (for the eight trials with multi-tree plots). An additional level of restriction of the randomisation occurred for those trials with incomplete blocks. Lastly for the seven trials without main plots (and hence genetic sets) the blocking structure was blocks and plots within blocks.

This process resulted in at least a single tree from each family in each block (see Table 1). Note that the randomisation of genetic sets to main plots within blocks is due to Schutz and Cockerham (1966) and was widely used by the RPBC for many years from the early 1970s. Use of this design, which is now discontinued, is likely to result in far greater accuracies for the comparison of families in the same genetic set as opposed to those in different genetic sets.

The total number of records in the MET dataset was 323,804 corresponding to 312,848 trees. Pedigree information was available on a total of 315,581 trees, which comprised 2,733 parental trees and the 312,848 progeny that were grown in the trials. No parental trees were grown in the trials. The aim of the analysis of the data is to explore additive genotype-by-environment (A × E) interaction and to obtain predicted breeding values (additive genetic effects) to enable selection of individuals for use as parents. The individuals considered for selection comprise existing parents (so-called backward selections) and for clonal trials, the clones themselves are of interest (forward selections).

Table 1 Summary of trial information

Expt	Date	Ftype	Plots	Trees	Families	Females	Males	Blocks	T/F/R	Sets	mean
E1	1987	OP	12,838	11,100	467	467		25	1	18	202.6
E2	1987	OP	10,165	8,788	467	467		25	1	18	192.5
E3	1971	OP	6,851	6,433	271	271		5	5	9	220.3
E4	1975	OP	3,827	3,243	101	101		10	3	4	191.9
E5	1975	CP	1,054	933	50	20	20	6	3	5	210.8
E6	1975	OP	4,525	3,856	107	107		10	4	4	149.7
E7	1975	CP	1,360	1,209	60	21	21	6	4	6	167.2
E8	2003	CP	3,883	3,615	128	62	55	30	1	4	201.0
E9	2003	CP	4,147	3,951	121	62	54	30	1	4	162.7
E10	2003	CP	2,482	2,360	125	62	55	24	1	4	192.6
E11	2003	CP	1,300	1,300	59	41	36	20	1	0	158.9
E12	2004	OP	3,355	3,258	127	126	5	30	1	4	204.7
E13	2004	OP	3,785	3,680	128	127	5	30	1	4	191.1
E14	2005	OP	3,279	2,879	121	119	5	30	1	4	213.7
E15	2005	OP	4,128	3,725	124	122	5	30	1	4	177.5
E16	2005	OP	2,462	2,462	84	50	51	30	1	0	161.3
E17	2005	OP	4,549	4,457	234	232	5	20	1	6	169.7
E18	1990	OP	1,209	584	19	19		35	1	0	202.4
E19	1990	OP	483	231	19	19		30	1	0	247.1
E20	1990	CP	3,820	3,370	133	63	80	32	1	5	209.5
E21	1990	CP	2,674	2,373	107	52	67	31	1	4	220.3
E22	1992	CP	3,075	2,740	562	125	5	30	1	4	196.9
E23	1992	CP	3,635	3,248	758	152	5	30	1	5	200.8
E24	1992	CP	4,642	4,164	741	152	5	30	1	5	173.5
E25	1992	OP	3,952	3,510	128	128		32	1	4	221.7
E26	1993	CP	1,946	1,814	84	46	66	25	1	3	204.0
E27	1993	CP	5,176	4,700	942	189	5	30	1	6	248.7
E28	1993	CP	4,496	4,121	813	165	5	30	1	5	187.4
E29	1993	CP	1,984	1,815	485	98	5	30	1	3	213.9
E30	1994	CP	3,930	3,469	80	17	17	32	1	2	159.4
E31	1994	CP	900	812	45	29	24	30	1	2	232.9
E32	1994	CP	1,282	1,183	46	29	24	30	1	2	222.7
E33	1994	CP	1,231	1,017	41	27	22	30	1	2	240.8
E34	1995	CP	701	589	26	26	21	30	1	0	202.1
E35	1995	CP	823	656	24	24	19	30	1	0	169.6
E36	1997	CP	1,469	319	33	25	18	6	7	10	231.3
E37	1997	CP	990	190	19	19	16	6	8	5	212.3
E38	1997	CP	2,008	330	33	25	18	6	9	10	198.6
E39	1997	CP	1,174	190	19	19	16	6	10	5	202.4
E40	1997	CP	1,437	1,238	229	52	6	30	1	2	219.4
E41	1997	CP	1,275	1,102	217	52	6	30	1	2	200.4
E42	1999	CP	2627	535	42	18	15	5	10	9	213.3
E43	1999	CP	2035	524	42	18	15	5	8	9	179.9
E44	1999	CP	2403	516	41	17	14	5	9	9	175.2
E45	1988	OP	7,480	6,888	224	224		33	1	9	161.4
E46	1988	OP	6604	6,092	224	224		32	1	9	211.4
E47	1988	OP	4,285	3,991	224	224		29	1	8	219.1
E48	2000	CP	1,983	1,930	116	56	59	26	1	6	173.3
E49	2000	CP	2,397	2,320	105	54	58	30	1	6	152.4

Table 1 continued

Expt	Date	Ftype	Plots	Trees	Families	Females	Males	Blocks	T/F/R	Sets	mean
E50	2000	CP	3,321	3,232	114	56	60	30	1	6	159.7
E51	1989	OP	11,253	10,283	329	329		32	1	12	190.4
E52	1989	OP	10,132	9,251	329	329		32	1	12	206.5
E53	1989	OP	9,560	8,759	329	329		32	1	12	171.7
E54	1975	OP	4,116	3,436	105	105		10	3	4	144.3
E55	1972	OP	4,931	4,583	104	104		10	5	4	233.5
E56	1972	CP	1,327	1,284	90	23	4	15	1	0	250.9
E57	1980	CP	6,354	5,698	203	86	96	46	1	4	234.0
E58	1981	OP	6,968	6,589	171	171		45	1	5	205.8
E59	1981	OP	5,754	5,438	170	170		35	1	5	90.7
E60	1983	OP	5,065	4,951	169	169		33	1	5	144.0
E61	1983	OP	5,043	4,933	169	169		33	1	5	156.6
E62	1985	CP	4,246	4,246	301	75	74	15	1	14	143.9
E63	1987	OP	14,077	12,097	540	540		25	1	21	189.2
E64	1972	CP	2,972	2,972	90	23	4	15	3	0	244.6
E65	1972	OP	2,692	2,512	104	104		7	4	4	225.8
E66	1975	CP	1,452	1,305	50	20	20	6	5	5	234.6
E67	1975	CP	1,085	963	50	20	20	6	3	5	204.4
E68	1975	OP	5,433	4,623	106	106		10	5	4	226.0
E69	1980	CP	8,174	7,274	203	86	96	50	1	4	179.8
E70	1975	CP	1,796	1,600	60	21	21	6	5	6	253.5
E71	1968	OP	8,121	8,121	372	372		5	5	10	240.1
E72	1969	OP	14,544	13,803	588	588		5	5	16	237.7
E73	1971	OP	7,847	7,361	298	298		5	5	10	248.4
E74	1975	CP	1,157	1,023	50	20	20	6	4	5	166.0
E75	1975	OP	3,766	3,217	100	100		10	3	4	114.6
E76	1975	CP	1,296	1,150	50	20	20	6	4	5	191.7
E77	1969	OP	13,206	12,733	564	564		5	5	15	228.8

Planting date; family type (open or closed pollinated); total number of plots; number of non-control trees, families, females and males; number of blocks; median number of trees per family per block; number of sets (or blocks); mean DBH

Statistical methods

Single-trial analysis

We commence by considering the analysis of a single trial. Let $\boldsymbol{y}$ denote the $n \times 1$ vector of data, where n is the number of plots in the trial. We assume there is pedigree information available on m individuals. The model for $\boldsymbol{y}$ can be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}_g\boldsymbol{u}_g + \boldsymbol{Z}_p\boldsymbol{u}_p + \boldsymbol{e} \qquad (1)$$

where $\boldsymbol{\tau}$ is a vector of fixed effects with associated design matrix $\boldsymbol{X}$; $\boldsymbol{u}_g$ is the $m \times 1$ vector of random genetic effects with associated design matrix $\boldsymbol{Z}_g$; $\boldsymbol{u}_p$ is a vector of random non-genetic (or peripheral) effects with associated design matrix $\boldsymbol{Z}_p$ and $\boldsymbol{e}$ is the vector of residuals for the trial. In the simplest case the vector $\boldsymbol{\tau}$ comprises an overall mean (intercept) for the trial but may include other effects as necessary. The vector $\boldsymbol{u}_p$ comprises sub-vectors associated with the blocking structure of the experimental design for each trial. Examples include, block effects, main plot within block effects and incomplete block within main plot within block effects.

We assume that the $\boldsymbol{u}_g, \boldsymbol{u}_p$ and $\boldsymbol{e}$ vectors of random effects are mutually independent, and distributed as multivariate Gaussian, with zero means. The variance matrix for $\boldsymbol{u}_p$ is given by $\boldsymbol{G}_p = \oplus_{k=1}^{b} \sigma^2_{p_k} \boldsymbol{I}_{q_k}$ where b is the number of components in $\boldsymbol{u}_p$ and q_k is the number of effects in (length of) $\boldsymbol{u}_{p_k}$. The variance matrix for the residuals is assumed to be $\boldsymbol{R} = \sigma^2 \boldsymbol{I}_n$. Note that in the analysis of field trials for cereal crops, say, it is usual that the spatial co-ordinates of the plots are readily available. This allows the use of spatial analysis techniques [see Stefanova et al. (2009), for example] which involve non-identity forms for $\boldsymbol{R}$. The spatial co-ordinates of the trees in our example were not readily available so the traditional complete and incomplete block models have been used for the non-genetic effects and residuals.

Full animal model

We consider a simple model for $\boldsymbol{u}_g$ given by

$$\boldsymbol{u}_g = \boldsymbol{u}_a + \boldsymbol{u}_e$$

where the two terms represent the additive and non-additive (or residual) genetic effects. More complex models including ones which partition the non-additive into dominance and residual genetic (that is, non-additive and non-dominance) can also be considered if applicable. Thus the model in Eq. (1) becomes

$$\boldsymbol{y} = \boldsymbol{X\tau} + \boldsymbol{Z}_g\boldsymbol{u}_a + \boldsymbol{Z}_g\boldsymbol{u}_e + \boldsymbol{Z}_p\boldsymbol{u}_p + \boldsymbol{e} \tag{2}$$

We assume that the variance matrices of the vectors of additive and non-additive genetic effects are given by

$$\begin{aligned} \mathrm{var}(\boldsymbol{u}_a) &= \sigma_a^2\boldsymbol{A} \quad \text{and} \\ \mathrm{var}(\boldsymbol{u}_e) &= \sigma_e^2\boldsymbol{I}_m \end{aligned}$$

where $\boldsymbol{A}$ is the $m \times m$ numerator relationship matrix.

The maintenance of sparsity is an important property for efficient estimation in any software package (see section on "Software") and hence we therefore consider a partitioning of the vectors of genetic effects into two sub-vectors, the first representing those trees with progeny (i.e. parental trees), and the latter representing those trees without progeny (i.e. non-parental trees). This partition is denoted by $\boldsymbol{u}_s^T = (\boldsymbol{u}_{s_p}^T,\ \boldsymbol{u}_{s_n}^T), s = g, a, e$. There is a conformal partitioning of the columns of $\boldsymbol{Z}_g$ given by $\boldsymbol{Z}_g = [\boldsymbol{Z}_{g_p},\ \boldsymbol{Z}_{g_n}]$. We let m_p and m_n denote the number of trees with and without progeny so that $m = m_p + m_n$. Typically m_p is substantially less than m. The numerator relationship matrix is then partitioned as

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{pp} & \boldsymbol{A}_{pn} \\ \boldsymbol{A}_{np} & \boldsymbol{A}_{nn} \end{bmatrix}$$

Now note that

$$\boldsymbol{u}_{a_n} = \boldsymbol{T}_{np}\boldsymbol{u}_{a_p} + \boldsymbol{u}_{m_n} \tag{3}$$

where $\boldsymbol{u}_{m_n}$ represents the so-called Mendelian variation for the non-parental trees, and the matrix $\boldsymbol{T}_{np}$ is a parent indicator matrix given by

$$\boldsymbol{T}_{np} = \tfrac{1}{2}(\boldsymbol{F}_{np} + \boldsymbol{M}_{np})$$

where $\boldsymbol{F}_{np}$ and $\boldsymbol{M}_{np}$ are $m_n \times m_p$ female and male parent indicator matrices, respectively. Note that if the female (male) parent of an individual is unknown the row of $\boldsymbol{F}_{np}$ ($\boldsymbol{M}_{np}$) corresponding to that individual will consist entirely of zeros so that missing parental information does not present any difficulties. It follows that

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{pp} & \boldsymbol{A}_{pp}\boldsymbol{T}_{np}^T \\ \boldsymbol{T}_{np}\boldsymbol{A}_{pp} & \boldsymbol{T}_{np}\boldsymbol{A}_{pp}\boldsymbol{T}_{np}^T + \boldsymbol{D}_{nn} \end{bmatrix}$$

where $\mathrm{var}(\boldsymbol{u}_{m_n}) = \sigma_a^2\boldsymbol{D}_{nn}$ is a diagonal matrix, with elements of $\boldsymbol{D}_{nn}$ given by $\frac{1}{2}(1 - f_{ai})$, f_{ai} being the mean of the inbreeding coefficients of the parents of the *ith* non-parental CP or clonal individual (tree) or $0.75 - 0.25f_{ai_f}$, f_{ai_f} being the inbreeding coefficient of the female parent for the *ith* non-parental OP individual. Using the properties of the inverse of a partitioned matrix it follows that for $\boldsymbol{Q}_{nn} = \boldsymbol{D}_{nn}^{-1}$

$$\boldsymbol{A}^{-1} = \begin{bmatrix} \boldsymbol{A}_{pp}^{-1} + \boldsymbol{T}_{np}^T\boldsymbol{Q}_{nn}\boldsymbol{T}_{np} & -\boldsymbol{T}_{np}^T\boldsymbol{Q}_{nn} \\ -\boldsymbol{Q}_{nn}\boldsymbol{T}_{np} & \boldsymbol{Q}_{nn} \end{bmatrix}$$

Thus, by ordering the genetic effects as parental trees followed by their progeny, the partition of the inverse of the numerator relationship matrix relating to progeny (that is, $\boldsymbol{Q}_{nn}$) is diagonal (maximally sparse). This ordering will allow the sparsity of the MME (Henderson 1950) to be maintained during the absorption process so as to provide a computationally efficient scheme. It also provides the basis for the reduced animal (RA) model introduced by Quass and Pollack (1980).

Reduced (and approximate reduced) animal model

Using the partitioning of the additive genetic effects into parental and non-parental, then substituting Eq. (3) into Eq. (2), gives the so-called reduced animal (RA) model:

$$\boldsymbol{y} = \boldsymbol{X\tau} + (\boldsymbol{Z}_{g_p} + \boldsymbol{Z}_{g_n}\boldsymbol{T}_{np})\boldsymbol{u}_{a_p} + \boldsymbol{Z}_{g_n}\boldsymbol{u}_{m_n} + \boldsymbol{Z}_g\boldsymbol{u}_e + \boldsymbol{Z}_p\boldsymbol{u}_p + \boldsymbol{e} \tag{4}$$

The genetic effects comprise the additive effects for parental trees, that is, $\boldsymbol{u}_{a_p}$ with associated variance matrix $\sigma_a^2\boldsymbol{A}_{pp}$; the Mendelian effects for non-parental trees, that is $\boldsymbol{u}_{m_n}$ with associated variance matrix $\sigma_a^2\boldsymbol{D}_{nn}$ and the residual genetic effects for all trees, that is $\boldsymbol{u}_e$ with associated variance matrix $\sigma_e^2\boldsymbol{I}_m$.

A more simple form for the design matrix for the additive effects for parental trees in Eq. (4) is now derived. Consider a partition of the data vector into two components, $\boldsymbol{y}_p$ and $\boldsymbol{y}_n$, corresponding to parental and non-parental trees. The components have length n_p and n_n respectively, so that $n = n_p + n_n$. In many cases there are no data on parents so that $n_p = 0$ but for generality, and for the prediction of forward selections, we allow for $n_p > 0$ here. For ease of illustration, and without loss of generality, we assume that the data vector is ordered as $\boldsymbol{y} = (\boldsymbol{y}_p^T, \boldsymbol{y}_n^T)^T$. Using this ordering the design matrix for $\boldsymbol{u}_{a_p}$ in Eq. (4) is given by

$$\begin{aligned} \boldsymbol{Z}_{g_p} + \boldsymbol{Z}_{g_n}\boldsymbol{T}_{np} &= \begin{bmatrix} \boldsymbol{Z}_{g_{pp}} \\ \boldsymbol{0} \end{bmatrix} + \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{Z}_{g_{nn}}\boldsymbol{T}_{np} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{Z}_{g_{pp}} \\ \boldsymbol{0} \end{bmatrix} + \begin{bmatrix} \boldsymbol{0} \\ \frac{1}{2}(\boldsymbol{Z}_{f_{np}} + \boldsymbol{Z}_{m_{np}}) \end{bmatrix} \end{aligned}$$

where $\boldsymbol{Z}_{f_{np}} = \boldsymbol{Z}_{g_{nn}}\boldsymbol{F}_{np}$ and $\boldsymbol{Z}_{m_{np}} = \boldsymbol{Z}_{g_{nn}}\boldsymbol{M}_{np}$ are the design matrices for females and males for the non-parental trees. Hence it follows that

$$\begin{aligned} \boldsymbol{Z}_{g_p} + \boldsymbol{Z}_{g_n}\boldsymbol{T}_{np} &= \begin{bmatrix} \frac{1}{2}(\boldsymbol{Z}_{g_{pp}} + \boldsymbol{Z}_{g_{pp}}) \\ \frac{1}{2}(\boldsymbol{Z}_{f_{np}} + \boldsymbol{Z}_{m_{np}}) \end{bmatrix} \\ &= \tfrac{1}{2}(\boldsymbol{Z}_f + \boldsymbol{Z}_m) \end{aligned} \tag{5}$$

where

$$\mathbf{Z}_f = \begin{bmatrix} \mathbf{Z}_{g_{pp}} \\ \mathbf{Z}_{f_{np}} \end{bmatrix} \qquad \text{and} \qquad \mathbf{Z}_m = \begin{bmatrix} \mathbf{Z}_{g_{pp}} \\ \mathbf{Z}_{m_{np}} \end{bmatrix}$$

are the design matrices for female and male parents, where the female and male parents are coded as themselves for parental trees. The RA model can then be written as

$$\boldsymbol{y} = \boldsymbol{X\tau} + \tfrac{1}{2}(\mathbf{Z}_f + \mathbf{Z}_m)\boldsymbol{u}_{a_p} + \mathbf{Z}_{g_n}\boldsymbol{u}_{m_n} + \mathbf{Z}_g\boldsymbol{u}_e + \mathbf{Z}_p\boldsymbol{u}_p + \boldsymbol{e} \tag{6}$$

Thus the elements of the design matrix are simply computed as the average of the elements of the female and male parent design matrices. Using the results in the previous section, it is easily shown that the variance of the data vector for the RA model is identical to that for the full animal model.

The MME for the RA model in Eq. (6) have the same number of equations ($2m$) for the genetic effects as does the full animal model. However, the model was originally proposed purely for the prediction of breeding values, that is, assuming known estimates of variance parameters. In this case, the Mendelian and residual genetic effects can be combined to form a composite term with the resultant model given by:

$$\boldsymbol{y} = \boldsymbol{X\tau} + \tfrac{1}{2}(\mathbf{Z}_f + \mathbf{Z}_m)\boldsymbol{u}_{a_p} + \mathbf{Z}_g\boldsymbol{u}_e^* + \mathbf{Z}_p\boldsymbol{u}_p + \boldsymbol{e} \tag{7}$$

where $\boldsymbol{u}_e^* = (\boldsymbol{u}_{e_p}^T, \boldsymbol{u}_{e_n}^{*T})^T$ and $\boldsymbol{u}_{e_n}^* = \boldsymbol{u}_{e_n} + \boldsymbol{u}_{m_n}$ with

$$\text{var}\left(\boldsymbol{u}_e^*\right) = \begin{bmatrix} \sigma_e^2\boldsymbol{I}_{m_p} & \mathbf{0} \\ \mathbf{0} & \sigma_e^2\boldsymbol{I}_{m_n} + \sigma_a^2\boldsymbol{D}_{nn} \end{bmatrix}$$

This formulation results in a reduction in the number of equations to be solved (there are now only $m_p + m$ equations for the genetic effects compared with $2m$ for the full animal model).

When there is a requirement to estimate variance parameters in addition to obtain E-BLUPs we can consider approximating the diagonal matrix $\boldsymbol{D}_{nn}$ for the Mendelian effects by a scaled identity matrix, given by either $\frac{1}{2}(1-\bar{f})\boldsymbol{I}_{m_n}$ where $\bar{f}$ is the mean of the f_{ai} values for trials containing the non-parental CP/clonal individuals, or by $(0.75 - 0.25\bar{f})\boldsymbol{I}_{m_n}$ where $\bar{f}$ is the mean of the f_{ai_f} values for trials containing the non-parental OP individuals. This is a similar approach to that of White et al. (2006) who, for their simpler setting, ignore inbreeding, so use $\boldsymbol{D}_{nn} = \frac{1}{2}\boldsymbol{I}_{m_n}$. The use of $\frac{1}{2}(1-\bar{f})\boldsymbol{I}_{m_n}$ will be a reasonable approximation if there is little heterogeneity in the f_{ai} values for either CP, clonal or OP trials containing the non-parental individuals. In this case, we can use the model as in Eq. (7) but with $\text{var}\left(\boldsymbol{u}_{e_n}^*\right) = \sigma_m^{2*}\boldsymbol{I}_{m_n}$ where $\sigma_m^{2*} = \sigma_e^2 + \frac{1}{2}(1-\bar{f})\sigma_a^2$ for CP or clonal trials or $\sigma_m^{2*} = \sigma_e^2 + (0.75 - 0.25\bar{f})\sigma_a^2$ for OP trials. We will call the resultant model the approximate reduced animal (ARA) model.

Forward selections

As discussed in the description of the motivating example, the aim of the analysis of the RPBC data is to obtain E-BLUPs of breeding values (additive genetic effects). One of the potential drawbacks of the ARA (and RA) model is that it does not allow direct prediction of the additive genetic effects for non-parental individuals (forward selections) since the Mendelian sampling effect for these individuals is confounded with the residual genetic effect. Quass and Pollack (1980) provide details for obtaining back-solutions for non-parental individuals in the RA model. We could use a similar approach for the ARA model but propose a simpler, more direct method. This involves augmenting the parental partition of the genetic effects to include both the (true) parents and the individuals for which forward selections are required. Operationally this involves the expansion of $\boldsymbol{A}_{pp}$ to include the additional individuals and the alteration of $\mathbf{Z}_f$ and $\mathbf{Z}_m$ to indicate that the parents of the forward selection individuals are no longer their true parents but are the individuals themselves. In this way, E-BLUPs (and associated accuracies) of the additive genetic effects for all the required individuals are obtained directly from the fit of the mixed model.

Special case of no data on parental individuals

The example dataset is typical of many tree breeding datasets in that parental trees are not grown in the trials so there are no data for these individuals. In this case, there are simplifications to both the full and reduced animal models due to the fact that $\mathbf{Z}_{g_p} = \mathbf{0}$, $\mathbf{Z}_{g_{pp}}$ does not exist and $\mathbf{Z}_{g_{nn}} = \mathbf{Z}_{g_n}$ so that $\mathbf{Z}_f = \mathbf{Z}_{f_{np}}(= \mathbf{Z}_{g_n}\boldsymbol{F}_{np})$ and $\mathbf{Z}_m = \mathbf{Z}_{m_{np}}(= \mathbf{Z}_{g_n}\boldsymbol{M}_{np})$.

Clonal trials The above simplifications mean that the ARA model for clonal trials with no data on parents can be written as

$$\boldsymbol{y} = \boldsymbol{X\tau} + \tfrac{1}{2}(\mathbf{Z}_{f_{np}} + \mathbf{Z}_{m_{np}})\boldsymbol{u}_{a_p} + \mathbf{Z}_{g_n}\boldsymbol{u}_{e_n}^* + \mathbf{Z}_p\boldsymbol{u}_p + \boldsymbol{e} \tag{8}$$

The full animal model for clonal trials with no data on parents remains as in Eq. (2).

OP/CP trials With these trials there is the additional simplification that $\mathbf{Z}_{g_n} = \boldsymbol{I}_n = \boldsymbol{I}_{m_n}$. In the context of the full animal model of Eq. (2) this means that the variances of the residual genetic and residual effects are not identifiable, so the model simplifies to

$$\boldsymbol{y} = \boldsymbol{X\tau} + \mathbf{Z}_g\boldsymbol{u}_a + \mathbf{Z}_p\boldsymbol{u}_p + \boldsymbol{e}_e^* \tag{9}$$

where $\boldsymbol{e}_e^* = \boldsymbol{u}_{e_n} + \boldsymbol{e}$ with variance matrix $\sigma_e^{2*}\boldsymbol{I}_n$ where $\sigma_e^{2*} = \sigma_e^2 + \sigma^2$.

The ARA model simplifies to

$$\boldsymbol{y} = \boldsymbol{X\tau} + \tfrac{1}{2}(\boldsymbol{F}_{np} + \boldsymbol{M}_{np})\boldsymbol{u}_{a_p} + \mathbf{Z}_p\boldsymbol{u}_p + \boldsymbol{e}_a^* \tag{10}$$

where $\boldsymbol{e}_a^* = \boldsymbol{u}_{m_n} + \boldsymbol{u}_{e_n} + \boldsymbol{e}$ with variance matrix $\sigma_a^{2*}\boldsymbol{I}_n$ where $\sigma_a^{2*} = \frac{1}{2}(1-\bar{f})\sigma_a^2 + \sigma_e^2 + \sigma^2$ for CP trials or $\sigma_a^{2*} = (0.75 - 0.25\bar{f})\sigma_a^2 + \sigma_e^2 + \sigma^2$ for OP trials.

Multi-environment trial analysis

Here we extend the models for the analysis of single trials to series of trials. We now let $\boldsymbol{y}$ denote the $n \times 1$ combined vector of data across all trials in the MET, so for t trials we write $\boldsymbol{y} = (\boldsymbol{y}_1^T, \boldsymbol{y}_2^T, \ldots \boldsymbol{y}_t^T)^T$ where $\boldsymbol{y}_j$ is the $n_j \times 1$ vector of data for the jth trial and n_j is the number of plots in that trial. Note then that $n = \sum_{j=1}^{t} n_j$. The model for $\boldsymbol{y}$ can be written as

$$\boldsymbol{y} = \boldsymbol{X\tau} + \boldsymbol{Z}_g\boldsymbol{u}_g + \boldsymbol{Z}_p\boldsymbol{u}_p + \boldsymbol{e}$$

where $\boldsymbol{\tau} = (\boldsymbol{\tau}_1^T, \boldsymbol{\tau}_2^T, \ldots \boldsymbol{\tau}_t^T)^T$ is a vector of fixed effects with associated design matrix $\boldsymbol{X}$ (assumed to have full column rank); $\boldsymbol{u}_g = (\boldsymbol{u}_{g_1}^T, \boldsymbol{u}_{g_2}^T, \ldots \boldsymbol{u}_{g_t}^T)^T$ is the $mt \times 1$ vector of genetic effects with associated design matrix $\boldsymbol{Z}_g$; $\boldsymbol{u}_p = (\boldsymbol{u}_{p_1}^T, \boldsymbol{u}_{p_2}^T, \ldots \boldsymbol{u}_{p_t}^T)^T$ is a vector of random non-genetic (or peripheral) effects with associated design matrix $\boldsymbol{Z}_p$ and $\boldsymbol{e} = (\boldsymbol{e}_1^T, \boldsymbol{e}_2^T, \ldots \boldsymbol{e}_t^T)^T$ is the combined vector of residuals from all trials.

As in the case of single-trial analysis we assume that $\boldsymbol{u}_g, \boldsymbol{u}_p$ and $\boldsymbol{e}$ are mutually independent, and distributed as multivariate Gaussian, with zero means. The variance matrix for $\boldsymbol{u}_p$ is given by $\boldsymbol{G}_p = \oplus_{k=1}^{b} \sigma_{p_k}^2 \boldsymbol{I}_{q_k}$ where b is the number of components in $\boldsymbol{u}_p$ and q_k is the number of effects in (length of) $\boldsymbol{u}_{p_k}$. The variance matrix for the residuals is assumed to be $\boldsymbol{R} = \oplus_{j=1}^{t} \sigma_j^2 \boldsymbol{I}_{n_j}$.

We partition the $mt \times 1$ vector of genetic effects $\boldsymbol{u}_g$ into additive effects $\boldsymbol{u}_a$ and residual effects $\boldsymbol{u}_e$, so that the MET model is given by

$$\boldsymbol{y} = \boldsymbol{X\tau} + \boldsymbol{Z}_g\boldsymbol{u}_a + \boldsymbol{Z}_g\boldsymbol{u}_e + \boldsymbol{Z}_p\boldsymbol{u}_p + \boldsymbol{e} \tag{11}$$

As before the vectors of genetic effects are partitioned into sub-vectors with the first representing trees with progeny and the second representing trees without progeny. This partition is denoted by $\boldsymbol{u}_s^T = (\boldsymbol{u}_{s_p}^T,\ \boldsymbol{u}_{s_n}^T), s = g, a, e$. There is a conformal partitioning of the columns of the $n \times mt$ design matrix $\boldsymbol{Z}_g$ given by $\boldsymbol{Z}_g = [\boldsymbol{Z}_{g_p},\ \boldsymbol{Z}_{g_n}]$ where the dimensions of the two sub-matrices are $n \times m_pt$ and $n \times m_nt$ respectively. We note that for this partition to apply, $\boldsymbol{u}_s$ must be ordered as trials within individuals. This ordering was adopted by Beeck et al. (2010) and is the reverse of the order considered in Smith et al. (2001). We assume that the variance matrices of the vectors of additive and residual genetic effects are given by

$$\mathrm{var}(\boldsymbol{u}_a) = \boldsymbol{A} \otimes \boldsymbol{G}_a \quad \text{and}$$
$$\mathrm{var}(\boldsymbol{u}_e) = \boldsymbol{I}_m \otimes \boldsymbol{G}_e$$

where, as before, $\boldsymbol{A}$ is the $m \times m$ numerator relationship matrix.

The matrices $\boldsymbol{G}_a$ and $\boldsymbol{G}_e$ are $t \times t$ symmetric positive (semi)-definite matrices and are generally referred to as the between environment additive and residual genetic variance matrices. The most general form for $\boldsymbol{G}_a$ (or $\boldsymbol{G}_e$) is a so-called unstructured form that contains $p = t(t+1)/2$ parameters to be estimated. Clearly as t increases p becomes prohibitively large and this influences both the ability to fit the unstructured variance model, as well as the reliability of the estimated parameters. The unstructured variance model has been frequently used for the analysis of tree breeding MET datasets. Baltunis et al. (2010), Apiolaza (2012), Raymond (2011) used the model in this context, but they mostly fitted the model to subsets of two trials from the full set of trials. Such an approach avoids the difficulty in fitting the model, but is statistically inefficient and can lead to an overall estimate of $\boldsymbol{G}_a$ (or $\boldsymbol{G}_e$) which is not positive (semi)-definite.

Apart from the lack of parsimony of the unstructured variance model, the other difficulty with fitting this variance model to MET datasets is the poor connectivity between trials. The example MET dataset has relatively good parental connectivity but many larger MET datasets, particularly those found in tree breeding, have poor parental connectivity. This has, in part, led to the approach adopted in the papers cited above, where the unstructured variance model is fitted to those pairs of trials having good connectivity.

We have found that the factor analytic (FA) variance model (Smith et al. 2001) provides a good approximation to the unstructured variance model (Kelly et al. 2007) and is both parsimonious and informative. The FA model can be viewed as arising from a multiplicative (or regression) model for the genetic effects in each environment. If we consider the additive genetic effect for individual i and trial j, we write

$$u_{a_{ij}} = \lambda_{a_{1j}} f_{a_{1i}} + \lambda_{a_{2j}} f_{a_{2i}} + \ldots \lambda_{a_{kj}} f_{a_{ki}} + \delta_{a_{ij}} \tag{12}$$

which involves a sum of k multiplicative terms. Each term is the product of a genetic effect ($f_{a_{ri}}$), which is known as a factor score, and an environment effect ($\lambda_{a_{rj}}$), which is known as a loading. The "order" (k) of the FA model is the number of factors (multiplicative terms) and we denote an FA model with k factors as an FA(k) model. The final term $\delta_{a_{ij}}$ represents a lack of fit of the regression model, and so will be termed a genetic regression residual.

The model in Eq. (12) can be written in vector notation as

$$\boldsymbol{u}_a = (\boldsymbol{I}_m \otimes \boldsymbol{\Lambda}_a)\boldsymbol{f}_a + \boldsymbol{\delta}_a \tag{13}$$

where $\boldsymbol{\Lambda}_a$ is the $t \times k$ matrix of environment loadings, $\boldsymbol{f}_a$ is the $mk \times 1$ vector of additive genetic scores and $\boldsymbol{\delta}_a$ is the $mt \times 1$ vector of genetic regression residuals. We assume

$\boldsymbol{f}_a$ and $\boldsymbol{\delta}_a$ are independent and are distributed as multivariate Gaussian with zero means and variance matrices given by

$$\text{var}(\boldsymbol{f}_a) = \boldsymbol{A} \otimes \boldsymbol{I}_k \quad \text{and} \quad \text{var}(\boldsymbol{\delta}_a) = \boldsymbol{A} \otimes \boldsymbol{\Psi}_a$$

where $\boldsymbol{\Psi}_a$ is a $t \times t$ diagonal matrix with a variance (known as a specific variance) for each environment. These assumptions lead to

$$\text{var}(\boldsymbol{u}_a) = \boldsymbol{A} \otimes \left(\boldsymbol{\Lambda}_a \boldsymbol{\Lambda}_a^T + \boldsymbol{\Psi}_a\right)$$

so that $\boldsymbol{G}_a = \left(\boldsymbol{\Lambda}_a \boldsymbol{\Lambda}_a^T + \boldsymbol{\Psi}_a\right)$.

To determine an appropriate value of k, we use a measure similar to an R^2 goodness-of-fit value for a multiple regression. This is partly driven by the fact that predicted breeding values for individual trials are based on the multiplicative terms alone, that is, the genetic regression residuals are excluded (see Cullis et al. 2010 for details). Thus we define, for each trial, the percentage of additive genetic variance accounted for by the k multiplicative terms in Eq. (12):

$$v_{a_j} = 100 \sum_{r=1}^{k} \lambda_{a_{rj}}^2 / \left(\sum_{r=1}^{k} \lambda_{a_{rj}}^2 + \psi_j \right)$$

In addition, we define an overall (that is, across trial) percentage variance accounted for as $\bar{v}_a = 100\text{tr}\left(\boldsymbol{\Lambda}_a \boldsymbol{\Lambda}_a^T\right)/\text{tr}(\boldsymbol{G}_a)$. We choose an order such that the overall percentage variance accounted for is high and the number of trials with low individual v_{a_j} values is small.

As before we write the non-parental additive genetic effects for each environment as:

$$\boldsymbol{u}_{a_n} = \boldsymbol{T}_{np} \boldsymbol{u}_{a_p} + \boldsymbol{u}_{m_n} \quad \text{and} \quad \boldsymbol{T}_{np} = \tfrac{1}{2}\left(\boldsymbol{F}_{np} + \boldsymbol{M}_{np}\right) \qquad (14)$$

where $\boldsymbol{u}_{m_n}$ is the $m_n t \times 1$ vector of Mendelian sampling effects for each environment and $\boldsymbol{F}_{np}$ and $\boldsymbol{M}_{np}$ are $m_n t \times m_p t$ female and male parent indicator matrices across all trials (ordered as trials within individuals), respectively. In terms of variance matrices we have

$$\text{var}\left(\boldsymbol{u}_{a_p}\right) = \boldsymbol{A}_{pp} \otimes \boldsymbol{G}_a \quad \text{and} \quad \text{var}\left(\boldsymbol{u}_{m_n}\right) = \boldsymbol{D}_{nn} \otimes \boldsymbol{G}_a$$

where $\boldsymbol{A}_{pp}$ and $\boldsymbol{D}_{nn}$ are as defined for the analysis of single trials.

Now we consider the full and approximate reduced animal models for two scenarios, namely when all the trials in the MET are clonal trials and when all are OP/CP trials. When there is a mixture of the two types, as is the case with our example, the models can be formulated as an amalgamation of those presented for the individual types and so for brevity are not detailed here. We assume the typical case in which there are no data on parental trees.

Clonal METs

The full animal model for clonal METs with no data on parents is as given in Eq. (11). The ARA is given by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \tfrac{1}{2}(\boldsymbol{Z}_{f_{np}} + \boldsymbol{Z}_{m_{np}})\boldsymbol{u}_{a_p} + \boldsymbol{Z}_{g_n}\boldsymbol{u}_{e_n}^* + \boldsymbol{Z}_p\boldsymbol{u}_p + \boldsymbol{e} \qquad (15)$$

where $\boldsymbol{Z}_{f_{np}} = \boldsymbol{Z}_{g_n}\boldsymbol{F}_{np}$ and $\boldsymbol{Z}_{m_{np}} = \boldsymbol{Z}_{g_n}\boldsymbol{M}_{np}$ are the $n \times m_n t$ design matrices for the female and male parents and $\boldsymbol{u}_{e_n}^* = \boldsymbol{u}_{e_n} + \boldsymbol{u}_{m_n}$ with variance matrix given by

$$\text{var}\left(\boldsymbol{u}_{e_n}^*\right) = \boldsymbol{I}_{m_n} \otimes \left(\boldsymbol{G}_e + \tfrac{1}{2}(1 - \bar{f})\boldsymbol{G}_a\right) \qquad (16)$$

Recall that in the analysis of a single trial, the variance matrix of this composite term could be written as a scaled identity matrix ($\sigma_m^{2*}\boldsymbol{I}_{m_n}$) where the scalar was the simple sum of two components ($\sigma_e^2 + \tfrac{1}{2}(1 - \bar{f})\sigma_a^2$). However, in Eq. (16) we have the sum of two variance matrices, each of which will either be modelled using an FA model (possibly with different orders) or assumed to have an unstructured form. In the analysis of crop breeding METs, experience has shown that the variance matrices for each component of the composite genetic residual can be quite different. Given this, it may not be sensible to use the ARA model for the analysis of clonal MET datasets. In practice, however, we often include clonal trees in the analysis as parents, that is, as forward selections. The problem is then avoided since the model reverts to the full animal model. This may not be an option for larger clonal MET datasets in which it is computationally prohibitive to fit the full animal model.

OP/CP METs

Since there is no true replication in OP/CP METs and individuals within a trial occur once and only once in the full MET dataset, then $n = m_n$ and

$$\boldsymbol{Z}_g(\boldsymbol{I}_m \otimes \boldsymbol{G}_e)\boldsymbol{Z}_g^T = \oplus_{j=1}^{t}\sigma_{e_j}^2\boldsymbol{I}_{n_j}$$

where $\sigma_{e_j}^2, j = 1, \ldots, t$ are the diagonal elements of $\boldsymbol{G}_e$. These variances and the residual variances are therefore not identifiable so that Eq. (11) becomes

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}_g\boldsymbol{u}_a + \boldsymbol{Z}_p\boldsymbol{u}_p + \boldsymbol{e}_e^* \qquad (17)$$

where $\boldsymbol{e}_e^* = \boldsymbol{u}_{e_n} + \boldsymbol{e}$ with

$$\begin{aligned}\text{var}\left(\boldsymbol{e}_e^*\right) &= \oplus_{j=1}^{t}(\sigma_{e_j}^2 + \sigma_j^2)\boldsymbol{I}_{n_j} \\ &= \oplus_{j=1}^{t}\sigma_{e_j}^{2*}\boldsymbol{I}_{n_j} \\ &= \boldsymbol{R}_e^*\end{aligned}$$

The ARA is given by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \tfrac{1}{2}(\boldsymbol{F}_{np} + \boldsymbol{M}_{np})\boldsymbol{u}_{a_p} + \boldsymbol{Z}_p\boldsymbol{u}_p + \boldsymbol{e}_a^* \qquad (18)$$

where $\boldsymbol{e}_a^* = \boldsymbol{u}_{m_n} + \boldsymbol{u}_{e_n} + \boldsymbol{e}$ with variance matrix $\oplus_{j=1}^{t} \sigma_{a_j}^{2*} \boldsymbol{I}_{n_j}$ where $\sigma_{a_j}^{2*} = \frac{1}{2}(1-\bar{f})\sigma_{a_j}^2 + \sigma_{e_j}^2 + \sigma_j^2$, for clonal or CP trials, or $\sigma_{a_j}^{2*} = (0.75 - 0.25\bar{f})\sigma_{a_j}^2 + \sigma_{e_j}^2 + \sigma_j^2$ for OP trials and $\sigma_{a_j}^2, j = 1, \ldots, t$ are the diagonal elements of $\boldsymbol{G}_a$.

Software

The fitting of the mixed models presented in this paper involves the estimation of the variance parameters using residual maximum likelihood (REML) and the estimation and prediction of the fixed and random effects. All models in this paper were fitted using the ASReml-R package (Butler et al. 2009) within the R statistical environment (R Core Team 2012) which uses the average information algorithm (Gilmour et al. 1995) for REML estimation of variance parameters. Note that the implementation for FA models in ASReml-R (Butler et al. 2009), which is due to Thompson et al. (2003) and Kelly et al. (2009), handles the (not uncommon) case where at least two of the REML estimates of the specific variances are zero and the overall estimate of $\boldsymbol{G}_a$ is of less than full rank.

The average information algorithm is an iterative procedure and, for any iteration, a key step is the solution of the MME using the process of absorption [see Gilmour et al. (1995)]. At convergence, that is, once the REML estimates of the variance parameters have been obtained, the solution of the MME can be used to provide empirical best linear unbiased estimates (E-BLUEs) of the fixed effects and empirical best linear unbiased predictions (E-BLUPs) of the random effects (Gilmour et al. 2004). The convention in this paper is to represent the REML estimate of a variance parameter (or matrix) using an overstrike caret symbol (for example, $\hat{\sigma}_a^2$) and the E-BLUP of a random effect (or vector of random effects) using an overstrike tilde symbol (for example, $\tilde{\boldsymbol{u}}_a$).

The syntax used to fit the FA-3 model in ASReml-R is presented in the Appendix. This syntax has been annotated to explain some of the arguments used in the call to the ASReml-R package.

Results

Prior to undertaking the analysis of a MET dataset it is crucial to assess the degree of genetic connectivity, since poor connectivity may prohibit reliable estimation of key genetic parameters, in particular genetic correlations between trials. Typically, in the analysis of MET datasets for inbred crops, such as most cereal grain crops, this is measured using the variety (in this case individual tree) concurrence matrix which gives the number of varieties in common for each pair of trials.

In the case of trials involving genetically unique individuals, such as trials using trees from OP or CP families, we may consider connectivity in terms of the parents, rather than the individual trees. This measure has been recently used by Baltunis et al. (2010) and Apiolaza (2012). Figure 1 displays the parental concurrence matrix for the example as a heatmap (R Core Team 2012). The diagonal elements of the matrix are the number of parents used in the trial and the off-diagonal elements are the number of parents in common, considering both female and male parents together. Note that the original MET dataset included an additional 5 trials but these had no parents in common with the other 77 trials and so were excluded from the final dataset. The heatmap in Fig. 1 has been ordered from top to bottom (and left to right) as the best to worst connected trials, determined on the basis of the percentage of zero concurrences for each trial. In an unpublished simulation study it was shown that when an FA model was used for genotype-by-environment effects, there was little bias in the estimated genetic correlation for a pair of trials with poor concurrence (even zero concurrence), provided there was sufficient linkage through other trials. Thus, even though Fig. 1 shows there is a number of pairs of trials with no parents in common, the pattern of connectivity appears to be sufficient for fitting the FA model for A × E effects.

The analysis conducted for these data corresponded to the ARA model described in the section for multi-environment trials. In the notation of that section, the example dataset relates to $t = 77$ trials and $m = 315{,}581$ trees. Predicted additive genetic effects were required both for backward selections (2,733 parental trees) and forward selections (1,062 clonal trees). Thus the parental partition of the genetic effects was augmented to include the clones. The first sub-vector of genetic effects related to these $m_p = 3{,}795$ individuals and the second to the remaining $m_n = 311{,}786$ (non-parental and non-clonal) trees. The matrix $\boldsymbol{A}_{pp}$ was expanded in a corresponding manner to relate to both backward and forward selections.

Recall that a key determinant of the suitability of the ARA model was the homogeneity of the f_{ai} values of the m_n non-parental (and in our case, non-clonal) individuals. For the example dataset, 99.7% of these values were zero so we have chosen to use $\bar{f} = 0$. This implies that if we were able to fit the full animal model to these data, then it is most likely that the REML estimates of the additive variance parameters from the ARA model and the full animal model would be quite similar. The combined analysis of OP, CP and clonal trials implies that there is no simple and direct equivalence between the two sets of estimates for other variance parameters. This is an area for future research.

The linear mixed model fitted to the data included a fixed main effect for each trial, and terms associated with the

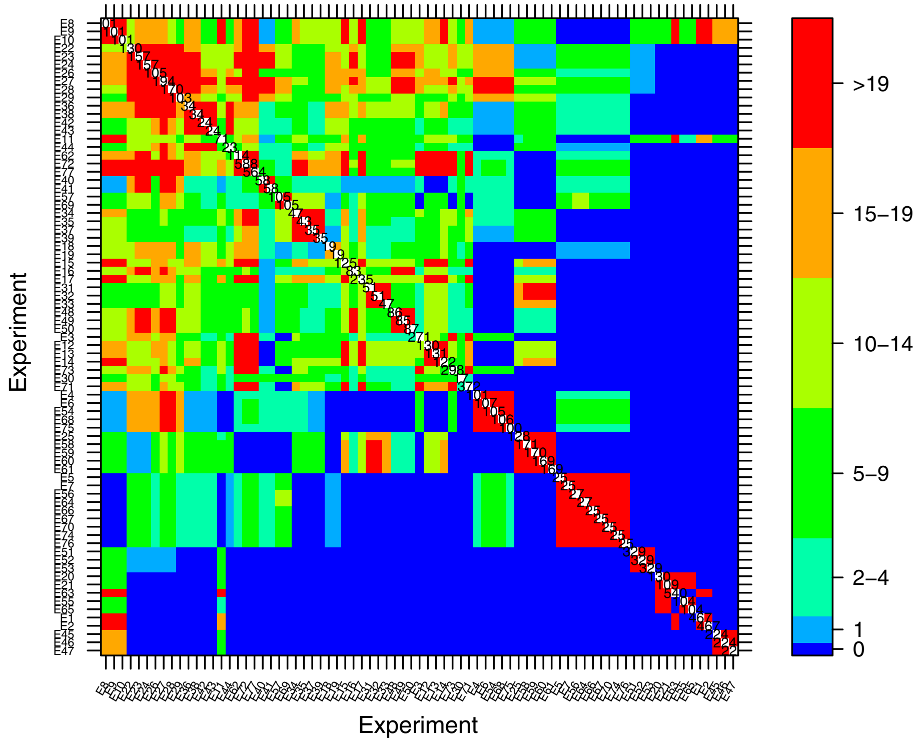


Fig. 1 Heatmap of the parental concurrence between all pairs of experiments. The *boxes* along the diagonal show the number of parents used in individual trials. The *boxes* on the off-diagonal are coloured according to the number of parents in common between trials as described in the key

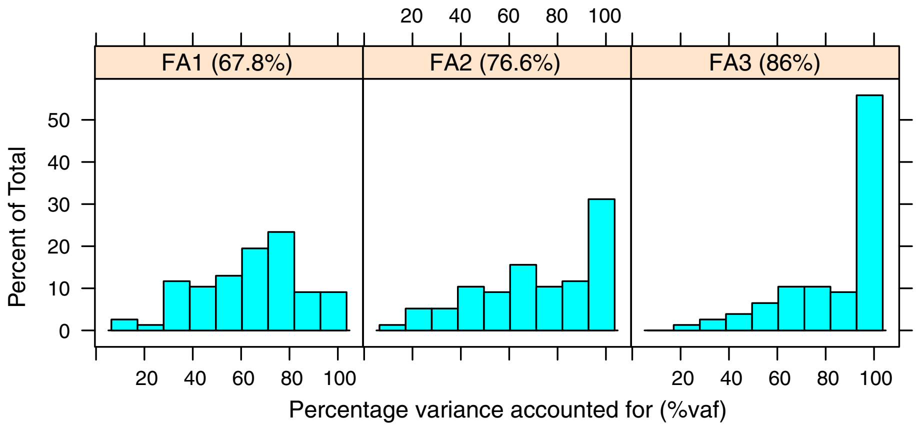


Fig. 2 Distributions of percentage variance accounted for in FA models fitted to between environment additive genetic variance matrix. Overall percentage for each FA model is given in *parentheses*

blocking structure for trials as appropriate. The variances of any of these random effects were allowed to differ between trials. Commensurate with the ARA model for a mixture of OP/CP and clonal trials, the model included additive genotype by trial effects for parents and clones (across all trials) and residual genotype by trial effects for clones (across clonal trials only). The variance matrix of the former was given by $\boldsymbol{A}_{pp} \otimes \boldsymbol{G}_a$ where $\boldsymbol{G}_a$ is a $t \times t$ matrix and the variance matrix of the latter by $\boldsymbol{I}_{m_c} \otimes \boldsymbol{G}_e$ where $m_c = 1{,}062$ is the total number of clones in the dataset and $\boldsymbol{G}_e$ is a $t_c \times t_c$ matrix where $t_c = 7$ is the number of clonal trials. FA models were fitted for $\boldsymbol{G}_a$ but $\boldsymbol{G}_e$ was assumed to have a diagonal form since the estimated non-additive genetic variances were found to be relatively small (see later). The residual variance was allowed to differ between trials.

FA models of order $k = 1, 2$ and 3 were fitted for $\boldsymbol{G}_a$. The distributions of the individual trial percentage variances accounted for, together with the overall values are shown in Fig. 2. The overall percentage variance accounted for by the FA3 model was 86 % and 66 trials had an individual value greater than 60 %. Higher order FA models could have been fitted, but given the parental connectivity, the size of the dataset and the fact that over 85 % of the total additive genetic variance has been explained by the FA3 model, we chose the FA3 model as providing an adequate fit to the data.

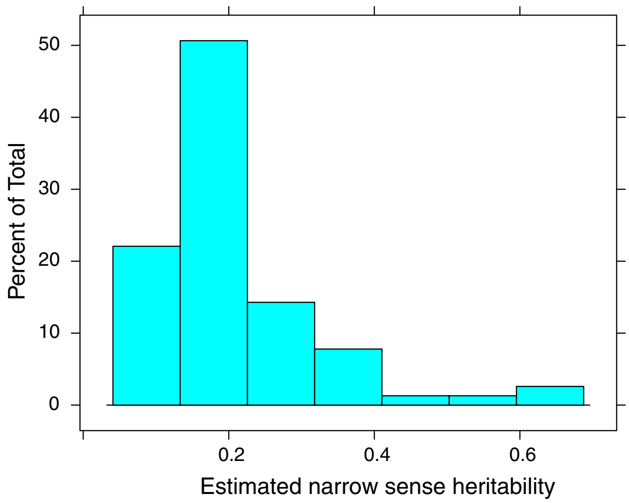


Fig. 3 Distribution of estimated narrow sense heritabilities for individual trials

A narrow sense heritability was computed for each trial as the ratio of the estimated additive genetic variance for the trial (appropriate diagonal element of $\hat{\boldsymbol{G}}_a$) to total variance. The latter was obtained as the sum of the estimated additive genetic and residual variances for OP/CP trials, whilst for clonal trials it also included the estimated non-additive genetic variance (appropriate diagonal element of $\hat{\boldsymbol{G}}_e$). Note that for clonal trials, the estimated non-additive genetic variance was relativey small, with an average contribution to total genetic (additive plus non-additive) variance of only 13 %. The distribution of individual trial heritabilities is shown in Fig. 3.

One of the key advantages of the FA approach is that it provides an estimate of the between environment additive genetic correlation matrix as a whole, rather than in a piecemeal manner. The matrix is obtained as $\hat{\boldsymbol{C}}_a = \hat{\boldsymbol{D}}_a\hat{\boldsymbol{G}}_a\hat{\boldsymbol{D}}_a$ where $\hat{\boldsymbol{D}}_a$ is a diagonal matrix with elements given by the inverse of the estimated additive genetic standard deviations for individual trials (inverse of square roots of diagonal elements of $\hat{\boldsymbol{G}}_a$). The estimated additive genetic correlation matrix measures the extent of crossover A × E interaction, that is, the level of disagreement between trials in terms of the rankings of individuals' additive genetic effects. An examination of $\hat{\boldsymbol{C}}_a$ is therefore crucial for the selection of individuals for use as parents for particular environments. The matrix may be depicted using a heatmap with the rows and columns of the matrix ordered in some way that aids with the visualisation of patterns of interaction. We have used an ordering based on the dendrogram from an agglomerative-nested hierarchical clustering algorithm as implemented in the agnes package in R (R Core Team 2012). In this way, trials that are highly correlated (so exhibit little crossover interaction) are located close together, whereas trials that are poorly correlated will be located further apart. The resultant heatmap is shown in Fig. 4. In general, the estimated additive genetic correlations between trials are reasonably high, with an average pairwise correlation of 0.54. However, there are still a large number of low correlations, with 25 % being lower than 0.37, a value that translates to substantial crossover A × E interaction.

Examination of the estimated additive genetic correlation matrix allows characterisation of environments according to their patterns of crossover A × E interaction. It is equally important to consider interaction from the dual perspective of the varieties. We need to know how individuals respond to a change in environment. The regression interpretation of the FA model [Eq. (12)] provides a natural framework for exploring this so-called variety stability. The factor scores in Eq. (12) can be thought of as regression coefficients and they reflect the additive genetic responses of individuals to the environmental covariates (factor loadings). Individuals that have near-zero estimated scores for all factors are stable in the sense of having little response to changes in the loadings.

The regression interpretation of the FA model is often most meaningful for loadings that have been rotated to a principal component solution. In this case, the first loading accounts for the maximum amount of genetic variance in the data, the second accounts for the next largest amount and is orthogonal to the first, and so on. When applied to the estimated loadings from the FA3 model fitted to the example, this resulted in the rotated loadings accounting for 59.1, 14.5 and 12.4 % of the additive genetic variance. (Note that these sum to the overall percentage variance accounted for by the FA3 model as previously reported).

Using the notation of Cullis et al. (2010) we let $\hat{\lambda}^*_{a_{rj}}$ be the rotated REML estimate of the loading for environment j in factor r and $\tilde{f}^*_{a_{ri}}$ be the associated rotated E-BLUP of the score (slope) for individual i. We then write the E-BLUP of the additive genetic effect for individual i in environment j as

$$\begin{aligned}\tilde{u}_{a_{ij}} &= \hat{\lambda}^*_{a_{1j}}\tilde{f}^*_{a_{1i}} + \hat{\lambda}^*_{a_{2j}}\tilde{f}^*_{a_{2i}} + \ldots \hat{\lambda}^*_{a_{kj}}\tilde{f}^*_{a_{ki}} + \tilde{\delta}_{a_{ij}} \\ &= \tilde{\beta}_{a_{ij}} + \tilde{\delta}_{a_{ij}}\end{aligned}$$

where $\tilde{\beta}_{a_{ij}}$ is the predicted A × E regression component. The splitting of the predicted additive genetic effect into a regression component and a lack of fit term is crucial for selection. Cullis et al. (2010) discuss the choice between using predictions that are marginal or conditional to the lack of fit term, so using $\tilde{\beta}_{a_{ij}}$ or $\tilde{u}_{a_{ij}}$, respectively. They conclude that the use of the marginal predictions provides the most consistent approach for selection

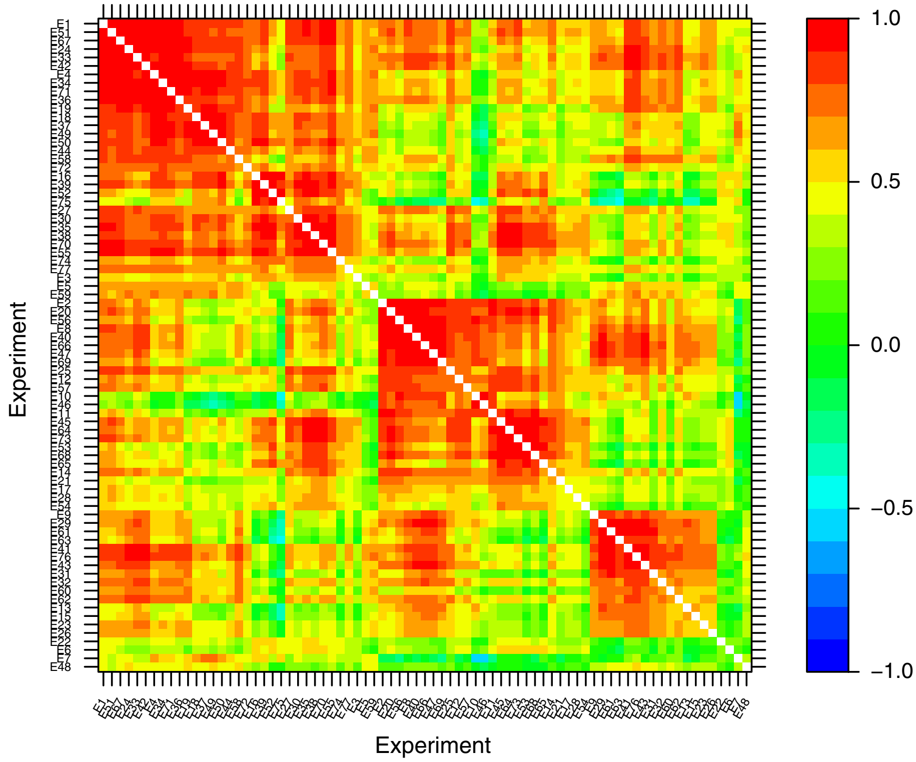


Fig. 4 Heatmap of the estimated between environment additive genetic correlation matrix

in a MET. We adopt this approach here for the selection of individuals to be used as parents and consequently refer to $\tilde{\beta}_{a_{ij}}$ as the predicted breeding value for individual i in environment j.

Varietal stability may be best viewed using latent regression plots which show genetic responses to each set of (rotated) factor loadings. For an individual i of interest, we consider a series of plots that are similar to added variable plots for a dependent variable (in our case the predicted breeding values $\tilde{\beta}_{a_{ij}}$) against a series of independent variables (in our case the rotated estimated factor loadings $\hat{\lambda}^*_{a_{rj}}, \ r = 1, \ldots, k$). The difference here is that there is a natural ordering of the independent variables from $1, \ldots, k$ (which is in decreasing order of percentage genetic variance explained), so, in the sequence of plots, we condition both the dependent and independent variable on the preceding factors. Thus the $y-$ and $x-$ axes for the sequence of plots are defined as follows:

Plot 1: $y_j = \tilde{\beta}_{a_{ij}}$ and $x_j = \hat{\lambda}^*_{a_{1j}}$

Plot 2: $y_j = \tilde{\beta}_{a_{ij}} - \hat{\lambda}^*_{a_{1j}} \tilde{f}^*_{a_{1i}}$ and $x_j = \hat{\lambda}^*_{a_{2j}}$[...]

Plot k: $y_j = \tilde{\beta}_{a_{ij}} - \sum_{r=1}^{k-1} \hat{\lambda}^*_{a_{rj}} \tilde{f}^*_{a_{ri}} = \hat{\lambda}^*_{a_{kj}} \tilde{f}^*_{a_{ki}}$ and $x_j = \hat{\lambda}^*_{a_{kj}}$

We consider the latent regression plots for a subset of eleven of the most heavily deployed parents from the RPBC production population nursery (provided they were also used in a reasonable number of trials in this MET dataset) and the parent P12. Table 2 presents the percentage representation in the production population and the E-BLUPs of the factor scores (rotated) for these 12 individuals. The parent P12 was included as it is very highly regarded in terms of DBH, but has recently been discarded due to an issue with spiral grain. Figs. 5, 6 and 7 show the regression plots for the 12 individuals. The data points

Table 2 Predicted (rotated) factor scores for 12 parents from the RPBC production population nursery

Parent	Factor 1	Factor 2	Factor 3	Experiments	% representation
P1	2.24	−0.11	−1.31	11	10.5
P2	1.61	−2.23	−1.90	24	7.2
P3	1.02	−1.16	−0.93	10	6.7
P4	1.26	0.03	−1.29	28	6.1
P5	2.52	0.07	−0.71	12	5.0
P6	0.14	−0.38	0.33	22	4.6
P7	0.92	0.68	−0.15	25	4.5
P8	−0.31	−0.79	0.50	14	4.2
P9	0.37	1.94	−0.21	11	4.0
P10	1.85	−0.03	−1.00	13	2.8
P11	0.56	−0.20	−1.17	27	2.2
P12	3.20	−0.03	0.95	26	0.6

Also shown is the number of experiments in the MET in which they were used as parents and the percentage representation in the production population

Fig. 5 Latent additive genetic regression plot for the first factor for 12 highly deployed individuals

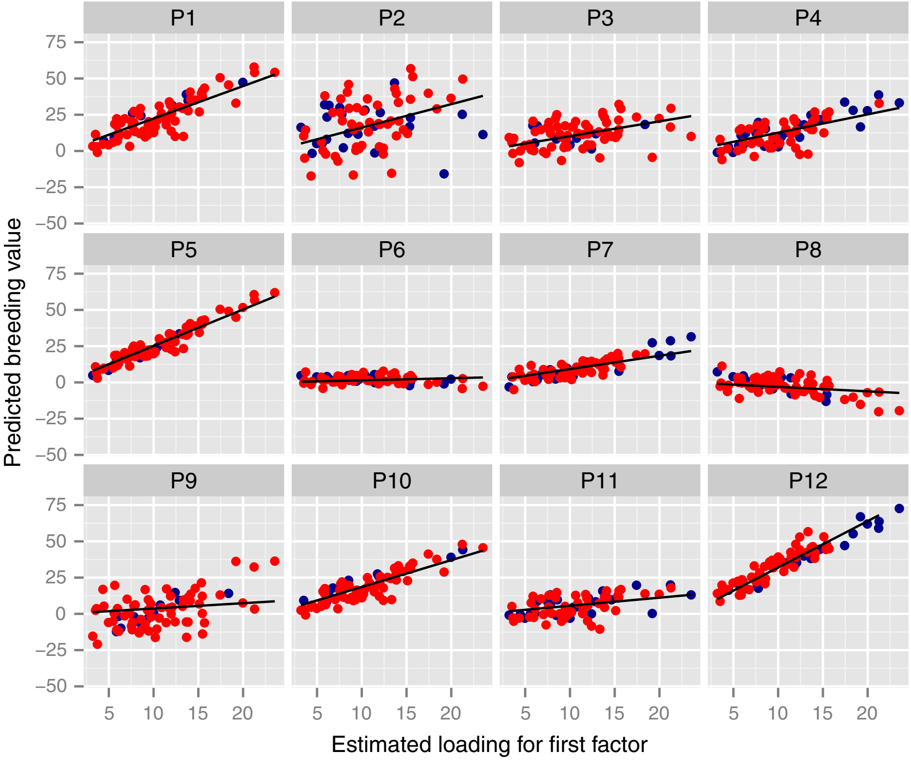
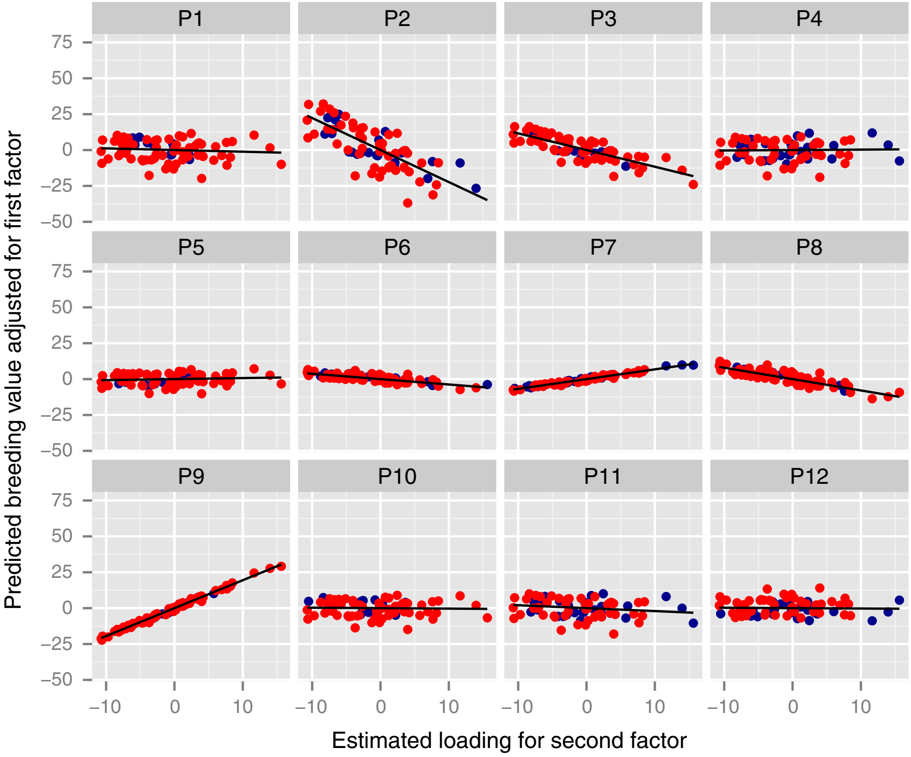


Fig. 6 Latent additive genetic regression plot for the second factor for 12 highly deployed individuals



on each plot are either coloured blue if the individual was used as a parent in the associated trial and red otherwise. The line drawn on each plot has a slope given by the predicted score for the individual andfactor concerned, that is, as given in Table 2. Recall that in this example, the first factor accounted for the majority (59.1 %) of A × E

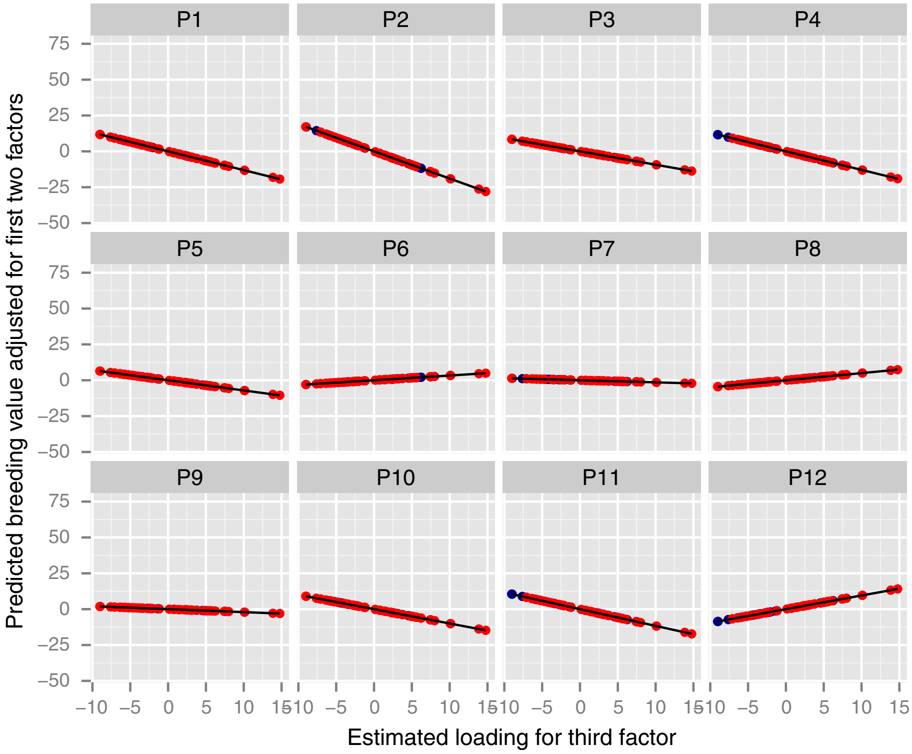


Fig. 7 Latent additive genetic regression plot for the third factor for 12 highly deployed individuals

variation so that the regressions on this factor have the greatest impact on predicted breeding values. Since all the estimated loadings for this factor are positive (see Fig. 5), this then means that large positive regression coefficients for this factor are desirable for high DBH. Of the individuals listed in Table 2, parents P12, P5 and P1 have the highest predicted scores for the first factor. Figure 5 shows that the predicted breeding values for these parents are always positive, and, as suggested by the regression, they increase substantially for environments with high estimated loadings. None of these three parents is sensitive to the second factor, but in terms of the third factor, theslopes for P12 and P1 have opposite signs (and P5 has a slope closer to zero). Parent P2 is more sensitive to the second and third factors than the first and parent P6 is relatively stable across all 3 factors.

It is interesting to consider the range of stabilities amongst the full set of parents and forward selections (3,795 individuals). Fig. 8 presents the distributions of the predicted rotated scores for each factor and for the parents currently in the production population nursery (97 individuals) and the remaining parents and forward selections (3,698 individuals). These plots demonstrate the wide range in stability for each factor. The most obvious feature is that there has been an upward shift in the mean of the distribution of scores for the production population compared with the other parents for the first factor, but no change in terms of the second and third factors.

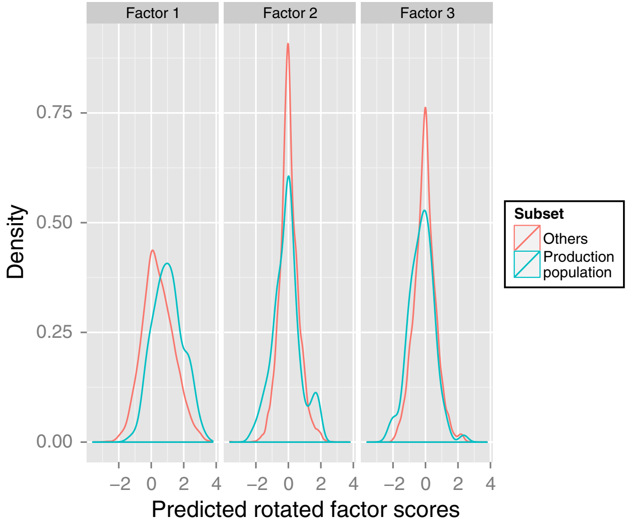


Fig. 8 Distributions of predicted (*rotated*) factor scores for parents in the RPBC production population nursery (97 individuals) and other parents and forward selections (3,698 individuals)

The variability is similar for both sets of individuals for all factors. Interestingly, parent P12, with a predicted score of 3.20 for the first factor (see Table 2) is near the top of the distribution for the production populationfor this factor, providing support for its credentials as a parent that can produce high DBH progeny in many environments.

Discussion

This paper has presented an approach to the analysis of tree breeding multi-environment trial (MET) datasets with moderate to large numbers of environments. The approach utilises a factor analytic (FA) model to capture the variance structure of the additive genotype-by-environment (A × E) effects in a parsimonious and holistic manner and uses an approximate reduced animal model to overcome computational issues. Using this approach we have demonstrated the existence of substantial A × E interaction for DBH in *Pinus radiata* in New Zealand and New South Wales (Australia). Alternative approaches based on subsetting either trials or genotypes to simplify the models or to reduce the computational burden would have failed to characterise the extent and nature of the A × E. The use of the ARA model reduced the number of equations associated with the A × E effects from approximately $21e6$ to $0.29e6$ which makes this analysis feasible for these data and other datasets of this size and complexity. The elapsed time (in seconds) per iteration for the FA3 model was approximately 300. The hardware platform used was a laptop using an Intel i7-4700MQ CPU @ 2.4 GHz and with 16GB RAM running with a 64-bit operating system.

The results from our analysis have potentially significant implications for both breeding and deployment strategies. Current strategies in the RPBC programme are based on overall performance. Since 2000 the philosophy has been to plant mixed seedlots with the aim of achieving a relatively constant average performance over a range of site types. In the MET dataset under study, this overall selection index was (weakly) related to parent sensitivity to the first factor from the FA model. It is therefore clear that much progress has been made. However, this factor only accounted for 59.1 % of the A × E variation, leaving a substantial amount of interaction that may be exploited for specific adaptation. Thus the practice of planting mixed parent seedlots may not be optimal and there could be substantial gains from planting the correct germplasm on the appropriate site types.

The challenge now is to develop an implementation strategy to capture this specific adaptation. The strategy clearly depends on the ability to predict the breeding value of an individual for environments other than those used in the current MET dataset. The obvious route to a solution is the so-called site matching approach, where new target environments can be matched with environments in the MET dataset. This is a non-trivial problem as the genetic response to the environment is high-dimensional, complex and potentially non-additive. Costa e Silva et al. (2006) demonstrated that some of the genotype-by-environment interaction in eucalyptus globulus may be due to changes in mean minimum temperature, precipitation and solar radiation for the warmest quarter of the year, and length of time when there is low soil moisture, but the responses varied according to the sub-races. Such a study for *Pinus radiata* is the subject of future research.

Author contributions BC conceived the statistical approach, derived the relevant algebra, conducted the analysis of the data and provided critical review of the drafted manuscript. PJ provided the data and information relating to the RPBC breeding programme. RT conceived the approach used to obtain forward selections and provided a review of the statistical approach. AS conceived the algebra for the latent regression plots, wrote the text for the draft paper and prepared the final manuscript. All authors have read and approved the final manuscript.

Acknowledgments Brian Cullis and Alison Smith gratefully acknowledge the financial support of the Grains Research and Development Corporation of Australia.

Conflict of interest The authors declare that they have no conflict of interest.

Appendix

Below is the ASReml-R syntax to fit the FA3 model. The ASReml-R function has numerous arguments which are described in detail in the user manual which is distributed with the package. The approach we use to fit an FA3 model is to use the REML estimates of the variance parameters from the FA2 model as starting values for the iterative fitting process for the FA3 model. We have found that this improves the chance of convergence without manual intervention. The first call to ASReml-R sets up a template which can then be populated with the appropriate starting values.

Note the formation of the design matrix for the additive effects of the parental and forward selection trees using the `and` constructor function.

The additional terms in the random model formula are terms which relate to the blocking structure of the trials. For example, the set of trials which were multi-tree plot trials is found in the data vector `pplt`, while the set of trials which were incomplete block designs is found in the data vector `pblk`. The relationship matrices for additive effects is provided in the `ginverse` argument, we require a very large amount of workspace and the term which is fitted as a sparse term is a factor with K levels where K is one more than the number of trees whose female parent was a control tree. Lastly the factor `TExpt` is a copy of the trial factor for those trials which are clonal trials else is it set to missing value indicator (NA).

```
> s3.ram.sv <- asreml(dbh ~ Expt, random = ~fa(Expt, 3):zero:ped(Fcln) +
+     and(fa(Expt, 3):half:ped(Fcln)) + and(fa(Expt, 3):half:ped(Mcln)) +
+     at(Expt):Replicate + at(Expt, s3.pset):Replicate:Setgroup +
+     at(Expt, s3.pplt):Trueplot + at(Expt, s3.pblk):Replicate:Iblk +
+     diag(TExpt):Tclone, rcov = ~at(Expt):units, data = s3.df,
+     ginverse = list(Fcln = s3.parents.ainv, Mcln = s3.parents.ainv),
+     na.method.X = "include", start.values = T, workspace = 1.8e+08,
+     sparse = ~Check)
> k <- 3
> fa2.gam <- matrix(summary(s3.ram.asr2, nice = T)$nice[[1]], ncol = k -
+     1 + 1)
> dimnames(fa2.gam) <- list(levels(s3.df$Expt), c("psi", "lam1",
+     "lam2"))
> ne <- length(levels(s3.df$Expt))
>
> all.gam <- s3.ram.asr2$gammas
> temp <- s3.ram.sv$gammas.table
> others <- rep(T, nrow(temp))
> others[grep("ped", temp$Gamma)] <- F
> these <- rep(T, length(all.gam))
> these[grep("ped", names(all.gam))] <- F
> temp$Value[others] <- all.gam[these]
> temp$Constraint <- as.character(temp$Constraint)
>
> temp$Value[grep("fa1", temp$Gamma)] <- fa2.gam[, "lam1"]
> temp$Value[grep("fa2", temp$Gamma)] <- fa2.gam[, "lam2"]
> temp$Value[grep("fa3", temp$Gamma)] <- c(0, 0, rep(0.02, ne -
+     2))
> temp$Value[grep("ped.*var", temp$Gamma)] <- fa2.gam[, "psi"] *
+     0.8
>
> s3.ram.asr3 <- asreml(dbh ~ Expt, random = ~fa(Expt, 3):zero:ped(Fcln) +
+     and(fa(Expt, 3):half:ped(Fcln)) + and(fa(Expt, 3):half:ped(Mcln)) +
+     at(Expt):Replicate + at(Expt, s3.pset):Replicate:Setgroup +
+     at(Expt, s3.pplt):Trueplot + at(Expt, s3.pblk):Replicate:Iblk +
+     diag(TExpt):Tclone, rcov = ~at(Expt):units, data = s3.df,
+     ginverse = list(Fcln = s3.parents.ainv, Mcln = s3.parents.ainv),
+     na.method.X = "include", R.param = temp, G.param = temp,
+     workspace = 4.8e+08, sparse = ~Check)
```

References

Apiolaza L (2012) Basic density of radiata pine in new zealand: genetic and environmental factors. Tree Genet Genomes 8:87–9

Bailey R (2008) Design of comparative experiments. Cambridge University Press, Cambridge

Baltunis B, Gapare W, Wu H (2010) Genetic parameters and genotype by environment interaction in radiata pine for growth and wood quality traits in australia. Silvae Genet 59:2–3

Beeck C, Cowling W, Smith A, Cullis B (2010) Analysis of yield and oil from a series of canola breeding trials. Part I: Fitting factor analytic models with pedigree information. Genome 53:992–1001

Burgueno J, Crossa J, Cotes J, Vincente F, Das B (2011) Prediction assessment of linear mixed models for multi-environment trials. Crop Sci 51:944–954

Butler DG, Cullis BR, Gilmour AR, Gogel BJ (2009) ASReml-R reference manual, release 3. Technical report, Queensland Department of Primary Industries

Costa e Silva J, Potts B, Dutkowski G, (2006) Genotype by environment interaction for growth of eucalyptus globulus in Australia. Tree Genet Genomes 2:61–75

Cullis B, Smith A, Beeck C, Cowling W (2010) Analysis of yield and oil from a series of canola breeding trials. Part II: Exploring VxE using factor analysis. Genome 53:1002–1016

Core Team R (2012) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0

Gilmour AR, Thompson R, Cullis BR (1995) AI, an efficient algorithm for REML estimation in linear mixed models. Biometrics 51:1440–1450

Gilmour AR, Cullis BR, Welham SJ, Gogel BJ, Thompson R (2004) An efficient computing strategy for prediction in mixed linear models. Comput Stat Data Anal 44:571–586

Hardner C, Dieters M, Dale G, DeLacy I, Basford K (2010) Patterns of genotype-by-environment interaction in diameter at breast height at age 3 for eucalypt hybrid clones grown for reafforestation of lands affected by salinity. Tree Genet Genomes 6:833–851

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

Kelly A, Smith A, Eccleston J, Cullis B (2007) The accuracy of varietal selection using factor analytic models for multi-environment plant breeding trials. Crop Sci 47:1063–1070

Kelly AM, Cullis BR, Gilmour AR, Eccleston JA, Thompson R (2009) Estimation in a multiplicative mixed model involving a genetic relationship matrix. Genet Select Evol 41:1286–1297

Kempton RA (1984) The use of biplots in interpreting variety by environment interactions. J Agric Sci Camb 103:123–135

Oakey H, Verbyla A, Cullis B, Pitchford W, Kuchel H (2006) Joint modelling of additive and non-additive genetic line effects in single field trials. Theor Appl Genet 113:809–819

Oakey H, Verbyla A, Cullis B, Wei X, Pitchford W (2007) Joint modelling of additive and non-additive (genetic line) effects in multi-environment trials. Theor Appl Genet 114:1319–1332

Quass RL, Pollack E (1980) Mixed model methodology for farm and ranch beef cattle testing programs. J Animal Sci 51:1277–1287

Raymond C (2011) Genotype by environment interactions for pinus radiata in new south wales, australia. Tree Genet Genomes 7:819–833

Schutz W, Cockerham CC (1966) The effect of field blocking on gain from selection. Biometrics 22:843–863

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

Stefanova K, Smith A, Cullis B (2009) Enhanced diagnostics for the spatial analysis of field trials. J Agric Biol Environ Stat 14:1–19

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

White I, Rainer PR, Knap Brotherstone S (2006) Variance components for survival of piglets at farrowing using a reduced animal model. Genet Select Evol 38:359–370

Zapata-Valenzuela J (2012) Use of analytical factor structure to increase heritability of clonal progeny tests of pinus taeda l. Chil J Agric Res 72:309–315