

# Exploring opportunities for reducing complexity of genotype-by-environment interaction models

Craig Hardner 

Received: 17 May 2017 / Accepted: 5 September 2017 / Published online: 17 October 2017  
© Springer Science+Business Media B.V. 2017

**Abstract** Heterogeneity in genetic effects among environments ( $G \times E$ ) is a common phenomenon in crop plants and can arise from heterogeneity in variance (scale effects) and/or crossover interaction. Here, a study of yield of macadamia progeny in 15 trials established at 9 locations and assessed for yield at 7 years is used to explore the impact on prediction of clonal values (additive + dominance effects) from (i) scaling observations by phenotypic standard deviation of each trial, and (ii) reducing complexity of the pattern of genotype-by-environment interaction. The initial fit of an unconstrained  $G \times E$  model to unscaled observations indicated significant  $G \times E$ , which was supported by the fit of the same model to scaled data. Scaling observations reduced heterogeneity of genetic parameter estimates among locations. Clustering of the additive and dominance genetic-by-environment covariance matrices from the fit of  $G \times E$  models to scaled observations and log-likelihood testing was used to identify reduced models where locations with apparent homogeneous genetic effects (genetic variance not significantly different, and genetic correlations not significantly different from 1) were grouped into single environments. Complexity reduction condensed the additive genetic-by-environment covariance matrix to 3

environments, and 4 environments for the dominance matrix, and the accuracy of parameter estimates increased, although accuracy of prediction as assessed by generalised heritability only improved for a few locations. On the other hand, accuracies of clonal values predicted from a main effects only  $G + E$  model were lower. Nevertheless, correlations of the averages of predicted clonal values across locations from different models were very high suggesting models are robust to parameter estimates. These results support the use of scaling by the phenotypic standard deviation to reduce heterogeneity in parameter estimates, and complexity reduction to improve accuracy of estimating parameters required to predict genetic effects.

**Keywords** Macadamia · Quantitative genetics · Multi-environment trials · Mixed-models

## Introduction

Genotype-by-environment interaction ( $G \times E$ ) is a common phenomenon in crop plants (Allard and Bradshaw 1964). Genotype-by-environment interaction may be expressed as heterogeneity in genetic variance among environments (scale changes) and/or heterogeneity in ranking of individuals (crossover interaction—COI) (Burgueno et al. 2008; Baker 1988). The presence of  $G \times E$  complicates genetic

---

C. Hardner (✉)  
Queensland Alliance for Agriculture and Food Innovation,  
University of Queensland, St Lucia, QLD 4072, Australia  
e-mail: craig.hardner@uq.edu.au

selection (Allard and Bradshaw 1964; Cooper and Delacy 1994; Comstock and Moll 1963). On one hand, if patterns in  $G \times E$  exist that can be explained by some repeatable factor, and there is sufficient advantage, elite genotypes may be selected for specific environments (Allard and Bradshaw 1964; Matheson and Cotterill 1990). Alternatively, if no repeatable factor can be identified that predict patterns,  $G \times E$  is treated as noise, reducing the repeatability of the genetic potential across the production environment. Commonly, field trials established across multiple environments (METs) are used to detect the presence and patterns of  $G \times E$ .

Linear mixed models, where genetic effects are treated as random factors, have become the standard approach for genetic evaluation, and are readily extended to modelling of METs (Smith et al. 2005). This approach provides a frame-work for combining all available information to predict the best (lowest variance) linear unbiased (BLUP) genetic effect (Henderson 1975; Patterson and Thompson 1971; Thompson and Meyer 1986; Hardner et al. 2016). However, prediction of genetic effects requires knowledge of their variance and covariance architecture, which is often estimated directly from available data (E-BLUP, Kackar and Harville 1981), adding an additional source of uncertainty to predictions.

A simple main effect and interaction model (i.e.  $G + G \times E$ ) may be used to detect  $G \times E$ , but may not be sufficiently flexible to identify patterns in  $G \times E$  (Smith et al. 2005). A more flexible approach is to treat the performance in each environments as a separate trait (Falconer 1952; Burdon 1977; Yamada 1962). This model enables estimation of unique genetic variances for each environment to account for scale effects, and pair-wise genetic covariances to model the degree of interaction (or re-ranking) among environments. Arief et al. (2015) demonstrated that the average of the off-diagonals of the most general unstructured genetic-by-environment covariance matrix is equal to the estimate of the genetic main effects variance in the simple  $G + G \times E$  model, and the average of the diagonals of the unstructured matrix is equal to the sum of the main effects and interaction variances of the  $G + G \times E$  model. However, the complexity of the an unstructured covariance matrix increases with number of environments ( $n$ ), as the number of parameters is given by  $n(n + 1)/2$ , leading

to over-parameterisation and estimation difficulties (Silva et al. 2009; Kelly et al. 2009; Hardner et al. 2010; Smith et al. 2001).

A factor analytic (FA) parameterisation of the genetic-by-environment covariance matrices has been developed to reduce the number of parameters requiring estimation (Smith et al. 2001; Piepho 1997) and provides a parsimonious approximation to the fully unstructured genetic-by-environment covariance matrix (Kelly et al. 2007). The FA parameterisation models the genetic effect of an individual at each location as the sum of: (i) the product over  $k$  hypothetical orthogonal factors of location loadings for each factor by the genetic scores of the individual for the respective factors, and (ii) a remaining genetic effect for each location not explained by the factor model (specific location deviation). The genetic-by-environment covariance matrix is estimated as the sum of the square of the vector/matrix of factor loadings and the variance in specific deviations (specific variance).

The complexity of  $G \times E$  models may be reduced through transformation of observations if  $G \times E$  is in part a consequence of variance heterogeneity. A common transformation is to scale the observation by the raw phenotypic standard deviation of each environment so that the raw phenotypic variance in each environment is 1.0 (Hill 1984; White et al. 2007). On this scale, no  $G \times E$  would be observed if the underlying heritabilities of the traits in different environments are equal and there is no re-ranking of individuals. This approach may support a reduction in the dimensionality of multi-variate models by enabling environments to be grouped into ‘mega-environments’, within which  $G \times E$  is not significant. By reducing the number of parameters requiring estimation, parameters may be more accurately estimated, possibly increasing the accuracy of predictions compared to more complex models.

This study examines the impact on prediction of genetic values of: (i) scaling observations by the phenotypic standard deviation for each specific trial prior to modelling fitting; and (ii) reducing the complexity of  $G \times E$ . The motivating example is a series of trials of macadamia progeny established at nine locations across the main production area in eastern Australia that have been assessed for yield at 7 years. Macadamia is a perennial sub-tropical tree that produces highly valued kernels that are consumed as

snack food, enrobed in chocolate, ingredients in bakery products and ice-cream, or edible oil (Hardner et al. 2009). Detailed interpretation and relevance for genetic improvement of macadamia of the  $G \times E$  patterns for this, and other traits, will be undertaken in a subsequent paper.

## Methods

### Motivating example

Field trials were established across the macadamia production area of Northern NSW and South-East Queensland, Australia. Observations for yield of NIS (nut-in-shell at 1.5% moisture content) 7 years after planting were available for 2068 macadamia individuals established across 14 trials at 9 locations (Tables 1, 2). All locations except East-Bundaberg (Bq) were grower farms. 2040 of the individuals were seedling progeny grown on own-roots produced from crossing 44 parents over 3 years (1999, 2000 and 2001) to produce 101 families. Progeny were established in field trials 2 years after crossing had been made (i.e. 2001, 2002 and 2003). Field trials also included individuals from 28 of the 44 parents, planted as scions grafted onto seedling rootstocks from open-pollinated seeds of the cultivar “H2”. Grafted parents were not replicated within each trial except for limited replication at the two trials at East Bundaberg (Bq1 and Bq3) and the 2003 Baffle Creek trial (Bc3). Individuals were laid out in incomplete blocks of single-tree plots designed with family as the treatment effect. The spatial arrangement of trials was defined as planting rows (*Row*) by planting spaces along planting rows (*Sp*) with distance between rows varying from 5 to 8 m (Table 1), but 4 m between plants within a row, except at Nn (5 m).

NIS per tree was assessed at 7 years after planting at each trial (i.e. 2006, 2007 and 2008). Fruit on the ground prior to March were considered immature (i.e. incomplete oil accumulation) and were not included in yield assessment. In general, four ground harvests of abscised fruit were undertaken throughout the season at approximately 6 weekly intervals: the first in mid-April, the second at the end of May or start of June, mid-July and late August, early September. Immediately following the last ground harvest, all remaining fruit were stripped from the tree. Harvested fruit were

transported to a central location and dehusked. Total wet nut-in-shell (WNIS) was assessed for each tree-by-harvest collection, from which 100 nuts were randomly sampled, dried to 1.5. % moisture content, weighed and used to convert WNIS to NIS for each tree-by-harvest collection. To examine the effect of missing neighbouring trees on prediction of genetic potential, status of neighbouring trees along the planting row was recorded as 0 neighbours, 1 neighbour or both (2) neighbours alive at the age of assessment.

### Statistical methods

Consistent with Falconer (1952) and others (Smith et al. 2001, 2005; Cullis et al. 2014; Hardner et al. 2010) performance in different environments was treated as a unique attribute. The genetic model assumed the total genetic effect of the  $i$ th individual in the  $q$ th environment was composed of additive and dominance genetic effects, i.e.

$$g_{iq} = a_{iq} + d_{iq}$$

The general  $G \times E$  model for the additive and dominance genetic effects of  $m$  individuals, assessed at  $t$  trials in  $w$  environments was:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}_u\mathbf{u} + \mathbf{Z}_a\mathbf{a} + \mathbf{Z}_d\mathbf{d} + \mathbf{r}$$

where  $\mathbf{y}$  was the vector of observation,  $\mathbf{b}$  was a vector of unknown fixed effects including the general mean, trial, and linear *Row* and linear *Sp*, neighbour status (NI) and propagation method (PM) for each trial,  $\mathbf{X}$  was a design matrix that mapped the observations onto the unknown fixed effects,  $\mathbf{u}$  was a vector of unknown random non-genetic effects including *Block*, *Row* and *Sp* effects for each attribute-by-trial,  $\mathbf{Z}_u$  was a design matrix that mapped the observations onto the unknown random effects,  $\mathbf{a}$  was a vector of unknown additive genetic effects for the  $i$ th individual (including ancestors) for the  $q$ th environment,  $\mathbf{Z}_a$  was a design matrix that mapped the observations onto the additive genetic effects,  $\mathbf{d}$  was vector of unknown dominance genetic effects for the  $i$ th individual (including ancestors) for the  $q$ th environment,  $\mathbf{Z}_d$  was a design matrix that mapped the observations onto the non-additive genetic effects, and  $\mathbf{r}$  was a vector of unknown random residual effects for each observation.

Variance of  $\mathbf{y}$  was defined as

**Table 1** Design of macadamia progeny trials analysed in this study including; location and trial code (Loc, Trial), year of planting (Year), number of planting rows (nRow), planting spaces (nSp), blocks (nBlk), parents that contributed to progeny (nPPar), families (nPFam), progeny individuals (nPInd), grafted parents planted in trial (nCv) and observations (nobs), and mean (mn) and variance (v) of raw (y) and square-root transformed (y) observations of NIS per tree at 7 years of age (kg). Also shown is the distance between planting rows at the different trials (Row(m))

Region	Site	Location	Trial	Year	nRow	Row(m)	nSp	nBlk	nPPar	nPFam	nPInd	nCvs	nobs	mn(y)	v(y)	mn(ty)	v(ty)
Bundaberg	Baffle Creek	Bc	Bc2	2002	6	8	36	6	40	46	119	20	141	2.84	6.07	1.47	0.68
Bundaberg	Baffle Creek	Bc	Bc3	2003	8	8	32	8	38	44	141	21	175	5.55	16.85	2.16	0.91
Bundaberg	N. Bundaberg	Bn	Bn1	2001	1	8	183	6	38	60	159	24	183	4.86	15.54	1.97	0.97
Bundaberg	E. Bundaberg	Bq	Bq1	2001	30	5	15	14	38	66	359	25	411	5.21	10.95	2.11	0.76
Bundaberg	E. Bundaberg	Bq	Bq3	2003	9	5	28	4	38	45	104	20	134	3.79	9.20	1.76	0.70
Bundaberg	Alloway	Bs	Bs2	2002	2	8	93	6	40	59	148	24	173	4.91	8.54	2.08	0.60
Gympie	Amamoor	Ga	Ga2	2002	7	7	28	6	40	59	162	23	186	3.37	6.73	1.66	0.62
Gympie	Amamoor	Ga	Ga3	2003	5	7	28	4	39	46	101	22	124	3.00	4.58	1.59	0.49
Gympie	E. Gympie	Ge	Ge1	2001	19	5	11	6	38	58	141	20	161	0.36	0.39	0.39	0.20
Gympie	Yandina	Gy	Gy1	2001	11	5	31	6	38	62	150	24	174	0.49	0.73	0.45	0.28
Gympie	Yandina	Gy	Gy2	2002	6	5	34	6	39	55	127	20	148	0.49	0.46	0.53	0.21
NNSW	Dunoon	Nd	Nd1	2001	10	5	21	6	38	65	178	24	202	3.39	8.66	1.61	0.79
NNSW	Dunoon	Nd	Nd3	2003	8	5	16	4	38	34	61	12	73	5.02	10.12	2.10	0.63
NNSW	Newrybar	Nn	Nn2	2002	6	8	33	6	40	45	90	22	112	1.48	2.75	1.03	0.44

**Table 2** Congruence of progeny families (upper triangle) and all (progeny + grafted parents) individuals (lower triangle) assessed for NIS per tree at 7 years of age at 14 trials (see Table 1) across major production regions in Australia

	Bc2	Bc3	Bn1	Bq1	Bq3	Bs2	Ga2	Ga3	Ge1	Gy1	Gy2	Nd1	Nd3	Nn2
Bc2	139\46	16	25	26	15	45	45	17	23	24	43	25	12	37
Bc3	18	162\44	22	26	41	21	22	42	23	22	19	24	34	16
Bn1	19	21	183\60	59	21	31	30	20	54	58	31	60	15	26
Bq1	19	21	24	384\66	25	36	36	24	57	61	34	64	18	29
Bq3	17	18	20	20	124\45	20	21	41	22	21	18	23	33	15
Bs2	18	20	23	23	19	172\59	56	22	31	32	54	35	16	43
Ga2	16	18	21	21	17	22	185\59	23	31	31	53	34	17	45
Ga3	17	19	22	22	18	21	19	123\46	21	20	20	22	33	17
Ge1	16	17	20	20	18	19	17	18	161\58	55	30	56	17	26
Gy1	19	21	24	24	20	23	21	22	20	174\62	32	61	16	26
Gy2	15	17	19	19	16	19	18	17	15	19	147\55	34	15	41
Nd1	19	21	24	24	20	23	21	22	20	24	19	202\65	16	28
Nd3	10	12	12	12	9	12	10	12	9	12	9	12	73\34	12
Nn2	18	19	21	21	17	21	19	19	17	21	17	21	11	112\45

$$\text{var}(\mathbf{y}) = \mathbf{V} = \mathbf{Z}_u \mathbf{G}_u \mathbf{Z}'_u + \mathbf{Z}_a \mathbf{G}_a \mathbf{Z}'_a + \mathbf{Z}_d \mathbf{G}_d \mathbf{Z}'_d + \mathbf{R}$$

where  $\mathbf{G}_u$  was the variance–covariance matrix of the random non-genetic effects among trials,  $\mathbf{G}_a$  was variance–covariance matrix of the random additive genetic effects among environments,  $\mathbf{G}_d$  was the variance–covariance matrix of the random dominance genetic effects among environments, and  $\mathbf{R}$  was the variance–covariance of the residual effects of the observations.

Non-genetic random factors at each trial were considered independent, hence  $\mathbf{G}_u$  was diagonal, with sub-blocks for each unique trial-by-non-genetic random effect combination.

The variance–covariance matrix of additive genetic effects among environments,  $\mathbf{G}_a$ , was modelled as a two-way separable process:

$$\mathbf{G}_a = \mathbf{A} \otimes \mathbf{E}_a$$

where  $\mathbf{A}$  was the matrix of additive genetic relationships among individuals (including ancestors) formed from historical pedigree relationships (Henderson 1975) and  $\mathbf{E}_a$  was the additive genetic-by-environment covariance matrix.

Similarly, the variance–covariance matrix of dominance genetic effects among environments,  $\mathbf{G}_d$ , was modelled as:

$$\mathbf{G}_d = \mathbf{D} \otimes \mathbf{E}_d$$

where  $\mathbf{D}$  was matrix of dominance genetic relationships among individuals (including ancestors, also estimated from the historical pedigree, ignoring inbreeding, Henderson 1985) and  $\mathbf{E}_d$  was the dominance genetic-by-environment covariance matrix.

The variance–covariance of residual effects,  $\mathbf{R}$ , was modelled as a block diagonal matrix with each block representing the variance–covariance matrix of residual effects for each trial,  $\mathbf{R}_j$ .  $\mathbf{R}_j$  was modelled as an anisotropic separable first-order autoregressive correlation structure in the two spatial dimensions (*Row* and *Sp*) with an independent error (nugget) term (Gilmour et al. 1997; Cullis et al. 1998; Costa e Silva et al. 2001; Costa e Silva and Graudal 2008; Smith et al. 2001; Dutkowski et al. 2002):

$$\mathbf{R}_j = \mathbf{R}_{\epsilon_j} + \mathbf{R}_{\eta_j}$$

where  $\mathbf{R}_{\epsilon_j}$  was the variance–covariance among spatially dependent residual effects, and  $\mathbf{R}_{\eta_j}$  was the variance–covariance among spatially independent residual (nugget) effects, at the *j*th trial.  $\mathbf{R}_{\epsilon_j}$  was modelled as

$$\mathbf{R}_{\epsilon_j} = \left[ \mathbf{P}_{Row_j}(\rho_{Row_j}) \otimes \mathbf{P}_{Sp_j}(\rho_{Sp_j}) \right] \sigma_{\epsilon_j}^2$$

where  $\mathbf{P}_{Row_j}(\rho_{Row_j})$  and  $\mathbf{P}_{Sp_j}(\rho_{Sp_j})$  were first-order autoregressive correlation matrix along the *Row* and

$Sp$  dimensions, respectively, and  $\sigma_{e_j}^2$  was the spatially dependent residual variance at the  $j$ th trial. The matrix  $\mathbf{R}_{\eta_j}$  was modelled as a two-way separable process:

$$\mathbf{R}_{\eta_j} = \mathbf{I}_{\eta_j} \sigma_{\eta_j}^2$$

where  $\sigma_{\eta_j}^2$  was the spatially independent residual variance for the  $j$ th trial.

Parameters of the mixed model were estimated using Restricted Maximum Likelihood approaches implemented in the statistical software ASReml (Gilmour et al. 2009). Estimation of standard errors of parameters, testing and estimation of fixed effects, and prediction of genetic effects was also undertaken using this package.

#### Individual trial analyses of transformed observations

Independent univariate models were initially fitted to raw observations from each trial to evaluate the assumption of normality of residuals, and identify significant sources of non-genetic variation for inclusion in multi-variate models. In these individual trial analyses, the additive and dominance genetic-by-environment, and non-genetic- and residual-by-trial covariance matrices described above were reduced to within trial variances. Analyses indicated that a square-root transformation of the raw observations ( $y$ ) was required to normalise the residual distributions. Following model fitting, significance of fixed effects was evaluated using Wald tests (Kenward and Roger 1997). Significance of random effects were evaluated using likelihood ratio test (Wilks 1938) that was adjusted for the case where the null hypothesis was that the parameter was at the boundary of the estimation space (Stram and Lee 1994). Akaike Information Criterion (Akaike 1974) (AIC) was used to more generally compare the goodness-of-fit among models with common fixed effects with lower AIC indicating a more parsimonious fit.

#### Fit of unconstrained $G \times E$ model to unscaled observations

An unconstrained  $G \times E$  model (M01, i.e. individual locations treated as environments for the additive— $\mathbf{E}_a$  and dominance genetic by environment— $\mathbf{E}_d$ ,

covariance matrices in the general  $G \times E$  model) was fitted to square-root transformed observations ( $y$ ) to describe patterns of genetic effects across locations for unscaled observations. Factor analytic order 1 (FA1) parameterisations (Smith et al. 2001) were used to model both the additive and dominance genetic-by-environment covariances matrices. As locations were treated as unique environments, genetic variances at trials within locations were constrained to be equal with a genetic correlation of 1. Non-genetic effects and residuals were modelled at the trial level. Only non-genetic effects significant in individual trial analyses were fitted.

To evaluate the accuracy of estimated FA1 model parameters, a  $z$ -score was estimated as the ratio of the parameter estimate to its standard error. Estimated parameters for the FA1 model of the additive and dominance genetic-by-environment covariance matrices were used to estimate the full additive and dominance genetic-by-location covariance matrices (Smith et al. 2001). These were summed to estimate the total (additive + dominance) genetic-by-location covariance matrix. Coefficient of variation of the estimates was calculated to quantify the heterogeneity in estimates among locations.

Additive and dominance genetic effects for each individual in the pedigree were predicted for each location using the estimated parameters from the model fit, and summed to obtain clonal values for each location ( $g_{M01.y}$ ). Average individual clonal values across locations were also obtained.

To quantify the accuracy of predicted additive effects, generalised narrow sense heritability at the  $q$ th location (Piepho and Mohring 2007; Cullis et al. 2006; Oaquet et al. 2006) was estimated as:

$$\hat{h}_q^{2*} = 1 - \frac{\bar{\sigma}_{\Delta A,q}^2}{2 \times \hat{\sigma}_{A,q}^2}$$

where  $\bar{\sigma}_{\Delta A,q}^2$  was the mean variance of the difference of additive predictions at the  $q$ th location, estimated from the prediction error variance matrix of additive effects and  $\hat{\sigma}_{A,q}^2$  was the estimated additive genetic variance at the  $q$ th location. A similar expression was used to estimate the generalised dominance heritability ( $\hat{d}_q^{2*}$ ). Generalised heritability is a function of accuracy, and, as such is different from classical individual heritability estimates and are only comparable for balanced designs (Cullis et al. 2006).

A singular value decomposition (SVD) of the table of predicted clonal values of individuals-by-location, standardised by location, was undertaken to study patterns in the interaction. Biplots (Kempton 1984) of the first and second location and individual singular vectors was undertaken with the location vectors displayed as lines from the origin and individuals as points. The angle between two location vectors on the biplot represents the correlation in clonal values among locations explained by the first and second vectors.

#### Scaling observations by within trial phenotypic variation

To evaluate the impact of scaling observations prior to model fit on the prediction of genetic values, the same unconstrained  $G \times E$  model used above (*M01*) was also fitted to the square-root transformed raw observations standardised by the phenotypic standard deviation of these observations for the respective trial (*sy*) so that observations within each trial were scaled to a unit scale (i.e. phenotypic variance of each trial equal to 1). As described above, *z*-scores were estimated to evaluate the accuracy of estimated FA1 model parameters, and these parameters were used to estimate the full additive and dominance genetic-by-location covariance matrices (which were summed to estimate the total genetic-by-location covariance matrix), and predict additive and dominance effects on the unit scale for each location. Predicted additive and dominance genetic effects were summed to predict clonal values by location on the unit scale (*sg\_M01.sy*) and the PEVs of these predictions were used to estimate generalised narrow sense (and dominance) heritability for each location. De-scaled clonal values (*g\_M01.sy*, i.e. predicted clonal values from the analysis of scaled observation on the square-root scale) were obtained by multiplying *sg\_M01.sy* by the phenotypic standard deviation of the respective location (obtained by taking the square-root of the average of the phenotypic variances of each trial at the respective location). Biplot of the SVD of the table of de-scaled clonal values were used to examine  $G \times E$  patterns as described above.

#### Reduced dimension $G \times E$ models

To evaluate the impact of reducing the complexity of  $G \times E$  models on prediction of genetic effects, sequential analyses of scaled square-root transformed observations (*sy*) were undertaken with models in which multiple locations of either the additive or dominance genetic-by-location covariance matrices were constrained to be the same environment if genetic effects were homogenous (i.e. equal variance, and genetic correlation of 1) across these locations. To detect groups of locations for which genetic effects appeared homogenous (i.e. environments), cluster analyses were undertaken of the genetic (either additive or dominance)-by-environment covariance (not correlation) matrices estimated from an unreduced model. Environments were grouped using Ward's minimum distance and Gruvaeus and Wainer algorithm (Hahsler et al. 2008) was used to order nodes of the dendrogram so that an environment at the edge of a cluster was adjacent to environments to which it was most similar in the neighbouring cluster. A reduced model was then constructed so that the genetic variance of environments that appeared most similar were constrained to be equal, and genetic correlation among these to one, with all other terms equivalent to the unreduced model. Following the fit of the reduced model, log-likelihood testing ( $\alpha = 0.05$ ) was undertaken and AIC examined to identify the most parsimonious model between the reduced and unreduced models. Where the reduced model was found to be more parsimonious than the unreduced model, the constraint applied in the reduced model was maintained in subsequent model fits. The initial unreduced model examined for complexity reduction was the un-constrained  $G \times E$  model described above (*M01*) where locations represented the environments of the additive and dominance genetic-by-environment covariance matrices. The most parsimonious reduced model (*M13*) was defined as the model for which any reduced model was a significantly poorer fit to the data.

As above, *z*-values of the estimated FA1 model parameters were used to evaluate the accuracy of model parameters. In addition, the total genetic-by-location covariance matrix was estimated by summing the additive and dominance genetic-by-location covariance matrices estimated from the FA1 model parameters. Clonal values for each individual by

location were predicted on the unit scale ( $sg\_M13.sy$ ), generalised narrow sense (and dominance) heritability estimated, trial phenotypic standard deviations were used to de-scale clonal values, biplots of the SVD of de-scaled predicted clonal values by locations were used to examine  $G \times E$  patterns, and average de-scaled clonal values across locations were obtained.

To evaluate the effect of not accounting for  $G \times E$  on the prediction of genetic effects, a main effects only  $G + E$  model that assumed no interaction in additive and dominance genetic effects with location (M00, i.e. common additive and dominance genetic variance and genetic correlation of 1 among locations) was fitted to the scaled square-root observations ( $sy$ ). This model maintained the same terms for individual trial non-genetic and residual variation as the previous models. Generalised heritabilities on the unit scale and predictions of unscaled clonal values ( $g\_M01.sy$ ) were estimated as above.

## Results

### Data characteristics

Phenotypic variance in raw untransformed observations of NIS at 7 years ranged from 0.39 at the East Gympie trial (Ge1) to 16.8 at the Baffle Creek 2003 planting (Bc3) (Table 1), a magnitude of over 40 fold. However, phenotypic variance of transformed (square-root) observations, undertaken to normalise the distribution of within trial residuals, ranged from 0.20 to 0.97, less than five fold difference. There was strong association between trial average and variance on both the transformed and untransformed scale.

### Fit of unconstrained $G \times E$ model (M01) to unstandardised data ( $y$ )

Across all locations, the loadings of the FA1 model of the additive genetic-by-environment covariance matrix of the unconstrained  $G \times E$  model (M01) explained on average 94% of the additive genetic variation in unscaled square-root transformed observations for NIS at age seven (Tables 3, 4). Loadings for the FA1 model of the dominance genetic-by-environment covariance matrix explained 73%. Additive genetic loadings explained 100% of the additive genetic variance at most locations, except for at Ga

(71%), Ge (81%) and Gy (96%) (Tables 3, 4). Dominance genetic loadings incompletely explained dominance variation at Bn (56%), Bq (84%), Bs (45%), Ge (99%) and Nn (1%).

Additive genetic variance of unscaled NIS on the square-root scale estimated using the unconstrained  $G \times E$  model (M01) ranged from 0.00 to 0.20, with a coefficient of variation of 0.78 (Tables 3, 4). Coefficient of estimated dominance variation was of similar magnitude (0.71) and coefficient of variation of estimates of total genetic variance was 0.59. At most locations (Bc, Bn, Bq, Bs, Ga and Gy), the estimated percentage of total genetic variance that was additive genetic was less than the percentage due to dominance (Tables 3, 4), with an average of 41%. At Bs, all genetic variance was estimated to be dominance, and at Ge all genetic effects were additive. However, estimated generalised heritability of additive genetic effects was higher than estimated generalised heritability of dominance effects for all locations.

In agreement with parameter estimates for the FA models, the first dimension of the singular value decomposition of the location standardised clonal (additive + dominance) values predicted from the fit of the M01 model to unscaled observations explained a large proportion of variance (84%), with the second dimension explaining 14% (Fig. 1a). The locations Bc, Bn, Bq, Ga, Gy and Nd formed a group within which clonal values were highly correlated among locations. Clonal values at Nn and Ge were also highly correlated between these locations, but less so with clonal values at Bc, Bn, Bq, Ga, Gy and Nd. Clonal values at Bs were not highly correlated with clonal values at any other location.

### Fit of unconstrained $G \times E$ model (M01) to observation scaled by the phenotypic standard deviation ( $sy$ )

The percentage of additive genetic and dominance genetic variance explained by loadings of the FA1 models of the respective genetic-by-environment covariance matrices estimated from the fit of the unconstrained  $G \times E$  model (M01) to observations scaled by the phenotypic standard deviation of the respective trial ( $sy$ ) was virtually identical to that for the fit of the same model to unscaled observations (Tables 3, 4). Similarly, there was no large differences in the  $z$ -value of the FA1 parameter estimates between



**Table 3** Estimated loadings of common hypothetical factor ( $\lambda$ ) and specific variances ( $\psi$ ) and z-scores ( $z$ ) at each location for a factor analytic (FA) order 1 model of the additive (A) and dominance (D) genetic-by-environment covariance matrices fitted to square-root observations ( $y$ ), or square-root transformed observations standardised by the trial phenotypic standard deviation ( $sy$ ), for NIS yield per tree at 7 years (kg) from macadamia progeny assessed at 9 locations (see Table 1 for details) for a fully unconstrained model that treated each location as an unique environment in the parameterisation of the additive and dominance genetic-by-environment covariance matrices (M01), or the most parsimonious model for which locations with similar genetic variance and high genetic correlation were constrained to be the same environment (M13, see Table 5)

Trait	Mod	Param	Bc		Bn		Bq		Bs		Ga		Ge		Gy		Nd		Nn	
			est	z	est	z	est	z	est	z	est	z	est	z	est	z	est	z	est	z
		n	316		183		545		173		310		161		322		275		112	
y	M01	$\psi_A$	0.00	NA	0.00	NA	0.00	NA	0.00	NA	0.01	0.39	0.01	0.47	0.00	0.05	0.00	0.00	NA	NA
sy	M01	$\psi_A$	0.00	NA	0.00	NA	0.00	NA	0.00	NA	0.02	0.34	0.05	0.46	0.00	0.06	0.00	0.00	NA	NA
sy	M13	$\psi_A$	0.00	NA	0.00	NA	0.00	NA	0.00	NA	0.00	NA	0.00	NA	0.00	NA	0.00	0.00	NA	NA
y	M01	$\lambda_A$	0.28	3.62	0.44	4.41	0.31	4.46	-0.04	-0.37	0.15	2.38	0.22	3.78	0.11	2.50	0.45	6.17	0.32	3.76
sy	M01	$\lambda_A$	0.31	3.62	0.45	4.40	0.36	4.44	-0.05	-0.38	0.20	2.33	0.50	3.84	0.22	2.56	0.53	6.22	0.49	3.78
sy	M13	$\lambda_A$	0.27	4.94	0.47	7.40	0.27	4.94	-0.05	-0.39	0.27	4.94	0.47	7.40	0.27	4.94	0.47	7.40	0.47	7.40
y	M01	$\psi_D$	0.00	NA	0.15	0.66	0.04	0.82	0.16	1.15	0.00	NA	0.00	NA	0.00	NA	0.00	NA	0.08	0.50
sy	M01	$\psi_D$	0.00	NA	0.15	0.67	0.06	0.79	0.26	1.13	0.00	NA	0.00	NA	0.00	NA	0.00	NA	0.18	0.47
sy	M13	$\psi_D$	0.00	NA	0.10	1.48	0.10	1.48	0.25	1.03	0.00	NA	0.00	NA	0.00	NA	0.00	NA	0.00	NA
y	M01	$\lambda_D$	0.39	4.40	0.43	3.87	0.47	6.69	0.36	2.91	0.31	4.13	0.00	0.05	0.22	4.14	0.30	3.51	0.03	0.26
sy	M01	$\lambda_D$	0.43	4.36	0.45	3.95	0.55	6.80	0.47	2.94	0.41	4.03	0.01	0.05	0.43	4.00	0.35	3.51	0.02	0.14
sy	M13	$\lambda_D$	0.41	7.63	0.48	5.80	0.48	5.80	0.49	3.00	0.41	7.63	0.01	0.12	0.41	7.63	0.41	7.63	0.01	0.12

**Table 4** Estimated additive (vA), dominance (vD) and total (vG) genetic variance, proportion of variance explained by loadings (%λ) and genetic variance due to additive genetic variation (%A) derived from estimated parameters presented in Table 3, and generalised narrow sense ( $h^{2*}$ ) and dominance ( $d^{2*}$ ) heritability at each location

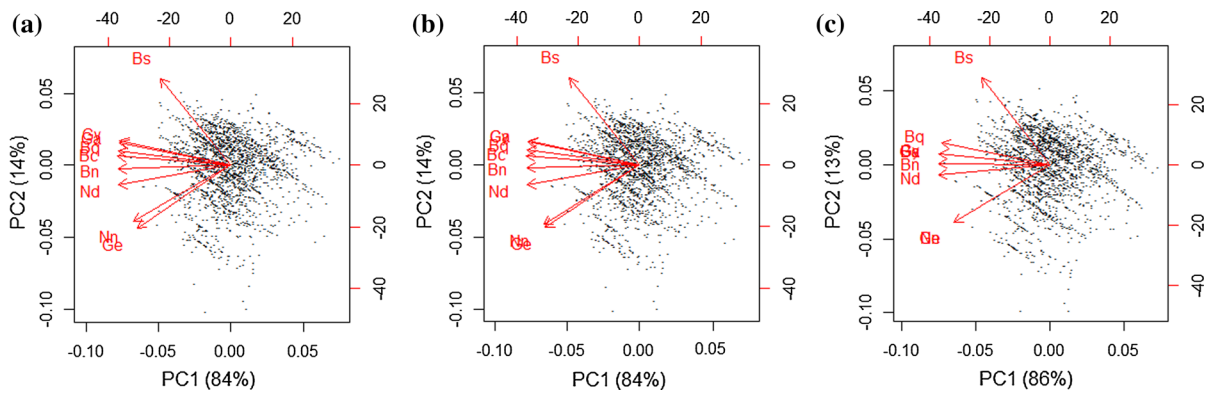
Trait	Mod	Param	Bc	Bn	Bq	Bs	Ga	Ge	Gy	Nd	Nn
y	M01	vA	0.08	0.19	0.10	0.00	0.03	0.06	0.01	0.20	0.11
sy	M01	vA	0.09	0.20	0.13	0.00	0.05	0.30	0.05	0.28	0.24
sy	M13	vA	0.08	0.22	0.08	0.00	0.08	0.22	0.08	0.22	0.22
y	M01	%λ <sub>A</sub>	1.00	1.00	1.00	1.00	0.71	0.81	0.96	1.00	1.00
sy	M01	%λ <sub>A</sub>	1.00	1.00	1.00	1.00	0.73	0.82	0.96	1.00	1.00
sy	M13	%λ <sub>A</sub>	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
y	M01	vD	0.15	0.33	0.26	0.29	0.10	0.00	0.05	0.09	0.09
sy	M01	vD	0.19	0.35	0.36	0.49	0.17	0.00	0.19	0.12	0.18
sy	M13	vD	0.17	0.34	0.34	0.49	0.17	0.00	0.17	0.17	0.00
y	M01	%λ <sub>D</sub>	1.00	0.56	0.84	0.45	1.00	0.99	1.00	1.00	0.01
sy	M01	%λ <sub>D</sub>	1.00	0.56	0.84	0.46	1.00	0.97	1.00	1.00	0.00
sy	M13	%λ <sub>D</sub>	1.00	0.69	0.69	0.49	1.00	1.00	1.00	1.00	1.00
y	M01	vG	0.23	0.52	0.36	0.29	0.13	0.06	0.06	0.29	0.19
sy	M01	vG	0.28	0.55	0.49	0.49	0.22	0.30	0.24	0.40	0.43
sy	M13	vG	0.24	0.56	0.41	0.49	0.24	0.22	0.24	0.39	0.22
y	M01	%A	33	37	26	0	26	100	19	69	55
sy	M01	%A	34	36	26	0	25	100	21	70	57
sy	M13	%A	31	40	18	0	31	100	31	57	100
y	M01	$h^{2*}$	0.46	0.46	0.46	0.45	0.35	0.42	0.44	0.46	0.46
sy	M01	$h^{2*}$	0.46	0.46	0.46	0.46	0.36	0.42	0.44	0.46	0.46
sy	M13	$h^{2*}$	0.45	0.45	0.45	0.44	0.45	0.45	0.45	0.45	0.45
y	M01	$d^{2*}$	0.30	0.21	0.30	0.19	0.30	0.30	0.30	0.30	0.05
sy	M01	$d^{2*}$	0.31	0.22	0.30	0.20	0.31	0.30	0.31	0.31	0.04
sy	M13	$d^{2*}$	0.29	0.29	0.29	0.20	0.29	0.29	0.29	0.29	0.29

the fit of the unconstrained G × E model to scaled (sy) or unscaled (y) data.

Coefficients of variation of additive, dominance and total genetic variance components among locations estimated from the fit of the unconstrained G × E model (M01) to scaled data (0.68, 0.61, 0.30, Tables 3, 4) were lower than from the fit of the same model to the unscaled scaled data (see above). However, there was only a negligible difference in estimated proportion of total genetic variation explained by additive effects between the fits of the M01 model to scaled or unscaled data (Tables 3, 4), with an average of 41%. Generalised heritabilities of additive and dominance effects predicted from the fit of the M01 model to scaled observations were only

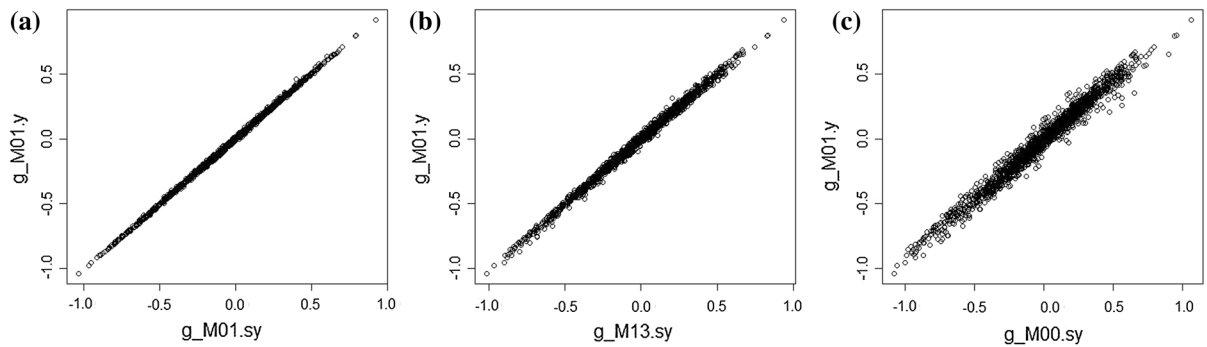
slightly higher compared to the fit of the same model to unscaled observations, except for a slightly smaller dominance variance at Nn (Tables 3, 4).

The structure of the biplot of de-scaled clonal values from the fit of M01 to the scaled observations (Fig. 1b), was virtually the same as that for the plot of the fit of the same model to unscaled data. The only small differences are that the genetic correlation between Nn and Ge, and Gy and Ga, were slightly higher. Average de-scaled clonal values across locations predicted from the fit of M01 to scaled data, was also almost perfectly correlated (0.9996) with average clonal values across locations predicted from the fit of the same model to unscaled observations (Fig. 2a).



**Fig. 1** Biplot of first 2 dimensions of singular value decomposition of location standardised clonal values (additive + dominance values) (a) predicted from the fit of an unconstrained  $G \times E$  model (M01, see Table 5) to square-root transformed NIS yield per tree ( $y$ ) at age 7 at 9 locations, (b) de-scaled predictions from the fit of M01 to square-root transformed

observations scaled by the standard deviation of the transformed observations at each location ( $sy$ ) and (c) de-scaled predictions from the fit of a reduced model (M13, Table 5) to square-root transformed observations scaled by the standard deviation of the transformed observations at each location ( $sy$ )



**Fig. 2** Plot of average of clonal value for NIS yield per tree at 7 years of age (kg) predicted from a fit of a unconstrained  $G \times E$  to unscaled observations ( $g_{M01.y}$ ) by (a) average of de-scaled clonal values predicted from the fit of a unconstrained  $G \times E$  model to observations scaled by the phenotypic standard deviation of the respective trial ( $g_{M01.sy}$ ), (b) average of de-

scaled clonal values predicted from the fit of the most parsimonious reduced  $G \times E$  model to observations scaled by the phenotypic standard deviation of the respective trial ( $g_{M13.sy}$ ), and (c) de-scaled clonal values predicted from the fit of a unconstrained  $G \times E$  model to observations scaled by the phenotypic standard deviation of the respective trial ( $g_{M00.sy}$ )

#### Fit of most parsimonious reduced $G \times E$ model (M13) to scaled observations ( $sy$ )

The most parsimonious reduced  $G \times E$  model (M13, Table 5) fitted to scaled observations ( $sy$ ) identified three groups of locations (environments) for the additive genetic-by-location covariance matrix, and four environments for the dominance genetic-by-location covariance matrix, within which genetic variances for each locations were constrained to be equal and genetic correlations were constrained to 1.0. This model required estimation of 37 parameters

compared to 53 for the fully unconstrained model (M01).

Factor loadings for the FA1 model of the additive genetic-by-environment covariance matrix for the fit of M13 to scaled data ( $sy$ ) explained 100% of the additive genetic variance at each location, compared to the 94% for M01 fitted to unscaled and scaled observations (Tables 3, 4). Similarly, loadings for the M13 model of the dominance genetic-by-environment covariance matrix explained 88% of the dominance variance at each location, compared to 76% for M01. In particular, 100% of dominance genetic variance in scaled observations at Nn was explained by the M13

**Table 5** Model alias, structure of additive (AE) and dominance (AE) genetic variance–covariance by environment matrices, log-likelihood (Logl), and degrees of freedom (df) for alternative models of the genetic by environment variance–covariance matrices in the analysis of squared NIS at 7 years

for macadamia seedling progeny and grafted parents planted in 14 trials in 9 locations (see Table 1 for full description) across 3 regions of the Australian macadamia production zone and scaled by the within trial phenotypic standard deviation

Model	AE	DE	Logl	df
M01	Bc, Bn, Bq, Bs, Ga, Ge, Gy, Nd, Nn	Bc, Bn, Bq, Bs, Ga, Ge, Gy, Nd, Nn	−940.20	53
M02	Bc, Bn, Bq, Bs, Ga, Ge, Gy, Nd, Nn	Bc = Gy, Bn, Bq, Bs, Ga, Ge, Nd, Nn	−940.20	52
M03	Bc, Bn, Bq, Bs, Ga, Ge, Gy, Nd, Nn	Bc = Ga = Gy, Bn, Bq, Bs, Ge, Nd, Nn	−940.22	51
M04	Bc, Bn, Bq, Bs, Ga, Ge, Gy, Nd, Nn	Bc = Ga = Gy = Nd, Bn, Bq, Bs, Ge, Nn	−940.30	51
M05	Bc, Bn, Bq, Bs, Ga, Ge, Gy, Nd, Nn	Bc = Ga = Gy = Nd, Bq = Bn, Bs, Ge, Nn	−940.64	47
M06	Bc, Bn, Bq, Bs, Ga, Ge, Gy, Nd, Nn	Bc = Bq = Bn = Ga = Gy = Nd, Bs, Ge, Nn	−943.66	47
M07	Bc, Bn, Bq, Bs, Ga, Ge, Gy, Nd, Nn	Bc = Ga = Gy = Nd, Bq = Bn, Ge = Nn, Bs	−940.82	46
M08	Nd = Nn, Bc, Bn, Bq, Bs, Ga, Ge, Gy	Bc = Ga = Gy = Nd, Bq = Bn, Ge = Nn, Bs	−940.22	45
M09	Ga = Gy, Nd = Nn, Bc, Bn, Bq, Bs, Ge	Bc = Ga = Gy = Nd, Bq = Bn, Ge = Nn, Bs	−940.25	44
M10	Bc = Bq, Ga = Gy, Nd = Nn, Bn, Bs, Ge	Bc = Ga = Gy = Nd, Bq = Bn, Ge = Nn, Bs	−940.34	43
M11	Bc = Bq, Ga = Gy, Bn = Nd = Nn, Bs, Ge	Bc = Ga = Gy = Nd, Bq = Bn, Ge = Nn, Bs	−941.67	41
M12	Bc = Bq, Ga = Gy, Bn = Ge = Nd = Nn, Bs	Bc = Ga = Gy = Nd, Bq = Bn, Ge = Nn, Bs	−941.88	39
M13	Bc = Bq = Ga = Gy, Bn = Ge = Nd = Nn, Bs	Bc = Ga = Gy = Nd, Bq = Bn, Ge = Nn, Bs	−943.72	37
M14	Bc = Bn = Bq = Ga = Ge = Gy = Nd = Nn, Bs	Bc = Ga = Gy = Nd, Bq = Bn, Ge = Nn, Bs	−952.47	35
M00	Bc = Bn = Bq = Ga = Ge = Gy = Nd = Nn = Bs	Bc = Bn = Bq = Ga = Ge = Gy = Nd = Nn = Bs	−970.90	30

The equality symbol (=) indicates locations constrained to be the same in each respective analysis

model compared to 0% for M01. However, percentage of dominance variance at the Bq location explained by the FA1 model in the most parsimonious reduced  $G \times E$  model (69%) was lower than that for M01 (84%). Under the M13 fit,  $z$ -values (indicating accuracy of parameters) for estimated parameters for the FA1 models of the additive and dominance genetic-by-environment covariance matrices were generally considerably higher than those for the fit of M01 to scaled data (Tables 3, 4), although the magnitude of this effect was not consistent across all locations.

Coefficient of variation among estimates of additive genetic variation from the fit of M13 to scaled observations was 0.63, less than that for the fit of the unconstrained  $G \times E$  model to scaled data, but coefficient of variation among dominance variance estimates (0.74), and those for total genetic variance (0.37), were larger. Reflecting the constraints applied in the reduced model, estimates of additive genetic variance in scaled observations from the fit of M13 were equal at Bn, Ge, Nd and Nn (0.22), as were additive genetic variances at Bc, Bq, Ga and Gy (0.08). In agreement with previous models, there was virtually no variance among additive genetic effects at Bs. Similarly, estimates of dominance genetic variance at

Bn, Bq were equal (0.34), as where estimates at Bc, Ga, Gy and Nd (0.17). Dominance genetic variance estimates from the fit of the reduced  $G \times E$  model to scaled observations were virtually 0 at Nn (in contrast with results from the fit of the unconstrained  $G \times E$  model) and Ge.

Total genetic variance was only homogenous among Bc, Ga and Gy (0.24), and at Ge and Nn (0.22) as the locations for which additive genetic effects were constrained to be homogenous were different from the locations for which dominance genetic effects were constrained to be homogenous. Percentage of total genetic variance explained by additive effects estimated from the fit of M13 to scaled data was greater than the percentage estimated from the fit of the unconstrained model to the same data at Bn, Ga, Gy and Nn, but lower at Bc, Bq, and Nd, with an average of 45%. The only location for which generalised heritability of additive effects was relatively higher for the fit of the reduced  $G \times E$  model to scaled data compared to the fit of the unconstrained model was Ga. Estimated of generalised heritability of dominance effects was slightly smaller for the fit of the M13 model compared to the fit of M01 for most

locations (Bc, Bq, Ga, Ge, Gy and Nd), but was relatively larger at Bn and Nn.

Again, the structure of the biplot of de-scaled clonal values from the fit of M13 to scaled data (Fig. 1c) was very similar to the biplots of clonal values predicted from the fit of M01 to unscaled (Fig. 1a) and scaled data (Fig. 1b). Under the M13 model, de-scaled clonal predictions at Ga, Gy and Bc were perfectly correlated, as were clonal predictions at Ge and Nn. The consistency between de-scaled clonal predictions from the fit of M13 and predictions from the fit of M01 to scaled and scaled data is demonstrated by the high correlation between average de-scaled clonal values across locations predicted from M13 and average clonal values across locations predicted from the fit of M01 to unscaled observations (0.9976, Fig. 2b).

Fit of main effects only  $G + E$  model (M00) to scaled observations (sy)

The estimate of additive genetic variance in scaled observations estimated from a non- $G \times E$  model (M00) was 0.12 (40% of total genetic variance, similar to that for more general models), and dominance variance was 0.18. Generalised heritability of additive genetic effects was 0.40, lower than the estimate for the fit of any additive  $G \times E$  model to both scaled and unscaled data, and 0.28 for dominance genetic effects, which was similar in magnitude to that from the dominance  $G \times E$  models to unscaled or scaled data. Nevertheless, de-scaled clonal values of individuals predicted across locations from this analysis was still highly correlated (0.9876) with average clonal values predicted from the fit of the unconstrained  $G \times E$  model to unscaled observations (Fig. 2c).

## Discussion

This study has demonstrated that FA models of both additive and dominance genetic-by-environment matrices can be successfully fitted to multi-environment data, as shown by others (e.g. Oakey et al. 2007). This is achieved by separating the variance-covariance of genetic effects among environments into a correlation matrix of genetic effects and the genetic-by-environment covariance matrix (Oakey et al. 2007; Kelly et al. 2009; Smith et al. 2001). Here a dominance relationship matrix was estimated from the historical

pedigree to predict dominance genetic effects of an individual. Other studies that include non-additive effects have often employed a family effect term, where all families are assumed unrelated, in the statistical model (Hardner et al. 2012; Costa e Silva et al. 2006; Cappa et al. 2012), or have been undertaken with clonally replicated material where a total non-additive genetic effect is used (Paget et al. 2014; Kelly et al. 2009). Although not demonstrated here, accuracy of predictions are higher from multivariate models compared to that for predictions of effects from univariate models by leveraging correlated information (Thompson and Meyer 1986; Hardner et al. 2016). Kelly et al. (2007) demonstrated the mean square error of prediction using an FA models was similar to that from an unstructured parameterisation of the same matrix, and in some cases may be lower for trials containing relatively few genetic entries (i.e. 80 compared to 200 and 500). FA models are particularly appealing for reduced rank genetic-by-environment covariance matrices (Thompson et al. 2003; Burgueno et al. 2011). This may occur when there is a relatively low number of genetic treatments per dimension. These models are also superior to pairwise bivariate estimations of between environment correlations that may produce unrealistic estimates of the covariance matrices (i.e. absolute correlations  $>1$ ) (Cullis et al. 2014; Hill and Thompson 1978).

The linear mixed model framework used here supports an unbiased approach to describing and testing the significance of  $G \times E$  patterns (Burgueno et al. 2008). Estimated environmental loadings of the FA model can be rotated to a principal component representation (Smith et al. 2001). Importantly, the mixed model approach readily accommodates missing and unbalanced data (Henderson et al. 1959; Thompson 1973) such that the effect of all genotypes can be predicted in each environments by leveraging information from relatives tested in those environments. As uncertainty increases (e.g. correlation among environments less than 1.0), predictions are shrunk towards zero (Henderson 1977). This fits neatly with the biplot of the decomposed predictions of clonal value by location table where individuals near the centre of the biplot (i.e. values of zero) represent performance that is not well explained by the singular vector dimensions displayed in the graph.

In the example examined here, the failure of the fit of a full unstructured matrix is a consequence of the limited data particularly for locations where only a single trial was established (i.e. Bn, Bs, Ge and Nn). The relative low values for the specific variances at most sites for the FA1 parameterisations of the additive, and to lesser extent dominance, genetic-by-location covariance matrices indicates that the loadings of the factor analytic parameterisation captured most of the structure of the genetic-by-location covariance matrices, and higher order FA models were unnecessary. The higher location specific variances for the unconstrained dominance genetic-by-location FA1 model may be a consequence of the sparser dominance relationship matrix. The presence of a negative (and small) environment loading for only additive genetic effects at the Bs location suggests that the FA model has not detected large cross-over interaction for additive and dominance effects (Burgueno et al. 2008; Callister et al. 2013) and much of the  $G \times E$  interaction is due to scale differences.

The near perfect correlation between the average of clonal values predicted from the fit of the unconstrained  $G \times E$  model to unscaled observations, and average of descaled clonal values predicted from the fit of the full model to observations scaled by the phenotypic variation for the respective trials, confirms that little bias is introduced by scaling by the phenotypic standard deviation of each trial. In addition, the lower coefficients of variation of genetic parameters estimated from the fit of the unconstrained model to the scaled observations, compared to those estimated from the fit of the same model to unscaled observation, supports recommendations by Hill (1984), Visscher et al. (1991) and White et al. (2007) that phenotypic scaling of observations can be used to reduce variance heterogeneity among environments. Nevertheless, there is evidence of heterogeneity in genetic variances among locations as the fit of the  $G \times E$  models to the data was significantly better than that of the model that did not account for  $G \times E$ .

This study has also demonstrated that the accuracy of estimated model parameters (i.e. components of the FA model) and complexity of  $G \times E$  patterns can be reduced by scaling observations by trial phenotypic standard deviation and combining locations among which genetic variance was homogenous and genetic correlations were one into single environmental

dimension. The very high correlation between average clonal effects for an individual across locations predicted from the fit of the unconstrained  $G \times E$  model to unscaled data, and the average of the descaled predictions of clonal effects from the fit of the most parsimonious constrained  $G \times E$  model, suggests that predictions are not greatly biased by the complexity reduction undertaken here. However, the accuracy of genetic predictions was only improved in a few cases (dominance effects at Bn and Nn, and additive effects at Ga) when the accuracy of estimated model parameters was improved through a reduction in complexity (i.e. estimation required for a fewer number of parameters). This may be a result of the robustness of predictions to variability in genetic parameter estimates (Kennedy 1981; Kackar and Harville 1981). The failure to find significant differences between the unconstrained  $G \times E$  model and restricted parsimonious models suggests that the hypothesis that the greater heterogeneity in genetic parameters estimated from the unrestricted model was due to sampling could not be rejected.

The complexity reduction approached undertaken here is similar to the that undertaken by Burgueno et al. (2008) to identify groups of environment within which genetic correlations among environments was perfect, indicating absence of cross-over interaction. The difference is that Burgueno et al. (2008) estimated a genetic-by-environment dissimilarity matrix using only the loadings, whereas the current study estimated the dissimilarity matrix from the complete (loadings + specific variance) genetic-by-environment covariance matrix, as the aim of the current study was to reduce the dimensions of the covariance matrix. In the current study, scaling was used to reduce variance heterogeneity, whereas heterogeneity of variance among environments would have inflated the specific variance term of the FA parameterisation in Burgueno et al. (2008). Presumably, the Burgueno et al. (2008) approach could be extended to test for homogeneity of specific variances among environments. In addition, an independent relationship model was fitted in the models in Burgueno et al. (2008), whereas here both additive and dominance relationship matrices were fitted. Both approaches provide formal methods to test for heterogeneity among environments and identifying mega-environments, rather than an arbitrary approach employing analysis of visual summaries of the data (Yang et al. 2009). However,

knowledge of  $G \times E$  patterns are only useful for selection if some repeatable factor can be identified which explains the detected patterns and can be used to predict patterns in the target deployment environments (Allard and Bradshaw 1964). Reducing the influence of noise on the complexity of  $G \times E$  patterns by complexity reduction may aid identification drivers of these patterns and hence their utilisation.

Hill (1984) demonstrates that when variances are heterogeneous across environments, but heritabilities are similar and genetic effects are highly correlated, selection based on a main effects only model will tend to select more individuals that have been evaluated in more variable environments. In this study, estimated generalised heritability was relatively similar across environments for the most parsimonious  $G \times E$  model. However, the high correlation of clonal effects predicted from the fit of the main effects only  $G + E$  model to scaled observations with average of clonal effects predicted from the fit of a fully unconstrained  $G \times E$  model (to the same data) suggests that clonal effects predicted from the  $G + E$  only model are not greatly biased. This is apparent even though the  $G \times E$  models were a significantly better fit to the data, and the biplot of the first two singular location vectors indicated  $G \times E$ . It may be that the relatively high correlation of clonal values among all locations, except Bs, overwhelms the interaction in clonal values between the Bs and the other locations. However, accuracy of additive genetic values predicted from  $G \times E$  models are greater than for the main genotypic effect model (Smith et al. 2015). A bias may also arise when modelling  $G \times E$  using the simplified main effect and interaction model ( $G + G \times E$ , which treats genetic variance and genetic correlations as homogeneous among environments). In addition, modelling of  $G \times E$  provides more flexibility in that weights can be used to target selection for specific environments (Smith et al. 2015; Cooper and Delacy 1994). In this study, genetic variance at different trials at the same location were assumed to be homogeneous with perfect genetic correlation as the number of records for most trials was low possibly resulting in over-parameterisation, difficulty in estimation and decreased accuracy (Wolak 2012), and because planting year causes variation which is difficult to control.

**Acknowledgements** The production of study germplasm, establishment and assessment of field trials, and analyses was

funded in part by Horticulture Australia Limited, CSIRO and the Queensland Government. The support of the landowners in providing the land and trial maintenance is also acknowledged. The initial idea for this study was generated following discussion with Dr Alison Kelly, Department of Agriculture and Fisheries, Queensland, Australia.

## References

- Akaike H (1974) New look at model identification. *IEEE Trans Autom Control* AC19, pp 716–723
- Allard RW, Bradshaw AD (1964) Implications of genotype-environment interaction in applied plant breeding. *Crop Sci* 4:503–508
- Arief VN, DeLacy IH, Crossa J, Payne T, Singh R, Braun HJ, Tian T, Basford KE, Dieters MJ (2015) Evaluating testing strategies for plant breeding field trials: redesigning a CIMMYT International Wheat Nursery. *Crop Sci* 55:164–177
- Baker RJ (1988) Tests for crossover genotype-environment interactions. *Can J Plant Sci* 68:405–410
- Burdon RD (1977) Genetic correlation as a concept for studying genotype-environment interaction in forest tree breeding. *Silvae Genet* 26:168–175
- Burgueno J, Crossa J, Cornelius PL, Yang RC (2008) Using factor analytic models for joining environments and genotypes without crossover genotype  $\times$  environment interaction. *Crop Sci* 48:1291–1305
- Burgueno J, Crossa J, Cotes JM, San Vicente F, Das B (2011) Prediction assessment of linear mixed models for multi-environment trials. *Crop Sci* 51:944–954
- Callister AN, England N, Collins S (2013) Predicted genetic gain and realised gain in stand volume of *Eucalyptus globulus*. *Tree Genet Genomes* 9:361–375
- Cappa EP, Yanchuk AD, Cartwright CV (2012) Bayesian inference for multi-environment spatial individual-tree models with additive and full-sib family genetic effects for large forest genetic trials. *Ann For Sci* 69:627–640
- Comstock RE, Moll RH (1963) Genotype-environment interactions. In: Hanson WD, Robinson HF (eds) *Statistical genetics and plant breeding*. Publication 982. National Academy of Sciences-National Research Council, Washington, USA, pp 164–196
- Cooper M, Delacy IH (1994) Relationships among analytical methods used to study genotypic variation and genotype-by-environment interaction in plant-breeding multi-environment experiments. *Theor Appl Genet* 88:561–572
- Costa e Silva J, Gaudal L (2008) Evaluation of an international series of *Pinus kesiya* provenance trials for growth and wood quality traits. *For Ecol Manag* 255:3477–3488
- Costa e Silva J, Dutkowski GW, Gilmour AR (2001) Analysis of early tree height in forest genetic trials is enhanced by including a spatially correlated residual. *Can J For Res* 31:1887–1893
- Costa e Silva J, Potts BM, Dutkowski GW (2006) Genotype by environment interaction for growth of *Eucalyptus globulus* in Australia. *Tree Genet Genomes* 2:61–75

- Cullis B, Gogel B, Verbyla A, Thompson R (1998) Spatial analysis of multi-environment early generation variety trials. *Biometrics* 54:1–18
- Cullis BR, Smith AB, Coombes NE (2006) On the design of early generation variety trials with correlated data. *J Agric Biol Environ Stat* 11:381–393
- Cullis BR, Jefferson P, Thompson R, Smith AB (2014) 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. *Theor Appl Genet* 127:2193–2210
- Dutkowski GW, Silva JCE, Gilmour AR, Lopez GA (2002) Spatial analysis methods for forest genetic trials. *Can J For Res* 32:2201–2214
- Falconer DS (1952) The problem of environment and selection. *Am Nat* 86:293–298
- Gilmour AR, Cullis BR, Verbyla AP (1997) Accounting for natural and extraneous variation in the analysis of field experiments. *J Agric Biol Environ Stat* 2:269–273
- Gilmour AR, Gogel B, Cullis BR, Thompson R (2009) ASReml User Guide Release 3.0. VSN International Ltd, Hemel Hempstead, UK
- Hahsler M, Hornik K, Buchta C (2008) Getting things in order: an introduction to the R package seriation. *J Stat Softw* 25:1–34
- Hardner CM, Peace C, Lowe AJ, Neal J, Pisanu P, Powell M, Schmidt A, Spain C, Williams K (2009) Genetic resources and domestication of macadamia. In: Janick J (ed) *Horticultural Reviews*, vol 35. Wiley, Hoboken, New Jersey, pp 1–125
- Hardner CM, Dieters M, Dale G, DeLacy I, Basford KE (2010) Patterns of genotype-by-environment interaction in diameter at breast height at age 3 for eucalypt hybrid clones grown for reforestation of lands affected by salinity. *Tree Genet Genomes* 6:833–851
- Hardner CM, Bally ISE, Wright CL (2012) Prediction of breeding values for average fruit weight in mango using a multivariate individual mixed model. *Euphytica* 186:463–477
- Hardner CM, Healey AL, Downes G, Herberling M, Gore PL (2016) Improving prediction accuracy and selection of open-pollinated seed-lots in *Eucalyptus dunnii* Maiden using a multivariate mixed model approach. *Ann For Sci* 73:1035–1046
- Henderson CR (1975) Best linear unbiased estimation and prediction under a selection model. *Biometrics* 31:423–447
- Henderson CR (1977) Best linear unbiased prediction of breeding values not in model for records. *J Dairy Sci* 60:783–787
- Henderson CR (1985) Best linear unbiased prediction of non-additive genetic merits on noninbred populations. *J Anim Sci* 60:111–117
- Henderson CR, Kempthorne O, Searle SR, Vonkrosigk CM (1959) The estimation of environmental and genetic trends from records subject to culling. *Biometrics* 15:192–218
- Hill WG (1984) On selection among groups with heterogeneous variance. *Anim Prod* 39:473–477
- Hill WG, Thompson R (1978) Probabilities of non-positive definite between group of genetic covariance matrices. *Biometrics* 34:429–439
- Kackar RN, Harville DA (1981) Unbiasedness of 2-stage estimation and prediction procedures for mixed linear-models. *Commun Stat A-Theor* 10:1249–1261
- Kelly AM, Smith AB, Eccleston JA, Cullis BR (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 Sel Evol* 41:33
- Kempton RA (1984) The use of biplots in interpreting variety by environment interactions. *J Agric Sci* 103:123–135
- Kennedy BW (1981) Variance component estimation and prediction of breeding values. *Can J Genet Cytol* 23:565–578
- Kenward MG, Roger JH (1997) Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* 53:983–997
- Matheson AC, Cotterill PP (1990) Utility of genotype x environment interactions. *For Ecol Manag* 30:159–174
- Oakey H, Verbyla A, Pitchford W, Cullis B, Kuchel H (2006) Joint modeling of additive and non-additive genetic line effects in single field trials. *Theor Appl Genet* 113:809–819
- Oakey H, Verbyla AP, Cullis BR, Wei XM, Pitchford WS (2007) Joint modeling of additive and non-additive (genetic line) effects in multi-environment trials. *Theor Appl Genet* 114:1319–1332
- Paget MF, Alspach PA, Genet RA, Apiolaza LA (2014) Genetic variance models for the evaluation of resistance to powdery scab (*Spongospora subterranea* f. sp subterranea) from long-term potato breeding trials. *Euphytica* 197:369–385
- Patterson HD, Thompson R (1971) Recovery of interblock information when block sizes are unequal. *Biometrika* 58:545–554
- Piepho HP (1997) Analyzing genotype-environment data by mixed models with multiplicative terms. *Biometrics* 53:761–766
- Piepho HP, Moring J (2007) Computing heritability and selection response from unbalanced plant breeding trials. *Genetics* 177:1881–1888
- Silva JCE, Borralho NMG, Araujo JA, Vaillancourt RE, Potts BM (2009) Genetic parameters for growth, wood density and pulp yield in *Eucalyptus globulus*. *Tree Genet Genomes* 5:291–305
- Smith A, Cullis B, Thompson R (2001) Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. *Biometrics* 57:1138–1147
- Smith AB, Cullis BR, Thompson R (2005) The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. *J Agric Sci* 143:449–462
- Smith AB, Ganesalingam A, Kuchel H, Cullis BR (2015) Factor analytic mixed models for the provision of grower information from national crop variety testing programs. *Theor Appl Genet* 128:55–72
- Stram DO, Lee JW (1994) Variance-components testing in the longitudinal mixed effects model. *Biometrics* 50:1171–1177
- Thompson R (1973) Estimation of variance and covariance components with an application when records are subject to culling. *Biometrics* 29:527–550



- Thompson R, Meyer K (1986) A review of theoretical aspects in the estimation of breeding values for multi-trait selection. *Livest Prod Sci* 15:299–313
- 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
- Visscher PM, Thompson R, Hill WG (1991) Estimation of genetic and environmental variances for fat yield in individual herds and an investigation into heterogeneity of variance between herds. *Livest Prod Sci* 28:273–290
- White TL, Adams WT, Neale DB (2007) *Forest genetics*. CAB International Wallingford, Wallingford, UK
- Wilks SS (1938) The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann Math Stat* 9:60–62
- Wolak ME (2012) *nadiv*: an R package to create relatedness matrices for estimating non-additive genetic variances in animal models. *Methods Ecol Evol* 3:792–796
- Yamada Y (1962) Genotype by environment interaction and genetic correlation of the same trait under different environments. *Jpn J Genet* 37:498–509
- Yang RC, Crossa J, Cornelius PL, Burgueno J (2009) Biplot analysis of genotype x environment interaction: proceed with caution. *Crop Sci* 49:1564–1576